

# **Integrative *omics* approaches for new target identification and therapeutics development**

Austė Kanapeckaitė

Thesis submitted for the Degree of Doctor of Philosophy

**School of Chemistry, Food and Pharmacy**

**Department of Pharmacology**

**December 2021**



## **Declaration of original authorship**

I certify that the research presented in this thesis is my own work. All resources and contributions have been properly and fully acknowledged.

Austė Kanapeckaitė

## Following publications originated within the scope of the conferral of a doctorate:

\*included in this dissertation

### Research Articles

1. \*[Kanapeckaitė A](#), Burokienė N. **Insights into therapeutic targets and biomarkers using integrated multi-‘omics’ approaches for dilated and ischemic cardiomyopathies.** Integrative Biology. 2021 May;13(5):121-37; doi: 10.1093/intbio/zyab007. PMID: 33969404.
2. \*[Kanapeckaitė A](#), Beaurivage C, Hancock M, Verschueren E. **Fi-score: a novel approach to characterise protein topology and aid in drug discovery studies.** Journal of Biomolecular Structure and Dynamics. 2020 Dec 7:1-1; doi: 10.1080/07391102.2020.1854859; PMID: 33297860.
3. Beaurivage C, [Kanapeckaitė A](#), Loomans C, Erdmann KS, Stallen J, Janssen RA. **Development of a human primary gut-on-a-chip to model inflammatory processes.** Nature Scientific Reports. 2020 Dec 8;10(1):1-6; doi: 10.1038/s41598-020-78359-2.
4. \*[Kanapeckaitė A](#), Beaurivage C, Jančorienė L, Mažeikienė A. **In silico drug discovery for a complex immunotherapeutic target-human c-Rel protein.** Biophysical Chemistry. 2021 Sep 1;276:106593; doi: 10.1016/j.bpc.2021.106593.  
*Selected as the issue cover.*
5. \*[Kanapeckaitė A](#). **Fiscore: effective protein structural data visualisation and exploration.** Artificial Intelligence in the Life Sciences. 2021 Dec 1;1-1; doi: 10.1016/j.aillsci.2021.100016.
6. \*[Kanapeckaitė A](#). **OmicInt package: exploring omics data and regulatory networks using integrative analyses and machine learning.** Artificial Intelligence in the Life Sciences. 2021 Dec 1;1-1; doi: 10.1016/j.aillsci.2021.100025.

## Reviews

1. Kanapeckaitė A, Burokienė N, Mažeikienė A, Cottrell GS, Widera D. **Biophysics is reshaping our perception of the epigenome: from DNA-level to high-throughput studies**. Cell Biophysical Reports. 2021 Sep 29:100028; doi: /10.1016/j.bpr.2021.100028.

## Software packages

1. \*Kanapeckaitė A. **OmicInt: Omics Network Exploration**. CRAN. 2021 Oct. 15. Version 1.1.7; <https://cran.r-project.org/web/packages/OmicInt/index.html>
2. \*Kanapeckaitė A. **Fiscore: Effective Protein Structural Data Visualisation and Exploration**. CRAN. 2021 Sep. 02. Version 0.1.3; <https://cran.r-project.org/web/packages/Fiscore/index.html>
3. \*Kanapeckaitė A. **Chemexpy: Cheminformatics package for compound feature evaluation**. PyPi. 2021 Oct. 07. Version 1.0.10; <https://pypi.org/project/chemexpy/>

## Conference talks

1. Kanapeckaitė A, Burokienė N, Mažeikienė A, Cottrell GS, Widera D. **Integrated *in silico* strategies for discovery of therapeutics with a focus on complex immunotherapeutic targets**. UK Conference of Bioinformatics and Computational Biology. 2021 Sep 29.

## Abstract

The growing research and commercial pressures for novel therapeutics development accentuate why better strategies are needed for drug discovery. The costly nature of developing a pharmaceutical compound as well as the shrinking pool of ‘easy’ targets are some of the key reasons why there is a research paradigm shift towards integrative and systems biology driven approaches. Moreover, multifactorial aspects of many diseases require more innovative clinical strategies rather than just focusing on a single target. Cardiovascular diseases as well as associated immune components exemplify this complexity well. This thesis aimed to introduce a gradual and highly integrative analytical framework by incorporating a full range of studies from disease target selection to high-throughput virtual screening so that a cost-effective and efficient stratification of targets and associated compounds could be achieved. Heart failure served as a case study for complex diseases where the first in-depth *omics* study on cardiomyopathies helped to elucidate new therapeutic avenues. This research tied in with a development of a novel scoring function and integrated machine learning approach for multiple therapeutic target classification and exploration. Finally, all pieces of the introduced research were used to create a highly integrative *in silico* screening workflow. Some of the key results included the first reported molecular dynamics analyses for a complex immunotherapeutic target, c-Rel, as well as 15 new therapeutic compounds that could potentially modulate this transcription factor subunit. Thus, this dissertation provided several important improvements for target identification, validation, and drug discovery that could significantly advance current development strategies and accelerate new therapeutics production.

## **Acknowledgements**

I would first like to thank my supervisors Dr Darius Widera, Dr Graeme S. Cottrell, and Dr Asta Mažeikienė for their very kind and patient support throughout the writing of this dissertation. Their insightful feedback and guidance not only encouraged me to bring my work to the higher level, but also helped me gain valuable research and writing experience.

I would also like to express my sincere gratitude to Dr Farhad Forouhar for his kind help, valuable advice, and mentorship during my time as a graduate researcher. Dr Forouhar's support meant a lot and the learnt lessons gave me new perspectives to successfully complete my dissertation.

I am also very grateful to all the researchers who shared their advice and insights throughout the various stages of writing this thesis. All the very in-depth discussions helped me to better crystallise research objectives.

## Table of contents

<b>Abstract</b> .....	6
<b>Acknowledgements</b> .....	7
<b>Table of contents</b> .....	8
<b>Abbreviations</b> .....	10
<b>Summary</b> .....	12
<b>1. Introduction</b> .....	14
1.1. Drug discovery and development: a historical perspective on how global R&D trends changed and shaped therapeutics development.....	14
1.2. Shifting paradigms in drug discovery and development: from high-throughput target-specific approaches to searching for new multi-network strategies.....	17
1.3. Novel R&D framework for complex diseases: rethinking therapeutics development with case studies on cardiomyopathies and inflammatory disease components.....	24
1.4. Development of a network-centric and highly integrative discovery process: addressing R&D challenges and creating new opportunities.....	31
1.5. Biophysical and computational chemistry method development: streamlining complex target evaluation and therapeutics discovery.....	35
<b>2. Insights into therapeutic targets and biomarkers using integrated multi-‘omics’ approaches for dilated and ischemic cardiomyopathies</b> .....	41
<b>3. <i>OmicInt</i> package: exploring <i>omics</i> data and regulatory networks using integrative analyses and machine learning</b> .....	59
<b>4. <i>Fi</i>-score: a novel approach to characterise protein topology and aid in drug discovery studies</b> .....	74
<b>5. <i>Fiscore</i> package: effective protein structural data visualisation and exploration</b> .....	86
<b>6. <i>In silico</i> drug discovery for a complex immunotherapeutic target - human c-Rel protein</b> .....	98



<b>7. General discussion</b> .....	116
7.1. Towards new R&D strategies: cardiomyopathies study revealed how to improve complex disease analyses and find new therapeutic avenues.....	116
7.2. Implementing a streamlined target evaluation and classification: new solutions for discovery pipelines.....	122
7.3. Highly integrative <i>in silico</i> screening pipeline: a better method to explore targets and capture potential hit compounds.....	124
7.4. Programmatic approaches and data management: maintaining and designing robust workflows.....	128
7.5. Thesis overview and conclusion.....	131
<b>8. Future work</b> .....	133
<b>9. References</b> .....	134
<b>10. Supplementary materials</b> .....	154
10.1. Insights into therapeutic targets and biomarkers using integrated multi-‘omics’ approaches for dilated and ischemic cardiomyopathies.....	155
10.2. Fi-score: a novel approach to characterise protein topology and aid in drug discovery studies.....	217
10.3. <i>Fiscore</i> package: effective protein structural data visualisation and exploration.....	223
10.4. <i>In silico</i> drug discovery for a complex immunotherapeutic target - human c-Rel protein.....	229
10.5. <i>Chemexpy</i> documentation.....	241

## Abbreviations

ACE	Angiotensin I-converting enzyme
ADME	Absorption, distribution, metabolism, and excretion
ADMET	Absorption, distribution, metabolism, excretion, and toxicity
ARB	Angiotensin II receptor blockers
cDNA	Complementary DNA
CVD	Cardiovascular disease
CYP	Cytochrome P450
DC	Dilated cardiomyopathy
DNA	Deoxyribonucleic acid
EF	Ejection fraction
FDA	US Food and Drug Administration
GCP	Good clinical practice
GLP	Good laboratory practice
GMM	Gaussian mixture model
GMP	Good manufacturing practice
GSEA	Gene set enrichment analysis
HF	Heart failure
HFpEF	Heart failure preserved ejection fraction
HF <sub>r</sub> EF	Heart failure reduced ejection fraction
HTS	High-throughput screening
HTVS	High-throughput virtual screening
IC	Ischemic cardiomyopathy
iPSC	Induced pluripotent stem cell
IKK	IκB kinase complex
ISH	<i>In situ</i> hybridisation
LVEF	Left ventricular ejection fraction
LC-MS	Liquid chromatography–mass spectrometry
miRNA	Micro RNA
MS	Mass spectrometry
mRNA	Messenger RNA
MRSA	Methicillin-resistant <i>Staphylococcus aureus</i>
NBE	New biological entity
NF-κB	NF-kappaB or nuclear factor kappa-light-chain-enhancer of activated B cells

NGS	Next generation sequencing
NIH	The National Institutes of Health
NME	New molecular entity
NMR	Nuclear magnetic resonance
PCR	Polymerase chain reaction
QC	Quality control
QSAR	Quantitative structure-activity relationships
R&D	Research and discovery
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
RT-PCR	Real time polymerase chain reaction
RT-qPCR	Reverse transcription-quantitative polymerase chain reaction
UHTS	Ultra-high-throughput screening
vdW	van der Waals

## Summary

Major advances in the pharmaceutical industry were primarily driven by the need to identify good therapeutic targets. However, it quickly became apparent that it is not enough to just screen multiple compounds or perform various genomic/phenotype screens for potential hit identification. As a result, there is an urgent need to develop new methods for biological data integration, network-centric target exploration, and *in silico* drug discovery. Moreover, paradigms in drug discovery are beginning to shift from target- to network-centric approaches so that better therapeutic options can be found for complex diseases. This was also addressed in the first experimental chapter of the thesis demonstrating that current strategies in treating multifactorial diseases are not sufficient and more integrative solutions are needed. As a case study, heart failure (specifically, dilated and ischemic cardiomyopathies) was investigated through the combination of bulk and single cell RNA-seq as well as the proteome and interactome datasets to reveal a high heterogeneity of various biological data resources. In addition, a scoring function was derived to better capture the interactome complexity and disease associations. This study also introduced a two-step machine learning pipeline to cluster and extract information on the targets that show favourable profiles for therapeutics development. The analysis revealed that despite the complex aetiology of heart failure, it is possible to elucidate metabolic and functional pathways that show therapeutic potential. This part of the research was accompanied by a development of a specialised software package to make such analyses more accessible to researchers (Chapter 3). The third experimental chapter of the dissertation aimed to address the need of better categorisation and exploratory approaches for multiple identified targets so that the data can be grouped based on physicochemical and structural features. For this purpose, a novel scoring system/methodology was devised helping to capture both local and distant topological as well as conformational features that can allow differentiating specific structural motifs in a protein. That is, the core goal of this analysis was to provide an effective method to characterise proteins prior to *in silico* screening by evaluating potentially dynamically active regions. The ability to categorise such highly dimensional data in an easy-to-store-and-retrieve way could significantly fast-track drug screening studies. Finally, the demonstrated machine learning approaches can expand the analysis of multiple targets by extracting and defining structural elements and motifs of various proteins. In order to aid with the implementation of the introduced scoring and machine learning methods, a special software

package was developed supplementing this experimental chapter (Chapter 5). The final experimental chapter of the thesis introduced a pipeline for an efficient development of therapeutic agents by building on the previous studies. The human c-Rel protein, as a challenging immunotherapeutic target, was chosen to demonstrate the existing hurdles and how they can be overcome using an integrative analytical approach starting with a careful selection of a target, followed by the evaluation of its druggability potential, and finally performing a stepwise *in silico* screening. A compound library of an unprecedented size (34 M) was prepared from an even larger set of compounds (659 M) which after gradual screening led to the identification of 15 high-scoring drug-like structures that could be used for preclinical screens as potentially highly selective c-Rel inhibitors/modulators. In addition, state-of-the-art molecular modelling and dynamics analyses provided for the first time some hints at how the target protein might interact with the DNA sequence. A cheminformatics software package was also created to help with screening compound selection and assessment (Supplementary materials).

Overall, the innovative biophysical, computational biology, bioinformatics, and cheminformatics methods presented here could significantly improve target selection and pre-screening analysis as well as accelerate pharmaceuticals development. Importantly, the developed highly integrative and network-centric approaches allow for a better understanding of pathological perturbations and can help deliver so much needed therapies faster and with a safer profile.

## 1. Introduction

### 1.1. Drug discovery and development: a historical perspective on how global R&D trends changed and shaped therapeutics development

The origins of the modern pharmaceutical industry can be traced back to the middle of the 19<sup>th</sup> century when companies, such as Eli Lilly, Merck, and Roche, moved into large-scale production of drugs. Moreover, newly established pharmaceutical businesses, such as Bayer, ICI, Sandoz, as well as Pfizer, started developing research labs to focus on medical applications and scale-up their chemical production<sup>1</sup>. Early in the 20<sup>th</sup> century, major pharmacological advances in synthetic organic chemistry and new compound exploration transformed the drug industry into large-scale manufacturing to meet the increasing demands of newly introduced drugs, such as analgesics and antibiotics<sup>1</sup>. The expanding pharmaceuticals market also prompted governments to undertake research and introduce necessary regulatory steps to establish safety and distribution policies. Furthermore, turbulent 20<sup>th</sup> century history and economic changes created a perfect environment for a small number of very large multi-national companies to dominate the market by the end of the century. Growing pharmaceutical businesses took advantage of the extraordinary scientific progress that allowed to associate a specific gene with a disease which in turn led to the emergence of new premises in research and discovery (R&D)<sup>1-3</sup>. One such novel concept was a ‘blockbuster’ drug that addressed a significant medical need and generated annual sales of at least \$1 billion<sup>1,4</sup>. Particularly, a ‘blockbuster’ is characterised by its market dominance for a specific indication, wide population use, and ability to achieve substantial profits. Such drugs typically represent a particular therapeutics class, e.g., statins<sup>1</sup>. In light of the growing number of potential therapeutic targets, this also marked the change from low-throughput studies to the development of high-throughput strategies as increasing production outputs became a necessity in both biology and chemistry to meet the demand of new drugs<sup>2,3,5-7</sup>. The abundance of data and funding allowed to identify obvious links between pathological phenotypes and the offending protein or proteins<sup>1,2</sup>. However, the era of 'low hanging fruit' in drug discovery started to dwindle towards the end of 20<sup>th</sup> century as accumulating costs, growing regulatory oversight, and the shrinking pool of ‘easy’ targets reduced companies’ outputs<sup>1-3,5,8</sup>. This became especially evident over the last twenty years when companies began to rethink old paradigms and search for innovative approaches to develop therapeutics<sup>2,8-12</sup>. Therefore, changing motivations of pharmaceutical companies as well as the

undercurrents that shaped today's R&D practices can be better understood when considering how drug discovery evolved during the turn of the century.

Revisiting the past 50 years of drug discovery, we can see how significant scientific advancements created both new opportunities and the vacuum space in R&D which may help explain current trends in the pharmaceutical industry. Since the 1980s, the scope, quality, and even the cost efficiency of the scientific and technological methods have markedly improved. Biopharmaceutical industry took full advantage of combinatorial chemistry by not only increasing the number of drug-like molecules that could be synthesised, but also scaling up considerably the size of chemical libraries<sup>7,13</sup>. Everything from sequencing to the fewer man-hours required to determine a three-dimensional protein structure facilitated the identification of lead compounds and targets<sup>13</sup>. High-throughput screening (HTS) of compound libraries against proteins of interest became ten times less expensive between mid-1990s and 2008<sup>5</sup>. Overall, the introduction of new discovery tools, such as computational analyses and transgenic mice to model pathologies, not only improved the scientific understanding of disease mechanisms, but also helped to form target-guided strategies<sup>13</sup>. While such improvements should have guaranteed a higher reliability in therapeutics development, the contrasting uneconomical R&D management and low numbers of new therapeutics pointed to many overlooked aspects in research. Particularly, shortcomings in industrial research organisation as well as an insufficient appreciation of the complex chemical and biological space stood out<sup>13,14</sup>. The exhaustion of obvious druggable targets have also often been used to explain the decreasing numbers of new molecular entity (NME) approvals and growing R&D expenditures<sup>14</sup>. This observation is based on the predicted druggable biological space which is approximated to contain around 600–1500 'drug targets' that could become the focus of industrial research<sup>14,15</sup>. Additional hurdles in improving NME outputs may also stem from the fact that the search for new targets often begins with basic academic research which is only later transferred to the industrial drug discovery setting<sup>14</sup>. Basic research in the context of pharmaceutical R&D is typically more risk-averse, while academic institutions tend to pursue novel and higher-risk targets with long-term investments in basic research<sup>13,14,16</sup>. This dichotomy may also explain the differences in the novelty of identified therapeutic targets and why pharmaceutical companies are now seeking to establish stronger partnerships with academia to identify potential therapeutic breakthroughs<sup>14,16</sup>. Finally, to better appreciate the changing landscape of R&D challenges and the low numbers of NMEs over the last decades, it is also important to consider the asymmetric situation caused by the patents system. Investments reaching billions in search for new drugs might

not bring in profit if the pipeline fails. Yet, pharmaceuticals that succeed to reach the market can be priced excessively to withstand commercial pressures and cover R&D past and future expenditures. The global patenting systems enabled and are still enabling pharmaceutical business to exploit various pricing schemes because once a therapeutic is out of patent it can be sold as a 'generic' at a considerably lower cost<sup>1,10</sup>. Thus, opposing business models in pharma companies pursuing new drugs and those capitalising on generic pharmaceuticals also add to the creation of an uneven risk and cost distribution which further increases R&D pressures and product costs<sup>1</sup>. All these factors contribute towards present day R&D issues, as core strategies in drug development programs evidently did not catch up with growing market constraints and the shrinking space of viable targets that can move quickly and successfully through the pipeline.

In order to better appreciate the business and research models that have emerged during the past 50 years, it is also necessary to consider the research timelines and expenditure dynamics since discovery and development of a new drug can not only take decades to reach patients, but can also require significant investments<sup>2,10,13,17,18</sup>. It is estimated that there was a steady rise in drug development costs since the 1950s with a linear increase on a logarithmic scale in R&D spending for every newly approved drug<sup>19</sup>. Such a steep growth in R&D spending to develop a single drug can result in much higher actual production costs than usually quoted 1.6 billion US \$ per drug successfully released into the market<sup>20</sup>. Considering not only the investment needed for drug development but also calculating in the attrition rates, i.e., failed compounds for a specific indication, the price for therapeutics production increases dramatically. Some estimates indicate that the cost of developing a drug and bringing it to the market was as high as 4 billion US \$ or more between 1997 and 2011<sup>17</sup>. Despite continuously increasing R&D expenditures, the number of new therapeutics has been in a steady decline<sup>20</sup>. Stagnation in NME development has been evident in the past decades with the overall approvals by the US Food and Drug Administration (FDA) remaining low since the 2000s after peaking in the mid-1990s<sup>14</sup>. These trends can be attributed to the high drop-out rates of candidate drugs during preclinical screening and later attrition in clinical testing phases due to safety or efficacy issues<sup>11,20,21</sup>. Late-stage attrition rates are estimated to be as high as 75%<sup>20,22</sup> representing one of the reasons for a high production and sale cost cycle. In addition, our prognostic abilities to predict the success or failure of a therapeutic candidate are not sufficient. In this context, only 13.8% of drug programs successfully progressed from Phase I to the final approval by the FDA (2000-2015)<sup>21,23-26</sup>. It was found that cancer drugs had the lowest success rate of 3.4%<sup>24-26</sup>. Similarly, drugs targeting the central nervous system also have poor success rates and



require longer development times when compared to other drug classes. Specifically, the success rate of neuropsychiatric drug candidates reaching the market is low (8.2%) and this trend can likely be explained by on average longer clinical development and trial time. Moreover, neurological agents typically fail during later clinical phases which also makes them a particularly expensive drug class to develop<sup>16</sup>. It is also important to note that clinical trial outcome tracking is dependent on the available information and statistical assessment methods which can introduce various discrepancies and biases in the reported statistics<sup>25</sup>. Such high failure rates lower investors' confidence in pursuing new drug discovery programs or alternative therapies, especially since the process might take years before a program's clinical potential is seen<sup>17,18,22</sup>. It has been argued that the industry needs to develop improved analytical frameworks for R&D pipelines; for example, early proof-of-concept screens, clinical risks identification, and a robust integration of risk evaluation with experimental medicine procedures have been listed as crucial factors to reform the discovery infrastructure<sup>2,10-13,18,27</sup>. However, it appears that boosting existing practices might not be enough as low success rates primarily reflect that the current understanding of selected targets and their tractability is insufficient<sup>2,8,10,22,28</sup>.

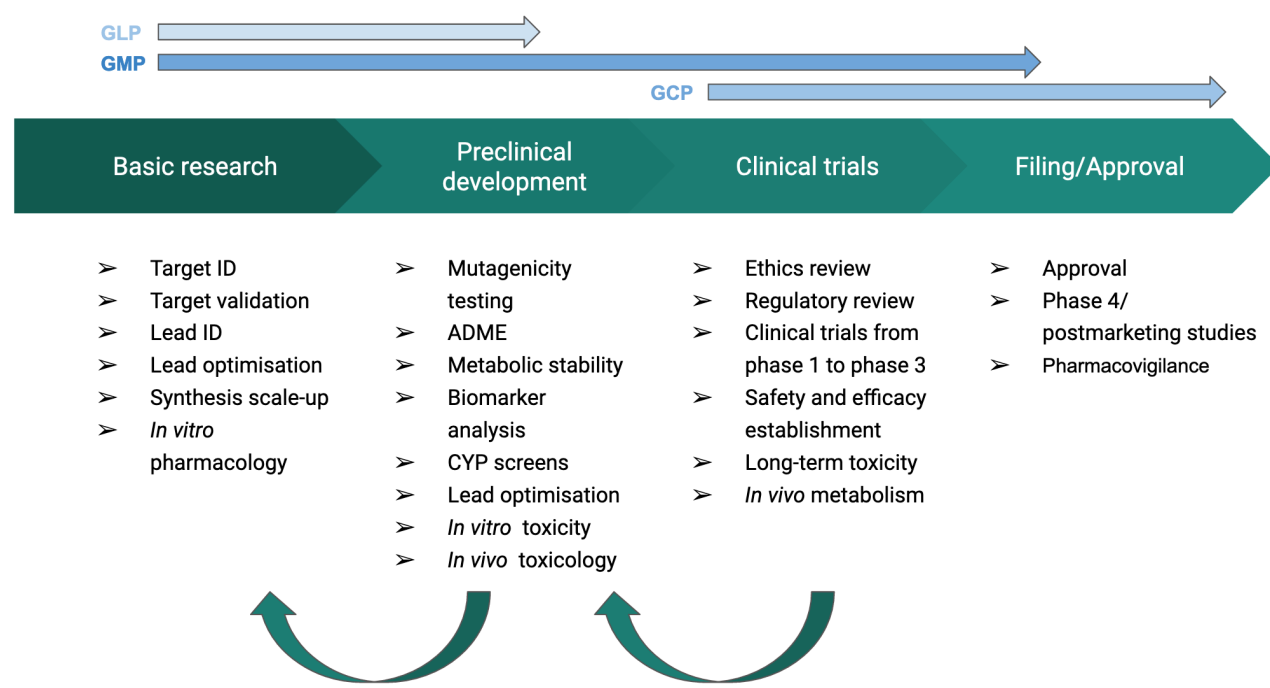
Considering the historical and economic context of the last half a century it becomes evident that innovating therapeutics development, accelerating R&D pipelines, and expanding the druggable target space can only be achieved if the fundamental approaches in drug discovery change<sup>11,13,17,18,22,27</sup>. Moreover, while one of the key research areas in current drug discovery remains finding better methods to identify unwanted toxicity or low efficacy as early in the pipeline as possible<sup>10</sup>, it is crucial to establish a more integrative approach towards drug discovery and take advantage of developments in the computational R&D space<sup>2,28</sup>. In other words, today's fractioned R&D space, despite the overall science progress, highlights why seeing the 'big picture' of discovery pipelines and focusing on integrative holistic approaches can help with the current challenges.

## **1.2. Shifting paradigms in drug discovery and development: from high-throughput target-specific approaches to searching for new multi-network strategies**

Developing new strategies for drug discovery rests on transforming the existing scientific and technological tools<sup>1</sup>. Specifically, rethinking some of the prevailing paradigms in target identification and tractability evaluation in early preclinical screens is believed to be necessary in order to improve the development of therapeutics and open the markets for more innovative

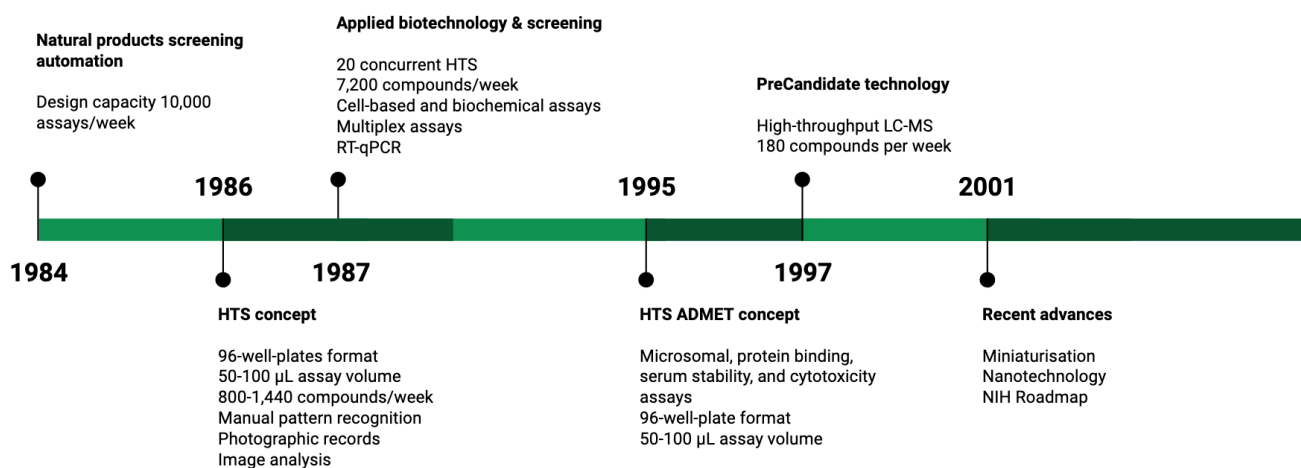
treatments<sup>9,13,18,19,27</sup>. Yet, this shift towards new approaches is gradual since recently introduced discovery and development protocols are also subject to regulatory validation<sup>1,8,10</sup>. Consequently, in order to understand the evolution from target-centric screening to disease network exploration, it is necessary to consider how preclinical strategies advanced during the past decades and how this affected R&D approaches<sup>1-3,6-8,29</sup>.

Broadly, drug development can be divided into several stages that build on early exploratory studies with increasingly complex assays and screens<sup>1-3,11</sup>. The process begins with disease-related genomics as well as target identification and validation studies. This is followed by lead discovery and optimisation studies which, if successful, conclude with clinical trials<sup>30</sup> (Fig. 1). However, before a complete pipeline is outlined and decided on, identifying a new drug starts by defining a particular disease of interest which might be studied in academia or in a pharmaceutical company's basic science division<sup>2,5,14</sup>. Typically, only once the research narrows down a specific target whether it is a gene, protein, or any other biological element that can be modulated, can further discovery work begin to define chemical entities that have the potential to engage that target<sup>1</sup>. This is an iterative process with feedback loops from preclinical development and clinical trials where many discovery steps intertwine (Fig. 1). Although a preclinical drug discovery program aims to deliver at least one clinical candidate molecule with sufficient biological activity, most discovery programs are designed to generate more than one potential drug candidate to minimise safety, potency, kinetics, and other clinical risks<sup>1,2</sup>. Despite this R&D outline, there is no one blueprint for finding good clinical candidate molecules and an extensive collaboration of biology, chemistry, toxicology, and pharmacokinetics is paramount for tailoring a specific pipeline to ensure clinical success<sup>1,8,14,16</sup> (Fig. 1).



**Figure 1.** Drug discovery and development stages from the basic research to approval. This process can be focused on small molecules (NMEs) or biological molecules (NBEs). Throughout the process various feedback loops exist that are represented with green arrows. Blue arrows depict the timelines for quality assurance processes, such as good laboratory practice (GLP), good manufacturing practice (GMP), and good clinical practice (GCP). Other abbreviations: absorption, distribution, metabolism, and excretion (ADME); cytochrome P450 (CYP). Based on the information from Mohs et al., 2017 and Earm et al., 2014<sup>2,16</sup>.

Until the 1990s, drug discovery and development largely followed a phenotypic or observation-based approach which was quite problematic as it was difficult to predict toxicity or understand the mode of action of a drug<sup>2</sup>. Moreover, before 1985, screening capacity was low and traditional biochemical as well as pharmacological drug discovery methods operated with, by today's standards, large reaction volumes (1 ml) to test individual compounds<sup>3</sup>. Thus, laboratory assay capacity ranged from around 20 to 50 compounds per week and further limitations were imposed by a relatively small compound selection (averaging 3000) which could take 1-2 years to test<sup>3</sup>. The inadequacy of such approaches was further highlighted with the advent of the recombinant DNA technology that expanded new therapeutic target selection and underscored the need to quickly assess a more diverse chemical space<sup>3</sup>. Only after enough biological knowledge accumulated, was there a shift towards target-based approaches where screening became driven by selected targets<sup>2</sup>. Between 1985 and 2000, research pressures and the need for new technological solutions to automate, maximise, and multiplex screening capacity led to the development of HTS<sup>3</sup>. This period represents a fast growth in the screening scalability, cost reduction, and data integration as well as target-centric and toxicology-centric method creation<sup>3</sup> (Fig. 2). This need to innovate also helped to overcome the technical limitations of biology and chemistry research so that more viable targets and candidate compounds could be identified and screened<sup>1,6,8</sup>.



**Figure 2.** Chronological sequence of key breakthrough developments describing the origin and evolution of high-throughput screening (HTS). The timeline depicts the first emergence of important concepts, HTS scaling, and new advances. Other abbreviations: reverse transcription-quantitative polymerase chain reaction (RT-qPCR), liquid chromatography–mass spectrometry (LC-MS), the National Institutes of Health (NIH). Adapted from Pereira & Williams, 2007<sup>3</sup>.

Turn of the century was marked by a rapid adoption of new methods in R&D and the expanding cellular biology space that was studied. In order to appreciate the changing discovery philosophy, it is necessary to consider the technical innovations that created the basis for target-centric approaches. Pre-2000, target selection and validation depended on specifically designed assays to understand what cells express targets, what interactions exist, and what can be therapeutically exploited. Studying differential gene expression was one of the first crucial steps in drug discovery at that time since traditional techniques, such as Northern blot analysis, became gradually complemented and/or replaced by a number of newer methods, e.g., *in situ* hybridisation (ISH)<sup>1,6,8</sup>. Even with technical advancements, these methods were still very labour intensive. This pushed for further development of methods that could be expanded into a multi-assay format. Microarray gridding (GeneChip™) and TaqMan® polymerase chain reaction (PCR) became prominent in the high-throughput analysis of genes. In addition, microarrays, real time (RT)-PCR-based TaqMan assays, as well as Spotfire® data analysis helped to introduce a comprehensive framework where differential expression readouts were integrated and analysed<sup>1–3,7,8,31</sup>. While these methods could not pick-up low abundance genes and suffered from noise, next-generation sequencing (NGS) leveled out the field over the following 20 years with the earlier technologies paving the way for increased research robustness and speed<sup>31–34</sup>. Early in the post-Human Genome Project era, RNA-mediated post-transcriptional gene silencing (e.g., miRNA) has opened new possibilities for gene expression modulation in many organisms and cells<sup>35</sup>. These breakthroughs led to the formation of the functional genomics field that combined physiology and pharmacology allowing the integration of many new experimental approaches, e.g., *in vivo* imaging (i.e., magnetic resonance imaging), mass spectrometry (MS), and microarray hybridisation, to determine particular gene functions<sup>8</sup>. Consequently, around the 2000s, research began to change with growing capabilities to quickly parallelise and process multiple samples or experimental set-ups. As high-throughput drug discovery continued to progress with emphasis on the genome, it also began to expand into proteome and metabolome research<sup>1,8</sup>. This was primarily caused by the realisation that mRNA expression does not necessarily correlate with protein levels<sup>36</sup> and that post-translational modifications or proteins resulting from alternative splicing might have different biological activities<sup>8</sup>. As a result, proteomic and metabolomic analysis permitted the capture of specific

pathological profiles that could be modulated via therapeutic intervention. Even though microarrays dominated the assessment of gene expression via cDNA and RNA analysis, their applications were expanded to include, for example, protein arrays to capture enzyme-substrate, protein-protein, and DNA-protein interactions<sup>37,38</sup>. Similarly, metabolomics integration into discovery and development allowed to characterise new disease markers and metabolic patterns by nuclear magnetic resonance (NMR) spectroscopy, MS, and chromatographic analysis of cell extracts<sup>1,8,39</sup>. Thus, from mid-1995s to 2005, there was a steady replacement of laborious and less optimal methods with a growing HTS dominance allowing the discovery of novel molecular targets<sup>8,40</sup>.

This change in R&D throughput offered many new hit compounds (identified via HTS) and allowed the creation of more focused screening libraries to quickly progress from hit identification to lead generation<sup>8</sup>. Advances in robotics, automation, and data handling allowed applying diverse biochemical assays to large chemical libraries (50,000-100,000 samples in a day)<sup>1</sup>. With developments in ultra-high-throughput screening (UHTS), HTS efforts have shifted into high gear since 2010 as the processing power has increased to 1,000,000 samples a day<sup>1</sup>. Typically, any screening that generates lead compounds takes place in several stages from broad identification to more specialised assays to achieve better pharmacokinetic profiles, such as absorption, distribution, metabolism, and excretion (ADME). Thus, after hit identification and triage, the next step is lead optimisation to reduce the number of potential leads from around 10–15 to 3–4<sup>1,41</sup>. It is important to highlight that this type of research is not linear and various new assays are explored to establish efficacy, bioavailability, as well as interaction characteristics. Similarly, compound modifications are also tested. This might take around 2-3 years since this time is also needed to design the process chemistry (to produce trial batches) and outline potential clinical trials<sup>1,16,41,42</sup>. Once activity characterisation is complete, a lead compound or compounds will move into clinical testing to establish if there is a usable pharmaceutical<sup>2</sup>. As can be seen, crystallisation of drug discovery and development principles required both the technological innovation and the broadening of our scientific understanding. This process is well reflected through the HTS evolution (Fig. 2) since target-centric discovery programs became paramount in the industrial research<sup>2,3,5,31</sup>. Furthermore, as R&D became centred around disease-targeting paradigms, companies tried to counteract market pressures with increased investments in their discovery platforms to produce more new leads<sup>1,2,20,43</sup>.

Yet, the significant technological progress and growing research interdisciplinarity did not result in larger therapeutics outputs. Increasing failure rates in clinical phases and the reduced number of first-in-class targets or compounds highlighted that disease-specific processes are more complex than anticipated<sup>10,13,24,44</sup>. In addition, more than two decades of the target-based approach

did not boost productivity levels in new drug development and many selected targets failed to be druggable<sup>1,2</sup>. Poor disease linkage, off-target effects, and toxicity underscored that biological processes cannot be solely defined by a single gene or a protein<sup>2</sup>. While discovery methods and technology evolved, companies failed to diversify their approaches and look ahead beyond the classical framework of drug searching for a single-target disease which was largely dictated by the existing HTS practices<sup>3,5,24,31</sup>. In other words, the reductionist approach for new drug identification has been dominating R&D pipelines and the complexity of disease biology has just relatively recently forced companies to change such attitudes and embrace systems biology ideas<sup>2,45,46</sup>.

To address R&D challenges, there has been increasingly more reliance on *in silico* approaches to evaluate targets and select the most optimal pharmacological intervention options where integrative and systems biology-based methods began to guide pipeline design and therapeutic decisions<sup>2,31,47-51</sup>. Various computational and biotechnological advances, including next-generation sequencing, transcriptomics, metabolomics, and proteomics, started to be combined using systems biology principles. This new type of biological Big Data integration used to study complex biological interactions is known as ‘*omics*’<sup>2,46</sup>. Thus, ‘*omics*’ represents a new concept in research where the dynamic picture of the pathological mechanisms, genomic variability, pharmacological readouts, and drug screening outcomes become a prerequisite to decrease attrition rates. Here, bioinformatics, cheminformatics, systems biology, and computational biology have become critical in drug discovery and reforming R&D. These computational methods are now employed to provide cost-effective target and drug candidate selection, identify potential toxicity events early in the pipeline, and prepare large-scale information integration for present and future studies<sup>52</sup>. Furthermore, the growing data volumes in screening studies necessitate the development of rational selection and storage methods to organise hundreds of thousands of compounds, their targets, and associated activity readouts<sup>19,53-56</sup>. Similarly, historical data on its own can save resources in future screens by allowing mining of the existing data<sup>2,20</sup>. Bringing computational methods into drug development also helps to limit the use of animal models in pharmacological research and encourages the advancement of alternative high-throughput systems, such as organoids or induced pluripotent stem cell (iPSC) screening assays<sup>2,28</sup>. In other words, transferring aspects of medicinal chemistry and pharmacology into the computational space creates a flexible research environment where *in vitro* and *in vivo* studies can be complemented by *in silico* and fast data integration.

Seeing the obvious benefits of highly integrative computational approaches, many pharmaceutical companies have opted to integrate *in silico* discovery platforms into their pipelines. Such investments are expected to help accelerate therapeutics discovery efforts and improve the overall success<sup>2,52</sup>. However, in order to significantly boost current drug discovery and development strategies, the fast-developing field of computational drug discovery needs structured and well-defined methods to identify promising targets, characterise compound engagement, and store valuable information<sup>20,28,57–60</sup>. Successful drug target identification and prioritisation primarily depend on the establishment of a causal association between a target (or targets) and a specific disease<sup>1,15,24</sup>. Thus, *in silico* methods need to avoid repeating reductionist strategies that are still often seen in classical therapeutics research where the focus is directed towards a very narrow spectrum of targets or a single target is believed to completely alter the disease phenotype<sup>2</sup>. The reality is much more complicated because of complex disease aetiologies and the underlying pathological heterogeneity<sup>2,15,45,46,61,62</sup>. As a result, shifting towards network-centric approaches to better understand pathological perturbations, promoting early and advanced *in silico* screening, as well as the systematic analysis of the selected targets can be invaluable in addressing the current challenges in pharmaceuticals development<sup>2,45,46,63–67</sup>. In other words, the algorithms used in screening should move away from the notion of ‘a single target equals a disease’ to a network-centric approach in order to capture the complexity of the full disease interactome. This appreciation should lead to a better analytical framework for studying cellular perturbations in a pathological state and assessing potential off-target effects<sup>2,45,46,66–71</sup>. Similarly, it is important to consider how a compound interacts with a target or targets by exploring the energetics of the interactions, binding dynamics, and molecular movements. Finally, all this information must be integrated into a format that allows to parse and categorise multiple targets so that valuable insights are not lost<sup>2,24,28,52,72</sup>. There are a growing number of highly integrated research examples where computational methods in drug discovery have already proven their value in urgent situations, such as the COVID-19 pandemic crisis<sup>73</sup>, or the changing landscape of chronic as well as emerging infections, namely tuberculosis and methicillin-resistant *Staphylococcus aureus* (MRSA)<sup>47,74</sup>.

While computational methodologies allow the expansion of the analytical space which undoubtedly helps to select relevant targets and their modulation approaches, further development of methods is necessary to improve today’s *in silico* strategies and create a more regulated research ecosystem<sup>2,28,52,70,75</sup>. Only by focusing on these questions and embracing a holistic discovery approach, can we begin to untangle the key elements in health and disease.

### 1.3. Novel R&D framework for complex diseases: rethinking therapeutics development with case studies on cardiomyopathies and inflammatory disease components

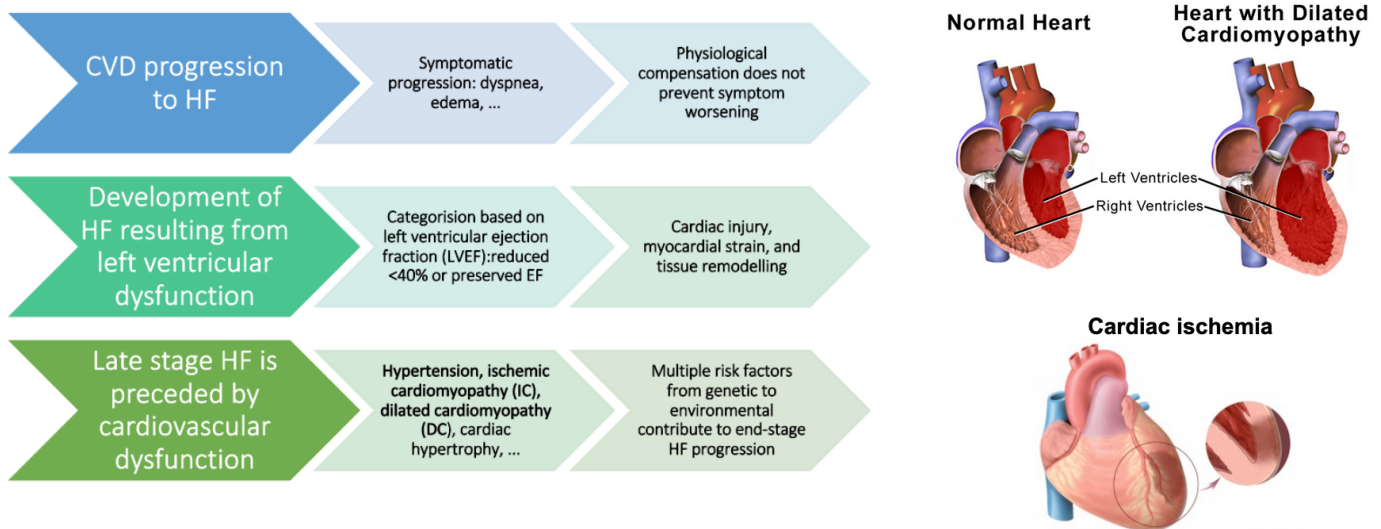
Understanding the full scope of the complex mechanisms underlying diseases is still challenging and a simplistic perspective of ‘one gene-one disease’ has proven to be unsuccessful<sup>2,45,76</sup>. Many diseases result not only from multiple genetic determinants, but also from regulatory and network interactions<sup>46,76</sup>. Thus, the multifactorial nature of many pathologies, such as depression, asthma, epilepsy, diabetes, rheumatoid arthritis, hypertension, or coronary artery disease, earn them a label of ‘complex disease’ where a combination of genetic, regulatory, or even environmental factors can all contribute at a varying degree<sup>77,78</sup>. Such stochastic aspects of diseases created new hurdles for drug discovery programs, especially considering what can be used as a good drug target<sup>2,15,78</sup>. While the human genome contains approximately 25,000 genes, only about a tenth of the expressed proteins are amenable to small-molecule modulation with less than a half of that subset believed to have any therapeutic potential<sup>15,60,61,79</sup>. Since the development of therapeutic compounds has a very low success rate with less than 2% of lead compounds reaching the market, generating effective pharmaceuticals might become especially challenging for immunotherapeutics or other complicated pharmacological categories<sup>1,25</sup>. These difficulties arise because a therapeutic entity can potentially have far reaching side effects through multiple interactions, such as homology-based or unspecific binding and conformation-dependent engagement<sup>61,80</sup>. Considering the growing need for methods to analyse intricate disease interactors and regulatory mechanisms<sup>2,28,45,46,52,53,78</sup>, this thesis will address how we can better investigate the underlying disease mechanisms and potential therapeutic targets when integrating different levels of biological data for both complex diseases and targets.

The first two experimental chapters (Chapter 2: Insights into therapeutic targets and biomarkers using integrated multi-‘*omics*’ approaches for dilated and ischemic cardiomyopathies and Chapter 3: *OmicInt* package: exploring *omics* data and regulatory networks using integrative analyses and machine learning) will introduce new strategies to study complex diseases and identify promising targets. In order to address the lack of integrative and network-centric approaches in R&D, these chapters will focus on the development of an integrative analytical strategy that combines *omics* analyses for robust target classification and assessment with a special focus on complex and unmet need diseases. The following chapters (Chapter 4: Fi-score: a novel approach to characterise protein topology and aid in drug discovery studies and Chapter 5: *Fiscore* package: effective protein structural data visualisation and exploration) will describe newly developed



methods for multiple target investigation in preparation for functional analyses, target validation, and drug screening. In addition, limited options in protein structural and topological exploration as well as machine learning, classification, and relational data storage prompted addressing these questions through a first-of-its-kind scoring function and a user-friendly software package. The final experimental chapter (Chapter 6: *In silico* drug discovery for a complex immunotherapeutic target - human c-Rel protein) will tie together all the pieces of the research presented in this thesis by demonstrating a highly-parallelised and integrative *in silico* screening platform that was developed for accelerated drug discovery. Seeing existing limitations in computational chemistry strategies, such as disjointed analyses, limited analytical protocols, and the lack of solutions for complex targets, encouraged to devise this analysis and screening methodology. Heart failure (HF) and a complex immunological target were selected as case studies to illustrate the present challenges in drug discovery and use these models to formulate potential solutions. Particularly, HF represents a multifactorial disease with the treatment targeting only the symptoms based on the severity of left ventricle dysfunction<sup>81</sup>. The exploration of immunological disease components through complex immune regulators can also help create new and broadly applicable therapeutic strategies<sup>82-86</sup>.

Despite cardiovascular disease (CVD) being the dominating global cause of death, investments and efforts in CVD drug development are declining<sup>81,87</sup>. CVD progresses to a clinical syndrome (or HF) which can be caused by a broad spectrum of diseases affecting the pericardium, endocardium, myocardium, heart valves, and vessels. The underlying structural and/or functional heart dysfunction in HF results in impaired ventricular filling or blood ejection<sup>88</sup>. The statistics for HF are worrying with approximately 2% of the adult population being affected world wide<sup>89</sup>. Since HF is an age-dependent clinical syndrome, fewer than 2% of HF sufferers belong to the population younger than 60 years; however, this proportion increases five-fold for those older than 75 years<sup>89</sup>. Patients with HF usually present with symptoms, such as reduced exercise tolerance, dyspnea, breathlessness, pulmonary crackles, and fluid retention which manifests through pulmonary and peripheral oedema (Fig. 3). Regardless of the physiological compensatory mechanisms in HF, such as increased muscle mass, cardiac filling pressure, and heart rate, this pathophysiological condition progressively worsens<sup>88,89</sup>.



**Figure 3.** Summary for cardiovascular pathology progression with key features described as well as the anatomical feature depiction. The diagram on the left depicts cardiovascular disease (CVD) progression to heart failure (HF) with specific subtypes, classification criteria, causes, symptoms, and main risk factors. Specific disease aspects, such as left ventricular dysfunction and late stage HF, are also summarised. On the right, the dilated cardiomyopathy (DC) and ischemic cardiomyopathy (IC) (or cardiac ischemia) cases are shown next to the normal heart. Clinical illustrations were adapted from [cidg.org.nz](http://cidg.org.nz) and [omicsonline.org](http://omicsonline.org).

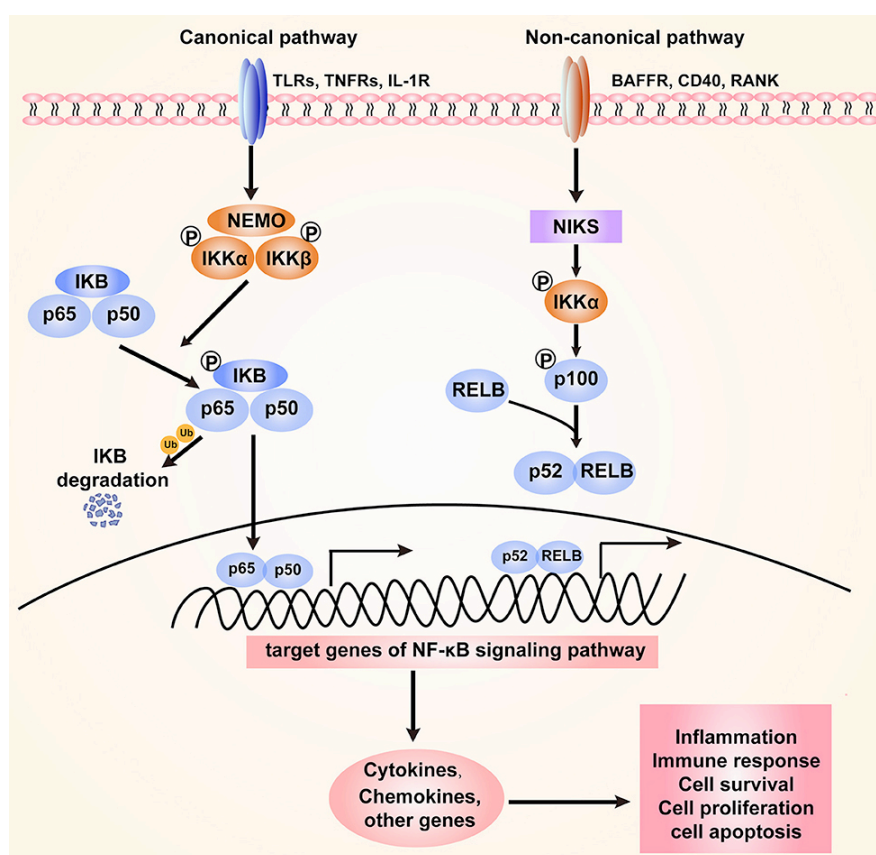
HF resulting from left ventricular dysfunction is further categorised according to left ventricular ejection fraction (LVEF) into HF with reduced ejection fraction (LVEF 40% or less), or HFrEF, and HF with preserved ejection fraction (HFpEF). While the exact definition of HFrEF is known to vary among different guidelines and studies, where LVEF cut-offs can be in a range of  $\leq 30\%$ ,  $\leq 35\%$ , and  $\leq 40\%$ , many clinicians, in their routine practice, would consider  $EF < 40\%$  as a significant systolic dysfunction to warrant the designation of HFrEF<sup>88</sup>. Hypertension, ischemic cardiomyopathy (IC), and dilated cardiomyopathy (DC) precede late-stage HFrEF which encompasses a diverse pathological spectrum<sup>81,88,89</sup>. The progression of HFrEF can be influenced by a number of risk factors resulting in cardiac injury and a subsequent development of myocardial dysfunction. The risk factors are shared with coronary artery disease where obesity, hypertension, hypercholesterolaemia, diabetes, as well as a familial history of HF, or exposure to cardiotoxic agents (e.g., alcohol, amphetamines, and cancer treatment) can promote cardiac injury (Fig. 3). Even though the initial pathophysiological condition is asymptomatic, it gradually worsens leading to end-stage HF<sup>81,88,89</sup>. Moreover, HFrEF exemplifies well how a limited investigation into a complex disease can lead to a long-standing paradigm formation of a single common pathway where only a limited number of genes are used to explain the observed pathological heterogeneity<sup>2,81,90</sup>. Early high-throughput studies did not account for the technical and analytical

limitations and focused on a very generalised explanation for HF, which inevitably led to the oversimplification of the processes at a molecular level<sup>2,91</sup>. With the growing evidence of complex regulatory networks in cardiopathologies<sup>92–94</sup>, it became clear that several genes (as it is the case with a ‘single common pathway’ theory in heart disease) did not provide an adequate measure of pathology development or progression. Furthermore, by simplifying the pathological premise, we lost opportunities to develop new therapeutics as evidenced by the fact that most current therapies for HFrEF did not specifically focus on disease aetiology or in-depth differentiation<sup>81,87,95</sup>. DC, IC, or HFrEF are managed with oral diuretics to treat hypervolemia. Angiotensin I-converting enzyme (ACE) inhibitors, angiotensin II receptor blockers (ARB), and statins have also shown benefit in the treatment of HF<sup>88,96–98</sup>. Thus, the existing symptomatic management highlights why there is a need for new therapeutic insights and why an improved analysis of underlying HF mechanisms is still urgently needed<sup>81,87,91,95</sup>.

Limited research in the therapeutic area of cardiopathologies is typically attributed to the low tolerance for side-effects and a lack of good biomarkers<sup>95</sup>. Thus, a more in-depth understanding of the disease aetiology on a molecular level beyond symptomatic treatment would allow for a better monitoring of the pathology progression, treatment efficacy evaluation, or even the discovery of new therapeutic targets<sup>81,89,95,99</sup>. In addition, the lack of systematic studies to uncover underlying heterogeneous mechanisms on the genomic, transcriptional, and expressed protein scale signifies the need to shift the analytical paradigm towards network-centric and data mining approaches<sup>81,87,91,95,100,101</sup>. A growing number of RNA-seq and metabolomics studies create an excellent resource for an in-depth look into cardiomyopathies where gaps in datasets can be enriched with the information collected from similar studies<sup>81,102,103</sup>. Moreover, multi-omics approaches can help to uncover the intricate biological mechanisms of pathological processes by recreating the complex interactome. These techniques can also be applied to many different indications<sup>2,45,104,105</sup>. Current HF treatment options rely on targeting the symptoms associated with the left ventricular failure without taking into account the heterogeneity of underlying mechanisms<sup>81,87,88</sup>. Due to this lack of therapeutic diversity and the urgent need for improvements in HF treatment, the reported research (Chapter 2) focused on human left ventricular dysfunction (a clinically significant reduction in LVEF) and the development of a new methodology to uncover disease-associated genes.

The study of cardiomyopathy signatures and targets (Chapter 2) hinted at the immune system alterations via the NF- $\kappa$ B pathway (or NF-kappaB, nuclear factor kappa-light-chain-

enhancer of activated B cells, pathway) (Fig. 4). The identified leukocyte migration signatures, such as *CXCL10*, and other inflammatory markers in IC suggested the involvement of major regulatory networks, such as the NF- $\kappa$ B pathway. The NF- $\kappa$ B pathway had already been linked to various hypertrophic, remodelling, and ischemic heart conditions<sup>86,106–108</sup>. Moreover, the exemplified need for novel immunotherapeutics that could be used for HF to improve treatment specificity and tolerability<sup>108–110</sup> motivated to focus on the immune system and the exploration of potential pharmacological strategies for relevant target modulation. As a result, the NF- $\kappa$ B pathway was selected as an excellent opportunity to develop discovery pipelines because of a significant unmet need for drugs that could effectively target this complex (Chapter 6)<sup>86,111</sup>. Importantly, formulated pipelines and methods can be widely applied to other problematic targets.



**Figure 4.** Simplified schematic representation of NF- $\kappa$ B signalling showing the canonical and non-canonical pathways; the illustration was adapted from Peng et al., 2020<sup>112</sup>. NF- $\kappa$ B hetero- or homo-dimers are formed by the Rel transcription factor family members: p50, p52, Rel A (p65), Rel B, and c-Rel. The canonical pathway (p65/p50) is inducible through TLRs, TNFRs, and IL-1R leading to the phosphorylation and degradation of the inhibitory protein I $\kappa$ B. This occurs primarily via the activation of the I $\kappa$ B kinase (IKK). IKK is composed of the catalytic IKK $\alpha$  and IKK $\beta$  subunits and a regulatory protein termed NEMO (NF- $\kappa$ B essential modulator) or IKK $\gamma$ . After NF- $\kappa$ B is released from the I $\kappa$ B-containing complex the activated NF- $\kappa$ B complex translocates into the nucleus. The non-canonical pathway (p52/RelB) is activated by BAFFR, CD40, and RANK. This cascade results in the phosphorylation of the NF- $\kappa$ B inducing kinase (NIK) and IKK $\alpha$ . This is followed by the translocation of the activated p52-RelB heterodimer into the nucleus. NF- $\kappa$ B signalling regulates various cellular processes that may involve inflammation, apoptosis, and immune response.

The NF- $\kappa$ B pathway illustrates well how far reaching immune-modulatory effects can be and why creating better immunotherapeutics can have a significant impact in many pathologies. NF- $\kappa$ B encompasses a broad spectrum of activities realised through the regulation of key genes in pro-survival and pro-apoptotic pathways (Fig. 4). NF- $\kappa$ B hetero- or homo-dimers are formed by the Rel transcription factor family members: p50, p52, Rel A (p65), Rel B, and c-Rel. It is important to note that due to post-translational processing the p50 and p52 proteins have no intrinsic ability to activate transcription as they lack the C-terminal transactivation domain in contrast to the other family members<sup>113,114</sup>. This multimeric transcription factor is regulated through the binding of  $\kappa$ B inhibitor proteins, which are subjected to proteosomal degradation after the activation of the I $\kappa$ B kinase complex (IKK) leading to the release of NF- $\kappa$ B<sup>111-114</sup>. The complexity of this master gene regulator lies in the fact that different multimer compositions exist and that NF- $\kappa$ B can be activated either through the canonical or non-canonical pathway. The canonical pathway (p50 and p65 or p50 and c-Rel heterodimers) controls multiple cellular functions including immune system activation and cellular survival, while the non-canonical pathway (mostly p52-RelB) is primarily involved in lymphoid organogenesis (Fig. 4)<sup>114-116</sup>. In addition, the NF- $\kappa$ B complexes containing either p65 or c-Rel are known to be involved in distinct biological roles, where multimers with p65 maintain cellular metabolism and inflammatory response regulation and c-Rel containing transcription factors play a role in a more specialised immune response and lymphoid development<sup>117,118</sup>. Even though NF- $\kappa$ B is at the nexus of multiple regulatory pathways and metabolic processes, so far no significant therapeutic advancements have been achieved to offer optimal pharmacological engagement<sup>119-121</sup>. Difficulties in establishing good drug candidates for NF- $\kappa$ B might be linked to the ubiquitous expression of NF- $\kappa$ B in multiple tissues, complex interaction dynamics, and a lack of understanding regarding various oligomer functions<sup>114-116,120,122</sup>. Nevertheless, these aspects of NF- $\kappa$ B signalling can be exploited to advance our drug discovery efforts<sup>114,115,123</sup>. As NF- $\kappa$ B is assembled from different dimers that vary between tissues and pathologies, this signalling feature can be strategically capitalised on to increase specificity and reduce off-target effects. The efficacy of this approach has recently been demonstrated through the inhibition of c-Rel function to delay melanoma growth by impairing effector Treg-mediated immunosuppression<sup>121,124</sup>. Furthermore, as we begin to better understand NF- $\kappa$ B function, it becomes apparent that dimer-forming proteins are not equivalent and possess different characteristics which can be utilised to accommodate new drug design<sup>124</sup>. These observations are especially relevant for cardiopathologies because NF- $\kappa$ B activity is also increased in such states<sup>106,107,123,125,126</sup> and there is evidence that the c-Rel subunit stimulates cardiac hypertrophy and fibrosis<sup>123</sup>. Gaspar-Pereira and colleagues demonstrated that c-Rel-

deficient mice have smaller hearts and do not develop cardiac hypertrophy and fibrosis during chronic angiotensin II infusion. The authors also reported for the first time that c-Rel is highly expressed and localised in the nuclei of diseased adult human hearts, whereas in normal hearts c-Rel was restricted to the cytoplasm<sup>123</sup>. Other studies have also hinted at the complex regulatory network of NF- $\kappa$ B showing that transcriptional regulation can have far reaching effects, including the promotion of global changes in the chromatin landscape to control cellular calcium regulating genes and cardiac function<sup>106</sup>. The growing evidence strongly suggests that NF- $\kappa$ B plays a role in heart disease, where the development and progression of inflammation and cardiac as well as vascular damage seem to be orchestrated by this transcription factor<sup>127,128</sup>. Furthermore, the reported c-Rel-dependent signalling in cardiac remodelling and hypertrophy presents an interesting opportunity to explore a novel therapeutic strategy that could be expanded to other diseases with abnormal tissue growth, such as cancer and fibrosis. c-Rel can be used as a target model and the developed screening blueprint can be directly applied to any Rel family member if, for example, another subunit of NF- $\kappa$ B needed to be targeted to achieve a therapeutic effect. As a result, c-Rel, a promising target in many human inflammatory and oncological pathologies, was selected as a case study for the development of an effective *in silico* screening platform since the NF- $\kappa$ B transcription factor currently has no successful therapeutic inhibitors or modulators (Chapter 6). In addition, the established analytical and screening pipeline can be transferred and adapted to any therapeutically relevant target.

Cardiovascular diseases underpin the development of HF and are a leading cause of death worldwide; thus, there is an undeniable need to rethink therapeutic protocols and search for novel treatment options<sup>81,88,89,123,126-129</sup>. In order to formulate a novel discovery framework for complex diseases, cardiomyopathies and an inflammatory component/target were selected as case studies to develop and test new methodologies. As demonstrated in the present thesis (Chapters 2 and 3), employing multi-*omics* centred approaches allows to explore multifactorial diseases in-depth and identify new clinical avenues. In the case of cardiomyopathies, the introduced integrative strategy enabled capturing a subtle differentiation between ischemic and hypertrophic states. Moreover, recent reports suggesting a strong involvement of NF- $\kappa$ B in cardiac remodelling<sup>123,126-130</sup> also motivated to find new disease targeting strategies to offer better clinical management options for patients (Chapter 6). The reported studies bridged multi-*omics*, computational biology, structural bioinformatics, as well as computational chemistry and helped to create an adaptable premise for future research since developed methodologies are robust and widely transferable.

#### **1.4. Development of a network-centric and highly integrative discovery process: addressing R&D challenges and creating new opportunities**

The growing research and commercial pressures for novel therapeutics accentuate why better strategies are needed for R&D and drug discovery<sup>2,11,13,17,18</sup>. The costly nature of developing a therapeutic compound as well as the shrinking pool of ‘easy’ targets are some of the key reasons why pharmaceuticals companies, research institutions, and researchers are shifting their focus towards integrative and systems biology driven approaches<sup>10,19,28,45,46,67,131</sup>. Moreover, multifactorial aspects of many diseases require more innovative treatment solutions rather than just focusing on a single target<sup>2,46,76,78</sup>. CVD as well as HF associated immune components demonstrate well how discerning network elements that contribute to a pathology might expedite the creation of better therapeutic solutions for patients<sup>86,89,95,101</sup>. As a result, to address major challenges in drug discovery, this thesis aimed to introduce a gradual and highly integrative analytical framework by incorporating a full range of studies from disease target selection to high-throughput virtual screening (HTVS) so that a cost-effective and efficient stratification of targets and associated compounds could be achieved.

Specifically, it was first necessary to develop a multi-*omics* based process to capture complex gene interaction patterns, establish disease association parameters, identify gene clusters of interest for the downstream analysis, and subsequently determine key interactors that could be used to build pathway maps (Chapters 2 and 3). In addition, creating a first-of-its-kind protein topology and conformational analysis function allowed not only to classify but also to identify therapeutically relevant features of selected targets for the downstream druggability analysis (Chapters 4 and 5). All this concluded with a demonstration of how existing drug discovery pipelines for *in silico* screening can be further improved with the expansion of compound screening strategies (Chapter 6). That is, the unification of molecular dynamics, modelling, topology, and physicochemical analyses provided solutions for challenging target investigation and led to the identification of potential therapeutic modulators. Thus, the outlined comprehensive and highly integrative analytical framework which builds on the network-centric and systems biology ideas offers new strategies for accelerating drug discovery and significantly reducing research costs and turnaround time.

In order to establish a network-centric premise for druggable target identification, it was necessary to build an integrative investigation framework. The developed methods will be introduced in the first experimental chapter of the thesis (Chapter 2) which will focus on the study

that showed for the first time how bulk and single cell RNA-sequencing as well as the proteomics analysis of the human heart tissue can be integrated to uncover specific networks. The explored HF regulome indicated how potential therapeutic targets or biomarkers can be studied for this multifactorial cardiac syndrome with limited therapeutic options<sup>81,87-89,95</sup>. Existing challenges in the *in silico* pharmacology field and the already laid out analytical groundwork in multi-network analyses also motivated to devise a highly integrative network-centric approach which could be used to build complex interaction pathways and extract information for shared expression patterns (Fig. 5). Moreover, the method introduced in this thesis is highly adaptable, which allows for further development as more data and algorithms become available.

Multi-omics research, that will be discussed in chapters 2 and 3, built on network pharmacology and systems biology where the disease causality is primarily believed to result from multiple players distributed unevenly throughout the transcriptome, expressome, and regulome<sup>45,67,132</sup>. Thus, phenotype modifications depend on a simultaneous modulation of multiple network nodes as outlined by the network biology theory<sup>2,45,105,132</sup>. The observed phenotypic robustness after gene deletion further confirms that polypharmacological modulation might be more successful than a highly selective drug engaging a single target since a disease phenotype is dependent on multiple genetic factors<sup>132-134</sup>. These observations were also supported by the network analysis studies where the exploration of links between drugs and drug targets revealed rich networks of polypharmacological interactions<sup>20,132-134</sup>. Moreover, similar systemic studies unveiled interesting patterns where drug targets were positioned between proteins that have overall more interactions than an average protein but less network connections than essential proteins<sup>132</sup>. Together, these findings encouraged researchers to consider how drug targets can be identified based on their position in the interactome<sup>20,132-135</sup>. These insights, however, also presented a challenge as merely mapping targets based on screening studies is unlikely to help recover more complex targets. Limitations in unifying the genomic, transcriptomic, proteomic, and metabolomic data can be traced back not only to analytical shortcomings, but also to the limited availability of high-quality data<sup>136,137</sup>. In other words, while the recent advances in multi-omics data acquisition introduced new platforms for studying complex diseases, comprehensive methods for multi-dimensional readout integration are still lacking<sup>138</sup>. This was a key motivator for chapters 2 and 3 to not only integrate existing methods, but also expand and improve currently employed techniques.

To better understand the analytical premise and challenges in network pharmacology and systems biology, it is essential to examine the fundamental analytical methodologies employed in these fields. Methods used to reconstruct the underlying relationships and dependencies from the

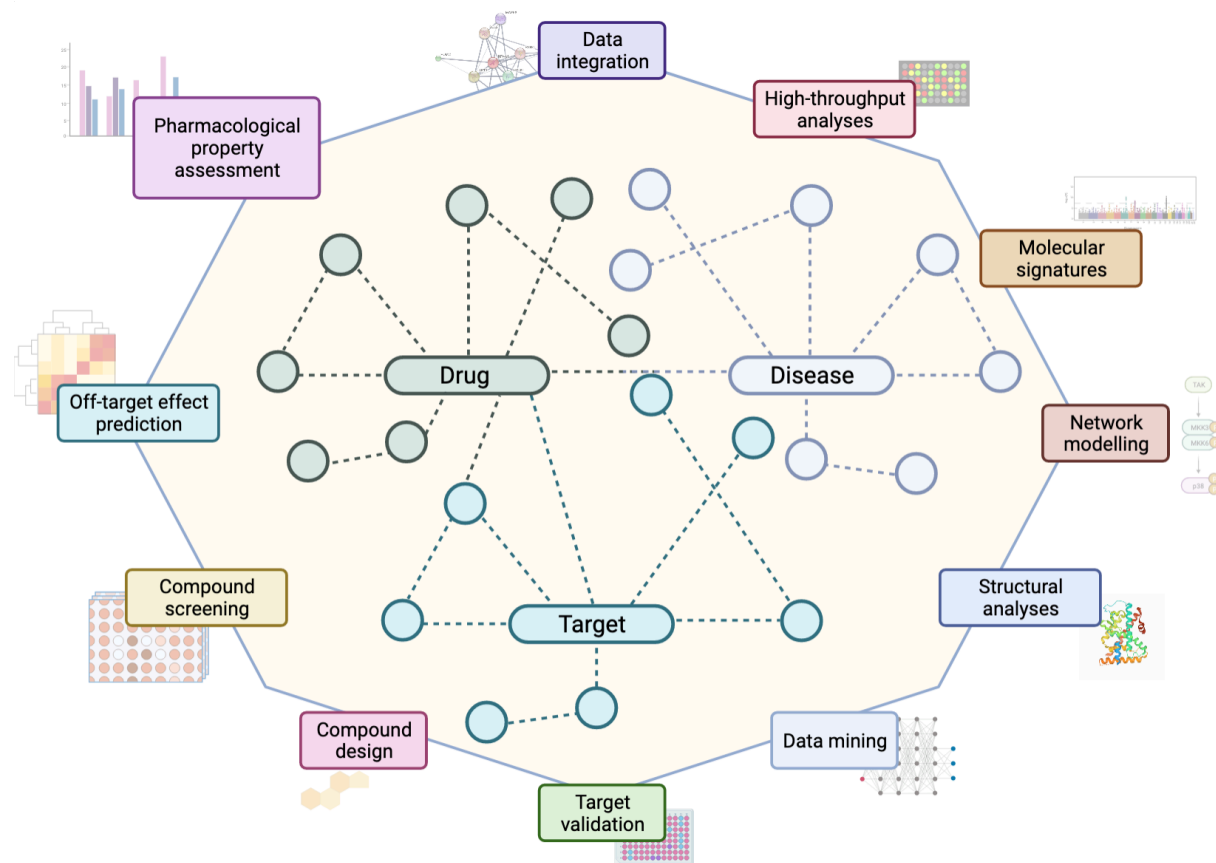


observed data range from relatively simple estimations using correlation or mutual information to probabilistic graph models, such as Bayesian network inference, and machine learning<sup>54,138,139</sup>. Furthermore, pathway and network analyses are the most common methods currently employed to assess cellular perturbation events where data might be generated from transcriptome studies; these studies represent high-level analyses aimed to elucidate disease associated processes (Fig. 5)<sup>138,140</sup>. Typically, these analytical techniques, also known as pathway enrichment analyses, can be split into several categories based on the underlying analytical principles: over-representation analysis, rank-based approaches, and topology-based methods<sup>138</sup>. Over-representation analysis is often listed as a first-generation approach and it is still widely used for various analyses because of its statistical simplicity achieved through hypergeometric distribution, chi-square, or Fisher's exact statistics<sup>138</sup>. However, at the same time, the method suffers from the assumed significance for all inputs, need for arbitrary thresholds that might not be optimal, and a considerable variation in significance<sup>138,141</sup>. In contrast, network enrichment methods using rank-based evaluation can account for the over-representation method limitations by including significance parameters in the calculations. Despite these improvements, this method is susceptible to the effects of a few highly significant markers and it also depends on the statistical analysis applied<sup>138,140-142</sup>. One of the more recent analytical techniques relies on the topology-based assessment where the pathway structure is an important component for the analysis. An example of this methodology is the EnrichNet tool which uses an enrichment score for every pathway via the estimated distance of that pathway to all other candidate genes in the network<sup>138,141</sup>. Overall, a shared shortcoming of all these techniques is the dependency on gene annotations to establish relevant associations where the available information can be influenced by the curation quality<sup>138</sup>. All this calls for new methodology that could help integrate several *omics* layers and subsequently incorporate the derived scores into machine learning or other classification pipelines.

To address this need, a highly integrative network-centric approach was developed (Chapters 2 and 3). One of the principal elements of this method is the determination of differentially expressed genes based on the negative binomial distribution to balance between the detection sensitivity and specificity<sup>143</sup>. Significantly changed genes in a selected condition are scaled based on their links to a specific disease where the association data is retrieved from database and text mining. Target-disease association consolidation was primarily calculated using the harmonic sum of scores dependent on data sources as previously described<sup>144,145</sup>. If *omics* datasets, such as protein expression levels or single cell expression data, are available for a specific study, they can also be incorporated to add additional weights to the score. Similarly, gene set

enrichment analysis (GSEA) was used as an intermediate quality control step to determine if any of the significantly changed genes show phenotypic or functional enrichment. GSEA estimates this by using ranked genes where the enrichment score is generated through a random walk using the weighted Kolmogorov-Smirnov-like statistic<sup>138,142</sup>. For each established gene that was significantly changed and potentially had a known association to a disease an interactor network was assigned. The interactor network was derived by retrieving threshold-regulated data from the STRING database that contains information on known protein interactions, indirect associations, as well as predicted links between proteins<sup>146,147</sup>. This new information layer allowed the integration of relevant data points, specifically the transcriptome, proteome, as well as regulome (Fig. 5). To identify meaningful clusters based on the network complexity (i.e., how many interactors a gene-protein is expected to have) and the adjusted expression score, Gaussian mixture models (GMMs) were selected as a primary machine learning classifier. The strength of GMMs lies in the probabilistic model nature where all data points are assumed to be derived from a mixture of a finite number of Gaussian distributions with unknown parameters<sup>148</sup>. It becomes evident that the soft classification of GMMs where a data point has a probability of belonging to a cluster is much more suitable to assess biological parameters compared to other hard classification techniques in machine learning, such as k-means, which provide a strict separation between classes. In other words, GMM clustered genes have a degree of membership for every specific category which could be especially helpful when using the derived probability values with downstream analyses or other machine learning techniques to find pathway convergence points as well as elements belonging to several networks/regulatory systems. Incorporation of information criterion (i.e., evaluating the quality of a statistical model for a given dataset) in model building also allows to fine tune the expected number of clusters. In addition, GMM in combination with the expectation-maximisation algorithm models parameters to maximise the likelihood of data point assignments<sup>148-151</sup>. Overall, the devised method provides a means to connect the expression, disease association, and network complexity values. Depending on the model used for the differential gene expression analysis and inclusion of additional weights (e.g., single cell readouts), the weighted expression scoring might provide a way to probabilistically differentiate gene expression values if, for example, an identified gene has a strong disease association. Moreover, this method can help evaluate what clusters are formed based on the local interactome for genes that changed significantly to become either upregulated or downregulated. This classification approach can link disease-associated genes with new candidates and help establish seed points around which a relevant pathway can be recreated. The developed

methodology could be particularly useful for target selection and evaluation during the preclinical development stage.



**Figure 5.** A schematic view of the integrative drug discovery process where different *omics* analyses are merged to establish disease, drug, and target links. The graph represents the considerations for the proposed integrative drug-discovery approach with different branches interlinked to capture relevant multi-*omics* aspects from target selection to pharmacological assessment.

### 1.5. Biophysical and computational chemistry method development: streamlining complex target evaluation and therapeutics discovery

Since target evaluation and rational drug design rely on identifying and characterising small-molecule binding sites on therapeutically relevant target proteins, developing a discovery process that incorporates both structural biology and computational chemistry becomes essential for the success of therapeutics screening<sup>15,152–154</sup>. In order to develop analytical solutions for target identification and successful screening implementation, this dissertation also introduced a newly developed method to investigate structural features and protein topology (Chapters 4 and 5). The presented approach can help to categorise multiple targets and extract core structural characteristics

during the pre-screening stage so that proteins of interest can be included in relational databases for a quick retrieval. Development of such approaches is critical for early research and discovery in a clinical pipeline as it is often possible to generate multiple potential targets that later need to be screened<sup>60,155,156</sup>. This was also reflected in the case study of HF (Chapter 2) showing that disease pathway and interactor investigation can generate a diverse set of therapeutic candidates. Such targets of interest would typically need to be further hierarchically ordered and prioritised based on their structural characteristics to facilitate the downstream screening and compound-based assays. Moreover, large compound library testing against a therapeutically relevant target poses a challenge of storing and keeping track of all the relevant readouts<sup>53,80,157</sup>. It is necessary not only to maintain the information of the physicochemical compound parameters or biochemical assay outputs but also to efficiently capture the key topological features for easier bi-directional clustering using compound and target information<sup>49,131,158</sup>. Seeing the existing limitations of structure-based data collection in the industry, one of the aims of the thesis was to address this need by introducing a topology and structure driven target categorisation (Chapters 4 and 5) that could be easily supplied to screening, data storage, or machine learning pipelines<sup>13,159,160</sup>.

To aid with pre-screening and screening preparation, a method was developed to classify multiple regions of interest within a target. It was hypothesised that having such information prior to the screening would enable the comparison and grouping of relevant topological characteristics. Such a classification system could be used to compare newly identified target proteins with a reference set of binding sites. If reference sites contained the information of known binders, then target biomolecules could be further classified based on the compound properties and identified pockets. In other words, this type of scoring provides an opportunity to easily integrate topological features of new proteins into relational databases<sup>68,161,162</sup>. Furthermore, in some instances a binding site might be conserved and it could be helpful to compare protein regions of interest across homologous and non-homologous protein sets<sup>163,164</sup>. Specifically, a topology-based scoring method could give insights into the conformation and not just the amino acid composition<sup>165-169</sup>. Finally, characterising protein sites through scoring could be used to compare proteins that have known drug binders with a newly identified target which has no known compounds. The described approach could be particularly useful in drug repurposing because protein sites that share similar characteristics could be used to infer drug binding in a new site based on already explored one<sup>55,69,70,105,170</sup>. Therefore, the established methodology to classify sites of interest could be extremely helpful in solving data organisation questions, reducing screening time and costs, as well as helping to achieve a faster turnaround<sup>28,49,50</sup>.

Target pre-screening and evaluation primarily depend on establishing protein-ligand interactions which are exploited by most of the currently marketed small-molecule drugs and such interaction information is typically based on the crystallographic analysis<sup>154</sup>. Thus, computational modelling primarily uses X-Ray-based data to evaluate energetics, cavity geometry, and physicochemical properties of a potential binding pocket<sup>171</sup>. Despite the growing number of computational chemistry tools, there is not one universal algorithm developed to incorporate sequence, structural, and conformational features that could be used for comparative studies<sup>28,172</sup>. As a result, combining multiple levels of analysis to capture the key structural features, such as B-factor values and dihedral angles, enabled establishing a comparative measure for physicochemical and spatial characteristics of a protein of interest. The established parameter could be used to analyse a single motif, binding site, or the whole protein. The usefulness of dihedral angle and B-factor values can be appreciated when considering the high information content that they provide. Specifically, a dihedral angle is the angle between two intersecting planes or half-planes and in the case of a protein this geometric representation is the internal angle of polypeptide backbone at which two adjacent planes meet<sup>173,174</sup>. Two dihedral angles per residue ( $\phi$ : C-N-C $\alpha$ -C, and  $\psi$ : N-C $\alpha$ -C-N) can be used to describe the conformation of the backbone since the polypeptide chain is locked between a pair of juxtaposing C $\alpha$  atoms in a single plane<sup>173,175</sup>. Consequently, protein dihedral angles contain the information on the local and global protein conformation as well as backbone restraints that result from the sequence composition<sup>175,176</sup>. B-factors, or oscillation amplitudes of the atoms around their equilibrium positions in the crystal structures, capture a decrease in the intensity of X-Ray diffraction because of the static as well as dynamic disorder where the latter is caused by the temperature-dependent atom vibrations. It has been shown that this parameter provides many additional layers of information, such as thermal motion paths, protein superimposition, packing, flexibility, and allows predicting the rotameric state of amino acids side-chains<sup>177-183</sup>. Considering the above, it becomes apparent that B-factors carry a lot of information on the complex intramolecular relationships. By incorporating B-factor estimates with protein dihedral angle values, we can capture both the local and global mobility of C $\alpha$  atoms as well as side chain influences. These observations led to the derivation of the Fi-score that fingerprints physicochemical and topological qualities of a region of interest taking into account conformation dependencies. Moreover, the scoring of any site can be subsequently visualised via distribution plots, 3D region visualisation, or integrated into machine learning to derive probability density distributions based on physicochemical properties.

A streamlined analytical process from therapeutically promising target identification to a detailed characterisation can enable the creation of a discovery platform that connects the information derived from biological assays and other studies with pharmacologically relevant compounds (Chapter 6). The identified hits can be subsequently improved in hit-to-lead phase to ensure optimal bioavailability and toxicity profiles<sup>152</sup>. In order to develop a holistic screening framework, the NF- $\kappa$ B pathway served as a model since it was implicated in the introduced HF study (Chapter 2) and has known links to cardiomyopathies<sup>116,123,126-130,184,185</sup>. Specifically, the c-Rel protein, as a complex immunotherapeutic target, was selected to model the screening pipeline (Chapter 6). Prior to the study reported in this thesis (Chapter 6), there were no in-depth reports of c-Rel structure models, interactions, or physicochemical analyses aside from the insights generated through X-Ray crystallography or sequence analysis studies<sup>186,187</sup>. As a result, an exhaustive computational analysis of likely and/or unusual binding sites in this target protein was performed to reveal therapeutically relevant characteristics (Chapter 6).

The cheminformatics and structural bioinformatics toolbox provides multiple methods to explore targets of interest from sequence based analysis to complex molecular modelling that can unveil important information about what structural elements could be susceptible to pharmacological modulation<sup>28,159,188-193</sup>. Broadly, the computer-assisted chemistry methods integrate ligand- and structure-based drug design strategies. Structure-based drug design relies on homology modelling, molecular dynamics, molecular docking, and structure-based virtual screening to evaluate potential ligand-target interactions. Ligand-based drug design focuses on pharmacophore modelling (i.e., abstractions of important molecular features), quantitative structure-activity relationships (QSAR), and ligand-based virtual screening to explore molecule databases where the focus of the analysis is to establish correlations between chemical features and pharmacological activity<sup>53,194,195</sup>. In the case of the c-Rel protein, the analytical process began with structure-based drug design where a focused analysis allowed to evaluate the physicochemical properties and determine potentially druggable sites. Using various molecular dynamics set-ups, comparative analyses, protein structure modelling, as well as GMMs<sup>148</sup> for Fi-scores enabled a computational characterisation of this NF- $\kappa$ B subunit. These techniques also helped to address the common issue when the crystal structures do not reflect protein native conformations or when a target does not have a good structure to analyse<sup>168,196</sup>. Normal mode analysis was employed to model the conformational changes in c-Rel since this method provides a fast and simple calculation of vibrational modes and protein flexibility. That is, atoms in a protein (or sometimes C $\alpha$  only) are

modelled as point masses connected by springs representing the interatomic force fields and this implementation (with possible variations in model types) is used to predict molecular motions<sup>197–199</sup>. In addition, a more in-depth molecular dynamics simulation was performed using GROMACS software tools where a selection of force fields, solvation models, temperature gradients, and other restrictions were customised to capture more intricate movements within the protein<sup>193</sup>.

Furthermore, ligand-based drug discovery was also utilised to select compounds from a large compound library (659 M drug candidates) and refine this diverse set of compounds based on their physicochemical features (34 M). This was achieved employing compound fingerprinting and classification where small molecules in a matrix-like representation were encoded with a fingerprint of the same type and length to create a searchable database of compound topological features<sup>158,200</sup>. All these analyses created a premise for a highly integrative screening platform conceptualisation (Fig. 5).

In order to create an analytical pipeline for binding site selection, compound docking, and interaction evaluation, computational chemistry analyses were done using Schrödinger cheminformatics suite<sup>201</sup>. This cheminformatics software offers the full range of HTVS options to screen hundreds of thousands of ligands and achieve higher enrichment of hits through GlideScore. Schrödinger's empirical scoring function is designed to maximise discovery of strong binders since GlideScore accounts for the physics of the binding process using multiple parameters, including a lipophilic-lipophilic term, hydrogen bond terms, a rotatable bond penalty, and contributions from protein-ligand Coulomb-vdW energies. In addition, GlideScore takes into account hydrophobic enclosure which is the displacement of water molecules by a ligand<sup>201</sup>. To accommodate the screening of an unprecedented library size<sup>80,157</sup>, a hierarchical *in silico* high-throughput screening was combined with the binding site selection, similar target analysis (e.g., p65<sup>111–118</sup>), and structural characterisation. This parallelisation led to the discovery of 15 hit compounds specific for the human c-Rel protein as well as the identification of potential drug-protein interaction mechanisms. Specifically, compound binding poses and protein subdomain movements were assessed using cutting-edge molecular dynamics methods to explore a wider spectrum of interactions. This strategy permitted to identify hit compounds and infer potential action mechanisms (e.g., disorder induced degradation). In addition, the inclusion of other homologous target screening data could be employed to develop multi-target approaches where a compound modulates several targets at a varying degree (this was explored as a control step with p65 that has high similarity with c-Rel)<sup>68,70,76,116–121,124,132,134,202</sup>. Moreover, the first in-depth structural modelling exploration of the

c-Rel subunit offered hints at how highly dynamically this protein might engage its target DNA. The hit compounds were additionally tested with a different docking and compound binding evaluation program/algorithms – Autodock Vina<sup>203</sup> and yielded similar results. The generated compounds and new target-ligand insights pave the way for the future development of highly selective human c-Rel inhibitors and/or modulators where therapeutics with novel action mechanisms could provide better options for pharmacological intervention in diseases, such as cardiomyopathies, since the current treatment is primarily based on the symptomatic management<sup>81,88,89,98,204</sup>. Broader applicability of this study also enables focusing not only on the druggable genome, but also on new target classes or polypharmacological approaches (i.e., working with complex targets).

Overall, creating a framework for a highly integrative target assessment and therapeutics development allowed highlighting that none of the R&D stages can be treated as separate entities but rather one step needs to inform the other<sup>2,28,45,202</sup>. To account for high attrition rates and the growing need to tackle complex targets, it is paramount to rethink present strategies and embrace holistic adaptable methods<sup>17,18,56,205,206</sup>. Moreover, the introduced analytical framework together with the screening pipeline showcases the potential of new network-centric methods where targets are seen as a part of the complex interactome with a multi-modulation potential<sup>2,105,202</sup>.



## **Integrative *omics* approaches for new target identification and therapeutics development**

### **2. Insights into therapeutic targets and biomarkers using integrated multi-‘*omics*’ approaches for dilated and ischemic cardiomyopathies**

**The experimental chapter is based on the following publication**

Kanapeckaitė A, Burokienė N. Insights into therapeutic targets and biomarkers using integrated multi-‘*omics*’ approaches for dilated and ischemic cardiomyopathies. Integrative Biology. 2021 May;13(5):121-37; doi: 10.1093/intbio/zyab007. PMID: 33969404.\*

\* The publisher’s error resulted in swapped Figures 2 and 3. The error has been reported and is being addressed by the publisher.

#### **Conclusion of this chapter**

Current strategies to treat heart failure mainly target symptoms based on the left ventricle dysfunction severity. There is a notable lack of systemic ‘*omics*’ studies for an in-depth analysis of heterogeneous disease mechanisms. This study, for the first time, demonstrated how bulk and single cell RNA-seq as well as the proteomics analysis of the human heart tissue can be integrated to uncover HF-specific networks and potential therapeutic targets or biomarkers for dilated and ischemic cardiomyopathies. Thus, my analysis allowed to reveal that despite a smaller number of samples which is often the case in some preclinical settings or smaller-scale studies, it is possible to discover new therapeutically relevant insights. Moreover, by devising a novel scoring system and applying machine learning methods, I was able to derive a method to untangle complex expression profiles to elucidate gene clusters that can be selected for downstream analyses. This study could be the first step towards a more systematic analysis that could be freely shared among researchers. Finally, my work helped to demonstrate that cardiopathology treatment can go beyond symptom management and that there are indeed distinct gene network and pathway profiles that could be of therapeutic interest.

#### **Contribution to this chapter (95%)**

- Methodology development which included equation and scoring function derivation as well as machine learning pipeline creation.
- Performed all the analytical, data mining, and experimental work as well as formulated conclusions.
- Conceptualised and wrote the manuscript, including the figure preparation.
- Corresponding author.

## ORIGINAL ARTICLE

# Insights into therapeutic targets and biomarkers using integrated multi-‘omics’ approaches for dilated and ischemic cardiomyopathies

Austė Kanapeckaitė<sup>1,\*</sup>, and Neringa Burokienė<sup>2</sup>

<sup>1</sup>Algorithm379, Laisvės g. 7, Vilnius LT-12007, Lithuania, and <sup>2</sup>Clinics of Internal Diseases, Family Medicine and Oncology, Institute of Clinical Medicine, Faculty of Medicine, Vilnius University, M. K. Čiurlionio str. 21/27, LT-03101 Vilnius, Lithuania

\*Corresponding author. E-mail: info@algorithm379.com

## Abstract

At present, heart failure (HF) treatment only targets the symptoms based on the left ventricle dysfunction severity; however, the lack of systemic ‘omics’ studies and available biological data to uncover the heterogeneous underlying mechanisms signifies the need to shift the analytical paradigm towards network-centric and data mining approaches. This study, for the first time, aimed to investigate how bulk and single cell RNA-sequencing as well as the proteomics analysis of the human heart tissue can be integrated to uncover HF-specific networks and potential therapeutic targets or biomarkers. We also aimed to address the issue of dealing with a limited number of samples and to show how appropriate statistical models, enrichment with other datasets as well as machine learning-guided analysis can aid in such cases. Furthermore, we elucidated specific gene expression profiles using transcriptomic and mined data from public databases. This was achieved using the two-step machine learning algorithm to predict the likelihood of the therapeutic target or biomarker tractability based on a novel scoring system, which has also been introduced in this study. The described methodology could be very useful for the target or biomarker selection and evaluation during the pre-clinical therapeutics development stage as well as disease progression monitoring. In addition, the present study sheds new light into the complex aetiology of HF, differentiating between subtle changes in dilated cardiomyopathies (DCs) and ischemic cardiomyopathies (ICs) on the single cell, proteome and whole transcriptome level, demonstrating that HF might be dependent on the involvement of not only the cardiomyocytes but also on other cell populations. Identified tissue remodelling and inflammatory processes can be beneficial when selecting targeted pharmacological management for DCs or ICs, respectively.

**Key words:** target identification; biomarker discovery; dilated cardiomyopathy; ischemic cardiomyopathy; omics data integration; machine learning for target prediction

## INSIGHT BOX

First report of an integrated multi-omics analysis for DCs and ICs led to the identification of metabolic and regulatory network differences for two types of cardiomyopathies. These findings revealed new therapeutic opportunities as well as highlighted the need to focus on genetic networks in disease development. To achieve this, a new scoring system was introduced allowing to evaluate genes/biomarkers based on the size of their network and disease association. Two-step machine learning pipeline employed the scoring system to uncover the potential therapeutic target clusters. Gained insights can be easily extended to other studies to take advantage of multi-omics approaches in therapeutic target investigation.

Received January 5, 2021; revised January 20, 2021; accepted April 7, 2021

© The Author(s) 2021. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

## INTRODUCTION

Cardiovascular disease (CVD) is the leading cause of death globally; however, both investment and efforts in CVD drug development are declining. This contrasts sharply with funding and drug approvals for other indications, such as oncology [1, 2]. While there are many factors contributing to this trend, low tolerance for side effects and lack of good biomarkers are some of the key challenges in implementing new therapies [2]. Thus, all of these call to revisit currently used approaches in the therapy development for CVD. Specifically, combining high-throughput RNA-sequencing (RNA-seq), proteome analysis and biological data mining could potentially facilitate the identification of new therapeutic targets by deconvoluting complex pathways involved in the pathological processes. Subsequently, gaining a better understanding of the disease aetiology on the molecular level could also be advantageous for a better monitoring of the pathology progress and treatment efficacy.

CVD leads to a clinical syndrome, known as heart failure (HF), which can be preceded by a structural and/or functional heart dysfunction. HF can be caused by a broad spectrum of diseases, involving the pericardium, endocardium, myocardium, heart valves and vessels; this heart function dysregulation leads to impaired ventricular filling or blood ejection [3].

HF affects approximately 40 million people globally as recorded in 2015, and an estimated 2% of the adult population is suffering from HF [4]. HF dominates in the elderly population, with the incidence rate being 6–10% for those over 65 years and more than 10% for the population older than 75 years [4, 5], with men showing a higher predisposition for CVD [6]. Most cardiomyopathies have complicated underlying causes where chronic or poorly controlled hypertension can lead to increased afterload resulting in higher cardiac workload, which in turn can precipitate the hypertrophy of the left ventricle. Decreased heart contractility and output in CVD can also be caused by a direct ischemic damage to the myocardium, which induces further scar formation and tissue remodelling [1]. Hypertension, ischemic cardiomyopathy (IC) and dilated cardiomyopathy (DC) precede later-stage HF with reduced ejection (HFrEF) [1, 4]. HFrEF encompasses a diverse pathologic spectrum and it is a good case example when long-standing paradigms of a single common pathway [1, 7] do not provide an adequate measure of the pathology development or progression. That is, most current therapies for HFrEF do not specifically focus on disease aetiology or in-depth differentiation [1, 2, 8]; thus, the heterogeneous nature of HF remains insufficiently addressed. As a result, the need of new therapeutic insights and an improved analysis of the underlying HF mechanisms was the divining force behind this study to develop a novel approach with integrated multi-‘omics’ and machine learning methods.

The dramatic expansion of RNA-seq and metabolomics screening capabilities provides an excellent resource for an in-depth look into cardiomyopathies. Moreover, cardiac sample collection cannot always be optimal and there are technical variations, and this can become especially problematic in clinical and small-scale studies when patient samples might be limited in number. However, a robust growth in novel statistical approaches allows researchers to better glean information from noisy datasets and clean the data from technical errors or batch effects. To address the discussed issues, we aimed to emulate scenarios when only a limited number of samples are available and to show that the statistical modelling and

enrichment with external resources can be a powerful method to compensate for a lower sample number or sample drop-out due to quality issues. It is, however, important to highlight that while we selected a small sample size for the analysis, it does not mean that small and large sample size groups can be regarded as equivalent. Moreover, this study also does not aim to provide a comparison between the outcomes of larger and smaller sample studies as there are so many great resources already addressing that [9–11]; the core aim is to demonstrate how researchers who have a limited number of samples can still successfully analyse their data to identify meaningful gene expression patterns and changes. Thus, with this study, we demonstrated how multi-‘omics’ approaches can help to uncover the intricate biological mechanisms of pathological processes.

As a result of the urgency to improve therapeutic solutions in HF, we selected the human left ventricle as a case study. Current HF treatment options rely on targeting the associated symptoms with left ventricular failure without taking into account the heterogeneity of the underlying mechanisms [1, 7] (Fig 1; Supplementary Tables S1 and S2). With our study, we introduced an approach to uncover new genes that might be important candidates in understanding the heterogeneous nature of HF. That is, we wanted to highlight the fact that not all patients with the same clinical condition share the same mutations and the disease progression might have multiple converging paths. Thus, using our proposed method to aggregate results, we can explore how these genes are associated with more dominant genetic factors which could lead to new therapeutic insights.

## METHODS

### Sample selection

Publicly available datasets were used to randomly select 12 human left ventricular RNA-seq samples (PRJNA477855, EBI: European Nucleotide Archive) [12] which were categorized to form non-failing (healthy), DC and IC groups; similar sets of samples were selected for the proteome analysis (PXD008934, EMBL-EBI: PRIDE) [13] (Supplementary Tables S1 and S2) with matched representation for all ages and sexes. RNA-seq and proteome analysis samples represent a small biological set for an independent analysis. Single cell RNA-seq of the murine non-myocyte cardiac cellulome (E-MTAB-6173) was downloaded from ArrayExpress database [14]. Human left ventricular myocardium was downloaded from publicly available Visium data from 10× Genomics [15].

### RNA-seq data pre-processing and exploratory analysis

The number of reads per sample averaged 59 million. Reads were filtered for quality and trimmed using Trimmomatic tool [16] and were aligned to the human genome reference GRCh37/hg19 [17] using HISAT2 [18] with a 95% average alignment rate. Ensembl GRCh37/hg19 [17] GTF was used for featureCounts [19] tool to count reads based on genomic features. Quality control was performed both pre- and post-alignment using MultiQC [20]. Single cell counts were acquired after the raw data were processed with Cell Ranger version 1.3 (10× Genomics) [15].

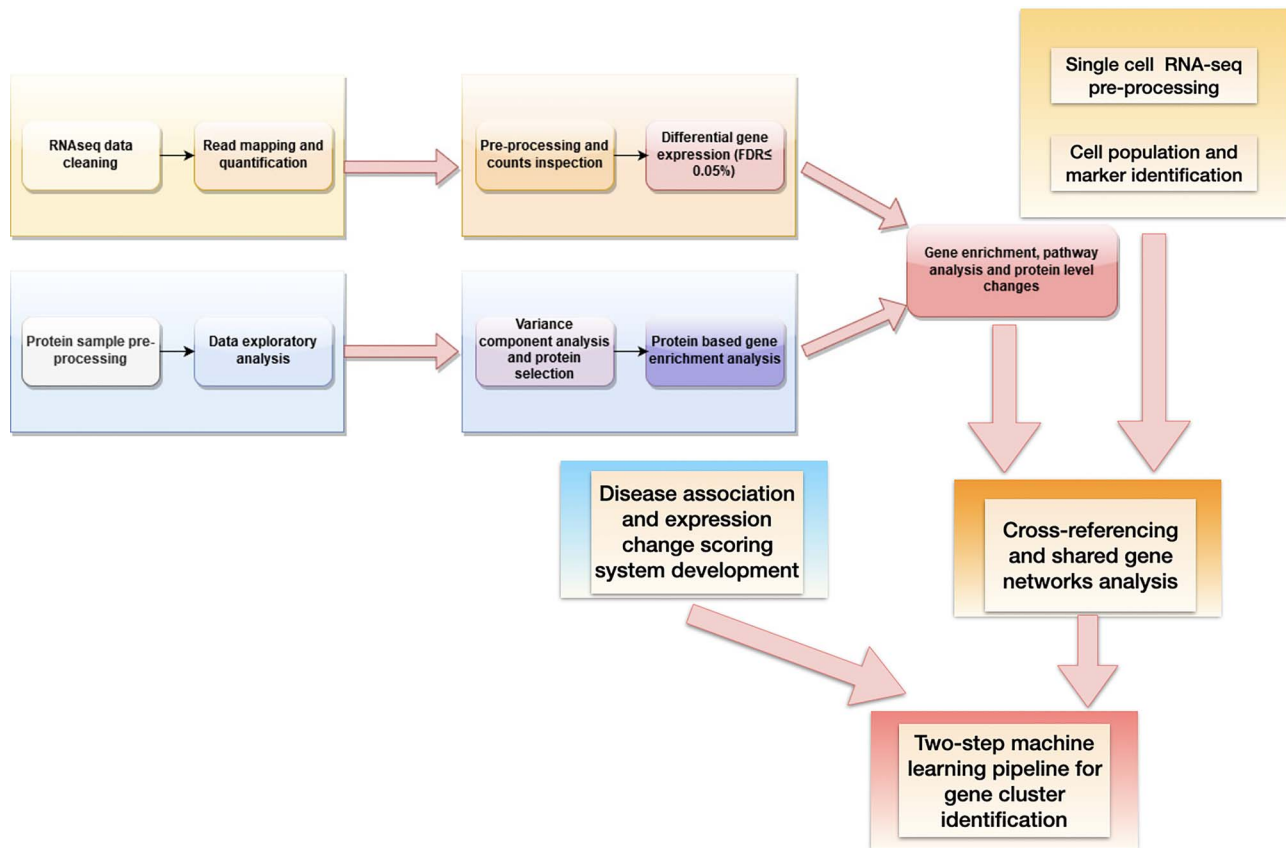


Figure 1. Diagram showing the steps for data processing and integration.

### Differential expression analysis

RStudio 3.6 [21] environment was used for raw RNA-seq counts pre-processing and quality control (Supplementary Fig. S2) and further analysis was done using package DESeq2 [22] as well as dependent packages for graphical processing and data manipulation. Seurat R package [23] was used to analyse single cell data maintaining mitochondrial DNA content at <5% for non-cardiomyocyte samples and <40% for cardiomyocytes. R packages: SingleR [24], CellDex [25] and Clustermole [26] were used to determine the cell types. Differential expression was established based on disease status while controlling for gender differences.

### Protein-level analysis

Protein abundance data were retrieved from earlier raw spectra analyses using MaxQuant version 1.5.3.30 [27]. Label-free quantification (LFQ) intensity values were used in lieu of protein abundance and were pre-processed to remove proteins with median distributions across all samples that were equal to 0 LFQ. LFQs were scaled by a factor of  $10^{-6}$  prior to DESeq2-based normalization and model fitting to find differences between conditions while controlling for gender effects. For the protein and gene set overlap per condition, only significantly changed ( $P_{adj} < 0.05$ ) genes and proteins were selected.

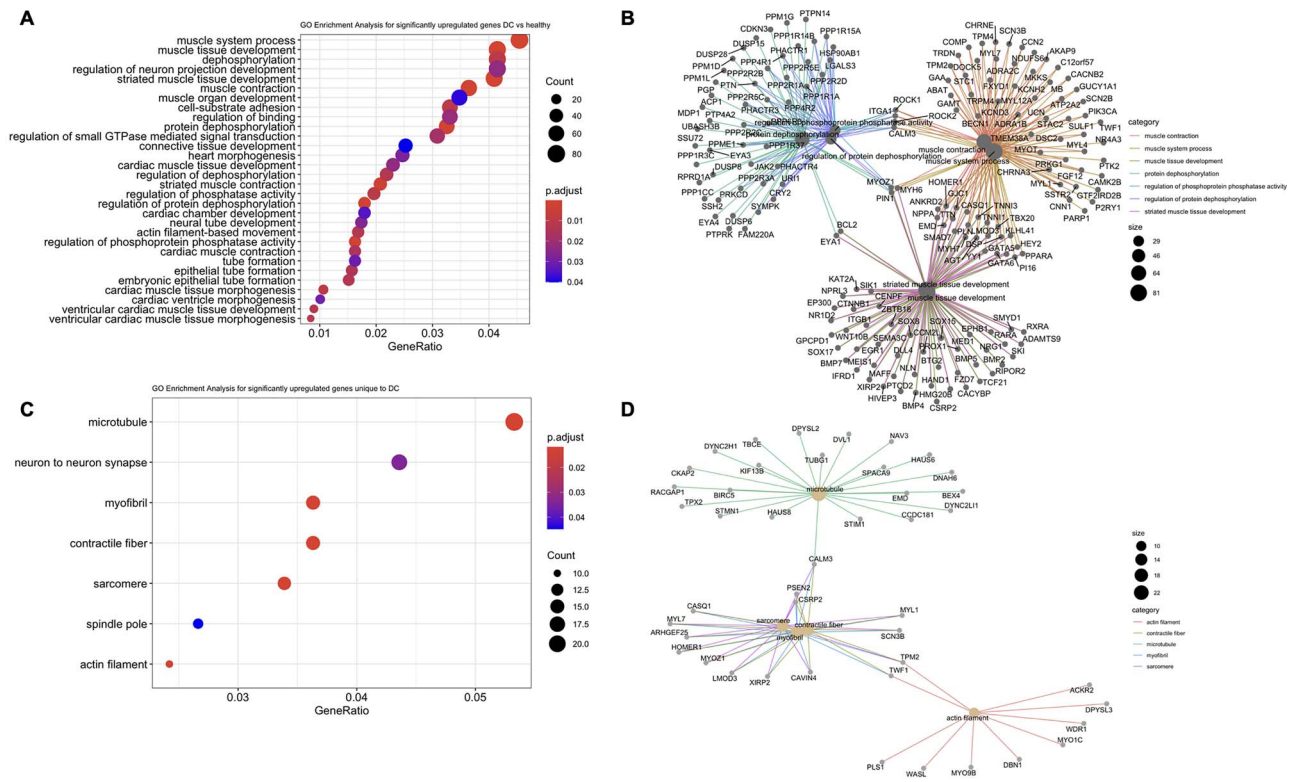
### Gene enrichment and pathway modelling

ClusterProfiler [28] and DEReport [29] as well as dependent packages were used for gene ontology and pathway analysis. Open Targets [30] and STRING (version 11, score\_threshold = 200)

[31] were used for data mining to build interactor networks. STRING database provides a source of known and predicted protein-protein interactions which may include direct (physical) and indirect (functional) associations, computational analysis-based predictions as well as other interaction data aggregated from primary databases [32]. Since STRING database does not provide disease-specific links, another database, namely Open Targets, was used to retrieve information on the human gene and disease associations for target identification and prioritization [30].

### Machine learning and disease-centric scoring

For the initial clustering, Gaussian mixture models (GMMs) were chosen since they function as a density estimator to establish cluster patterns. The probabilistic nature of GMM was best suited to perform parameter separation [33]. Identified clusters with GMM were isolated and subjected to agglomerative hierarchical clustering [34] since this method is the most optimal to find small sub-clusters determined by Silhouette and Elbow methods [35, 36]. GMMs (with the following parameters: max\_iter = 1000, covariance\_type = 'full' or 'spherical', tol = 0.001, random\_state = 0) were implemented to cluster genes based on their scaled log2 fold change (selected  $LFC_{score} > |1.5|$ ) and the number of interactors (i.e. expressed gene's degree) (Supplementary Equation (1)). Scaling factor was determined by the cumulative score of multiple mined resources (Open Targets) [30] where a gene was assigned a value (from 0 to 1) based on its probabilistic links to a specific disease (denoted as  $\alpha$  in Supplementary Equation (1)). The number of interactors was identified using



**Figure 2.** Human left ventricle bulk RNA-seq gene count clustering and distribution analysis showing Spearman correlation calculated distances (A) and Euclidean distances (B) for rlog-transformed counts; PCA plots provide grouping by condition (C) and gender (D).

the STRING database of known protein–protein interactions [31]. GMM clustering evaluation was performed using probabilistic statistical measures quantifying the model performance for the different number of clusters. Evaluation parameters were based on Akaike information criterion (AIC) [37] and the Bayesian information criterion (BIC) [37]. Python Scikit-Learn GMM (scikit-learn 0.22.2) [38] was used to determine and project the Gaussian mixture modelled density and distribution of selected gene parameters.

### Machine learning pipeline validation

Genome-wide association studies (GWAS) dataset of human genetic variants [39] was cross-referenced against the identified clusters retrieving the normalized association score for a CVD category (set size 5551). Open Targets platform search (target set screen: >28 000 genes) for cluster genes against any indications related to heart disease (e.g. hypertension, cardiomyopathy and HF) was also performed. Complete records of PubMed [40] (>30 million) were text-mined for CVD-associated terms retrieving the number of articles/studies where the gene is mentioned in the disease context. Scoring and machine learning analyses were validated with an independent dataset of biopsies for dilated and non-failing heart (GEO: GSE3585) [41] as well as diabetic HF and healthy samples (GEO: GSE26887) [41] by selecting significantly changed genes ( $P_{adj} > 0.05$ ) in the disease.

### Statistical analysis and graphs

Statistical analyses (including plots and graphs) were performed in RStudio [21] environment. Machine learning and GMM plots were done in Python [42] programming environment.

## RESULTS

### RNA-seq captured specific gene changes in dilated and ischemic heart conditions

Exploratory analysis of the human left ventricle bulk RNA-seq data (PRJNA477855) revealed that the sample count distribution and coverage depths were consistent (Supplementary Figs S1–S3) without any marked batch effects. However, clustering analysis (Fig. 2A and B) indicated that samples were relatively homogenous based on their gene expression, with only the non-failing (healthy) group showing the clearest separation. Moreover, dilated and ischemic groups were intertwined without minimal subdivision. This trend was also reflected in the principal component analysis (PCA) (Fig. 2C), where the disease groups not only had a marked overlap but the intra-sample variability was also higher when compared to the healthy group. Employing pre-processing quality controls, such as batch effect, count distribution, coverage depth and count correlation analyses as well as PCA, allows to assess if the expected high patient sample variability can be reasonably modelled with the downstream statistical models. It is advised to start the analytical pipelines with exploratory analyses as samples generated from patient tissues tend to have a high variability.

Despite high homology between samples, it is possible to identify biologically meaningful genes if they had a marked upregulation or downregulation. This assumption was confirmed by the proportion of significantly changed genes for dilated (11.1%) and ischemic (17.6%) pathological states when contrasted to the healthy tissue (Supplementary Fig. S4).

Interestingly, the genes that changed significantly for each investigated contrast showed some overlap, which can be

attributed to the complex nature of the regulatory pathways involved [1, 8]. Both unique and full sets of significantly changed genes per contrast group (Supplementary Fig. S5) were used to enrich for marker genes as it was necessary to examine the shared and disease-specific expression patterns in the pathology. Genes that had the most notable change based on P-adjusted value when comparing DC versus healthy samples showed a clear separation for these two conditions (Supplementary Fig. S6A). However, the same set of genes did not show such a pattern in ischemic disease. When selecting genes based on the lowest P-adjusted value for the contrast of IC versus healthy heart samples (Supplementary Fig. S6B), the ischemic heart sample genes formed a separate cluster, while the healthy and DC samples were dispersed and were relatively similar in their expression values. Further investigation of the most significantly changed genes in DC (Supplementary Fig. S6A) revealed that the ribosomal protein S17 (RPS17) expression is the most notably changed. While ribosomal proteins might be a leftover due to the sample preparation, there is emerging evidence of ribosomal protein expression and/or mutational changes being involved in numerous diseases [43]. Since there was no other over-representation for ribosomal genes, it is possible that the observed expression levels might be biologically meaningful. Other groups of genes, such as SLIT and NTRK-like family member 4 (SLITRK4) and glycosyltransferase 8 domain containing 2 (GLT8D2), have been reported to have links to tissue structural changes [30]. Upregulated myozenin-1 (MYOZ1), enolase 2 (ENO2) and bone morphogenetic protein 2 (BMP2) are all linked to heart tissue hypertrophy or were identified as potential biomarkers in the disease [30, 44, 45]. These findings are especially interesting when compared to the downregulated genes, specifically carbonic anhydrase 11 (CA11), intercellular adhesion molecule 3 (ICAM3) and ELOVL fatty acid elongase 2 (ELOVL2), as these molecules have been associated with HF, vascular injury and changes in tissue metabolism [30, 46, 47].

In contrast to the dilation of the heart, ischemic conditions were found to be dominated by the immune system, fibrosis- and cell proliferation-linked genes, namely, C-X3-C motif chemokine ligand 1 (CX3CL1), proto-oncogene c-Fos (FOS), transmembrane protein 259 (TMEM259), REC8 meiotic recombination protein (REC8) and formin homology 2 domain containing 1 (FHOD1), that were significantly expressed and, some of the genes, such as CX3CL1 and TMEM259, are candidate genes for novel biomarkers and/or therapeutic targets for the ischemic heart disease [30, 48–50]. The group of downregulated genes in ischemia, for example, TAO kinase 1 (TAOK1) and MINDY2 (lysine 48 deubiquitinase 2), is categorized as being involved in some inflammatory processes [30].

Exploring uniquely and significantly changed genes in DC or IC but ranking based on the fold change (Supplementary Fig. S7), we can immediately see that DC showed interesting metabolic patterns, such as the upregulation of 5-HT transporter (serotonin transporter, SLC6A4), with dependence on sodium and chloride movement across the membrane as well as an increase in CYP3A5 expression; RPS17 also belonged to this LFC ranked category. As in previous P-adjusted value categorization, the ischemic heart tissue had a more pronounced signature of immune process involvement, for example, major histocompatibility complex, class I, C (HLA-C) and immunoglobulin lambda variable 6-57 (IGLV6-57) (Supplementary Fig. S7). It was also further demonstrated that the significantly changed genes for the contrasts of interest showed no marked sex biases (Supplementary Figs S8 and S9); thus, the following analyses focused

on the biological processes driving the observed changes in the expression patterns.

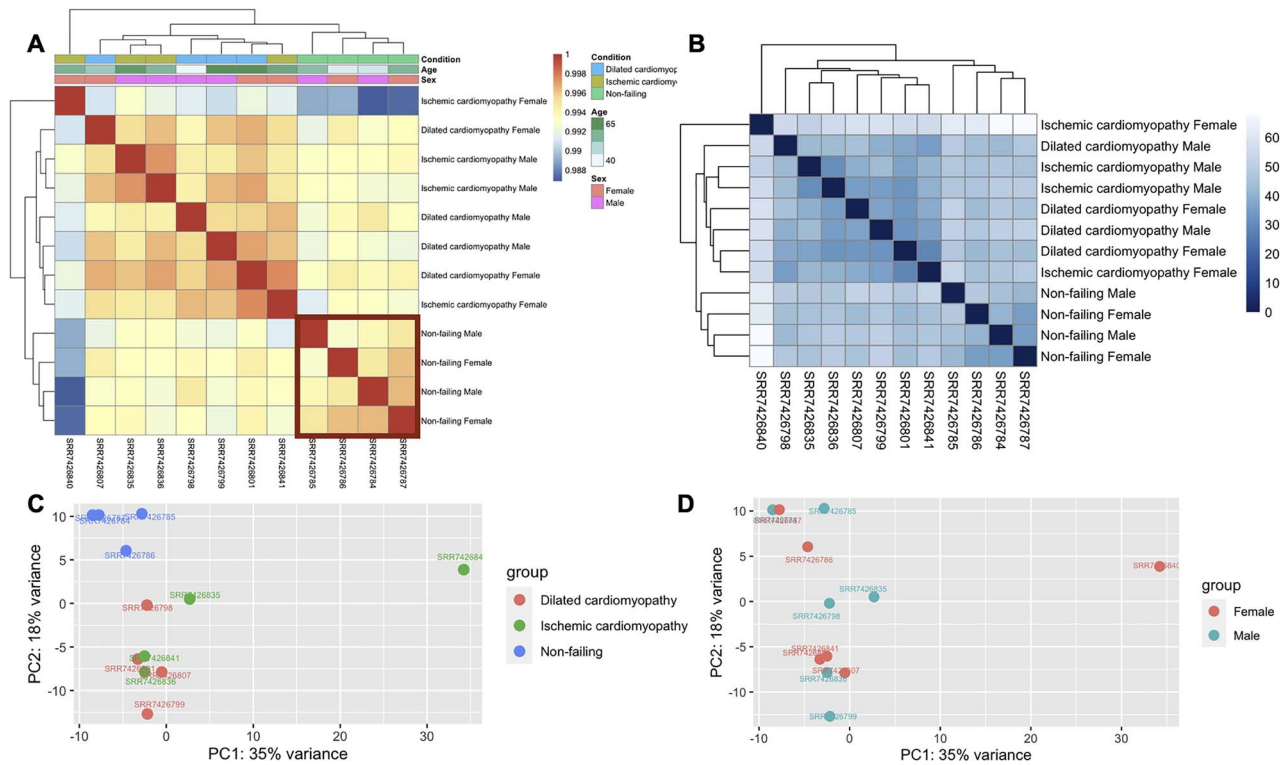
### RNA-seq revealed a clear pathological process bifurcation for DCs and ICs

Emerging differences between the ischemic and dilated heart were further cemented by exploring gene enrichment and the associated biological processes. Not surprisingly, enriched processes for the dilated heart (Fig. 3A and B) belonged to myocardium remodelling, ventricular cardiac muscle tissue morphogenesis and muscle tissue development. However, a specific set of enriched process was found for the genes that were only significantly changed in the dilated and not ischemic heart (Fig. 3C and D); these genes are involved in the microtubule, myofibril, sarcomere and contractile fibre processes. There are 64 genes (Supplementary Table S3) that were not only significantly changed when comparing the dilated heart state with a healthy sample but were also clustered into distinct cellular processes (Fig. 3C and D). Some of those genes, namely, myosin light chain 1 (MYL1), dynein axonemal heavy chain 6 (DNAH6), MYOZ1 and atypical chemokine receptor 2 (ACKR2), showed a significant upregulation in a disease state and could be of interest as potential therapeutic targets or biomarkers [30, 44].

Enrichment of the gene networks for ischemic conditions revealed a specific involvement in heart ventricular cardiac muscle tissue morphogenesis and broader metabolic functions, such as GTPase activity-linked processes (Supplementary Fig. S10A and B). While tissue remodelling is expectedly shared between ischemic and DC, there were more subtle differences in ischemic conditions that hint towards ER stress and inflammatory processes (Supplementary Fig. S10C and D). For example, spingomyelinase (SMPD3) has been previously implicated in Golgi vesicular protein transport where the inactivation of this enzyme disrupted proteostasis, leading to ER stress [30, 51]. At the intersection of the ER stress and immunological processes, there was another significantly upregulated gene, formyl peptide receptor 2 (FPR2), whose downregulation has been shown to alleviate the oxidative and inflammatory burden [52]. Intriguingly, there were a number of chemokine ligands (e.g. CXCL11, CXCL10 and CCL5) that were highly expressed as well as some chemokine receptors (e.g. CXCR3 and CCR7) and other markers, such as CD2 (Supplementary Table S4). While chemokine ligands can be expressed on a number of cells [30, 48, 53], the receptor role is more associated with T-cells and other lymphoid cells or tissues [48, 53]. CD2 marker expression is very clearly ascribed to T-cells and complex immune regulatory environment [54], and these findings likely point to a heterogeneous nature of the heart samples with other lymphoid cells infiltrating the affected tissues. Nevertheless, there is a clear shift in the ischemic tissue state with an increased inflammatory burden and with multiple regulatory mechanisms engaged (e.g. CX3CL1) [48].

### Proteome analysis highlighted underlying metabolic differences in ischemic and hypertrophic heart states

Correlation between the expression levels of mRNA and protein is relatively difficult to establish with poor predictive power for the protein levels based on the gene expression [55]. Despite that, it was necessary to establish if proteome from a myocardial tissue-rich left human ventricle (Supplementary Fig. S11) could complement the RNA-seq data.

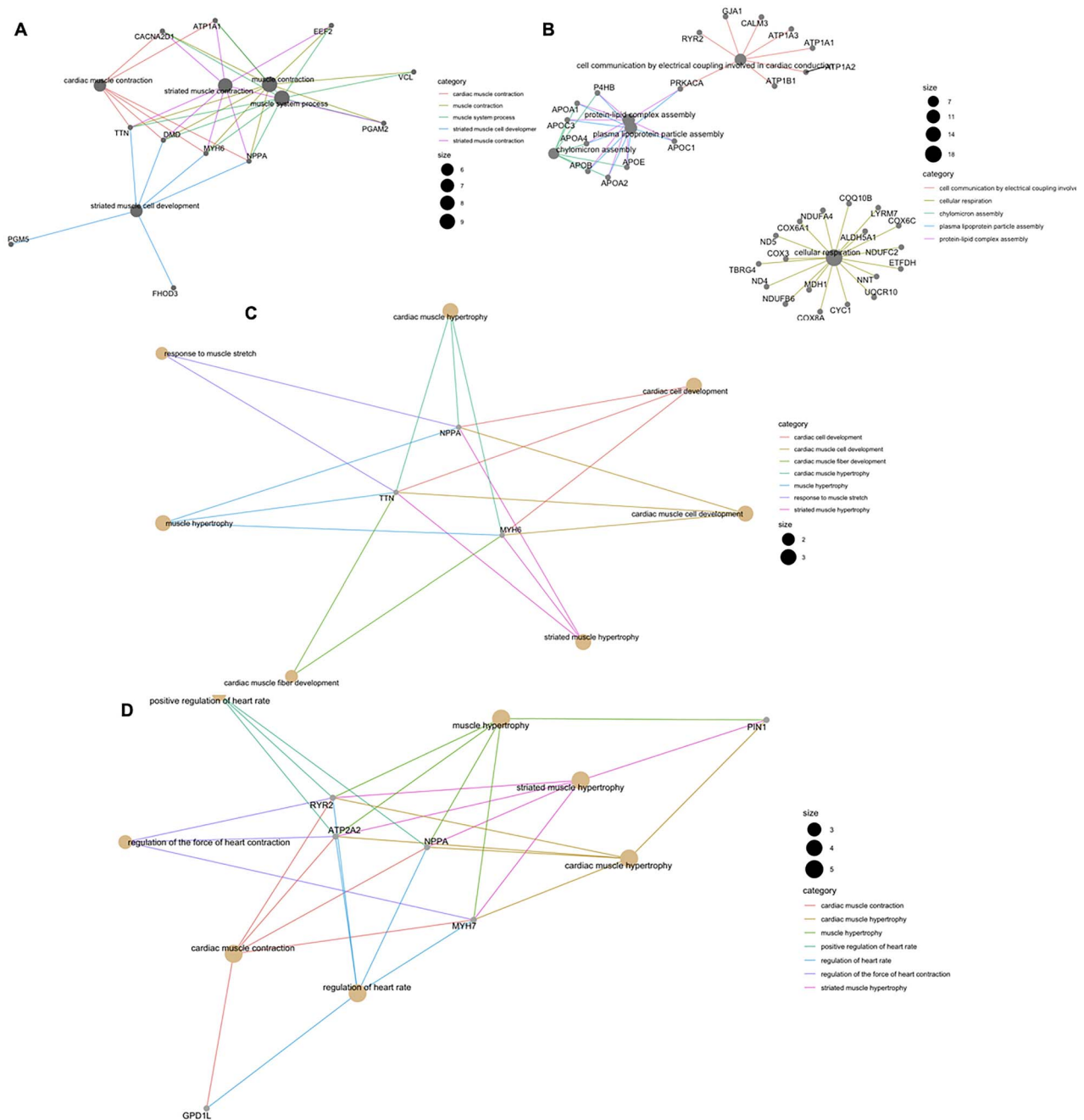


**Figure 3.** Enrichment analysis for all significantly changed genes in the DC versus healthy contrast group where enriched cellular processes (A) and the visualization of the top highest ranking processes and corresponding genes (B) are provided in the distribution plots and network maps, respectively. Enrichment analysis for genes that changed significantly in DC versus healthy but not in IC versus healthy are plotted as cellular processes distribution (C) and the visualization of the top highest ranking processes and corresponding genes are shown in network maps (D). Gene set size that was enriched and P-adjusted value provided with the plots.

While investigating protein abundances, it became clear that the samples were quite similar as was the case with RNA-seq data (Fig. 2); yet, it was possible to see some sub-divisions for DC, IC as well as the non-failing samples and pathological state samples varied less, showing a clear separation between ischemic and hypertrophic conditions with no gender-dependent effects (Supplementary Fig. S12).

The next step of the analysis was to investigate for protein enrichment and compare with the data from RNA-seq study. Proteome data had a substantially lower recovery of data points (close to 3000) when compared to nearly 19000 for RNA-seq (Supplementary Figs S4 and S13). As expected, heart dilation leads to not only increased strain over heart but also causes subsequent muscle tissue remodelling (Fig. 4A). There were 13 genes that showed a significant change in the RNA-seq samples as well as their matching counterparts on the protein level in the same contrast category (Supplementary Tables S5 and S6). For example, natriuretic peptides precursor A (NPPA), aortic carboxypeptidase-like protein (AEBP1) and collagen type XIV alpha 1 chain (COL14A1) genes as well as their corresponding proteins showed a significant upregulation in DC; in a similar fashion, myosin heavy chain,  $\alpha$  isoform (MYH6) and ADP-ribosyltransferase 3 (ART3) were downregulated. All of these genes point to the remodelling events within the tissue, however, only several genes that showed enrichment on the protein level could be clustered based on their cellular role (Fig. 4C). Interestingly, titin (TTN) expression levels dropped significantly, but the reverse was true when evaluating for its protein levels (Supplementary Tables S5 and S6; Supplementary Fig. S12). This bifurcation might likely occur due to multiple factors, namely, mRNA stability and protein half-life [55], which also demonstrates that gene or protein expression values cannot be used as sole measures, but rather a systematic approach is needed.

A completely different picture can be seen when looking into ischemic heart transcriptome and proteome (Supplementary Tables S7 and S8) and, while functional enrichment in the proteome study pointed towards lipid biogenesis and cellular respiration processes, the overlap between transcriptome and proteome only showed the enrichment for heart muscle hypertrophy, regulation of the heart rate as well as contraction force (Fig. 4B and D). Trying to compare protein versus gene expression further complicated the picture (Supplementary Tables S7 and S8; Supplementary Fig. S14) as there was less agreement in the expression changes. As a case example, myosin heavy chain 7 (MYH7) had a slight upregulation under ischemic conditions on the gene expression level, but this was markedly reduced on the protein level. This division in the expression values likely points to the complex regulatory mechanism for MYH7 under ischemic conditions as it is usually linked to the dilation and hypertrophy of the heart [56]. Several other genes, specifically, cytochrome C oxidase subunit 8A (COX8A) and coenzyme Q-binding protein COQ10 homologue B (COQ10B), which are linked to the ischemic injury and loss of mitochondrial integrity, [30, 46, 57] showed similar patterns in gene and protein LFC. Reverse was true for some of the genes that are reportedly involved in the HF kininogen 1 (KNG1) [58], retinol binding protein 4 (RBP4) [59] and, apolipoprotein B (APOB) [60] (Supplementary Tables S7 and S8; Supplementary Fig. S14). This time, no immune system-associated enrichment was found for myocardial tissue-rich samples as compared to RNA-seq data (Fig. 4), which likely confirms the complex composition of heart cellulome and the presence of other cells that might be infiltrating tissues at different time points as the disease progresses. Overall, enrichment data (Supplementary Fig. S10B and D) for ischemic cardiomyopathy demonstrated metabolic changes involving lipid generation and other proteins responsible for cellular respiration integrity.



**Figure 4.** Enrichment analysis for all significantly changed proteins in DC and IC when compared to healthy samples with enriched cellular processes for DC versus healthy (A) and enriched cellular processes for IC versus healthy in human left ventricle proteome (B). Gene names that are shared between significantly changed proteome and transcriptome for DC (C) and IC (D) contrasts versus healthy tissue.

### Single cell RNA-seq analysis of mice heart tissues revealed intricate cellulome composition that shared definitive markers with human heart RNA-seq data

To better appreciate the cellular composition of the heart, an available single cell study on the murine non-myocyte cardiac cellulome was analysed and integrated with earlier studies. While differences between species are a hurdle, this initial analysis aimed to get a better understanding of what cells can be found in the heart, their relative proportions and associated marker genes and to compare all of that with the findings in the human heart

samples. Earlier analyses hinted at the possibility of other cells infiltrating the heart; thus, it was necessary to explore further what cellular composition can be expected.

Mouse heart preparations with cardiomyocyte cell population mostly removed (Supplementary Fig. S15, and Table S9) split the remaining cells between the matrix fibroblasts and subtypes of fibroblasts (the largest proportion) as well as various types of lymphocytes and leukocytes. Several interesting subgroups, for example, axin2+ cells, displaying stem-like cell properties and involved in fibrotic and regenerative events [30] were found. Comparative analysis between human bulk



RNA-seq significantly changed genes (either up- or down-regulated) and mouse single cell RNA-seq markers revealed a number of matches (Table 1). Most notably, DC conditions were predominated by cardiomyocytes and fibroblast-like cells with some immune cell types. This was reversed in ischemic conditions with a high immune cell infiltration (Table 1). Comparing how different and non-cardiomyocyte-enriched cells cluster in mice heart tissues (Fig. 5), we can see that the separation was quite distinct where marker gene patterns (Supplementary Figs S16 and S17) allowed to differentiate this rich cellulome. Cross-referencing single cell sequencing data with proteome analysis as well as bulk RNA-seq data (DC) revealed two genes, ART3 and microfibril-associated glycoprotein 4 (MFAP4), to be also matched with mice heart cellulome markers. ART3 has been reported previously to be expressed in the heart [30], but MFAP4 has several strong links to the heart hypertrophy [30, 61]. Ischemic heart dataset analyses did not reveal such an overlap between bulk and single RNA-seq datasets as well as the proteome analysis.

### Single cell RNA-seq analysis of the human heart left ventricle indicated the existence of divergent cell types for hypertrophic and ischemic tissue conditions

Single cell sequencing of the human left ventricle revealed a complex mixture of cell types with expected cardiomyocytes, myoblasts and heart smooth muscle cells comprising nearly 65% of all cells and lymphoid cells adding up to more than a quarter of all cell populations combined (Supplementary Figs S18 and S19). These observations confirmed earlier findings (Figs 5 and 6) where gene expression patterns suggested the involvement of immune and other cell types that might contribute to fibrotic and remodelling events within the heart tissue. Specific marker genes for the human left ventricle showed varying expression patterns but a clear distinction between cardiomyocytes, heart smooth muscle cells or myofibroblasts and required an elaborate combination of multiple marker genes to differentiate the groups precisely (Supplementary Figs S20 and S21).

Further exploration of the DC genes that changed significantly and had corresponding markers in the human left ventricle bulk RNA-seq identified a change in the expression for genes likely involved in heart tissue remodelling; however, when this set was cross-referenced with matching proteome analysis, it did not return any hits. Haemoglobin subunit alpha and beta (HBA1/2, HBB) was significantly upregulated in human bulk RNA-seq (contrast: DC vs. healthy) but showed a moderate change in single cell myocyte/myoblast population when this cell group was compared against the rest of heart cells (Table 1). Alpha subunit expression of the said globins has been implicated in the vascular tone and function maintenance [30]; together, haemoglobin expression might suggest compensatory mechanisms for the tissue undergoing contractile and remodelling stress. Similarly, orphan nuclear receptor 4A1 (NR4A1) showed a marked upregulation in the hypertrophic state, while in a healthy left ventricle, it was low (Table 1). NR4A1 has recently been described to play a role in cardiac stress responses and hypertrophic growth [62]. As a complete contrast, DC showed a marked loss in adipocyte signatures (Table 1), for example, fatty acid-binding protein 4 (FABP4) has been shown to contribute to cardiac metabolism [30, 63]; thus, observed alterations might indicate a change in the energy metabolism of the heart.

Interestingly, ischemic heart tissue was very similar to the hypertrophic state when compared to human left ventricle cellulome. For example, a notable upregulation in most of the

genes, HBB, HBA1/2 and NR4A1, was matched between the different pathological states; however, lumican (LUM) and HBB were also found to be significantly upregulated in the ischemic heart proteome analysis. LUM has been shown to propagate the pro-fibrotic events in the HF [64]. The downregulated genes in ischemic conditions also followed similar patterns to the hypertrophic heart observed earlier, notably, FABP4 and glycerol-3-phosphate dehydrogenase 1 (GPD1) belong to the gene group involved in lipid and amino acid metabolism [30]. While IC downregulated genes are dominated by adipocyte-associated markers in this cross-reference analysis, TTN was also found to be downregulated under myoblast/myocytes group. TTN has been linked to remodelling and changes in the ischemic heart [65], which based on the present study findings, could be used to differentiate between hypertrophic and ischemic changes.

### New scoring system to evaluate genes using a two-step machine clustering approach revealed sets of disease-specific interactors

The richest data available are from bulk RNA-seq experiments; thus, a scoring system was devised to take the advantage of RNA-seq data and match with the data mined from multiple resources. That is, our derived scoring equation,  $LFC_{score}$  (Supplementary Equation (1)) scales LFC value for a given contrast (e.g. disease vs. healthy state) by a total association score (denoted as  $\alpha$  in Supplementary Equation (1)) retrieved from the Open Targets platform [66]. The score takes into consideration multiple data resources and evidence for a given gene (e.g. clinical precedence, reports in literature and/or known interactors).  $LFC_{score}$  introduces an important concept of adding weights to contrast LFC values based on known links to diseases or relevant phenotypes.

Two hundred and twenty-nine associations were retrieved for IC, and a far larger number of gene scores (3521) was downloaded for DC [30].

To identify the potential links between significantly changed genes in a given contrast, a two-step machine learning approach was employed using GMMs to identify gene clusters with the highest probability to share similar expression patterns, number of interactors (e.g. the gene's degree in our interaction network) and, subsequently, each cluster can be further analysed using agglomerative hierarchical clustering to achieve a better refinement between associations. To estimate the impact of expression changes as evaluated by  $LFC_{score}$  and the protein network size, an assumption was made that if a protein is known to have multiple interactions, then it is likely that more cellular processes will be perturbed when compared to a smaller and isolated network. GMM-based clustering revealed approximately the same number of features across DC and IC groups (Fig. 6). To test the impact of the  $LFC_{score}$ , the analysis was compared with a regular LFC. In the case of DC, there was a notable difference in the identified cluster distributions; by contrast, IC did not show such a noticeable difference primarily because the association scores were few and very low for this cardiopathology (IC mean for association score: 0.00023; max value: 0.01960; DC mean for association score: 0.07050; max value: 1). It became apparent that the more associations are used as weights, the better is the resolution in clustering that can be achieved.

This was followed by the extraction of identified clusters and a downstream hierarchical clustering. For example, a gene set from one of the bigger GMM clusters—cluster 0 (Fig. 7; Supplementary Table S10) for DC was probed further to reveal subtle variations between genes. A case example of

**Table 1.** Combined marker genes.

Significantly upregulated genes in DC versus healthy that had matching markers in human heart single cell RNA-seq										
Symbol	Base mean	Log2 fold change	lfcSE	Stat	P-value	P.adj	Single cell cluster	Type	P_val_adj	avg_logFC
ALAS2	9.880762	3.658381	0.9682434	3.778369	1.578586e-04	3.595666e-03	4	Myoblast/myocytes	1.768096e-28	0.1815624
CD74	15295.999951	1.396172	0.4761631	2.932130	3.366455e-03	3.194793e-02	5	Lymphoid cells/macrophages	1.959989e-17	0.3724019
HBA1	127.857722	4.147787	0.6848996	6.056051	1.395042e-09	4.125772e-07	4	Myoblast/myocytes	7.535268e-128	1.4673468
HBA2	321.183281	3.855863	0.5223493	7.381772	1.561965e-13	1.494525e-10	4	Myoblast/myocytes	2.312018e-135	1.4857391
HBB	1783.859396	4.697214	0.6945263	6.763191	1.349853e-11	8.444890e-09	4	Myoblast/myocytes	3.858435e-141	1.5070114
LUM	8659.295482	1.059213	0.3698468	2.863924	4.184283e-03	3.705038e-02	1	Smooth muscles/adipocyte- and fibroblast-like cells	8.399361e-30	0.2002449
NR4A1	1352.267936	1.607393	0.2756688	5.830884	5.513446e-09	1.318849e-06	2	Myofibroblasts	1.293227e-76	0.7037670
Significantly downregulated genes in DC versus healthy that had matching markers in human heart single cell RNA-seq										
Symbol	Base mean	Log2 fold change	lfcSE	stat	P-value	P.adj	Cluster	Type	P_val_adj	avg_logFC
FABP4	3699.5330	-1.2801910	0.20011132	-6.397394	1.580509e-10	6.121084e-08	6	Adipocytes	1.618486e-23	1.3383664
G0S2	1384.9720	-1.8636480	0.48482019	-3.843998	1.210458e-04	3.010598e-03	6	Adipocytes	3.875007e-09	1.0612945
GPD1	386.2192	-1.7924675	0.40377314	-4.439294	9.025461e-06	4.192406e-04	6	Adipocytes	3.141539e-26	0.8650561
S100A8	212.1838	-1.2542770	0.42169215	-2.974390	2.935715e-03	2.920633e-02	4	Myoblast/myocytes	1.000000e+00	0.1923214
Significantly upregulated genes in DC versus healthy that had matching markers in mouse heart single cell RNA-seq										
Symbol	Base mean	Log2 fold change	lfcSE	Stat	P-value	P.adj	Cluster	Names	P_val_adj	avg_logFC
CCN2	2930.87706	1.444716	0.4096894	3.526370	4.212979e-04	7.368635e-03	5	Fibr reticular cells/Con. tissue fibr	2.497010e-204	1.0816956
CD74	15295.99995	1.396172	0.4761631	2.932130	3.366455e-03	3.194793e-02	6	Macr activated/monocytes	0.000000e+00	2.7764279
COMP	357.97479	4.500956	1.1484004	3.919326	8.879678e-05	2.399283e-03	3	Activated fibr	3.683388e-232	1.9834490
CXCL2	141.19737	1.377606	0.4822002	2.856918	4.277768e-03	3.769349e-02	6	Macr activated/monocytes	5.346760e-71	2.1222650
EGR1	1125.64888	1.934639	0.4455200	4.342429	1.409161e-05	6.016120e-04	1	Con. tissue fibr/adipocytes	1.872001e-249	0.9102685
FMOD	2086.27650	2.208960	0.7861859	2.809718	4.958497e-03	4.160662e-02	3	Activated fibr	0.000000e+00	1.9923917
HSD11B1	77.58577	1.182886	0.3795059	3.116911	1.827567e-02	2.081737e-02	0	Matrix fibr	3.150651e-149	0.8352948
LPL	44.694.70260	1.321233	0.2591931	5.097484	3.441981e-07	3.413857e-05	0	Matrix fibr	2.054064e-163	0.6813763
MFAP4	3545.32414	1.184731	0.3634084	3.260055	1.113907e-03	1.496186e-02	5	Fibr reticular cells/Con. tissue fibr	1.956096e-207	1.4763358
P116	1678.73870	2.389334	0.5017015	4.762461	1.912466e-06	1.269721e-04	4	Skin-like fibr/axin2+ cells	1.567072e-116	0.7149607
VTN	1746.66613	1.226781	0.3230142	3.797917	1.459171e-04	3.434858e-03	10	Pericytes/cardiomyocytes	0.000000e+00	3.2261843
Significantly downregulated genes in DC versus healthy that had matching markers in mouse heart single cell RNA-seq										
Symbol	Base mean	Log2 fold change	lfcSE	Stat	P-value	P.adj	Cluster	Names	P_val_adj	avg_logFC
ART3	2301.6166	-1.0665936	0.3073363	-3.470444	5.195979e-04	8.695246e-03	10	Pericytes/cardiomyocytes	0.000000e+00	2.2127282
DBI	5957.5961	-1.1071120	0.1736465	-6.375667	1.821688e-10	6.891065e-08	15	Oligodendrocytes/glia-like cells	3.093461e-105	2.3971686
FABP4	3699.5330	-1.2801910	0.2001113	-6.397394	1.580509e-10	6.121084e-08	8	Vascular endo cells	0.000000e+00	3.6800042
S100A8	212.1838	-1.2542770	0.4216922	-2.974390	2.935715e-03	2.920633e-02	18	Lymphocytes/neutrophils	4.135267e-134	5.8014206
S100A9	553.8739	-1.2624115	0.4214594	-2.995333	2.741451e-03	2.794013e-02	18	Lymphocytes/neutrophils	4.414938e-164	5.5492214
Significantly upregulated genes in IC versus healthy that had matching markers in human heart single cell RNA-seq										
Symbol	Base mean	Log2 fold change	lfcSE	Stat	P-value	P.adj	Cluster	Type	P_val_adj	avg_logFC
CD74	15 296.0000	1.924495	0.4761500	4.041783	5.304634e-05	1.013193e-03	5	Lymphoid cells/macrophages	1.959989e-17	0.3724019
FOS	552.7522	2.433294	0.4614677	5.272945	1.342517e-07	8.238896e-06	2	Myofibroblasts	4.594938e-92	0.8547702
HBA1	127.8577	3.458750	0.6857723	5.043584	4.568919e-07	2.177241e-05	4	Myoblast/myocytes	7.535268e-128	1.4673468
HBA2	321.1833	3.467975	0.5225764	6.636303	3.216483e-11	8.491005e-09	4	Myoblast/myocytes	2.312018e-135	1.4857391
HBB	1783.8594	3.862919	0.6946055	5.561314	2.677515e-08	2.137861e-06	4	Myoblast/myocytes	3.858435e-141	1.5070114
JUNB	1004.2035	1.196043	0.3565928	3.354087	7.962727e-04	7.967078e-03	2	Myofibroblasts	3.138348e-93	0.7437075
LUM	8659.2955	1.082989	0.3698410	2.928255	3.408707e-03	2.320516e-02	1	Smooth muscles/adipocyte- and fibroblast-like cells	8.399361e-30	0.2002449
NR4A1	1352.2679	1.134481	0.2758204	4.113114	3.903568e-05	8.076799e-04	2	Myofibroblasts	1.293227e-76	0.7037670
PTN	1172.3506	1.240390	0.3404728	3.643141	2.693317e-04	3.524197e-03	1	Smooth muscles/adipocyte- and fibroblast-like cells	3.016167e-15	0.1970879

(Continued)

Table 1. Continued.

Significantly unregulated genes in DC versus healthy that had matching markers in human heart single cell RNA-seq										
Symbol	Base mean	Log2 fold change	lfcSE	Stat	P-value	P.adj	Single cell cluster	Type	P_val_adj	avg_logFC
Significantly downregulated genes in IC versus healthy that had matching markers in human heart single cell RNA-seq										
Symbol	Base mean	Log2 fold change	lfcSE	Stat	P-value	P.adj	Cluster	Type	P_val_adj	avg_logFC
FABP4	3699.53303	-1.0361608	0.19998439	-5.181208	2.204529e-07	1.255600e-05	6	Adipocytes	1.618486e-23	1.3383664
GPD1	386.21920	-1.7856639	0.40352890	-4.425120	9.638855e-06	2.593913e-04	6	Adipocytes	3.141539e-26	0.8650561
MGST1	331.26691	-1.6291823	0.53631060	-3.037759	2.383446e-03	1.795248e-02	6	Adipocytes	2.222223e-19	1.1258766
RBP4	51.50649	-3.0112881	0.80596479	-3.736253	1.867829e-04	2.658098e-03	6	Adipocytes	3.371234e-200	1.1091309
TTN	26 0372.84301	-1.1102638	0.14893528	-7.454673	9.009087e-14	6.590030e-11	4	Myoblast/myocytes	2.237662e-21	0.2690609
Significantly upregulated genes in IC versus healthy that had matching markers in mouse heart single cell RNA-seq										
Symbol	Base mean	Log2 fold change	lfcSE	Stat	P-value	P.adj	Cluster	Names	P_val_adj	avg_logFC
CCL3	21.12713	2.193338	0.8185624	2.679500	7.373228e-03	4.046710e-02	14	Lymphocytes	1.108071e-21	1.7594966
CCL5	120.60019	2.805093	0.7345592	3.818744	1.341331e-04	2.082883e-03	16	T/NK cells	1.515124e-71	4.2358648
CD74	15 295.99995	1.924495	0.4761500	4.041783	5.304634e-05	1.013193e-03	6	Macr activated/monocytes	0.000000e+00	2.7764279
CD79A	11.17099	4.595454	1.3925479	3.300033	9.667345e-04	9.224189e-03	12	B cells: memory, naive, mature	0.000000e+00	3.0975730
EGFL7	2239.74334	1.112055	0.1924802	5.777502	7.581777e-09	7.461096e-07	8	Vascular endo cells	0.000000e+00	2.4291359
EGR1	1125.64888	2.589524	0.4453014	5.815216	6.055571e-09	6.324236e-07	1	Con. tissue fibr/adipocytes	1.872001e-249	0.9102685
HCST	80.97730	1.980339	0.6233735	3.176809	1.489050e-03	1.279152e-02	16	T/NK cells	1.892737e-273	1.5809485
IGHM	1109.34675	6.381438	1.5602478	4.090015	4.313446e-05	8.611875e-04	12	B cells: memory, naive, mature	0.000000e+00	2.8022132
JUNB	1004.20349	1.196043	0.3565928	3.354087	7.962727e-04	7.967078e-03	1	Con. tissue fibr/adipocytes	8.491950e-212	0.9091541
MS4A1	12.48049	3.681749	1.2067727	3.050905	2.281524e-03	1.732604e-02	12	B cells: memory, naive, mature	0.000000e+00	2.0164658
NKG7	46.03626	2.189070	0.6864901	3.188786	1.428714e-03	1.239486e-02	16	T/NK cells	0.000000e+00	2.6245828
PI16	1678.73870	1.637016	0.5018133	3.262201	1.105506e-03	1.023131e-02	4	Skin-like fibr/axin2+ cells	1.567072e-116	0.7149607
SMOC2	1903.44520	1.314703	0.3295271	3.989667	6.616609e-05	1.205266e-03	0	Matrix fibr	9.428873e-297	0.9696528
TRBC2	120.39393	2.705592	0.7875605	3.435408	5.916620e-04	6.389565e-03	16	T/NK cells	0.000000e+00	2.0733703
Significantly upregulated genes in IC versus healthy that had matching markers in mouse heart single cell RNA-seq										
Symbol	Base mean	Log2 fold change	lfcSE	Stat	P-value	P.adj	Cluster	Names	P_val_adj	avg_logFC
ART3	2301.6166	-1.7021963	0.3075560	-5.534590	3.119575e-08	2.413100e-06	10	Pericytes/cardiomycocytes	0.000000e+00	2.2127282
FABP4	3699.5330	-1.0361608	0.1999844	-5.181208	2.204529e-07	1.255600e-05	8	Vascular endo cells	0.000000e+00	3.6800042

Filtering parameters: LFC < |1|; P.adj < 0.005. Bulk RNA-seq samples (PRJNA477855) were categorized to form: non-failing (healthy), DC and IC groups; similar sets of samples were selected for proteome analysis (PXD008934) with matched representation for ages and sexes. Single cell RNA-seq of the murine non-myocyte cardiac cellulose (E-MTAB-6173) was used to identify additional markers for the heart.

one of the sub-clusters, SMAD7, syntaxin-1B (STX1B) and transcription factor SOX-17 (SOX17), show how genes with similar expression and network profile can be grouped and, in this case, these genes belong to the different branches of a complex network regulating tissue morphogenesis, vesicle docking and growth [30, 31]. Some of the IC sub-cluster members, such as tumour necrosis factor receptor superfamily member 11B (TNFRSF11B) and serine/threonine-protein kinase pim-2 (PIM2), showed a convergence of two signalling branches via Myc proto-oncogene protein (MYC) [31] (Fig. 7; Supplementary Table S11).

To achieve a better organization of identified clusters and to cross-reference the findings, GWAS dataset of human genetic variants was searched against the identified clusters, retrieving associations for a CVD category (set size 5551). In order to expand the search for all known heart diseases, taking into account text mining, expression data as well as clinical evidence, Open Targets platform was used to retrieve the association scores for cluster genes; by parsing the platform, it was possible to retrieve the values for more than 28 000 genes [30]. Finally, for the most

up-to-date analysis of complete PubMed records PubMed [40] (>30 million), a text-mining based search was performed for any CVD-associated term, retrieving the number of articles/studies where the gene is mentioned in the disease context. This analysis returned two comprehensive tables for DCs and ICs (Supplementary Tables S10 and S11), where each cluster had a number of genes linked to cardiovascular pathologies either based on all or some of the parameters (i.e. GWAS association, Open Targets knowledge-base association or the number of publications where gene appears in the context of cardiopathology). What is especially useful is that genes which have sparser or even no known links to the disease belong to clusters with better-defined members. This could lead to the identification of new biomarkers or a better understanding of their function since they were classified based on their interaction network complexity and expression. For example, the mentioned SMAD7, SNCA and SOX17 have clearly established links to heart pathology; however, their cluster (number 0) has some less well-defined members, such as SLC6A12 or SPNS3, and these carriers/transporters could be interesting candidates for further exploration

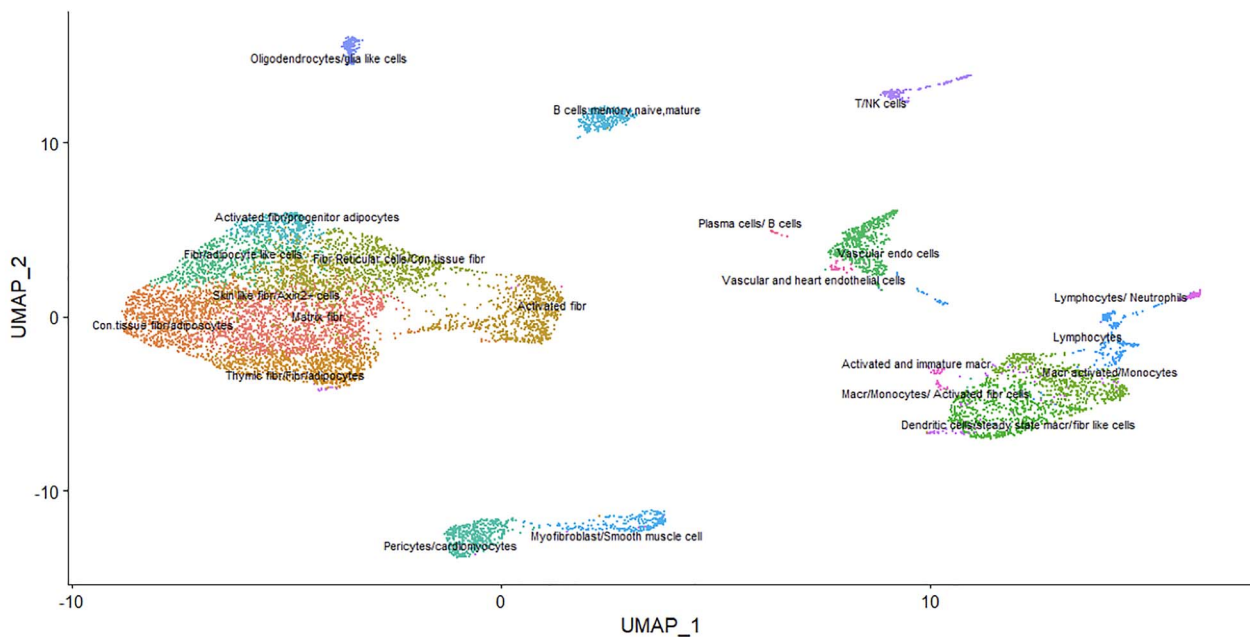


Figure 5. Mouse non-cardiomyocyte single cell RNA-seq cellulome UMAP decomposition showing relative distances and the uncovered clusters of different cells. Some longer names were abbreviated; for full names, please refer to Supplementary Table S9.

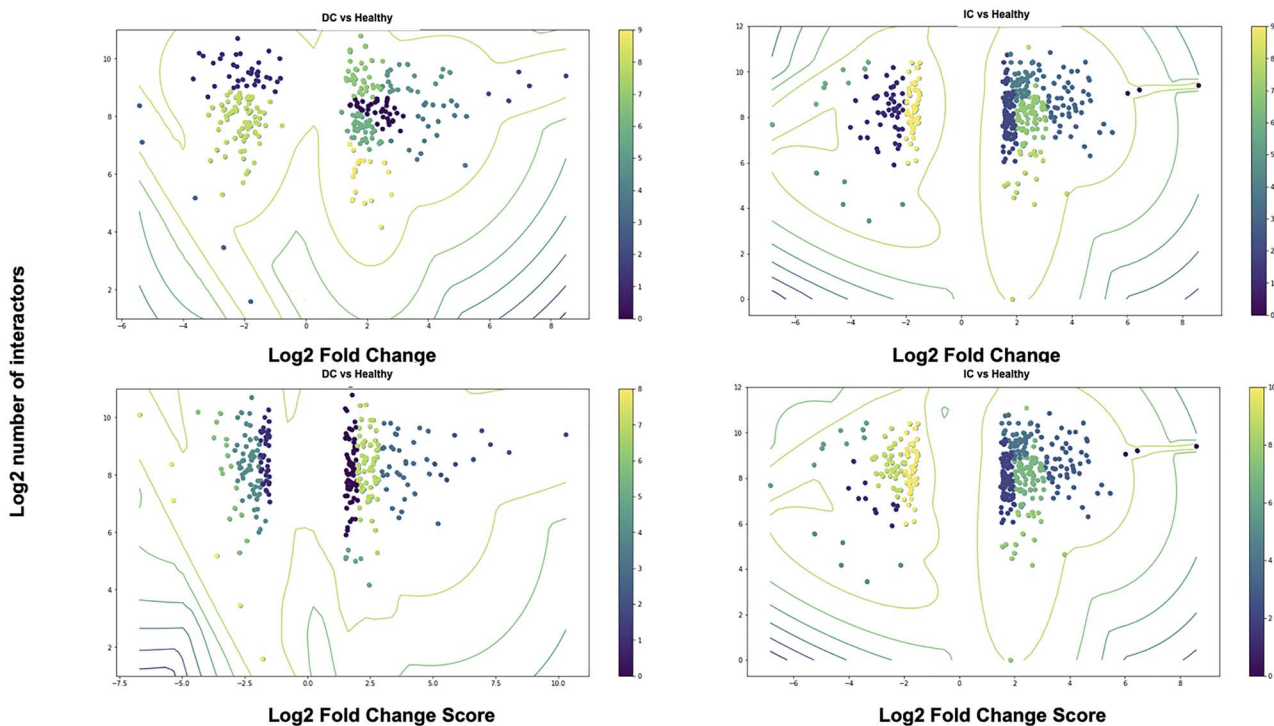
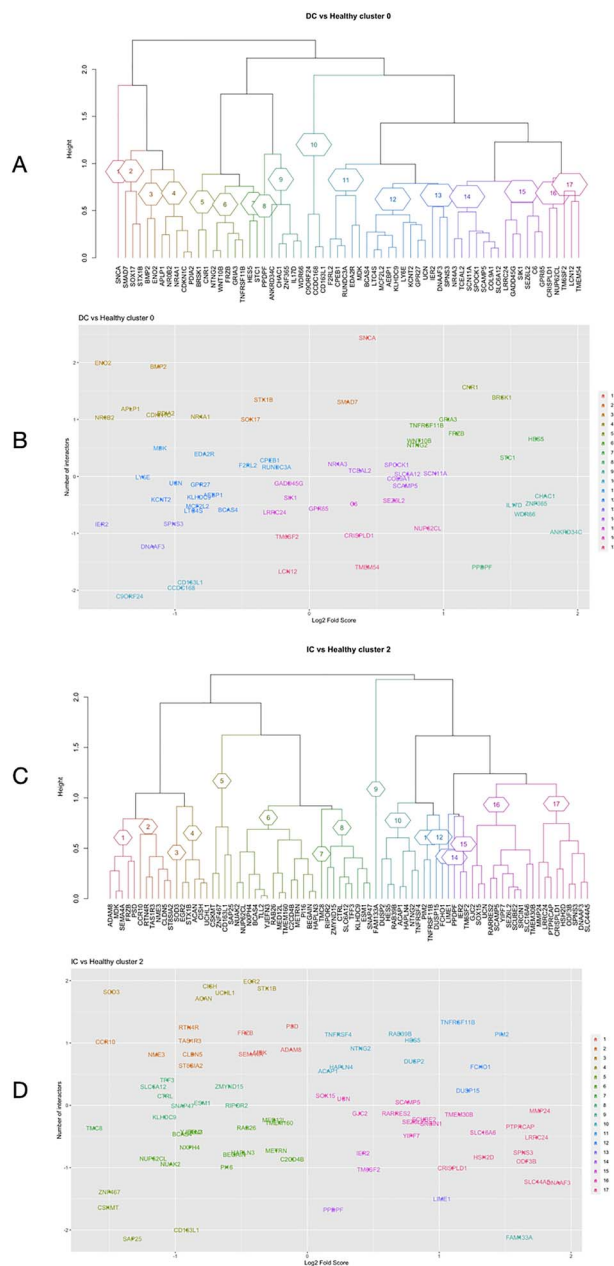


Figure 6. Human heart left ventricle bulk RNA-seq GMM clustering showing specific grouping based on either LFC or LFC<sub>Score</sub> against known or predicted number of interactions for that gene. Colour bar shows the specific cluster number and colour association.

based on their grouping. As can be seen, our approach helps to uncover new genes that might be important candidates in understanding the heterogeneous nature of HF; such findings point to the fact that not all patients with the same clinical condition share the same mutations, disease progression might have multiple converging paths and using aggregated results we

can explore how these genes are associated with more dominant genetic factors.

The genes that are shared between two conditions, namely DC and IC, when the initial clustering was performed via GMM were extracted (to find the overlap) and that overlap alone was clustered further to see underlying patterns



**Figure 7.** Human heart left ventricle bulk RNA-seq GMM analysis identified multiple clusters which were further subjected to hierarchical clustering (dendrogram panels). Representative clusters are shown where the gene distribution for DC versus healthy (cluster 0) can be seen in the dendrogram (A) and distribution plot (B), similarly IC versus healthy (cluster 2) gene distribution is shown in the dendrogram (C) and distribution plot (D). All dot plots (C and D) show grouped gene distribution for z-score scaled parameters on which the sub-clustering was performed.

(Supplementary Fig. S22). Some of the juxtaposed groups (coloured branches) are matched between two pathologies based on the gene-disease association, LFC and the degree number that the expressed protein has. For example, one such shared group of protein Wnt-9a (WNT9A), HBB and F-box and leucine-rich repeat protein 16 (FBXL16) belong to the same large network of interactors [31] likely playing a role in tissue function and local signalling events. The described analysis could be very useful in understanding the potential convergence points for diseases and how shared genes are grouped per disease profile.

## Validation of machine learning approach

In order to assess if our developed analysis can uncover gene groups that have similar expression and network size profiles, two independent RNA-seq studies were analysed to test the analytical pipeline and explore whether the identified gene clusters allowed to group well-defined genes with unknown new candidates or provided new insights based on the network size and expression changes.

Differentially expressed genes ( $P_{adj} > 0.05$ ) derived from the first dataset consisting of DC and healthy tissue biopsies [67] were introduced into the previously described pipeline. In the same manner, the second dataset was tested; the samples of this dataset were comprised of heart tissue from diabetic patients affected by post-ischemic HF as well as healthy tissue [68]. GMM-identified clusters (Supplementary Fig. S23) were subjected to cross-referencing with Open Targets, GWAS and PubMed records to retrieve the records associated with heart disease.

Interestingly, while selected pathologies have different underlying causes, every cluster had a number of genes associated with CVD when cross-referenced against different databases (Supplementary Tables S12 and S13). For example, significantly changed genes in DC biopsies formed nine clusters (Supplementary Fig. S23; Supplementary Table S12) via GMM of which some gene groups pointed to epigenetically active biomarkers, namely, H2AFZ and H1FO (cluster 0), that are relatively newly linked to the disease. However, when newly identified genes cluster closely with other more established candidate genes, it is possible to use that information either for targeted screens or for trying to deconvolute the involved pathways. Another example from genes that were significantly changed in diabetic patients affected by post-ischemic HF revealed similar patterns in terms of cluster formation (Supplementary Fig. S23; Supplementary Table S13).

By juxtaposing the complexity of the interaction network as well as expression changes, it is possible to establish groups of similar patterns which can be further hierarchically clustered to refine the relationships within a selected group. This refinement can aid when selecting specific genes for testing panels because selections can be spread out through clusters, avoiding picking all candidate genes from the same group. It is also worth mentioning that the choice to use a more diverse set of public records for validation was aimed at reducing any inherent biases and at representing a broader spectrum of information available on the heart disease.

## DISCUSSION

DC is an important cause of HF, which is characterized by the ventricular enlargement and subsequent systolic dysfunction. By contrast, IC is a clinical manifestation with a complex causality ranging from coronary artery disease to other changes in the heart muscle which decrease the nutrient and oxygen supply. A wide spectrum of etiologies, including inherited, inflammatory and/or infectious diseases, can predispose the heart to this pathological remodelling [1, 4, 30, 69, 70].

Studying the cardiac impairment resulting from heart dilation or ischemia is complicated by the mixture of known as well as idiopathic causes. Moreover, integrating a complex transcriptional landscape might be difficult as evident in the past studies reporting on experimental or meta-analyses [1, 71]; this is because, collected tissues for experiments differ to some extent and a sample population might introduce various other

confounding factors (e.g. treatment and co-morbidities). In addition, depending on the statistical assumptions made and the model selection, the results may vary. This becomes especially evident when analysing smaller sample sets, which is often the case in the clinical and smaller-scale studies. Our study goal was to emulate these scenarios and show that the statistical modelling and enrichment with external resources can be a powerful method to compensate for a lower sample number or sample drop-out due to quality issues. We also want to stress that while our study used a small sample size for the analysis, it does not mean that small and large sample size groups can be regarded as equivalent; there are many excellent works discussing the sample size effects and associated analytical complexities [9–11]. As a result, we wanted to demonstrate how researchers who have a limited number of samples can still successfully analyse their data to identify meaningful gene expression patterns and changes. Furthermore, we explored how the different 'omics' resources for cardiomyopathy can be used to study the differential gene expression and functional processes as well as what we could learn from integrating such datasets.

The first part of the analysis focused on the human left ventricle tissue bulk RNA-seq analysis for two indications: DC and IC. By analysing significantly changed genes, it was possible to see a subtle separation between hypertrophic and ischemic heart conditions. For example, DC tissue had a number of significantly upregulated genes (BMP2, MYOZ1 and ENO2) that showed strong associations with myocardial tissue remodelling and structural changes when compared to the healthy samples [30, 44, 45]. Some other genes, such as RPS17, SLITRK4 and GLT8D2, belong to newer additions of potentially valuable genes and have just recently been implicated in DC. These genes are involved in protein synthesis and post-translational modifications as well as cell growth control [30, 43]. An opposing group of genes that were downregulated (CA11, ICAM3 and ELOVL2) hints at the metabolic perturbations [30, 46, 47] spanning the spectrum from cellular respiration changes to the potential loss of the membrane integrity in the tissue that is actively being remodelled. When contrasting these findings with ischemic heart conditions, there was a notable change in the upregulation of the pro-inflammatory and pro-fibrotic genes. For example, CX3CL1 is an especially intriguing gene as it encodes an atypical chemokine which can exist in either a membrane-bound form or as a soluble chemokine; the membrane-integrated form is largely expressed on the endothelial cells in myocardial ischemia and HF [30, 48, 53]. Another candidate gene and potential biomarker of note, TMEM259, has some associations with ischemic conditions as well as ER protein degradation pathways [30, 72]. A number of other genes, FOC, REC8, FHOD1 as well as TAOK1 and MINDY2, are involved in the modulation of cell proliferation, immune signalling and protein turnover [30, 49, 73]. These findings provided the first hints of the potential exacerbation of ER stress as well as inflammation-induced damage propagating the ischemia and tissue fibrosis cycle. Managing the exacerbation of the inflammation might be helpful in preserving the heart tissue function. In addition, gene expression changes comprising the broad spectrum of cellular metabolism and growth processes were more pronounced in DC, and further exploration would help to determine if targeting energy metabolism could suppress the hypertrophy.

The differences between DC and IC were further highlighted when clustering genes based on their involvement in cellular processes. Myocardium remodelling, ventricular cardiac muscle tissue morphogenesis and muscle tissue development as

well as other tissue structure- and integrity-related processes were enriched for DC (Figs 3 and 4). With a further refinement—uniquely and significantly changed genes showed a specific clustering under microtubule, myofibril, sarcomere and contractile fibre process group (Figs 3 and 4). Some of those genes, MYL1, DNAH6, MYOZ1 and ACKR2, could be of a special interest as potential therapeutic targets or biomarkers because of their reported roles in heart muscle function. For example, MYL1, DNAH6 and MYOZ1 were named in various reports linking them to hypertrophy, changes in contractility and myocardium cell function [30, 44, 75]. Tissue overgrowth mediated by these genes could be targeted to reduce the excessive strain on the myocardium in the early stages of the disease development. ACKR2 has been demonstrated to reduce inflammation and vascular remodelling after myocardium injury, and this identified upregulation might indicate the compensatory mechanism for the tissue remodelling [30, 53]. Thus, enhancing or stimulating this protective signalling might be a valuable therapeutic option (Supplementary Table S3).

In contrast to hypertrophic heart muscle, ischemic heart-enriched gene networks had clear links to ER stress; for example, SMPD3, TMEM259, epidermal growth factor (EGF) and APOB have been shown to lead to ER stress when their normal function is perturbed [30, 51, 60, 72]. In addition, FPR2 as well as CX3CL1 interlink inflammatory processes with higher ER protein turnover burden [48, 52] (Supplementary Fig S6, Table S4 and Figure S10). Under ischemic conditions, perturbations in oxygen and nutrient supply as well as undergoing cellular stress can lead to mitochondrial and proteome stability changes which likely propagate fibrotic remodelling events [46, 69]. Thus, pharmaceutical management of poor tissue oxygenation and inflammation could be a useful therapeutic approach to limit tissue injury.

It was intriguing to find that the levels of chemokine ligands (e.g. CXCL11, CXCL10 and CCL5), chemokine receptors (e.g. CXCR3 and CCR7) as well as other markers, such as CD2, were significantly changed under myocardial ischemia (Supplementary Table S4). This, however, might be likely attributable to the T-cells and other lymphoid cells infiltrating heart tissue as can also be seen in a mouse left ventricle non-myocardial cellulose study (Supplementary Figs S15 and S16). A significant proportion of fibroblast and fibroblast-like cells can also be found in a healthy human heart (Supplementary Fig. S18; Table 1), and this population, under myocardial stress conditions, can change its proportions further by propagating pro-inflammatory and pro-fibrotic environment. Moreover, normal subpopulations of immune cells identified in the heart, such as monocytes, macrophages, mast cells, eosinophils, neutrophils B cells and T-cells, can also be activated and lead to a pro-inflammatory state [53, 75]. These considerations need to be taken into account when analysing data at different resolution levels where different types of cells can show a varying degree of contribution in the bulk transcriptome.

This was especially evident when juxtaposing enriched protein groups to the corresponding gene values from the RNA-seq studies. As proteome data had about six times lower recovery than bulk RNA-seq (Fig. 3; Supplementary Fig. S14), it became clear that it is only possible to identify genes and their networks that are above the detection level and show substantial abundance. Despite these limitations, important marker molecules associated with DC were found; that is, NPPA, AEBP1, MFAP4 and COL14A1 were upregulated both on the gene and protein levels. All of these genes and their readouts on a protein level could potentially be used as biomarkers since there is experimental and clinical evidence for their role in the dilated left ventricle

remodelling [30, 61, 76, 77]. MFAP4 was also matched to the mice heart cellulome fibroblast markers; this target is quite interesting as it reoccurred in all three types of 'omics' datasets and it not only has several strong links to the heart hypertrophy but has also been investigated as a potential therapeutic target [61].

In the case of downregulated genes and their proteins, MYH6 and ART3 form a unique group; while MYH6 mutations are linked to hereditary cardiomyopathies [30, 71, 77], ART3 function remains to be defined, but it was found in cardiac proteome profiling [30]. The present study also identified ART3 as a pericyte/cardiomyocyte marker from a single cell study for mouse heart cellulome with most myocytes removed. This not only confirms that ART3 expression allows it to be associated with cardiomyocytes and differentiated from other cells but could also suggest that the reversal of this downregulation might be a new therapeutic opportunity. In addition, TTN had a contrasting pattern where gene expression levels were decreased and the protein expression was upregulated (Supplementary Fig. S14). TTN mutations are well-documented for DC; however, while mutated and truncated TTN proteins lead to the disease parthenogenesis, the higher expression role is not clear [30, 65]. In addition, TTN was also found to be of low expression under myoblast/myocytes group in the human left ventricle single cell RNA-seq (Table 1). It is possible to hypothesize that as the heart muscle remodelling progresses, some of the compensatory mechanisms might increase the contractile fibre and associated protein production; at the same time, RNA expression levels drop by secondary regulatory mechanisms to reduce the protein production burden. More in-depth experimental studies investigating TTN and its expression dynamics are necessary to understand whether there is any prognostic or therapeutic value.

Ischemic heart transcriptome and proteome (Supplementary Tables S7 and S8; Supplementary Fig. S14) overlap only showed the enrichment for heart muscle hypertrophy, regulation of the heart rate as well as contraction force (Fig. 4B and D). For example, while MYH7, COX8A, COQ10B had a slightly increased gene expression, protein expression values were markedly suppressed. MYH7 is a well-known driver of cardiac tissue hypertrophy [30, 56, 77]; thus, lack of nutrients reaching heart might prevent tissue growth and dampen related pathways. Moreover, decrease in COX8A might be a protective mechanism to reduce oxidative metabolism [30]. However, other perturbations, such as the loss of COQ10B ensuring mitochondrial integrity [57] likely overcome measures against oxidative stress, leading to the ischemic tissue injury propagation (Supplementary Tables S7 and S8; Supplementary Fig. S14). Other gene products, namely, APOB, RBP4 and KNG1, playing the role in the HF were overexpressed despite reduced mRNA levels, which could give a glimpse into the perturbed energy metabolism and tissue blood perfusion [30, 58–60]. This sharp contrast could hint towards potential therapeutic avenues to inhibit RBP4-based signalling and APOB-induced ER stress that are likely contributing to further tissue injury and remodelling. While the heart left ventricle proteome did not capture strong immune associations as previously shown in bulk RNA-seq, it is noteworthy that lipid metabolism-associated proteins had a clear presence (Supplementary Tables S7 and S8; Supplementary Fig. S14). Moreover, cardiomyocytes are not a homogenous group of cells as can be seen in bulk and single cell RNA-seq, and this is also true on the protein level where a small subset of proteins show variability between cardiomyocytes in a mosaic pattern and can likely be further altered under pathological conditions

[30, 74, 77]. This analytical direction of comparing the RNA-seq and proteome set overlap could be further developed in the future studies to increase the analysis resolution; that is, a potential next avenue of such an analysis could be establishing the significance of the overlap to assess how the differentially expressed gene levels translate to the protein expression. This kind of evaluation could be tested by performing additional statistical tests to capture what LFC as well as P.adj value thresholds lead to the most significant overlap.

In parallel, all of the above findings were also compared to the human left ventricle single cell RNA-seq. While the majority of cells were mostly cardiomyocytes and other muscle tissue cells, more than a quarter was comprised of various immune cells (Table 1, Supplementary Figs S18–S21). One of the most interesting findings was a matched significant upregulation between bulk and single cell RNA-seq as well as the proteome data that returned LUM and HBB genes for the ischemic heart conditions. Experiments with LUM demonstrated its ability to increase the levels of lysyl oxidase, collagen type I alpha 2 and transforming growth factor- $\beta$ 1 and to decrease the activity of the collagen-degrading enzyme matrix metalloproteinase-9; thus, these profibrotic events are associated with a higher potential for HF [30, 64]. Targeting LUM might help control the fibrotic tissue transformation in the heart and it could also be used as a prognostic marker. Yet, the expression dynamics of LUM are not entirely clear as more recent reports indicate that LUM might be involved in compensatory and counterbalancing functions during active HF [78]. Such contradictions reaffirm the complexities of the underlying pathology mechanisms, and further research is needed to understand at what HF stages these expression changes occur and when it is best to have a pharmacological intervention. Furthermore, alpha subunit expression of the globins has been implicated in vascular tone and function maintenance [30, 79]; thus, it is possible that beta subunit expression might be involved in similar compensatory mechanisms for the tissue undergoing ischemic stress, but again, further research would help to establish the therapeutic potential of HBB and other globins.

Another interesting candidate target was identified for the DC, namely, NR4A1. This orphan receptor showed a marked upregulation in the hypertrophic state, while in a healthy left ventricle, its expression remained low (Table 1). NR4A1 has recently emerged as one of the key players in cardiac stress responses and hypertrophic growth [62]. Also, hypertrophic tissue state showed a marked loss in adipocyte signatures (Table 1), and some of the downregulated genes followed similar patterns in the ischemic heart observed earlier—notably, FABP4 and glycerol-3-phosphate dehydrogenase 1 (GPD1). FABP4 has been suggested to influence the cardiac size and myocardial function under pathological states, and it might indicate changes in the energy metabolism of the heart [63]. GPD1 is known to play a role in oxidative stress responses as well as affect lipid and amino acid metabolism [30], and it is possible that the observed downward shift in GDP1 expression is a compensatory mechanism and could be a valuable marker.

All of these observations clearly delineate the need to appreciate the different levels of 'omics' datasets. While bulk and single cell transcriptome as well as proteome analyses [1, 71, 74] provide us with varying degrees of resolution in cases of complex tissue and more so, in cases of wide spectrum pathologies, it might become difficult to integrate such variable datasets. Thus, to address the main challenge of biological data integration, a scoring system that would take the advantage of bulk RNA-seq

data and match with the data mined from multiple resources for each gene was devised and introduced in this study. As demonstrated earlier, the richest biological data are still only available from bulk RNA-seq experiments, and all other resources, such as proteomics or single cell RNA-seq, had only a very small overlap with the genes identified from bulk sequencing (Supplementary Tables S5–S8); moreover, regular RNA-seq is still a more universal research choice to untangle transcriptional profiles. As a result, a scoring system was used to capture the level of gene expression change (LFC) along with any mined disease associations for that gene so that it was possible to supply this information to machine learning pipelines and group existing data points to predict biologically meaningful gene expression patterns. Specifically, our devised LFCscore method allows to evaluate how a gene participates in the network and to what extent it can cause a perturbation if the gene function is disrupted. Such grouping is the first step to integrate LFC, differentially expressed genes and protein–protein interactions when recreating a signalling network. This could be especially useful if researchers enriched the scoring with additional weights to add new information for the clustering.

Our two-step machine learning approach returned multiple subgroups which showed similar multi-profile characteristics; for example, SMAD7, STX1B and SOX17 not only belonged to the same sub-cluster [31] (Supplementary Fig. S15) but are also a part of a complex network regulating tissue morphogenesis, vesicle docking and growth in DC significantly changed genes [30, 77]. Similarly, TNFRSF11B, PIM2 converged via MYC in the network [30] linked to angiogenesis and anti-apoptotic pro-growth effects in IC group (Supplementary Fig. S15). While genes belonging to the same cluster hint towards interesting target candidates, they are not necessarily direct interactors, and the observed degree of separation for these genes could be useful when building network models or using them as seed points to predict the extent of local network perturbations. Moreover, newly identified genes that could potentially be important candidates in driving the pathological state can be found using aggregated results where we can explore how these genes are associated with more dominant genetic factors. This strategy can be very useful when selecting genes for downstream screening studies and when prioritizing new targets. Another application of this method is to cluster genes that are shared between two pathologies. As we demonstrated in this study, gene subsets that show similar profiles in different conditions can be further clustered and assessed. For example, WNT9A, HBB and FBXL16 were both clustered to the same group for dilated and IC when a pool of 160 shared genes was subjected to agglomerative hierarchical clustering. This could be very useful in understanding the potential convergence points for diseases, establishing shared expression patterns and selecting therapeutic targets that are substantially unique.

To further verify the scoring and machine learning method, all of the identified clusters were extensively cross-referenced with the GWAS dataset of human heart disease genetic variants [39], clinical/experimental evidence from Open Targets platform [30] as well as complete PubMed [40] records for any cardiovascular pathologies. This analysis revealed that the proposed method allows to juxtapose rarer or newly discovered targets with more known genes linked to the DCs and ICs (Supplementary Tables S10 and S11). The extensive search on disease parameters (i.e. GWAS association, Open Targets knowledge-base association, PubMed records) allowed to capture genes with different levels of information. Moreover, the same trend was verified for two additional datasets of cardiopathologies where genes with

parser or even no known links to the disease belonged to clusters with better-defined members. This strategy could lead to the identification of new biomarkers or a better understanding of their function because the proposed analysis is based on the gene interaction network complexity and expression. Most importantly, researchers can adjust the scoring system based on their in-house data and known associations to perform a more focused analysis prior to selecting targets for downstream screens and to avoid selecting groups of genes that belong to the same effector network.

Current strategies to treat the HF mainly target symptoms based on the left ventricle dysfunction severity. There is a notable lack of systemic 'omics' studies for an in-depth analysis of heterogeneous disease mechanisms. This study, for the first time, demonstrated how bulk and single cell RNA-seq as well as the proteomics analysis of the human heart tissue can be integrated to uncover HF-specific networks and potential therapeutic targets or biomarkers for DCs and ICs. Thus, we showed that despite a smaller number of samples which is often the case in some pre-clinical settings or smaller-scale studies, it is possible to discover new therapeutically relevant insights. Moreover, by applying the novel scoring system and machine learning methods, we can untangle complex expression profiles to elucidate gene clusters that can be selected for downstream analyses. This study could be the first step towards a more systematic analysis that could be freely shared among researchers. Finally, it was demonstrated that cardiopathology treatment can go beyond symptom management and that there are indeed distinct gene network and pathway profiles that could be of therapeutic interest.

## SUPPLEMENTARY DATA

Supplementary data are available at *INTBIO Journal* online.

## AUTHORS' CONTRIBUTIONS

A.K. devised the methodology, performed the analyses and wrote the manuscript; N.B. provided the critical review, medical perspective and suggestions for the manuscript.

## CONFLICT OF INTEREST

None declared.

## DATA AVAILABILITY

All data are freely available and information is provided in the Methods section.

## REFERENCES

1. Sweet ME, Cociolo A, Slavov D et al. Transcriptome analysis of human heart failure reveals dysregulated cell adhesion in dilated cardiomyopathy and activated immune pathways in ischemic heart failure. *BMC Genomics* 2018;19:812. doi: 10.1186/s12864-018-5213-9.
2. Packer M. The imminent demise of cardiovascular drug development. *JAMA Cardiol* 2017;2:1293–4.
3. Malik A, Brito D, Chhabra L. Congestive Heart Failure. [Updated 2021 Feb 11]. In: *StatPearls [Internet]*. Treasure Island (FL): StatPearls Publishing, 2021. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK430873/>



4. Metra M, Teerlink JR. Heart failure. *Lancet* 2017;**390**:1981–95.
5. Christopher JLM, Vos T, Lopez AD et al. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the global burden of disease study 2015. *Lancet* 2016;**388**:1545–602.
6. Strömberg A, Mårtensson J. Gender differences in patients with heart failure. *Eur J Cardiovasc Nurs* 2003;**2**:7–18.
7. Bowles NE, Bowles KR, Towbin JA. The ‘final common pathway’ hypothesis and inherited cardiovascular disease: the role of cytoskeletal proteins in dilated cardiomyopathy. *Herz* 2000;**25**:168–75.
8. Fordyce CB, Roe MT, Ahmad T et al. Cardiovascular drug development: is it dead or just hibernating? *J Am Coll Cardiol* 2015;**65**:1567–82. doi: [10.1016/j.jacc.2015.03.016](https://doi.org/10.1016/j.jacc.2015.03.016).
9. Li CI, Samuels DC, Zhao YY et al. Power and sample size calculations for high-throughput sequencing-based experiments. *Brief Bioinform* 2017;**19**:1247–55.
10. Fumagalli M. Assessing the effect of sequencing depth and sample size in population genetics inferences. *PLoS One* 2013;**8**:79667.
11. Hart SN, Therneau TM, Zhang Y et al. Calculating sample size estimates for RNA sequencing data. *J Comput Biol* 2013;**20**:970–8.
12. PRIDE-Proteomics Identification Database. <https://www.ebi.ac.uk/pride/archive/projects/PXD008934>.
13. PRIDE-Proteomics Identification Database. <https://www.ebi.ac.uk/pride/>.
14. Array Express < EMBL-EBI. <https://www.ebi.ac.uk/arrayexpress/>.
15. Datasets-Spatial Gene Expression-Official 10x Genomics Support. <https://support.10xgenomics.com/spatial-gene-expression/datasets/>.
16. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;**30**:2114–20.
17. GRCh37- hg 19- Genome-Assembly-NCBI. [https://www.ncbi.nlm.nih.gov/assembly/GCF\\_000001405.13/](https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.13/).
18. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* 2015;**12**:357–60.
19. Liao Y, Smyth GK, Shi W. Feature counts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 2014;**30**:923–30.
20. Multi QC. <https://multiqc.info/>.
21. RStudio | Open Source & Professional Software for Data Science Teams - RStudio. <https://rstudio.com/>.
22. Bioconductor - DESeq2. <http://bioconductor.org/packages/release/bioc/html/DESeq2.html>.
23. Seurat. <https://satijalab.org/seurat/>.
24. Bioconductor-Single R. <http://bioconductor.org/packages/release/bioc/html/Single R.html>.
25. Bioconductor - CellDex. <https://bioconductor.org/packages/release/data/experiment/html/celldex.html>.
26. Introduction to Clustermole. <https://cran.r-project.org/web/packages/clustermole/vignettes/clustermole-intro.html>.
27. Max Quant. <https://www.maxquant.org/>.
28. Bioconductor - Cluster Profiler. <https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html>.
29. Bioconductor-DEGreport. <https://www.bioconductor.org/packages/release/bioc/html/DEGreport.html>.
30. Home-Open Targets. <https://www.opentargets.org/>.
31. STRING: Functional Protein Association Networks. <https://string-db.org/>.
32. Szklarczyk D, Gable AL, Lyon D et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* 2019;**47**:D607–13.
33. Reynolds D. Gaussian mixture models. In: *Encyclopedia of Biometrics*. US: Springer US, 2009, 659–63. doi:[10.1007/978-0-387-73003-5\\_196](https://doi.org/10.1007/978-0-387-73003-5_196).
34. Murtagh F, Legendre P. Ward’s hierarchical agglomerative clustering method: which algorithms implement Ward’s criterion? *J Classif* 2014;**31**:274–95.
35. Leonard K, Peter JR. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, Inc., 1990. doi:[10.1002/9780470316801](https://doi.org/10.1002/9780470316801).
36. Bholowalia P, Kumar A. EBK-means: a clustering technique based on elbow method and K-means in WSN. *International Journal of Computer Applications* 2014;**105**:17–24.
37. Vrieze SI. Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion BIC. *Psychol Methods* 2012;**17**:228–43.
38. scikit-learn: Machine Learning in Python — scikit-learn 0.23.2 Documentation. <https://scikit-learn.org/stable/>.
39. Li MJ, Wang P, Liu X et al. GWASdb: a database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res* 2012;**40**:D1047–54. doi: [10.1093/nar/gkr1182](https://doi.org/10.1093/nar/gkr1182).
40. PubMed. <https://pubmed.ncbi.nlm.nih.gov/>.
41. GEO Accession Viewer. <https://metadataplus.biothings.io/geo/GSE116250>.
42. Welcome to Python.org. <https://www.python.org/>.
43. Bohnsack KE, Bohnsack MT. Uncovering the assembly pathway of human ribosomes and its emerging links to disease. *EMBO J* 2019;**38**:e100278. doi: [10.15252/emboj.2018100278](https://doi.org/10.15252/emboj.2018100278).
44. Arola AM, Sanchez X, Murphy RT et al. Mutations in PDLIM3 and MYOZ1 encoding myocyte Z line proteins are infrequently found in idiopathic dilated cardiomyopathy. *Mol Genet Metab* 2007;**90**:435–40.
45. Rivera-Feliciano J, Tabin CJ. Bmp 2 instructs cardiac progenitors to form the heart-valve-inducing field. *Dev Biol* 2006;**295**:580–8.
46. Pletsch-Borba L, Grafetstätter M, Hüsing A et al. Vascular injury biomarkers and stroke risk: a population-based study. *Neurology* 2020;**94**:e2337–45.
47. Aspatwar A, Tolvanen MEE, Parkkila S. Phylogeny and expression of carbonic anhydrase-related proteins. *BMC Mol Biol* 2010;**11**:1–19.
48. Altin SE, Schulze PC. Fractalkine: a novel cardiac chemokine? *Cardiovasc Res* 2011;**92**:361–2.
49. Griffin J, Emery BR, Christensen GL et al. Analysis of the meiotic recombination gene REC8 for sequence variations in a population with severe male factor infertility. *Syst Biol Reprod Med* 2008;**54**:163–5.
50. Jinn S, Drolet RE, Cramer PE et al. TMEM175 deficiency impairs lysosomal and mitochondrial function and increases  $\alpha$ -synuclein aggregation. *Proc Natl Acad Sci U S A* 2017;**114**:2389–94.
51. Stoffel W, Hammels I, Jenke B et al. Neutral sphingomyelinase (SMPD3) deficiency disrupts the golgi secretory pathway and causes growth inhibition. *Cell Death Dis* 2016;**7**:e2488–8.
52. Liu H, Lin Z, Ma Y. Suppression of Fpr 2 expression protects against endotoxin-induced acute lung injury by interacting with Nrf 2-regulated TAK1 activation. *Biomed Pharmacother* 2020;**125**:109943.

53. Cochain C, Auvynet C, Poupel L et al. The chemokine decoy receptor D6 prevents excessive inflammation and adverse ventricular remodeling after myocardial infarction. *Arterioscler Thromb Vasc Biol* 2012;**32**: 2206–13.
54. CD2- An Overview | Science Direct Topics. <https://www.science-direct.com/topics/neuroscience/cd2>.
55. De Sousa Abreu R, Penalva LO, Marcotte EM et al. Global signatures of protein and mRNA expression levels. *Molecular Bio Systems* 2009;**5**:1512–26.
56. Bollen IAE, van der Velden J. The contribution of mutations in MYH7 to the onset of cardiomyopathy. *Netherlands Heart Journal* 2017;**25**:653–4.
57. Zhong X, Yi X, da Silveira E Sá RC et al. CoQ10 deficiency may indicate mitochondrial dysfunction in Cr(VI) toxicity. *Int J Mol Sci* 2017;**18**:816. doi: [10.3390/ijms18040816](https://doi.org/10.3390/ijms18040816).
58. Bellei E, Bergamini S, Monari E et al. Evaluation of potential cardiovascular risk protein biomarkers in high severity restless legs syndrome. *J Neural Transm* 2019;**126**: 1313–20.
59. Li X, Zhang KZ, Yan JJ et al. Serum retinol-binding protein 4 as a predictor of cardiovascular events in elderly patients with chronic heart failure. *ESC Heart Fail* 2020;**7**: 542–50.
60. Su Q, Tsai J, Xu E et al. Apolipoprotein B100 acts as a molecular link between lipid-induced endoplasmic reticulum stress and hepatic insulin resistance. *Hepatology* 2009;**50**: 77–84.
61. Wang HB, Yang J, Shuai W et al. Deletion of microfibrillar-associated protein 4 attenuates left ventricular remodeling and dysfunction in heart failure. *J Am Heart Assoc* 2020;**9**:e015307. doi: [10.1161/JAHA.119.015307](https://doi.org/10.1161/JAHA.119.015307).
62. Zhao Y, Bruemmer D. NR4A orphan nuclear receptors: transcriptional regulators of gene expression in metabolism and vascular biology. *Arterioscler Thromb Vasc Biol* 2010;**30**: 1535–41.
63. Rodríguez-Calvo R, Girona J, Alegret JM et al. Role of the fatty acid-binding protein 4 in heart failure and cardiovascular disease. *J Endocrinol* 2017;**233**:R173–84.
64. Engebretsen KVT, Lunde IG, Strand ME et al. Lumican is increased in experimental and clinical heart failure, and its production by cardiac fibroblasts is induced by mechanical and proinflammatory stimuli. *FEBS J* 2013;**280**: 2382–98.
65. Tharp CA, Haywood ME, Sbaizero O et al. The giant protein Titin's role in cardiomyopathy: genetic, transcriptional, and post-translational modifications of TTN and their contribution to cardiac disease. *Front Physiol* 2019;**10**: 1436.
66. Schmidtke P, Barril X. Understanding and predicting drug-gability. A high-throughput method for detection of drug binding sites. *J Med Chem* 2010;**53**:5858–67.
67. Barth AS, Kuner R, Bunes A et al. Identification of a common gene expression signature in dilated cardiomyopathy across independent microarray studies. *J Am Coll Cardiol* 2006;**48**:1610–7.
68. Greco S, Fasanaro P, Castelvechio S et al. MicroRNA dysregulation in diabetic ischemic heart failure patients. *Diabetes* 2012;**61**:1633–41.
69. Bhandari B, Quintanilla Rodriguez BS, Masood W. *Ischemic Cardiomyopathy*. [Updated 2020 Sep 17]. In: *StatPearls [Internet]*. Treasure Island (FL): StatPearls Publishing, 2021. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK537301/>.
70. Lakdawala NK, Winterfield JR, Funke BH. Dilated cardiomyopathy. *Circ Arrhythm Electrophysiol* 2013;**6**:228–37.
71. Alimadadi A, Munroe PB, Joe B et al. Meta-analysis of dilated cardiomyopathy using cardiac RNA-Seq transcriptomic datasets. *Genes (Basel)* 2020;**11**:60. doi: [10.3390/genes11010060](https://doi.org/10.3390/genes11010060).
72. Zhang M, Lu H, Xie X et al. TMEM175 mediates lysosomal function and participates in neuronal injury induced by cerebral ischemia-reperfusion. *Mol Brain* 2020;**13**:113.
73. Iskratsch T, Yu CH, Mathur A et al. FHOD1 is needed for directed forces and adhesion maturation during cell spreading and migration. *Dev Cell* 2013;**27**:545–59.
74. Wang TY, Lee D, Fox-Talbot K et al. Cardiomyocytes have mosaic patterns of protein expression. *Cardiovasc Pathol* 2018;**34**:50–7.
75. Strassheim D, Dempsey EC, Gerasimovskaya E et al. Role of inflammatory cell subtypes in heart failure. *J Immunol Res* 2019;**2019**:2164017. doi: [10.1155/2019/2164017](https://doi.org/10.1155/2019/2164017).
76. Tarazón E, Roselló-Lletí E, Rivera M et al. RNA sequencing analysis and atrial natriuretic peptide production in patients with dilated and ischemic cardiomyopathy. *PLoS One* 2014;**9**:e90157. doi: [10.1371/journal.pone.0090157](https://doi.org/10.1371/journal.pone.0090157).
77. Witt E, Hammer E, Dörr M et al. Correlation of gene expression and clinical parameters identifies a set of genes reflecting LV systolic dysfunction and morphological alterations. *Physiol Genomics* 2019;**51**:356–67.
78. Mohammadzadeh N, Lunde IG, Andenæs K et al. The extracellular matrix proteoglycan lumican improves survival and counteracts cardiac dilatation and failure in mice subjected to pressure overload. *Sci Rep* 2019;**9**:9206. <https://doi.org/10.1038/s41598-019-45651-9>.
79. Sangwung P, Zhou G, Lu Y et al. Regulation of endothelial hemoglobin alpha expression by Kruppel-like factors. *Vasc Med* 2017;**22**:363–9. doi: [10.1177/1358863X17722211](https://doi.org/10.1177/1358863X17722211).

## Integrative *omics* approaches for new target identification and therapeutics development

### 3. *OmicInt* package: exploring *omics* data and regulatory networks using integrative analyses and machine learning

The experimental chapter is based on the published software package and publication in preparation

1. [Kanapeckaitė A. OmicInt: Omics Network Exploration. CRAN. 2021 Oct. 15. Version 1.1.7; https://cran.r-project.org/web/packages/OmicInt/index.html](https://cran.r-project.org/web/packages/OmicInt/index.html)
2. [Kanapeckaitė A. \*OmicInt\* package: exploring \*omics\* data and regulatory networks using integrative analyses and machine learning. \*Accepted and in preparation.\*](#)

#### Conclusion of this chapter

My developed *OmicInt* package provides a unique combination of functions and tools for researchers to explore gene expression data sets. A special focus of the package is also making machine learning, specifically Gaussian mixture models, more accessible to the researchers that do not have a background in the ML/AI field. In addition, advanced functions for epigenomics analysis permit the exploration of the epigenetic regulatory layer. This might be helpful when identifying genes that may depend on the epigenetic regulation. Specifically, if a CpG island containing gene changed expression during treatment or disease progression, this might indicate a dependence on the epigenetic regulation. Similarly, exploring a gene's miRNA network could hint at other interacting genes which might not have been picked up by the differential expression analysis. Exploring miRNA networks could also help prepare for RNA interference studies. Moreover, miRNA interactome analysis provides the first in-depth look into what genes are controlled by the same set of miRNAs. Thus, *OmicInt* offers a comprehensive, evolving, and adaptable platform for gene expression analysis in the context of the transcriptome, proteome, and epigenome.

#### Contribution to this chapter (100%)

- Methodology development which included equation and scoring function derivation as well as machine learning pipeline creation.
- Developed new programmatic features to accompany the related publication.
- Performed software package development and testing.
- Conceptualised and wrote the documentation files, vignettes, and manuscript, including the figure preparation.
- Corresponding author and maintainer.



## Methods &amp; Protocols

## *OmicInt* package: Exploring *omics* data and regulatory networks using integrative analyses and machine learning

Auste Kanapeckaite<sup>a,b</sup><sup>a</sup> Algorithm379, Laisvės g. 7, Vilnius, Lithuania<sup>b</sup> University of Reading, School of Pharmacy, Hopkins Building, Reading RG6 6UB United Kingdom

## ARTICLE INFO

## Keywords:

Machine learning  
Gaussian mixture models  
Omics analyses  
Epigenomics  
Gene networks

## ABSTRACT

*OmicInt* is an R software package developed for a user-friendly and in-depth exploration of significantly changed genes, gene expression patterns, and the associated epigenetic features as well as the related miRNA environment. In addition, *OmicInt* offers single cell RNA-seq and proteomics data integration to elucidate specific expression profiles. To achieve this, *OmicInt* builds on a novel scoring function capturing expression and pathology associations. The developed scoring function together with the implemented Gaussian mixture modelling pipeline helps to explore genes and the linked interactome networks. The machine learning pipeline was designed to make the analyses straightforward for the non-experts so that researchers could take advantage of advanced analytics for their data evaluation. Additional functionalities, such as protein type and cellular location classification, provide useful assessments of the key interactors. The introduced package can aid in studying specific gene networks, understanding cellular perturbation events, and exploring interactions that might not be easily detectable otherwise. Thus, this robust set of bioinformatics tools can be very beneficial in drug discovery and target evaluation. *OmicInt* is designed to be freely accessible to involve a larger bioinformatics community and continuously improve the developed algorithmic methods.

## 1. Introduction

*OmicInt* is an R software package developed for an in-depth exploration of significantly changed genes, gene expression patterns, and the associated epigenetic features as well as the related miRNA environment. The package helps to assess gene clusters based on their known interactors (proteome level) using several different resources, e.g., UniProt and STRING DB [1–3]. Moreover, *OmicInt* provides an easy Gaussian mixture modelling [4–6] pipeline for an integrative analysis that can be used by a non-expert to explore gene expression data. Specifically, the package builds on a previously developed method to explore gene networks using significantly changed genes, their log-fold-change values (LFC), and the predicted interactome complexity [5]. This approach can aid in studying specific gene networks, understanding cellular perturbation events, and exploring interactions that might not be easily detectable otherwise [5]. To this end, the package offers many different utilities to help researchers quickly explore their data in a user-friendly way where machine learning is made easily accessible to non-experts (Figs. 1 and 2). It is also important to highlight that the lack of freely available tools to explore complex expressome data motivated the creation of this set of tools. For example, commercial solutions, such as Clarivate analytics [7], are almost inaccessible to individual users because of the very expensive software. Freely available tools, namely

GeneMANIA or Cytoscape platforms [8–11], while very useful, do not permit machine learning applications or complex regulome integration. Thus, seeing the existing need for *omics* dedicated tools that could evolve as more bioinformaticians get involved encouraged creating the *OmicInt* package.

Machine learning which offer effective methods to assess multi-dimensional biological data is also a very important part of the developed package. For the purpose of biological data evaluation, Gaussian mixture models (GMMs) were selected as they employ a probability based classification where each data point assignment has a different probability of belonging to one of the clusters [4–6]. The probabilistic nature of GMM relies on the assumption that the data can be explained by a finite mixture of Gaussian distributions with unknown parameters [4]. As a result, this is a soft classification method that is more suitable to assess biological parameters in comparison to hard classification techniques (e.g., k-means) [4–6]. This is because gene or protein interaction networks are dynamic systems and probabilistic feature separation allows for more flexibility in defining boundaries between groups [5]. Moreover, the extracted probability values can be incorporated into other analytical pipelines to further refine the data. The developed GMM pipeline automates the assessment of the information criterion to optimize the number of clusters for modelling and also predicts the best suited model for the expectation-maximisation (EM) algorithm which

E-mail address: [auste.kan@algorithm379.com](mailto:auste.kan@algorithm379.com)

<https://doi.org/10.1016/j.ailsci.2021.100025>

Received 25 November 2021; Accepted 26 November 2021

Available online 11 December 2021

2667-3185/© 2021 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

Example table with only LFC values

	A	B	C
1	Symbol	log2FoldChange	pvalue
2	SAR1A	-2.18777269502012	2.65652091646361E-05
3	C6orf62	-2.6742131704395	1.6915673694844E-07
4	AXL	-2.78650785919511	0.000173959539412
5	BICC1	-3.59855326113771	0.00027388866015
6	CAPZA1	-1.73278403064529	0.000232183462116
7	TXNIP	1.46062912017625	7.34720482186151E-05
8	HNRNPH1	-1.81995372081358	0.000592319156385
9	RAB31	-1.79837938364367	0.00041973140141
10	UBE2B	-2.06138280667398	0.000125275768063
11	PAFAH1B2	-1.56007182816026	0.001391919840109
12	EIF2S3	-1.74298250448705	0.001063319998638
13	YWHAG	-1.55076880524874	0.000815638927099
14	ENAH	-1.69808760167615	0.000269611887499
15	PPP3CA	-2.67216823748962	3.98459567587323E-06

Example table with LFC, beta, and gamma values

	A	B	C	D	E
1	Symbol	log2FoldChange	pvalue	beta	gamma
2	SAR1A	-2.18777269502012	2.65652091646361E-05	0.25	0
3	C6orf62	-2.6742131704395	1.6915673694844E-07	0	0.3
4	AXL	-2.78650785919511	0.000173959539412	0	0.56
5	BICC1	-3.59855326113771	0.00027388866015	0	0
6	CAPZA1	-1.73278403064529	0.000232183462116	0.4	0
7	TXNIP	1.46062912017625	7.34720482186151E-05	0	0.75
8	HNRNPH1	-1.81995372081358	0.000592319156385	0	0
9	RAB31	-1.79837938364367	0.00041973140141	0	0.02
10	UBE2B	-2.06138280667398	0.000125275768063	0.41	0
11	PAFAH1B2	-1.56007182816026	0.001391919840109	0	0
12	EIF2S3	-1.74298250448705	0.001063319998638	0.7	0
13	YWHAG	-1.55076880524874	0.000815638927099	0.8	0
14	ENAH	-1.69808760167615	0.000269611887499	0	0.015
15	PPP3CA	-2.67216823748962	3.98459567587323E-06	0	0

Metadata file example

	A	B
1	Sample_ID	Condition
2	CAD1	hypertension
3	CAD2	hypertension
4	CAD3	hypertension
5	CAD4	hypertension
6	CAD5	hypertension
7	CAD6	hypertension
8	CAD10	hypertension
9	CAD11	hypertension
10	N10	healthy
11	N12	healthy
12	N13	healthy
13	N14	healthy
14	N15	healthy
15	RF2	CKD

Normalised count table example

	A	B	C	D	E	F
1	Symbol	CAD1	CAD2	CAD3	CAD4	CAD5
2	MT-CYB	9384.55930132127	11923.5039503911	34985.4747081763	4216.00298751307	12402.4413183367
3	MT-ND4	10934.2964538035	12360.328040959	24509.2381415289	5360.27179089838	12871.5303032926
4	MT-CO1	9722.28571644071	11516.7510834493	12963.1580920346	5587.18655703799	11594.0474618033
5	MT-CO3	8205.9264136993	10957.0042717434	17443.7581673424	5740.20990972336	13001.314416671
6	FN1	192.371263022342	249.538399563515	211.313998100786	748.609107229636	466.544198419069
7	COL1A2	138.536021504896	178.317080449193	135.124152051885	504.13857973424	331.670511967011
8	ACTB	777.02198590181	690.05544741877	1249.03562947764	1889.20954257112	2393.7958689793
9	MT-ATP6	6347.89277812717	7713.00507741775	11965.5224075336	3373.85049516584	7503.72723493652
10	MT-CO2	5702.58768313804	4504.35275998582	8569.36665664999	1882.92091163884	5782.60327385963
11	MT-ND1	6299.08215915135	9499.34127387032	10046.9718242743	4573.4068454974	11624.5849002453
12	MT-ND5	3404.89957517343	3910.3142020397	5409.21359962516	2197.61448454146	2744.12814888307
13	MALAT1	6320.61625575833	6583.48771279727	2876.10032088432	2196.56637938608	3348.09082029103
14	SAR1A	3043.84455539642	4617.25173991519	9943.43858399882	3869.86625994907	4469.49342085468
15	MT-ND4L	2144.79602205507	3601.68848587763	3118.47429106776	1752.16979350528	3407.46917281709

Fig. 1. Examples for the required data formats which include the normalised gene expression values, log fold change (LFC) values, and the meta data file.



Fig. 2. Schematic representation of package functions and specific analyses.

helps to maximise the likelihood of data point assignments [4,12]. As a result, the users do not need to have an extensive knowledge to fine-tune their GMM parameters as the process is streamlined for them.

The key analytical parameter in the machine learning pipeline and exploratory analyses is a specific score, namely  $LFC_{score}$ , which can have a different derivation depending on the selected parameters Eqs. (1)–

(3). The user has several options to select from since the equations were expanded with additional data based on the earlier derivation of the multi-omics Eq. (5). The score  $\alpha$  values are downloaded automatically from curated database images which were generated via text mining to retrieve, update, and integrate data in an easier-to-use format (i.e., database image) for the analyses. Databases used include Disgenet,

Uniprot, and STRING DB [1,3,13]. For example,  $\alpha_{\text{asoc}}$  score allows to infer how strongly a gene is linked to a disease or pathological phenotype ranging from 0 (no link) to 1 (the strongest association) Eq. (1) [13]. Similarly,  $\alpha_{\text{spec}}$  captures how specific a gene is when describing the pathology Eq. (2) [13]. Association scores are based on different curated resources as described earlier [13]. The user can choose from different types of scores (“association\_score”, “specificity\_score”, or the geometric mean of both) when selecting the type of the equation for  $\text{LFC}_{\text{score}}$ . Scores  $\beta_{\text{cell}}$  and  $\gamma_{\text{prot}}$  are the scaled values for single cell and proteome data, respectively. That is,  $\beta_{\text{cell}}$  has to be provided by the user if they have such experimental information integrated where a gene value from a single cell data cluster is extracted using a pseudo-bulk differential gene expression approach. The LFC scores from pseudo-bulk data need to be scaled according to the Eq. (4). The same approach should be applied when calculating  $\gamma_{\text{prot}}$  for protein (corresponding gene) values.

$$\text{LFC}_{\text{score}} = \text{LFC}(1 + \alpha_{\text{asoc}} + \beta_{\text{cell}} + \gamma_{\text{prot}}) \quad (1)$$

$\text{LFC}_{\text{score}}$  equation where LFC - Log Fold Change, base 2;  $\alpha_{\text{asoc}}$  - a disease association score;  $\beta_{\text{cell}}$  - scaled single cell LFC;  $\gamma_{\text{prot}}$  - scaled proteome LFC.

$$\text{LFC}_{\text{score}} = \text{LFC}(1 + \alpha_{\text{spec}} + \beta_{\text{cell}} + \gamma_{\text{prot}}) \quad (2)$$

$\text{LFC}_{\text{score}}$  equation where LFC - Log Fold Change, base 2;  $\alpha_{\text{spec}}$  - a disease specificity score;  $\beta_{\text{cell}}$  - scaled single cell LFC;  $\gamma_{\text{prot}}$  - scaled proteome LFC.

$$\text{LFC}_{\text{score}} = \text{LFC}\left(1 + \sqrt{(\alpha_{\text{asoc}} \alpha_{\text{spec}})} + \beta_{\text{cell}} + \gamma_{\text{prot}}\right) \quad (3)$$

$\text{LFC}_{\text{score}}$  equation where LFC - Log Fold Change, base 2;  $\alpha_{\text{asoc}}$  and  $\alpha_{\text{spec}}$  are integrated using a geometric average score;  $\beta_{\text{cell}}$  - scaled single cell LFC;  $\gamma_{\text{prot}}$  - scaled proteome LFC.

$$\text{LFC}_{\text{scaled}} = \text{LFC}_{\text{gene}} / \text{LFC}_{\text{median}} \quad (4)$$

$\beta_{\text{cell}}$  or  $\gamma_{\text{prot}}$  scaling example where  $\text{LFC}_{\text{gene}}$  - a gene specific value and  $\text{LFC}_{\text{median}}$  - a median value for all available LFC values per specific condition and gene set.

*OmicInt* provides many other valuable tools to map the interactome using information on the target cellular location or protein class/function type. In addition, density functions allow for an exhaustive assessment of gene distributions which may hint at potential functions or dominant processes within a specific condition. Epigenetic feature (CpG islands, GC%) and miRNA exploration tools also provide additional information on the epigenome and non-coding regulome which might be relevant for some genes and conditions, especially if a higher enrichment of these patterns can be found. Currently, the analyses are only available for human data sets. The software package is freely distributed via Github and CRAN repositories to make the analyses accessible to researchers [14,15]. Github environment also provides opportunities to submit requests or suggestions and participate in further algorithm development [14].

## 2. Methods

*OmicInt* package architecture (Fig. 2) is divided into gene expression, gene cluster/pattern, and epigenetic feature/regulatory network analysis with a detailed vignette to guide the user [14,15]. Machine learning pipeline is based on Gaussian mixture models which is designed to include the optimal cluster number (Bayesian information criterion), automatic model fitting during the expectation maximisation phase of clustering, model-based hierarchical clustering, as well as density estimation and discriminant analysis [4,12]. The package enables advanced options to perform a user-specified clustering to use the data in other workflows. *OmicInt* also retrieves data from multiple databases by generating combined and curated database images for easier use [1,3,13]. The package was built using functional programming principles and the analyses were benchmarked using the following studies distributed via NCBI GEO database [16]: GSE160145, GSE3585, GSE26887, and GSE116250.

## 3. Results

### 3.1. Data preprocessing

Before starting the analysis the user must ensure that the supplied data is in the right format. There are several different options to prepare a data frame (CSV format) that contains all the relevant experimental information Fig. 1; Eqs. (1)–(4). Depending on the selection, the downstream analyses will provide interactive graphs and maps (Fig. 2). Consistent data preparation and integration allow for a stable processing workflow which enables an efficient organisation of data sets.

Data pre-processing relies on the *score\_genes* function that collects data from the STRING database and other disease association data sets to scale and prepare additional score integration [3,13]. Several key parameters should be provided; the *data* parameter requires a data frame containing gene names as row names and a column with LFC values. The example is provided in Fig. 1; the parameter *alpha* ( $\alpha$ ) has a default value set as “association” which gives a score from 0 to 1 based on how strongly a gene is associated with a pathological phenotype; other options are “specificity” - to give values based on how specific a gene is when describing a disease and “geometric” - to give a geometric mean score of both association and specificity. The  $\alpha$  score is calculated automatically for the genes in the data set. In addition, it is possible to add weighted single cell and proteomics data by selecting additional parameters. The parameter *beta* is set to have a default value as FALSE; if TRUE, the user needs to supply data with a column *beta* that contains information on gene associations from single cell studies. Similarly, parameter *gamma* has a default value FALSE; if TRUE, the user is required to supply data with a column *gamma* that contains information on gene associations from proteome studies. The function returns a data frame for the downstream analyses.

```
#Code example for data preprocessing
#data<-score_genes("data.csv")
#head(data)

# Symbol Log2FoldChange pvalue Interactors Association_score
#1 SAR1A -2.187773 2.656521e-05 24 0.000000
#2 C6orf62 -2.674213 1.691567e-07 0 0.000000
#3 AXL -2.786508 1.739595e-04 2 0.3230769
#4 BICC1 -3.598553 2.738887e-04 3 0.3000000
#5 CAPZA1 -1.732784 2.321835e-04 66 0.3789474
#6 TXNIP 1.460629 7.347205e-05 30 0.3000000

# Specificity_score LFCscore
#1 0.000 -2.187773
#2 0.000 -2.674213
#3 0.590 -3.686764
#4 0.751 -4.678119
#5 0.601 -2.389418
#6 0.631 1.898818
```

### 3.2. Exploratory analyses

Function *density\_plot* plots a density plot for gene expression data prepared by the *score\_genes* function. The plots can be used for a quick assessment and summarisation of the overall parameters (Fig. 3). Specifically, the plots allow the evaluation of how key parameters, such as LFC,  $\text{LFC}_{\text{score}}$ , and disease association or specificity scores, associate with the highest frequency protein classes and cellular locations. For example, the most frequent protein classes may have specific distribution patterns hinting at predominant cellular processes. Similarly, examining distributions for cellular locations might highlight the most involved and/or affected cellular structures.

```
#An example of a function call to get density plots
#density_plot(data)
```

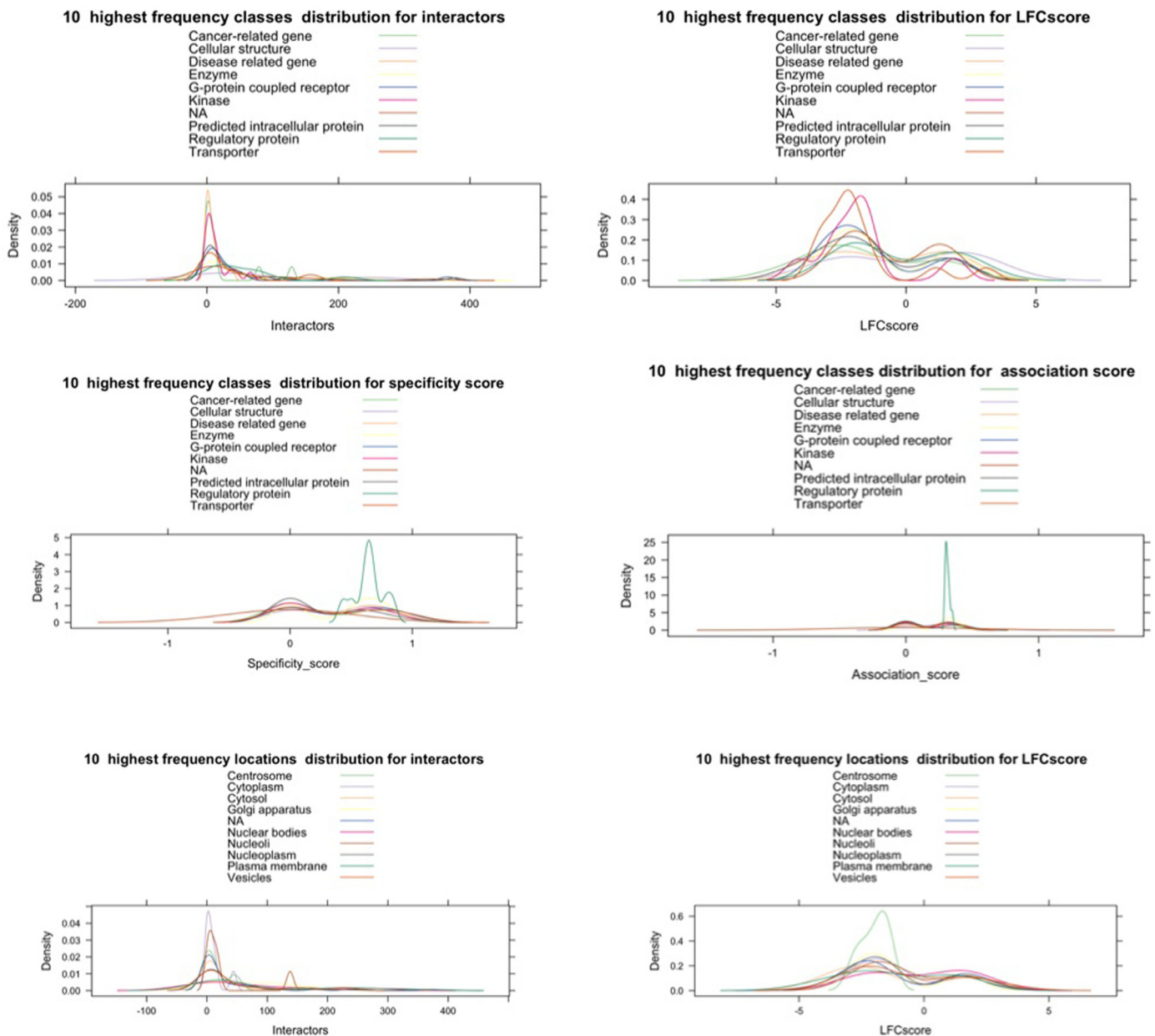


Fig. 3. Density plot examples for different parameters.

Function *feature\_distribution* also provides a way to visualise main feature distributions through density plots combined with LFC<sub>score</sub> and interactor number scatter plots (Fig. 4). These plots allow to quickly assess if there are any dependencies between LFC<sub>score</sub> and the interactor numbers. Such plots also help to see if any obvious gene clusters emerge. In early analyses this can aid in understanding whether the expression is dependent on any cellular site or protein class which could suggest a specific functional enrichment. This function might issue a warning if the data points were missing or too few for density plotting; however, it does not affect the overall visualisation.

```
#A simple call to implement the feature distribution analysis
#feature_distribution(data)
```

Function *plot\_3D\_distribution* allows to explore 3D distributions between the number of interactors, LFC<sub>score</sub>, and p.adj values. In addition to providing a data parameter, the user can select how to color data points depending on the association or specificity score (e.g., selecting “specificity”) (Fig. 5). This analysis can help identify specific clusters

for the expression patterns and interactors based on the significance of how the gene expression changed in a given condition. In addition, data point coloring based on gene association or specificity in the context of diseases can help capture additional patterns in the data.

```
#A function call example to explore the data in the interactive 3D plot
#plot_3D_distribution(data)
```

Function *class\_summary* provides analysis on main protein classes where a barplot helps to visualise the class distribution. Similarly, the function *location\_summary* summarises the location distribution data (Fig. 6). Assessing this information can highlight if there are any specific biases in data for target location or function which might indicate underlying cellular perturbations or changes in the function.

```
#Functions class_summary and Location_summary are called for the preprocessed data
frame

#class_summary(data)
#Location_summary(data)
```

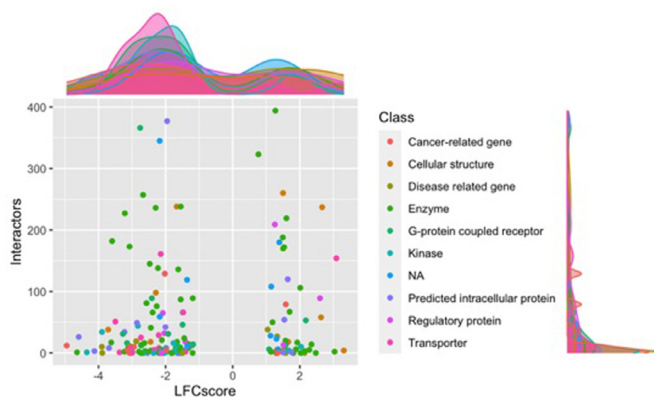
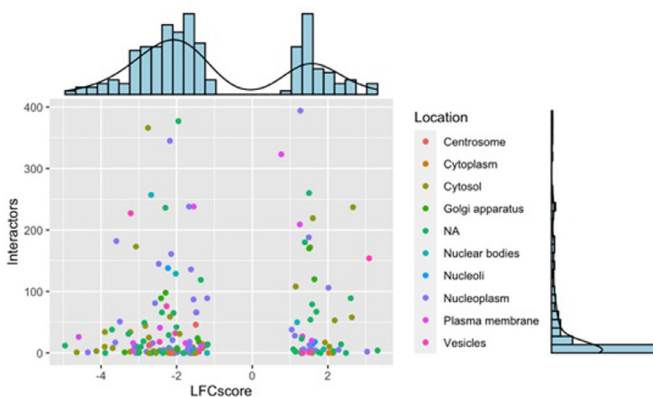
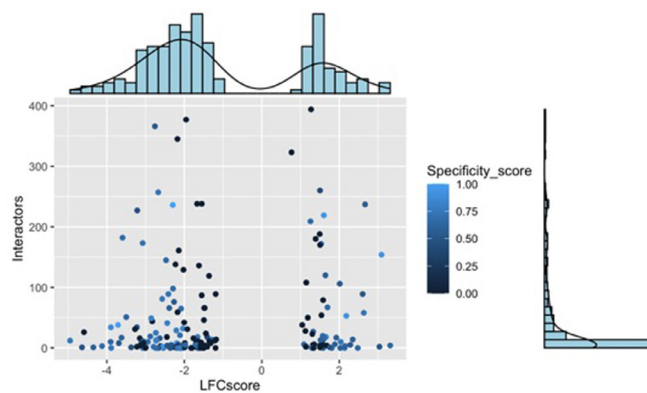
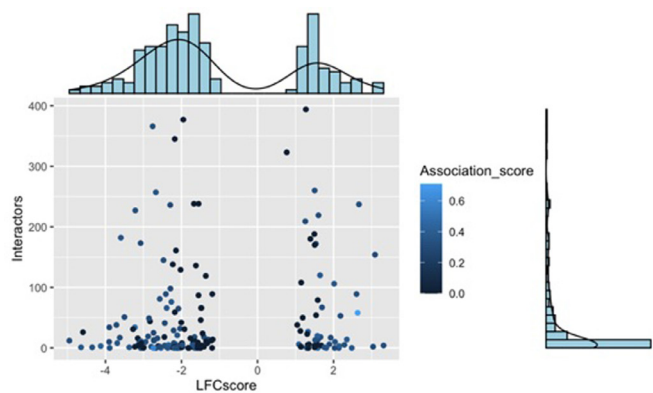
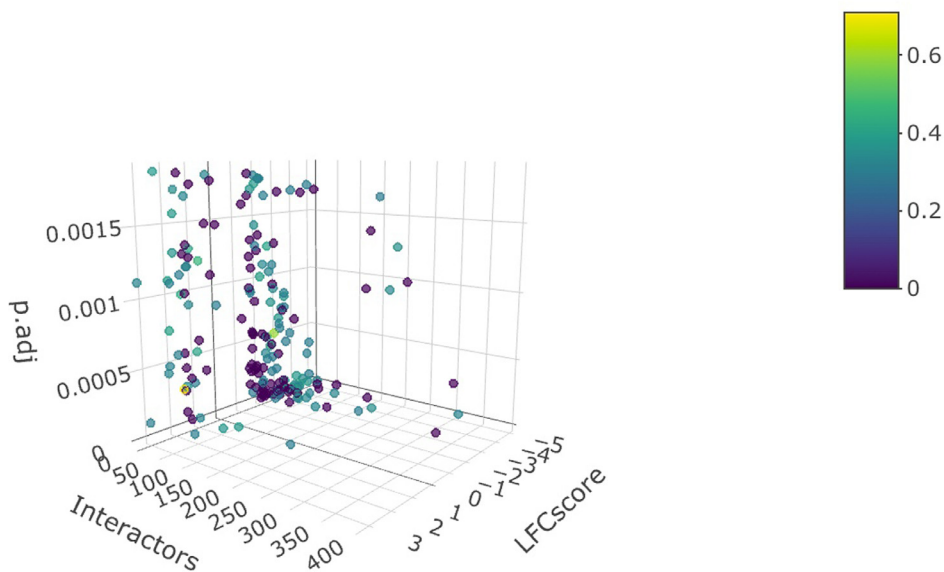


Fig. 4. Feature distribution plot examples.



Fig. 5. Interactive 3D feature distribution.



Function *location\_map* allows the visualisation of how the highest and lowest  $LFC_{score}$  genes cluster based on the protein cellular location data (Fig. 7). The user can specify the number of the top and lowest genes to consider. The function returns a dendrogram generated based on  $LFC_{score}$  values. The “euclidean” method is used for distance calculation and the “Ward.D2” method - for hclust generation. Gene labels are colored to indicate major clusters where the hclust generated cluster number is doubled to select for more subgroups. In addition, to achieve a finer separation of lower dendrogram branches the following equation is used to set the height for the color differentiation of different branches Eq. (5).

This equation takes the mean value for hclust function height calculation and multiplies by the dendrogram cluster number scaled twice. The plot also provides cellular location visualisation for each gene (Fig. 7).

$$H_{dendrogram} = (hclust_{height} / hclust_n) \cdot dendrogram_{cluster\_number} \cdot 2 \quad (5)$$

The height calculation for the color differentiation of different dendrogram branches.

```
#Location mapping is done using a single function call
#Location_map(data)
```



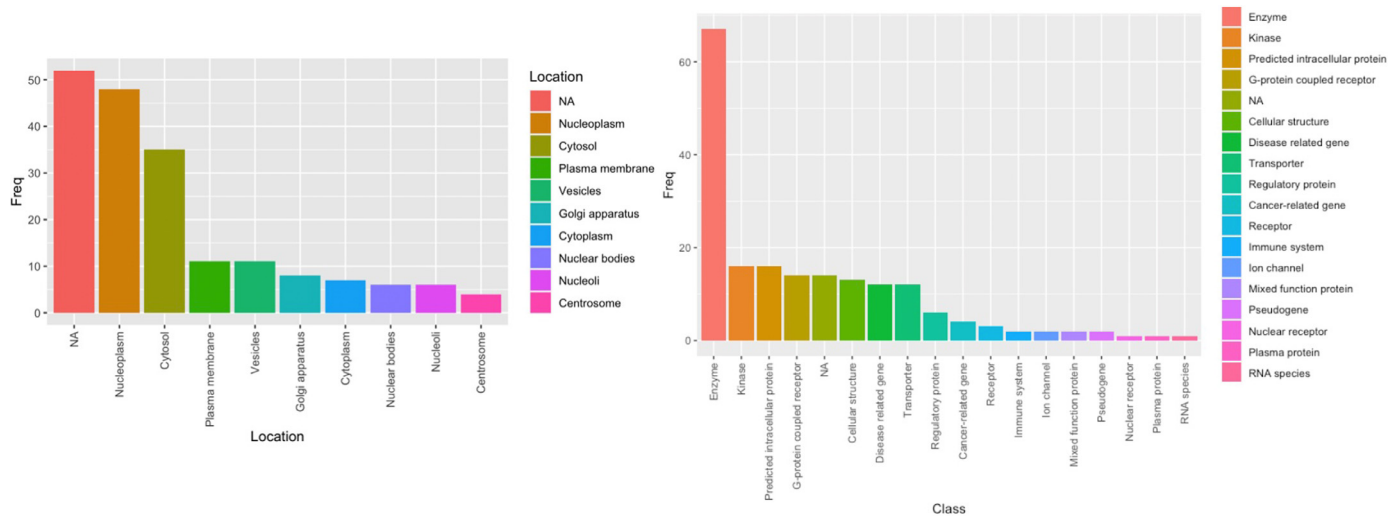


Fig. 6. Location and class summary plots; NA – no classification available.

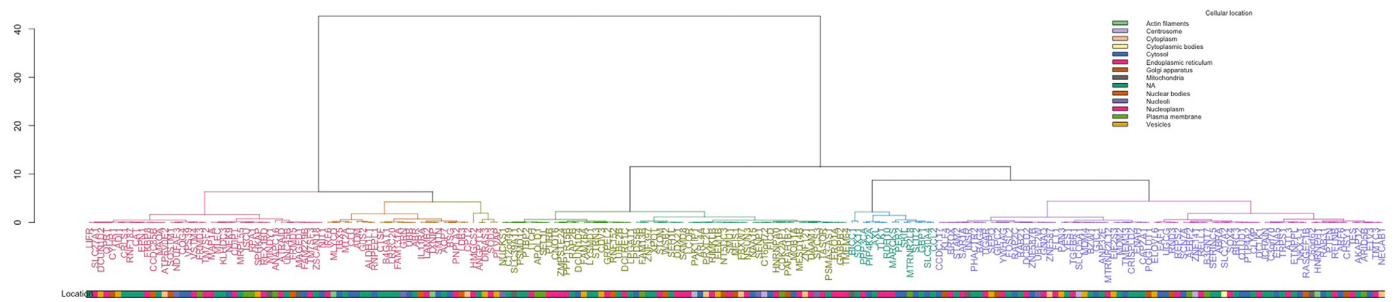


Fig. 7. Dendrogram with mapped cellular locations where coloured gene symbols represent the identified clusters and coloured branches show smaller subclusters. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Similarly, the function *class\_map* provides a visualisation of how the highest and lowest  $LFC_{score}$  genes cluster based on protein class. In addition to a data frame generated by *score\_genes*, the function also requires a *num* parameter to specify the number of genes to consider from the top upregulated and downregulated genes, if this option is not selected all genes will be used (Fig. 8).

```
#Class mapping is done using a single function call
#class_map(data, 20)
```

*HK\_genes* function provides a convenient overview of the housekeeping genes and allows to check if these genes varied throughout conditions. Depending on the number of conditions separate plots will be generated (Fig. 9). Inspecting housekeeping genes can help understand if there was any significant variation between sample groups which might have arisen from biological or technical variation.

```
# To retrieve housekeeping gene assessment only a single function call is required
#HK_genes(data)
```

### 3.3. Gene cluster and expression pattern analyses

Function *cluster\_genes* helps to select an optimal number of clusters and a model to be fitted during the EM phase of clustering for GMM. The function provides summaries and helps to visualise gene clusters based on generated data using *score\_genes* function. Weighted gene expression

is clustered based on the interactome complexity, i.e., the number of known interactors according to the STRING database [3], with a cut-off of 700 for the score threshold. The threshold is set automatically to control for the reliability of the interactions [2,3]. The function also provides scatter and dimension reduction plots to analyse the clusters and features in the data (Fig. 10). Required parameters include a data frame containing a processed expression file from *score\_genes* with  $LFC_{score}$  and a *max\_range* number for cluster exploration during the model selection (the default value is 20 clusters). The *clusters* parameter can be provided for the number of clusters to test when the cluster number estimation is not based on the best BIC output (the user then also needs to supply *modelNames*). This option allows users to perform GMM for a specific number of clusters. The *modelNames* parameter can only be supplied when the *clusters* value is also specified. This option will model the data based on the user parameters for the cluster assignment (Fig. 10) which can be helpful if a different number of clusters helps to explain the data better. The function not only provides a summarised modelling output and plots but also returns a data frame with assigned clusters which can be used by more advanced users in other machine learning pipelines or data comparison studies. For example, gene set clustering based on the interactome size provides insights on the emerging patterns for gene expression changes and the size of the involved network. Selecting specific genes can help build signalling networks based on the identified seed points. Feature distribution analysis also helps to assess the emerging trends in the data based on the variability. That is, gene variation patterns in the experiment might indicate functionally related groups which could be used to reconstruct relevant pathways (Fig. 10). The user is advised to set seed before using the function to get reproducible results.

```
#A machine learning pipeline is implemented using a simple call
#The output data frame can be stored in a new object for further analyses
#model_report<-cluster_genes(data)

# The function will automatically output the Bayesian information criterion (BIC) and
the type of the model for fitting
# Detailed explanation of the model selection is provided with a dependency package -
McLust.

#Best BIC values:
#          VVI,6          VVI,7          VVI,5
#BIC      -2481.986    -2483.544400    -2485.429861
#BIC diff    0.000      -1.558131      -3.443592
# An example of the model report summary output
# head(model_report)

#      Interactors  LFCscore Cluster Symbol
#CAPZA1         66  -2.389418         1 CAPZA1
#RAB31           0  -2.542398         2  RAB31
#UBE2B           2  -2.061383         2  UBE2B
#YWHAG           21  -2.111448         1  YWHAG
#ENAH            29  -2.207514         1  ENAH
#PPP3CA          51  -3.500322         1  PPP3CA
```

Function *cluster\_links* provides the same Gaussian mixture modeling pipeline as *cluster\_genes*; however, instead of the interactor number clustering, the user can select a specific disease score *type* (the default selection is “association”). This parameter can define either the association or specificity for a disease, i.e., if the gene has known links to disease

phenotypes and how specific it is when describing a pathology. The function also provides scatter and dimension reduction plots to analyse the clusters and features in the data. An additional output is a model report summarising the cluster assignments which can be used in other modelling analyses. This information can be used to compare association and network size influences for different clusters and gene expression patterns.

Function *pattern\_search* explores the occurrences of specific patterns in gene sets. That is, it searches each condition for emerging patterns (e.g., if multiple conditions are provided) to group genes that changed in a similar manner (Figs. 11 and 12). The search algorithm works by first generating potential patterns to search depending on the number of subclasses. For example, if a condition has several subclasses as in the case example, where Condition 1 has healthy, hypertensive, and chronic kidney disease (CKD) groups, then potential pattern scenarios are generated, e.g., “up-up-up” or “down-up-down”. Following this, the overall expression for each gene is calculated using geometric mean across all conditions, this gives a basal line against which an individual gene expression value is weighed to deduce if it is in a ‘up’ or ‘down’ state. Comparing against a baseline is a more universal approach than performing a pair wise comparisons which may not be effective for multiple subclasses or complex interactions. In addition, averaging expression using a geometric mean method provides a baseline for comparisons taking into account all the extreme values which might result either from biological or technical effects. It is important to note that taking a geometric mean might not be optimal in all cases, but in a balanced experiment it should provide additional information for the downstream analyses. The function returns a summary of how many genes are identified for each pattern type across conditions.

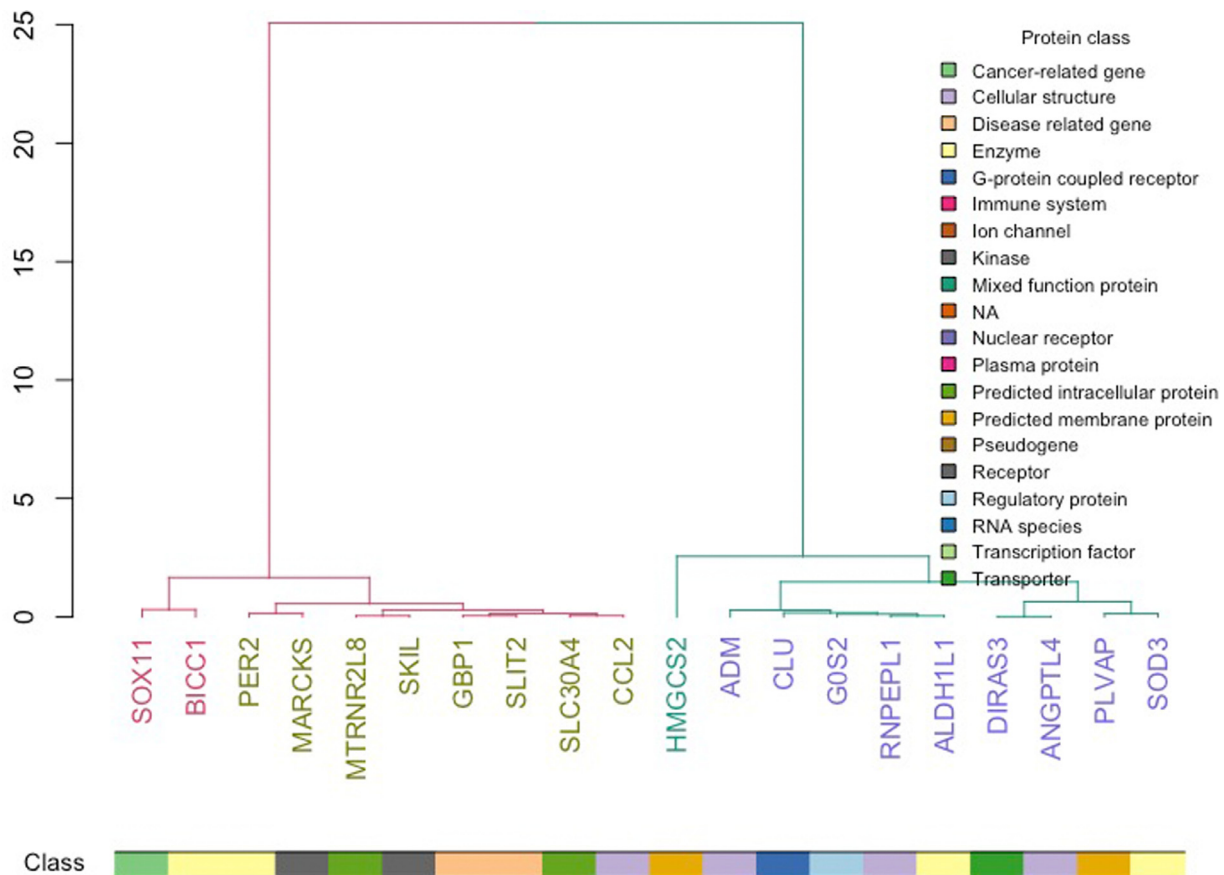
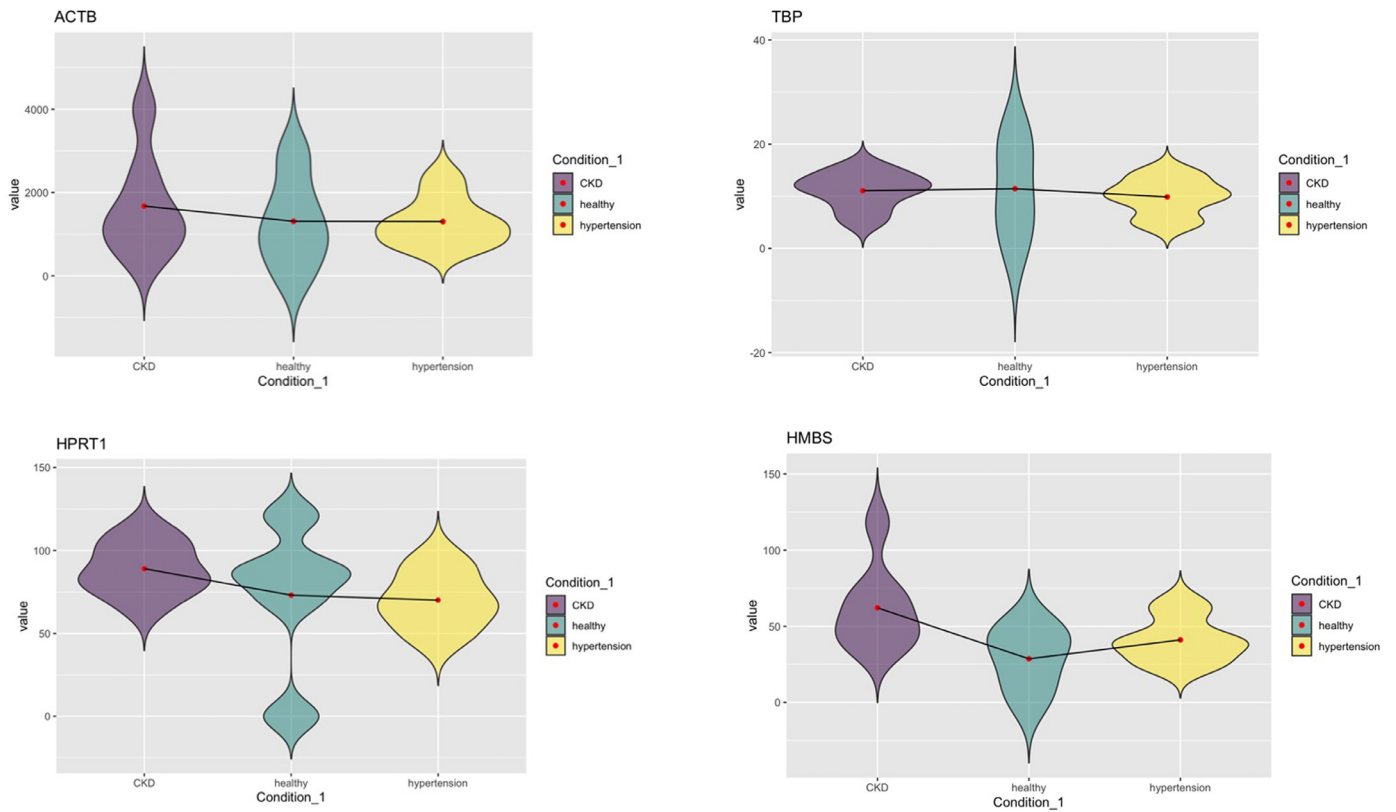
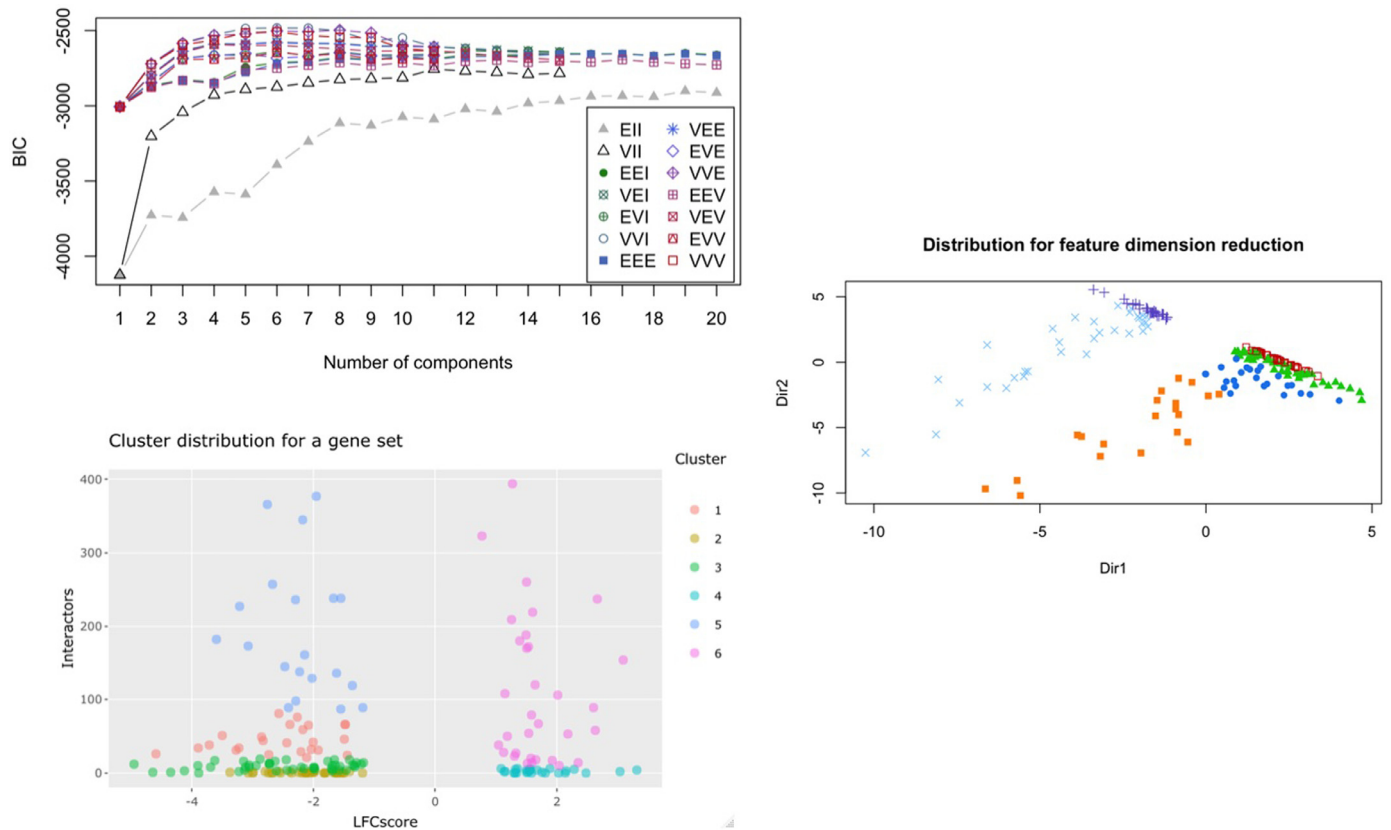


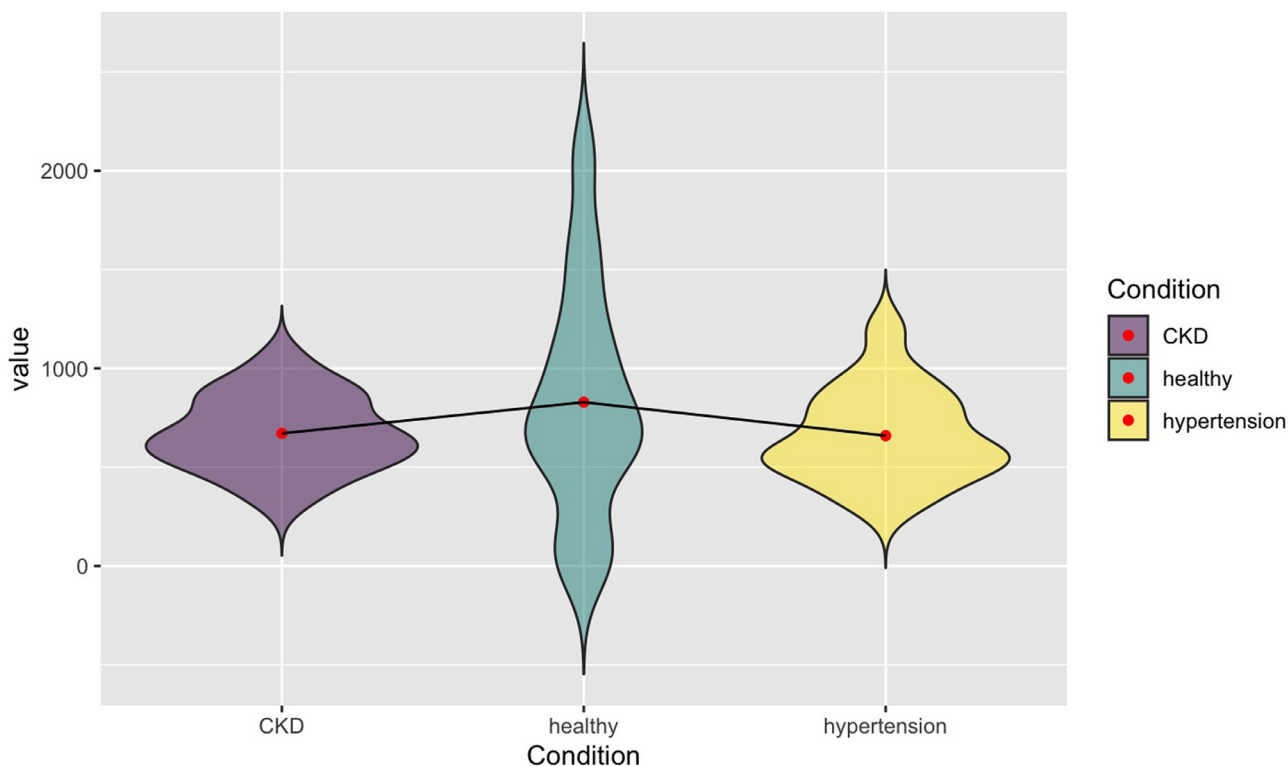
Fig. 8. Dendrogram with mapped protein functions where coloured gene symbols represent the identified clusters and coloured branches show smaller subclusters. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



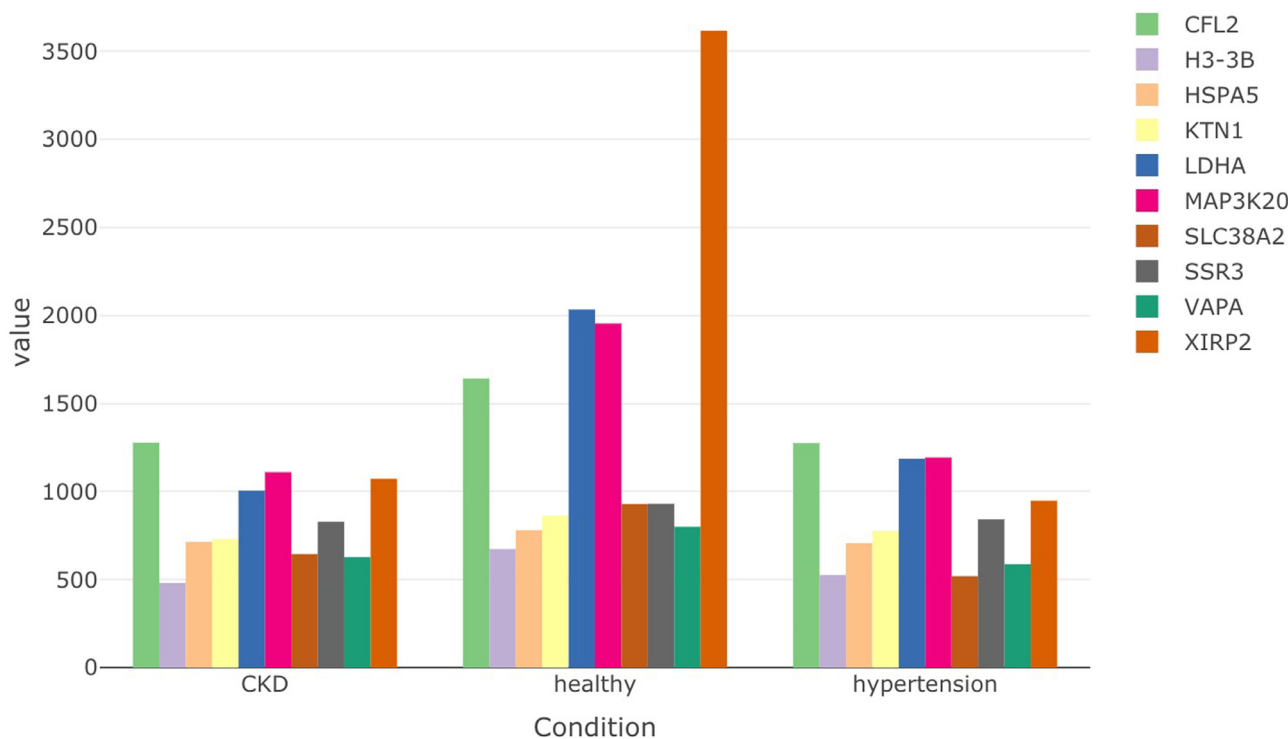
**Fig. 9.** An example of the housekeeping gene distribution. The red marker indicates the mean for the group and violin plots allow to assess global distribution patterns. CKD – chronic kidney disease patient group, healthy – healthy population group, and hypertension – hypertensive patient group. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



**Fig. 10.** GMM analysis examples showing the Bayesian information criterion (BIC) evaluation and model type prediction, clustering analysis, as well as the dimension reduction analysis based on the intrinsic variability within the data.



**Fig. 11.** Gene distribution patterns for a specific expression pattern subset where mean values (signified with a red point) are connected to highlight the pattern features with respect to the mean value (the example is from a “down-up-down” pattern group). CKD – chronic kidney disease patient group, healthy – healthy population group, and hypertension – hypertensive patient group. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



**Fig. 12.** Individual gene distributions when selecting a specific expression pattern and range (the example is from a “down-up-down” pattern group). CKD – chronic kidney disease patient group, healthy – healthy population group, and hypertension – hypertensive patient group.

```
# Condition subclasses
# "CKD"      "healthy"    "hypertension"

#
# The algorithm can be implemented using a single function call where a meta data file
# is also provided to extract information on the subclasses
# pattern_search(data, meta)
#
#
#           Gene count
#down_down_down      0
#down_down_up       2679
#down_up_down       670
#down_up_up        1076
#up_down_down      5503
#up_down_up       2550
#up_up_down       2856
#up_up_up         361
```

The returned gene list contains groups of genes for the different types of patterns. A pattern of interest can be selected to further explore the genes that changed their expression in a specific manner.

```
# Example pattern selection
# $up_up_down
# [1] "A4GALT"      "AASDHPPT"    "AATF"
# [4] "ABCC11"     "ABCC9"       "ABCG2"
# [7] "ABHD13"     "ABHD2"       "ABI3"
# [10] "ABI3BP"     "ABITRAM"     "ABO"
```

This analysis can be followed by *pattern\_plots* which allows to explore distributions for a selected pattern group. The user must provide a subsetted data frame and low/high parameters to select a specific range. The selection is needed because in some instances the expression values might differ significantly and visualising all data points will prevent exploring any meaningful subsets. The outputs allow to evaluate how genes distribute in a subset for different conditions (Fig. 11) and how individual gene values vary in a selected subgroup (Fig. 12).

Function *cluster\_heatmap* uses the information mined from the STRING database [3] to map experimental, referenced, and inferred interactions to see if there are any interactors in the set of significantly changed genes. This heatmap function provides a clustered visualisation of all the genes that have shared interactions (Fig. 13). This information allows to quickly assess how many genes in a specific condition that changed significantly might be part of the same regulatory cluster. Such data can help select specific targets depending on the therapeutic strategy.

```
#Finding interacting proteins and mapping them using a heatmap can be achieved via a
#single function call

#cluster_heatmap(data)
```

Function *interactor\_map* helps to visualise the information mined from the STRING database [3] and map direct and referenced interactions to see if there are any interactors in the set of significantly changed genes and how they are linked. This visual network is an alternative for a heatmap with additional information on the functional gene features (Fig. 14).

```
#Mapping interactors can be done through a single function call
#interactor_map(data)
```

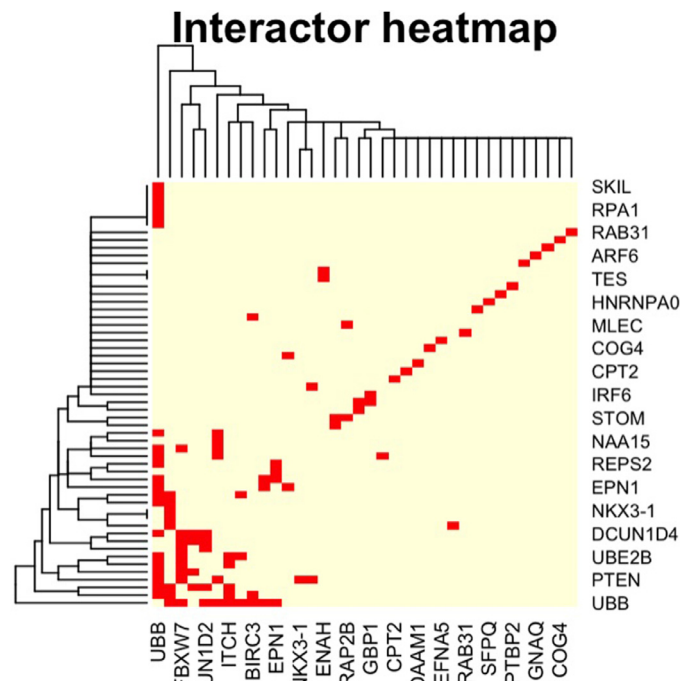


Fig. 13. Cluster heatmap examples where known interactors are connected via the red squares. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

### 3.4. Epigenomics data integration and analysis

Function *CpG\_summary* provides information on the gene CpG island and GC content. The function checks genes against known CpG islands and provides various plots to assess emerging data features. CpG islands were retrieved from the data available with the Genome Reference Consortium (Human Build 38) [17], this information was cross-referenced with the Ensembl database [18] to retrieve overlaps between CpG islands and genes. The function provides a number of analytical plots to assess whether the CpG profile (via GC %) has any influence on the gene expression, interactor number, disease specificity, and disease associations (Figs. 15 and 16). All this information is provided in the context of the assigned protein classes/functional groups. This analysis offers additional insights into the complex interplay between the genome, transcriptome, and epigenome [19]. In addition, the function outputs a data table that contains genomic locations and gene information based on the Ensembl database [18] so that the user can perform additional analyses.

```
#CpG summary requires a single function call to retrieve relevant information. The
#output can be assigned to a new R object for further analyses
```

```
#cpg_genes<-CpG_summary(data)
```

Function *miRNA\_summary\_validated* allows to check how many of the differentially expressed genes have known miRNAs (Figs. 17 and 18). The information on validated/known miRNAs is collected from mining multiple databases, namely *miRecords*, *TarBase*, *miRTarBase*, *PhenomiR*, *miR2Disease*, *Pharmaco-miR*. The function also returns a data table with miRNA information that can be used for designing RNA interference experiments.

```
#head(cpg_genes)
#Output example

# Symbol Log2FoldChange      pvalue Association_score
#1 SAR1A -2.187773 2.656521e-05 0.000000
#2 C6orf62 -2.674213 1.691567e-07 0.000000
#3 AXL -2.786508 1.739595e-04 0.3230769
#4 BICC1 -3.598553 2.738887e-04 0.3000000
#5 CAPZA1 -1.732784 2.321835e-04 0.3789474
#6 TXNIP 1.460629 7.347205e-05 0.3000000

# Specificity_score LFCscore CpG GC_content
#1 0.000 -2.187773 chr1:1211340:1214153 70.33
#2 0.000 -2.674213 NA NA
#3 0.590 -3.686764 chr1:1471765:1497848 58.83
#4 0.751 -4.678119 NA NA
#5 0.601 -2.389418 NA NA
#6 0.631 1.898818 NA NA

# Class
#1 Receptor
#2 NA
#3 Pseudogene
#4 Enzyme
#5 Enzyme
#6 Regulatory protein
```

```
#miRNA analysis example
#df<-miRNA_summary_validated(data)
```

Function *miRNA\_network* allows to examine if a gene set has shared regulatory miRNAs (Fig. 19). This function could be especially useful as it could help exploring the non-coding layer of the regulatory network. This information can aid in studying how some genes are controlled by several miRNAs and detect additional links between genes that changed expression. Moreover, using miRNA analyses can be applied in designing RNA interference studies to select the most optimal interference sequences. miRNA content information can be accessed through the function's output.

```
#An example of calling miRNA summary function for the predicted miRNAs based on the submitted data
#df<-miRNA_summary_predicted(data)
```

Function *miRNA\_summary\_predicted* is similar to the earlier function; however, it allows to check how many of the differentially expressed genes have predicted miRNAs. The information is collected from mining multiple databases that use algorithms to infer likely miRNAs. The databases include miRTarBase, PITA, PicTar, miRecords, miRanda, DIANA-microT, miRDB, TarBase, TargetScan, MicroCosm, and EIMMo. The function also returns a data table with miRNA information that can be used in designing RNA interference experiments.

#### 4. Discussion

*OmicInt* package provides a unique combination of functions and tools for researchers to explore gene expression data sets. A special focus of the package is also making machine learning, specifically Gaussian mixture models [4–6], more accessible to the researchers that do not have a background in the ML/AI field. In addition, the lack of tools for the exploration of the complex expressome data highlighted the need for such a set of bioinformatics tools. For example, commercial solutions, such as Clarivate analytics [7], are very expensive and cannot be easily used by individual researchers. Freely available tools, namely GeneMANIA or Cytoscape platforms [9–11,20], do not permit machine learning applications or complex regulome integration. As a result, the *OmicInt* package was developed for advanced and user-friendly *omics* analyses.

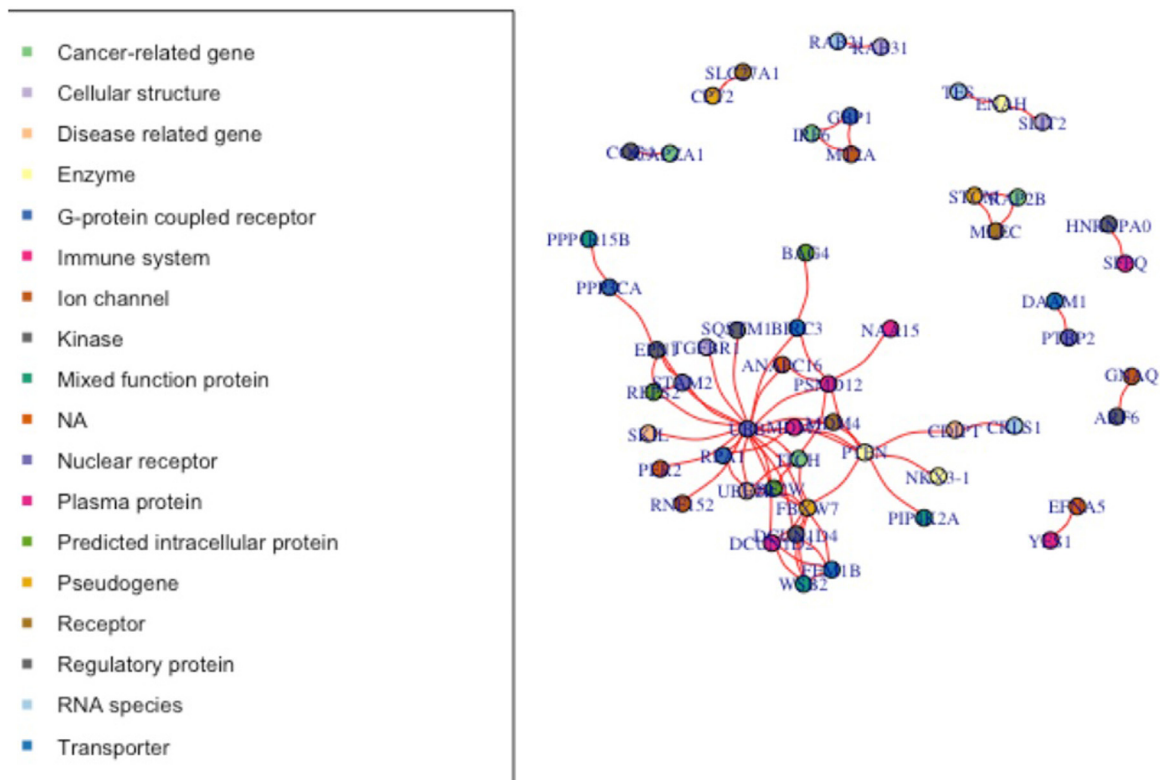


Fig. 14. Interactor map examples.

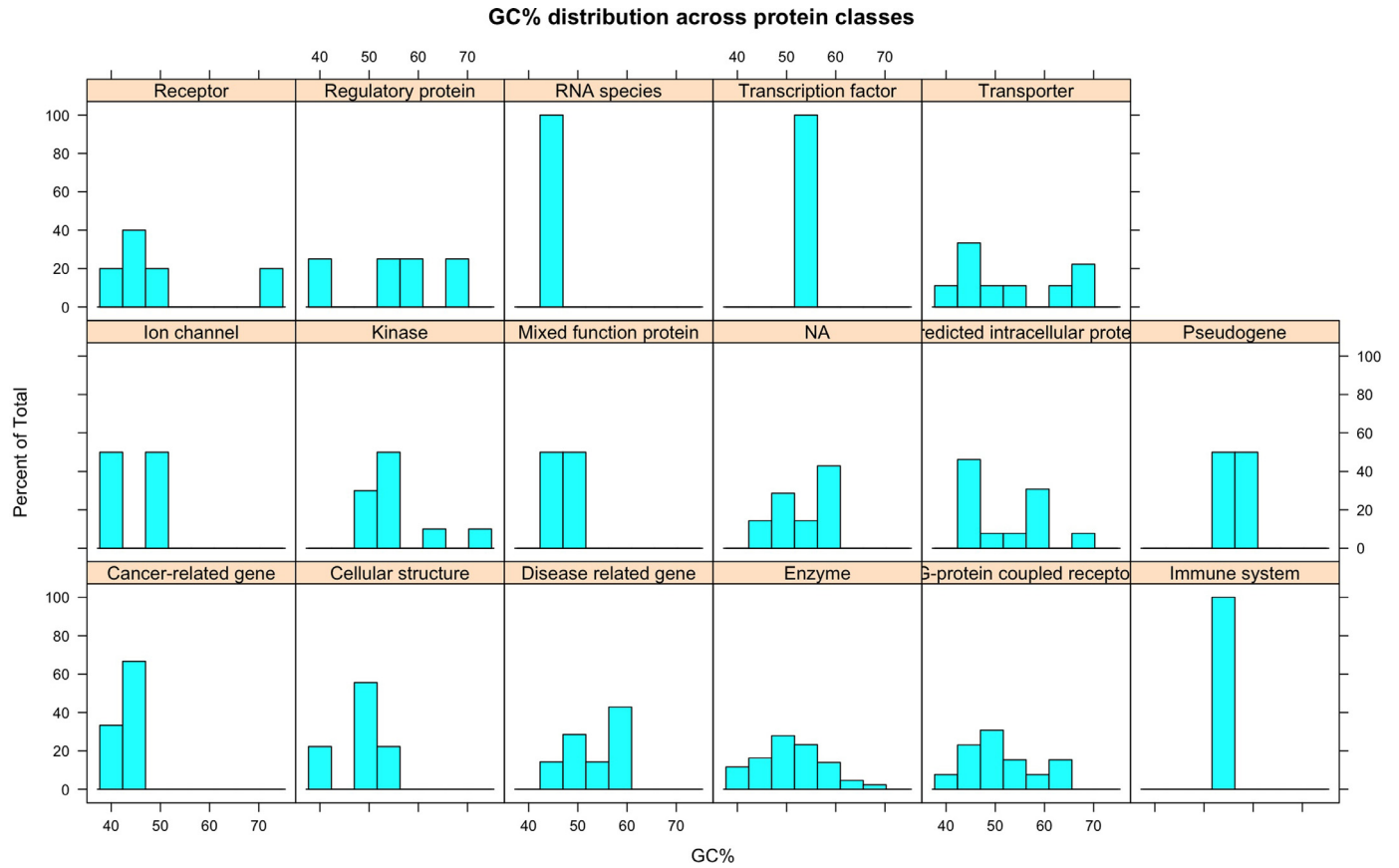


Fig. 15. CpG summary examples where the GC% content distribution is shown for different protein classes.

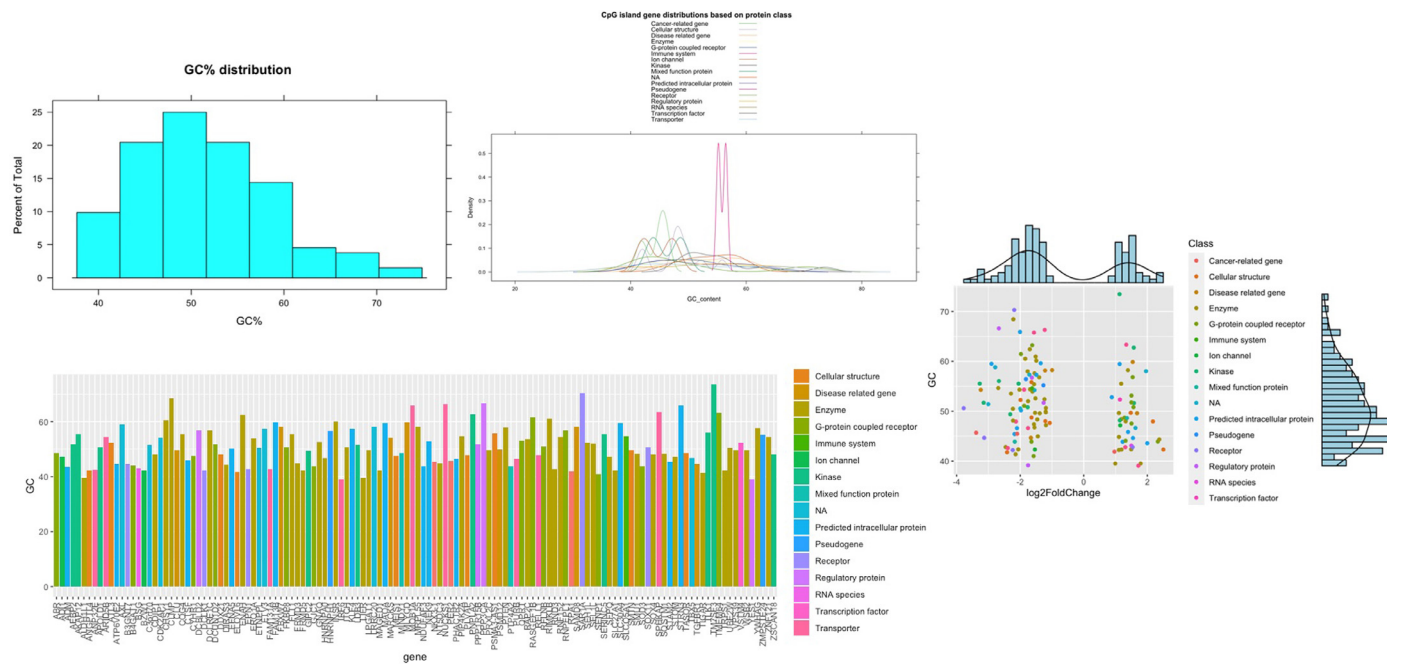


Fig. 16. CpG summary examples where GC% profiles are shown for different genes and their corresponding protein classes. Summary density plots and histograms are also shown for different parameters.

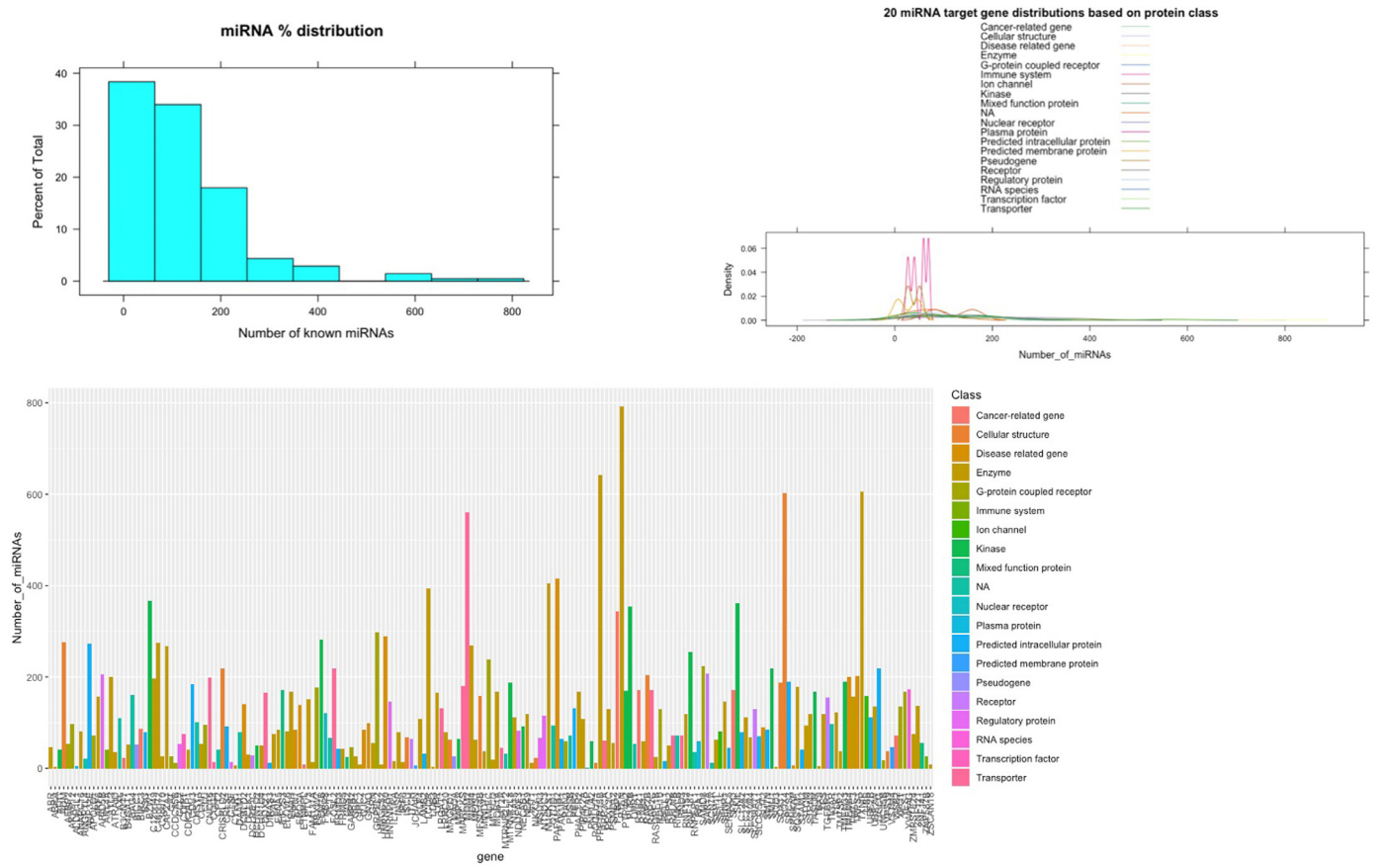


Fig. 17. Validated miRNA summary examples where distribution profiles are shown for different genes and their corresponding protein classes. Summary density plots and histograms are also shown for different parameters.

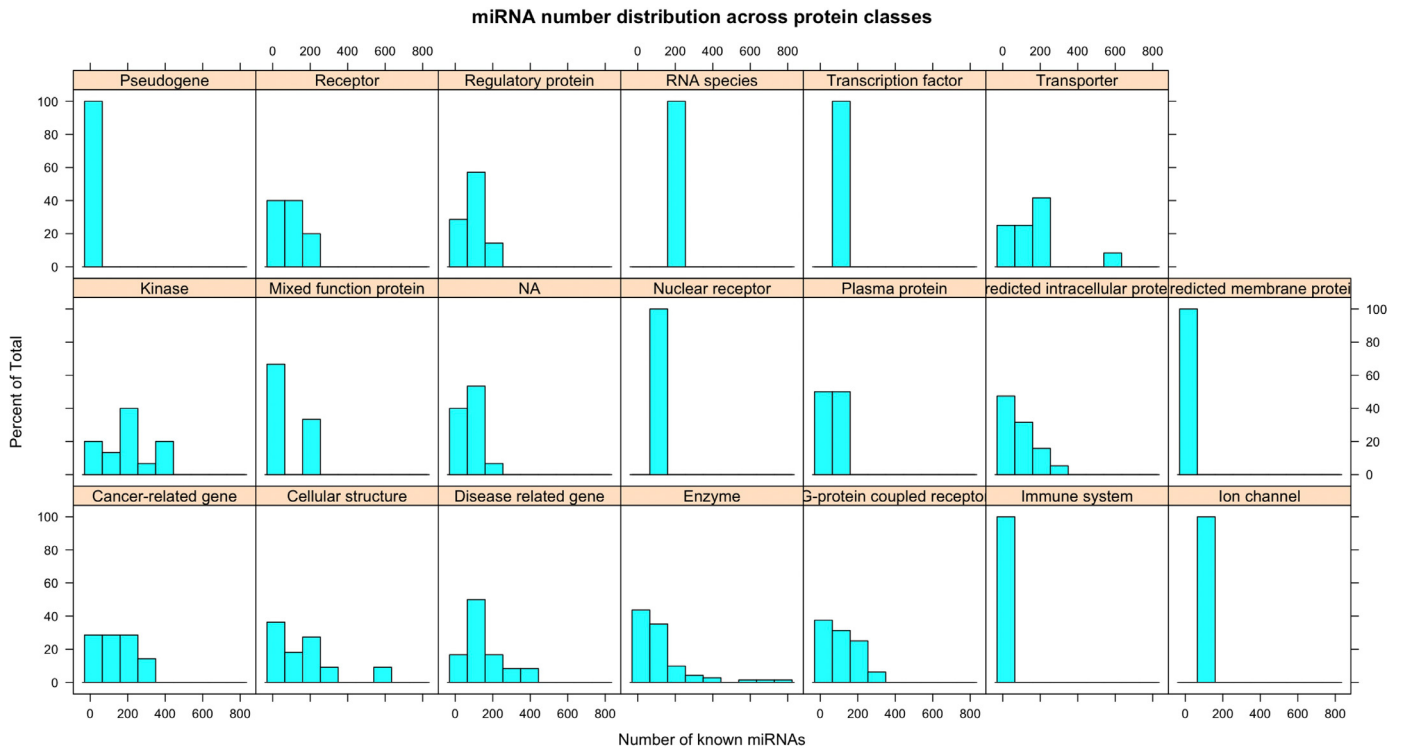


Fig. 18. Validated miRNA summary examples where miRNA content distribution is shown for different protein classes.



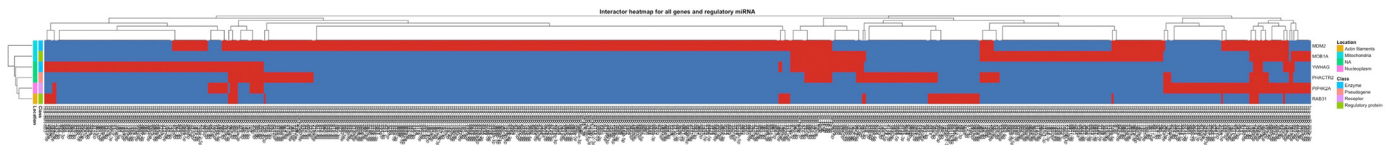


Fig. 19. miRNA network plot example where genes and miRNAs are mapped using a heatmap so that shared links are highlighted in red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The developed scoring functions and GMM pipeline enables exhaustive analysis of the expressome and the associated interactome complexity. The automated processing takes care of the machine learning model optimisation making this analysis easily adaptable to individual researcher's needs. The implementation of probabilistic modelling creates opportunities for new insights based on gene expression changes, disease associations, and the size of the network for a specific gene. Extracting this information can establish relevant seed points to recreate complex signalling pathways or use this data to select genes that should be subjected to downstream *in vitro* studies ensuring that a diverse selection is made.

In addition, advanced functions for epigenomics analysis permit the exploration of the epigenetic regulatory layer. This might be very helpful when identifying genes that may depend on epigenetic regulation [19]. Specifically, if a CpG island containing gene changed expression during treatment or disease progression, it might suggest that there is an epigenetic component controlling the expression levels. Similarly, exploring a gene's miRNA network could hint at other interacting genes which might not have been picked up by the differential expression analysis or help prepare for RNA interference studies. Moreover, miRNA interactome analysis provides the first in-depth look into what genes are controlled by the same set of miRNAs.

Additional functionalities of the package create an analytical environment to summarise gene functional classes or infer what cellular compartments are typically associated with the gene/protein. Such assessments in the context of expression changes or disease association can highlight emerging patterns in specific cellular states under the investigation. A specially designed function to extract gene pattern profiles can aid in a further refinement of causal gene networks when considering a specific phenotype or a condition.

Thus, *OmicInt* offers a comprehensive, evolving, and adaptable platform for gene expression analysis in the context of the transcriptome, proteome, and epigenome. The analyses are made freely available to all researchers where further contributions and algorithmic development are also made possible.

#### Declaration of Competing Interest

The author declares no conflict of interest regarding the publication of the manuscript.

#### Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ailsci.2021.100025.

#### References

- [1] UniProt [Internet]. Available from: <https://www.uniprot.org/> 2021.
- [2] Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 2019 Jan 8;47(D1):D607–13.
- [3] STRING: functional protein association networks [Internet]. Available from: <https://string-db.org/> 2021.
- [4] Reynolds D. Gaussian mixture models. *Encycl. Biom.* 2009;659–63.
- [5] Kanapekaitė A, Burokienė N. Insights into therapeutic targets and biomarkers using integrated multi-omics' approaches for dilated and ischemic cardiomyopathies. *Integr. Biol. (Camb.)* [Internet]. 2021 May 1;13(5):121–37. Available from: <https://pubmed.ncbi.nlm.nih.gov/33969404/>.
- [6] Liu Z, Song Y, Xie C, Tang Z. A new clustering method of gene expression data based on multivariate Gaussian mixture models. *Signal Image Video Process* 2015 Feb 8;10(2):359–68. 2015 102 [Internet]. Available from: <https://link.springer.com/article/10.1007/s11760-015-0749-5>.
- [7] Clarivate - data, insights and analytics for the innovation lifecycle - Clarivate [Internet]. Available from: <https://clarivate.com/> 2021.
- [8] Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrwi R, Chao P, et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.* [Internet] 2010 Jul 1;38(suppl\_2):W214–20.
- [9] GeneMANIA [Internet]. Available from: <https://genemania.org/> 2021.
- [10] Cytoscape: an open source platform for complex network analysis and visualization [Internet]. Available from: <https://cytoscape.org/> 2021.
- [11] Otasek D, Morris JH, Bouças J, Pico AR, Demchak B. Cytoscape automation: empowering workflow-based network analysis. *Genome Biol.* 2019 Sep 2;20(1).
- [12] Scrucca L, Fop M, Murphy TB, Raftery AE. mclust 5: clustering, classification and density estimation using gaussian finite mixture models. *R. J.* [Internet] 2016;8(1):289 Available from: [/pmc/articles/PMC5096736/](https://pmc/articles/PMC5096736/).
- [13] DisGeNET - a database of gene-disease associations [Internet]. Available from: <https://www.disgenet.org/> 2021.
- [14] AusteKan/OmicInt: OmicInt Package [Internet]. Available from: <https://github.com/AusteKan/OmicInt> 2021.
- [15] CRAN - Package OmicInt [Internet]. Available from: <https://cran.r-project.org/web/packages/OmicInt/index.html> 2021.
- [16] Home - GEO - NCBI [Internet]. Available from: <https://www.ncbi.nlm.nih.gov/geo/> 2021.
- [17] GRCh38 - hg38 - Genome - Assembly - NCBI [Internet]. Available from: [https://www.ncbi.nlm.nih.gov/assembly/GCF\\_000001405.26/](https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.26/) 2021.
- [18] Ensembl genome browser [Internet]. 2021. Available from: <https://www.ensembl.org/index.html>.
- [19] Cazaly E, Saad J, Wang W, Heckman C, Ollikainen M, Tang J. Making sense of the epigenome using data integration approaches [Internet]. *Front. Pharmacol. Front. Media S.A.* 2019;10:126.
- [20] Ali M. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. 2008;9:1–15. Available from: [papers2://publication/uid/D18A7677-333C-443F-B060-94EFCED7857C6](https://pubmed.ncbi.nlm.nih.gov/18187677/).

## **Integrative *omics* approaches for new target identification and therapeutics development**

### **4. Fi-score: a novel approach to characterise protein topology and aid in drug discovery studies**

**The experimental chapter is based on the following publication**

Kanapeckaitė A, Beaurivage C, Hancock M, Verschueren E. Fi-score: a novel approach to characterise protein topology and aid in drug discovery studies. *Journal of Biomolecular Structure and Dynamics*. 2020 Dec 7:1-1; doi: 10.1080/07391102.2020.1854859; PMID: 33297860.

#### **Conclusion of this chapter**

This chapter introduces a new method that I developed helping to characterise proteins prior to *in silico* screening by evaluating potentially dynamically active regions or predicting sites that share similar qualities in the side chain distribution and movement. Incorporating the Fi-score with other physicochemical parameters, such as hydrophobicity, could greatly improve detecting multiple functionally relevant sites within a target or capturing similar profiles across different proteins. The detected sites could be subjected to docking studies. Moreover, my developed analytical pipeline helped to show that using machine learning approaches expands the analytical scope by extracting and defining structural elements or motifs of various proteins. Thus, Fi-score focused analysis can aid in primary target selection studies and also advance drug or biologics formulation methods by evaluating potential binding sites or interaction surfaces. This innovative biophysical analysis method could significantly improve target selection, pre-screening analysis and speed up biologics engineering.

#### **Contribution to this chapter (95%)**

- Derived the scoring equation and machine learning pipeline.
- Performed all the analytical, data mining, and experimental work as well as formulated conclusions.
- Performed benchmarking and comparative analyses.
- Conceptualised and wrote the manuscript, including the figure preparation.
- Corresponding author.



## Fi-score: a novel approach to characterise protein topology and aid in drug discovery studies

Austè Kanapekaitė<sup>a</sup> , Claudia Beurivage<sup>b,c</sup>, Matthew Hancock<sup>a</sup> and Erik Verschueren<sup>a</sup>

<sup>a</sup>Galapagos NV, Mechelen, Belgium; <sup>b</sup>Galapagos BV, Leiden, The Netherlands; <sup>c</sup>Department of Biomedical Science, Faculty of Science, University of Sheffield, Sheffield, UK

Communicated by Ramaswamy H. Sarma

### ABSTRACT

Target evaluation is at the centre of rational drug design and biologics development. In order to successfully engineer antibodies, T-cell receptors or small molecules it is necessary to identify and characterise potential binding or contact sites on therapeutically relevant target proteins. Currently, there are numerous challenges in achieving a better docking precision as well as characterising relevant sites. We devised a first-of-its-kind *in silico* protein fingerprinting approach based on the dihedral angle and B-factor distribution to probe binding sites and sites of structural importance. Our derived Fi-score can be used to classify protein regions or individual structural subsets of interest and the described scoring system could be integrated into other discovery pipelines, such as protein classification databases, or applied to investigate new targets. We further demonstrated how our method can be integrated into machine learning Gaussian mixture models to predict different structural elements. Fi-score, in combination with other biophysical analytical methods depending on the research goals, could help to classify and systematically analyse not only targets but also drug candidates that bind to specific sites. The described methodology could greatly improve pre-screening stage, target selection and drug repurposing efforts in finding other matching targets.

### HIGHLIGHTS

- Description and derivation of a first-of-its-kind *in silico* protein fingerprinting method using B-factors and dihedral angles.
- Derived Fi-score allows to characterise the whole protein or selected regions of interest.
- Demonstration how machine learning using Gaussian mixture models on Fi-scores captures and allows to predict functional protein topology elements.
- Fi-score is a novel method to help evaluate therapeutic targets and engineer effective biologics.

**Abbreviations:** AIC: Akaike information criterion; BIC: Bayesian information criterion; HTS: high-throughput screening; PLI: target protein–ligand interactions

### ARTICLE HISTORY

Received 24 August 2020  
Accepted 17 November 2020

### KEYWORDS

Drug discovery; dihedral angles; B-factor; machine learning; Gaussian mixture models; conformation distal information; protein site characterisation


## Introduction

The identification of lead compounds showing pharmacological promise is the focal point of early-stage drug discovery. While large libraries of compounds against a therapeutically relevant target are subjected to high-throughput screening (HTS) to select new lead compounds, this method becomes more and more supplemented or preceded by *in silico* HTS within the pharmaceutical industry. This shift in the paradigm can be attributed to the high costs and time-consuming nature of the design and completion of HTS screens (Dias & de Azevedo, 2008). In contrast, early stage *in silico* screening offers not only a better understanding of relevant biological topology, potential active sites but also allows a progressive optimisation of the pharmacological properties and potency of selected compounds. Yet, structurally complex sites or sites with a wide dynamic range pose a

challenge; especially, when selecting between a family of targets or targets with similar topology (Gangadharan et al., 2017).

While the human genome contains approximately 25,000 genes, only about 10% of the expressed proteins are amenable to small-molecule modulation and less than a half of that subset has therapeutic potential. In addition, the development of therapeutic compounds have a very low success rate as less than 2% of lead compounds succeed to get to the market (Dias & de Azevedo, 2008; Gangadharan et al., 2017; Knapp, 2016; Santos et al., 2017). The picture gets even more complicated for immunotherapeutics development as lead compounds can have potentially far reaching side effects and the identification and validation of disease-specific targets is also complicated by the fact that numerous proteins can undergo significant conformational changes

**CONTACT** Austè Kanapekaitė  [austekan@gmail.com](mailto:austekan@gmail.com)  Galapagos NV, Zernikedreef 16, 2333CL, Mechelen, Belgium

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/07391102.2020.1854859>.

© 2020 Informa UK Limited, trading as Taylor & Francis Group

throughout their immune cycle (Gangadharan et al., 2017; Knapp, 2016). Consequently, we theorise that having the means to compare multiple regions of interest within the target prior to the screening would be extremely beneficial. For example, if a reference set of binding sites that are known to bind compounds can be classified based on their topology and physicochemical properties, this information could be used to compare and evaluate new sites of interest for the compound binding after these new sites are scored on the same parameters. That is, such scoring could be easily integrated into a relational database of protein targets in a discovery pipeline. Moreover, in some instances a binding site might be conserved and it could be useful to compare protein regions of interest between multiple homologous proteins using a scoring method that could give insights into the conformation and not just the amino acid composition. Another example could be characterising a protein site with a score that has known binders and comparing it to a score of a new target which has no known binding compounds. This could be especially helpful in drug repurposing because protein sites of similar characteristics would potentially allow to infer drug binding in a new site based on already explored one. The compound could then be docked *in silico* or subjected to *in vitro* studies and if the investigational pipeline has multiple new targets such a pre-screening strategy could help to prioritise. Thus, we believe that establishing an effective methodology to classify sites of interest could be extremely beneficial in terms of the screening cost reduction and faster turnaround.

Most currently marketed small-molecule drugs are developed to target protein–ligand interactions (PLI) (Fuller et al., 2009) and this information is primarily provided by the crystallographic analysis. Crystallographic structure analysis has revealed that PLI sites are hydrophobic pockets concave in shape with more complex topological features than those found on protein surfaces, but they can also be relatively flat and large (Buckle et al., 1996; Fauman et al., 2011; Fuller et al., 2009; Mann & Hermans, 2000; Pérot et al., 2010). As a result, computational analysis to probe potential binding sites of proteins exploits these features to evaluate energetics, cavity geometry and physicochemical properties of a potential binding pocket. However, there are additional challenges as the selected sites might be topologically constrained and because of growing computational costs broader conformational changes may not be incorporated into the binding grid analysis. Furthermore, there is not one universal algorithm developed that could be suitable for all scenarios; therefore, we aimed to combine multiple levels of analysis, capturing B-factor values and the dihedral angle structure to establish a comparative measure of physicochemical characteristics of a protein of interest that could be used to analyse a single motif, expanded to a site or the whole protein (Siglioccolo et al., 2010). We here describe a method to derive a score for a site of interest which could be visualised via distribution plots, 3D region visualisation or integrated into machine learning to derive probability density distributions based on physicochemical properties; all of

these applications of a site score could be used to infer characteristics of a region under investigation.

Protein dihedral angles contain information on the local protein conformation in such a way that a protein backbone conformation can be highly accurately rebuilt based on the native dihedral angles. Extracting this information can facilitate in narrowing down the conformational space, which in turn can be superimposed on specific physicochemical properties of the region of interest (De Juan et al., 2013; Faraggi et al., 2009; Heffernan, 2015; Schlessinger & Rost, 2005). While Ramachandran basin allows a holistic description of conformation, this approach lacks statistical description with a focus on the torsion angle distributions of specific sequence and thus, in consideration of the circular nature of angles, traditional parametric or non-parametric density estimation methods cannot work properly to approximate Ramachandran distributions; this is also supported by the findings of the current study. As a result, all of this calls for a more unified approach in analysing local protein regions and extracting the high information content from dihedral angle distribution. By extension, capturing sequence information content can facilitate current efforts to build improved predictive models for dihedral angle and protein three-dimensional structure determination as well as target evaluation for drug screens (Faraggi et al., 2009; Heffernan, 2015). To achieve this additional parameter, the oscillation amplitudes of the atoms around their equilibrium positions (B-factors) in the crystal structures were used; this relationship is described in the first equation.

$$B = 8(\pi^2)u^2 \quad (1)$$

Equation (1): B-factor evaluation, where oscillation amplitude is  $u$ .

While B-factors are used in the atomic form factor calculation to measure scattering amplitude (Eq. (2)), B-factors have a much more complex influence on atoms and the overall structure because of their dependence on conformational disorder, dynamic alterations of the sequence seen via the changes in the positional dispersion of B-factors (Fauman et al., 2011; Tang et al., 2019).

$$f = f_o \cdot \exp\left(-\frac{B \cdot \sin^2\theta}{\lambda^2}\right) \quad (2)$$

Equation (2): Scattering amplitude evaluation, where  $B$  is the B-factor,  $f_o$  is the atomic form factor, and  $\lambda$  is the X-ray wavelength.

In addition, B-factors provide means to gain insight into many aspects of molecular dynamics, such as thermal motion paths, protein superimposition and predict the rotameric state of amino acids side-chains (Carugo, 2018; Carugo & Argos, 1999; Carugo & Eisenhaber, 1997; Weiss, 2007). B-factors were shown to be related to protein packing and depend on the three-dimensional structure (Heffernan, 2015; Parthasarathy & Murthy, 1997; Vihinen 1987; Weiss, 2007; Yin et al., 2011) and there are numerous other studies investigating protein flexibility through B-factors (Bornot et al., 2011; Liu et al., 2014; Parthasarathy & Murthy, 1997; Vihinen et al., 1994). Moreover, B-factors allow the capture of differences

between crystal packing sites and biologically relevant protein-protein interaction sites (Liu et al., 2014). It becomes apparent that B-factors carry a lot of information on both local and distant protein topologies and by incorporating B-factor estimates we include additional information on the local mobility of a C $\alpha$  atom. This leads to our derived equation (Eq. (4)) that provides a fingerprint score or Fi-score through the cumulative sum of standard deviation normalised dihedral angles and scaled B-factors divided by the amino acid residue number of a selected region of interest. The fingerprint value captures physicochemical qualities of a region of interest dependent on conformation; moreover, by normalising and scaling we can effectively compare regions of different targets.

$$B_{i\text{-norm}} = \frac{B_i - \min(B)}{\max(B) - \min(B)} \quad (3)$$

**Equation (3):** Min-max normalisation and scaling of B-factor where  $B_{i\text{-norm}}$  is scaled B-factor,  $B_i$  is the B-factor for C $\alpha$ ,  $B_{\max}$  is the largest B-factor value for the total protein B-factors for all C $\alpha$ ,  $B_{\min}$  is the smallest B-factor value for the total protein B-factors for all C $\alpha$ . B-factor normalisation is based on the full length protein.

$$F_{i\text{score}} = \frac{1}{N} \sum_i \frac{\phi_i \psi_i}{\sigma_{\phi_i} \sigma_{\psi_i}} B_{i\text{-norm}} \quad (4)$$

**Equation (4):** Fi-score evaluation where  $N$  is the total number of atoms for which dihedral angle information is available,  $\phi$  and  $\psi$  values represent dihedral angles for an C $\alpha$  atom,  $\sigma_{\phi}$  and  $\sigma_{\psi}$  represent corresponding standard deviations for the torsion angles and  $B_{i\text{-norm}}$  is a normalised B-factor value for the C $\alpha$  atom. B-factor,  $\sigma_{\phi}$  and  $\sigma_{\psi}$  normalisation are based on the full length protein.

The described methodology could be of great pharmaceutical interest to identify families of targets that are affected by drug treatment and to characterise binding sites after mutational studies. For example, when a signalling protein family contains known drug targets, fingerprinting can define additional druggable family members without relying on the sequence similarity alone but actually measuring physicochemical parameters (Brazhnik et al., 2002). That is, Fi-score can be employed to capture a region of interest in a single value form and the generated scores of multiple such sites could be clustered to enrich based on the similarity between profiles. This could be useful in building relational databases since it is still difficult to capture protein region or domain information in a meaningful and concise way. Moreover, Fi-score visualisation can also aid to accurately evaluate the score distribution of different regions along the protein sequence. For example, domain alignment algorithms rely on the direct amino acid sequence, while dihedral angles and B-factors capture both local and distal information as their distribution is dependent not only on the immediate sequence but also on the steric hindrances as well as the conformation of other protein regions. As a result, Fi-score could be applied in machine learning to cluster Fi-scores so that dynamically similar sites can be grouped and evaluated prior to computationally expensive *in silico* HTS.

In summary, we aimed to devise an equation to capture selected protein site properties that could be used to evaluate structural motifs. This can be achieved either focusing on individual amino acids and inspecting Fi-score distributions or selecting a region of interest to generate a single value to classify sites.

## Methods

### Protein set selection and analysis

A total of 3352 proteins structures were downloaded directly from RCSB Protein Data Bank (RCSB PDB, n.d.) by first selecting proteins based on their features using Pfam 32.0 (Pfam, n.d.) and Structural Classification of Proteins (SCOP) databases (SCOP, n.d.) (Table 1, Supplementary material). This diverse set of randomly selected proteins was used for comparative studies of secondary structure elements (50,043 in total) (Figure 1, supplementary material). We then proceeded to select the representative examples (Table 1, PDB IDs in bold) which were analysed using protein BLAST (BLAST: Basic Local Alignment Search Tool, n.d.) to find good candidates to form protein pairs that showed a varying degree of similarity (Table 1, PDB IDs not highlighted). From this initial pool, candidate proteins were selected maintaining diversity of resolution and R-factor. All paired proteins were subjected to local alignment to identify regions with as much diversity as possible in their identity and similarity scores. These regions were extracted and sequences were globally aligned to get the final score on the identity, similarity and gaps since only that region of interest will be used for Fi-scoring. The alignment and testing was performed with the following tools and parameters MSA (MUSCLE algorithm, default parameters; UGENE software version 1.32 (Okonechnikov et al., 2012), pairwise alignment (Smith-Waterman algorithm-Water (EMBOSS), matrix: BLOSSUM62; gap opening: 10; gap extension: 0.5), global pairwise alignment (Needleman-Wunsch algorithm- Needle (EMBOSS), matrix: BLOSSUM62; gap opening: 10; gap extension: 0.5) (EMBOSS programs, EMBL-EBI, n.d.) and Protein-Blast/PSI-Blast analyses using default settings were employed to assess the sequences (BLAST: Basic Local Alignment Search Tool, n.d.).

### Protein dihedral angle analysis and site scoring

Protein dihedral angles were analysed using R package: Bio3D (Grant et al., 2006) with specific modifications to allow dihedral angle retrieval, fingerprint calculation and visualisation (R studio, version 1.1.463) (RStudio, n.d.). Additional functionalities were introduced to better capture dihedral angle and B-factor distribution. Hydrophobicity scoring for a selected site was calculated based on Kyte-Doolittle scale (R package: Peptides) (Osorio et al., 2015). We selected Kyte-Doolittle scale since it is a widely used hydrophobicity scale and has been previously successfully employed in various algorithms predicting protein secondary structure elements and their distribution (Kyte & Doolittle, 1982; Zhao & London, 2006).

**Table 1.** Characterisation and scoring of different target protein regions.

Protein name	Protein PDB ID	Chain	Amino acid number	Fi-score	Hydrophobicity score	Alignment scores	RMSD	$\Delta$ Fi-score
Human GABA-A receptor, subunit beta-2	6D6U	A	209–300	0.04467536	0.7630435	Identity: 58/96 (60.4%) Similarity: 75/96 (78.1%) Gaps: 4/96 (4.2%)	1.148	0.07702
Human glycine receptor alpha-3	5CFB	A	211–306	−0.03234241	0.8642857			
Interleukin-1 beta mutant F146Y	1TWM	A	5–100	−0.2749203	−0.5885417	Identity: 15/109 (13.8%) Similarity: 27/109 (24.8%) Gaps: 57/109 (52.3%)	11.407	0.58062
Therapeutical antibody fragment of canakinumab	4G5Z	H	151–215	−0.8555358	−0.08			
Therapeutical antibody fragment of canakinumab	4G5Z	H	12–214	−0.574754	−0.1655172	Identity: 64/210 (30.5%) Similarity: 95/210 (45.2%) Gaps: 22/210 (10.5%)	2.535	0.21784
Therapeutical antibody fragment of canakinumab	4G5Z	L	13–207	−0.7925909	−0.4451282	Identity: 109/110 (99.1%) Similarity: 109/110 (100.0%) Gaps: 0/110 (0.0%)	0.416	0.5890659
Catalytic antibody 21H3 with hapten	1UM4	L	103–212	−0.4737261	−0.5027273	Identity: 207/207 (100.0%) Similarity: 207/207 (100.0%) Gaps: 0/207 (0.0%)		
Therapeutical antibody fragment of canakinumab	4G5Z	L	101–210	−1.062792	−0.5027273	Identity: 207/207 (100.0%) Similarity: 207/207 (100.0%) Gaps: 0/207 (0.0%)		
Heat shock protein 90-HSP90	2QF6	A	20–220	−0.1866865	−0.2333333	Identity: 7/14 (50.0%) Similarity: 13/23 (56.5%) Gaps: 1/23 (4.3%)	0.023	0.04451
Heat shock protein 90-HSP90	2QF6	B	20–220	−0.2311956	−0.2333333			
p53 Tetramers, conserved DNA binding site	2AC0	A	249–262	−0.434457	0.3285714	Identity: 5/14 (35.7%) Similarity: 7/14 (50.0%) Gaps: 0/14 (0.0%)	0.679	0.3655604
HDM2 in complex with a beta-hairpin, SWIB/MDM2 domain	2AXI	A	29–42	−0.0688966	0.1857143			
Src homology 2 (SH2) domain	4EIH	A	239–261	−0.1131818	0.2391304		0.212	0.0608529
Rad18 ubiquitin ligase RING domain structure	2Y43	A	66–87	−0.0523289	−0.4409091			

\* $\Delta$ Fi-score—an absolute value of the difference in Fi-scores; PDB IDs in bold—proteins selected initially for screen that were matched to another protein in the pair based on different similarity and identity values.

### Protein visualisation and structural analysis

PyMOL (Molecular Graphics System, Version 2.0 Schrödinger, LLC) (Delano, 2002) was used for protein visualisation and superimposition studies (RMSD calculations) as well as structural analysis integrating python code for robust parsing.

### Protein feature capture

Gaussian mixture models (GMMs) (with the following parameters: max\_iter = 1000, covariance\_type='full' or 'spherical', tol = 0.001, random\_state = 0) were implemented to cluster Fi-score profiled protein sequences. Model selection and evaluation was performed using probabilistic statistical measures that are used to quantify the model performance. We opted for Akaike information criterion (AIC) (Vrieze, 2012) and the Bayesian information criterion (BIC) (Vrieze, 2012) since AIC provides an estimate of in-sample error prediction and information loss, while BIC helps to evaluate for potential overfitting using the likelihood function to estimate the number of parameters. The number of components for clustering and correction of the over-fitting was established using AIC and BIC where the smallest difference between YAIC and YBIC information criterion values was used to determine a component number (usually spanning the inflection point of both curves). Python Scikit-Learn GMM (scikit-learn 0.22.2) (scikit-learn, n.d.) was used for the above analyses where Gaussian mixture and expectation-maximisation algorithms were defined by Eqs. (5)–(7) to estimate the density and distribution of Fi-scores for amino acids.

$$p(X_n) = \sum_{k=1}^K p(X_n|Z)p(Z) = \sum_{k=1}^K \pi_k N(X_n|\mu_k, \Sigma_k) \quad (5)$$

Equation (5): Equation defining a Gaussian Mixture; where  $\Sigma_k$ -covariance for the Gaussian,  $K$  is the number of clusters of the dataset,  $\mu_k$ -cluster centre,  $\pi_k$ -mixing probability,  $z$  - a latent variable defining a probability that data point comes from the Gaussian.

$$Q(\theta^*, \theta) = E[\ln p(X, Z|\theta^*)] = \sum_z p(Z|X, \theta) \ln p(X, Z|\theta^*) \quad (6)$$

Equation (6): Expectation step defining the equation where the current value of the parameters  $\theta^*$  is used to find the posterior distribution of the latent variables given by  $P(Z|X, \theta^*)$ .

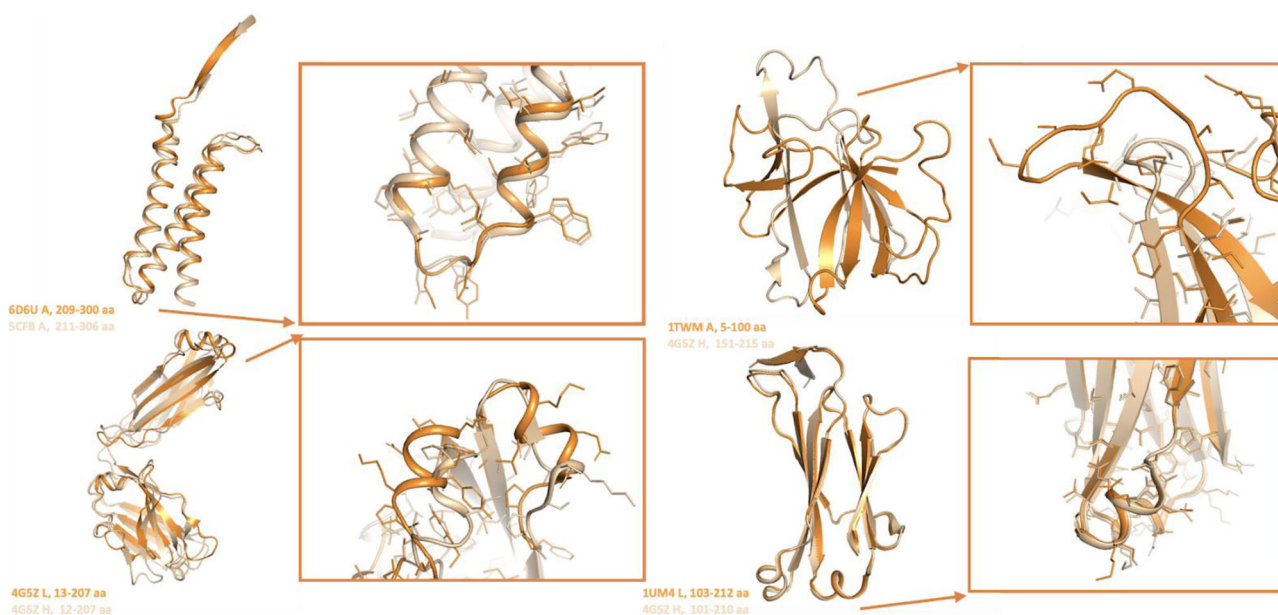
$$\theta^* = \arg \max_{\theta} Q(\theta^*, \theta) \quad (7)$$

Equation (7): Maximisation step defining the equation to find the expectation under the posterior distribution of the latent variables with a new estimate for the parameters.

## Results

### Fi-score derivation

We developed a method allowing to capture the side chain as well as the mean atomic displacement distribution in a single fingerprint score or 'Fi-score'. Fi-score equation through the use of standard deviation normalised dihedral angle values and scaled B-factor using min-max method allows to



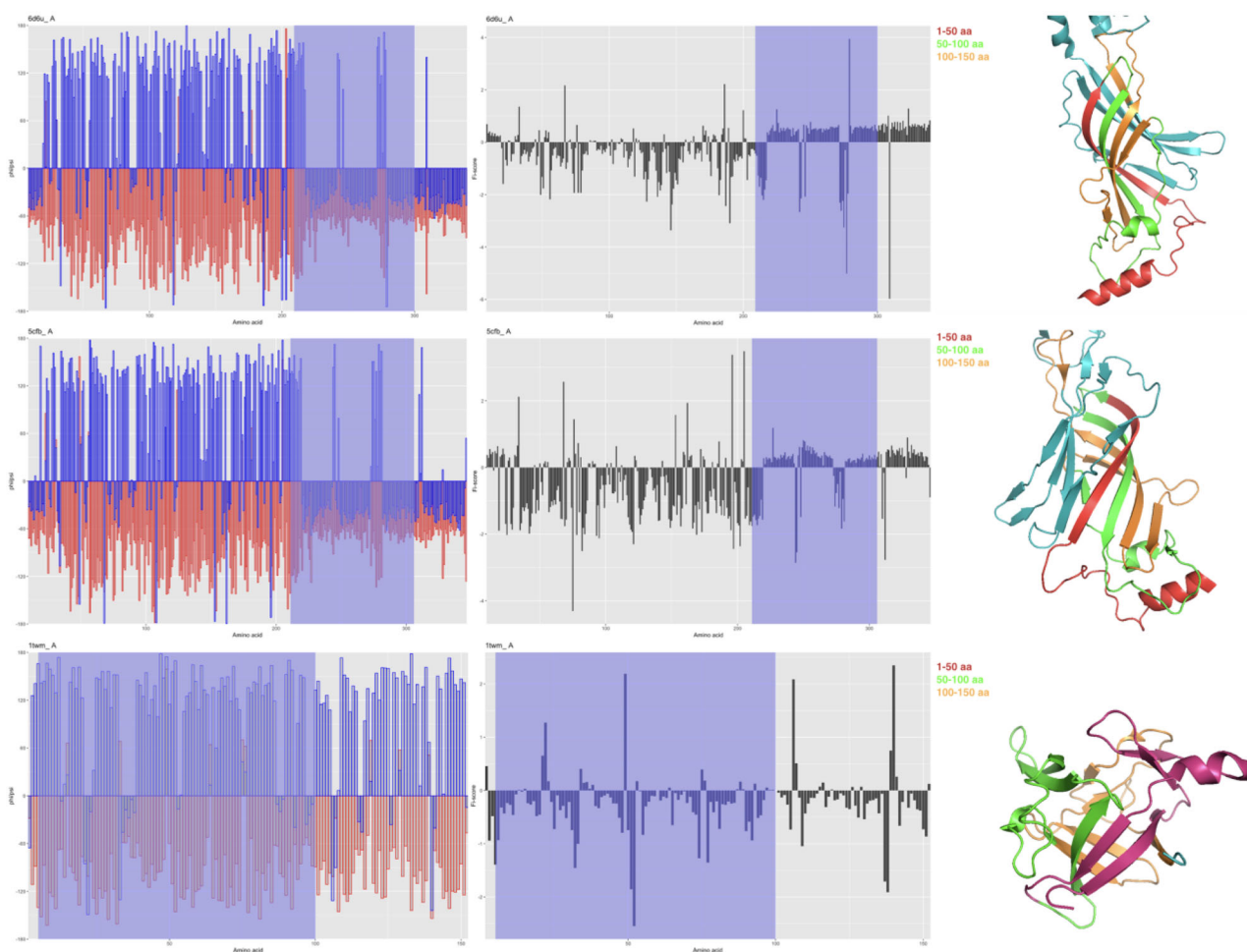
**Figure 1.** Superimposition of representative examples of human protein regions listed in Table 1 where PDB ID and colour is specified next to the structure. Images rendered with PyMol.

effectively compare the resulting score across different targets or sites. We looked into how conformational criteria can be extracted from the backbone torsion angles ( $\psi, \phi$ ) that follow a very specific local geometry to avoid steric clashes (Figure 2, Supplementary material); this led us to adopt a standard deviation normalisation for the observed torsional angles. While several different normalisation approaches exist for dihedral angles (Shen et al., 2018; Tosatto & Battistutta, 2007), the mathematical techniques directly depend on the parameter incorporation into further equations which, in our case, needed to be formulated in a way to preserve Ramachandran plot directionality based on positive and negative value so that multiplication operation allowed to predict either  $\beta$ -sheet/strand type of conformation (negative) or  $\alpha$ -helix (positive) for the most predominant secondary structure elements (Figures 1 and 2, Supplementary material). When the cumulative score is calculated the ultimate value can indicate the predominance of the said structures and in rarer situations a less dominant conformations, such as a left-handed  $\alpha$ -helix. Similarly, B-factor values needed to be scaled since values may be on different scales owing to dissimilar refinement procedures (Figures 3 and 4, Supplementary material) (Carugo, 2018; Carugo & Argos, 1998; Parthasarathy & Murthy, 1997; Yuan et al., 2003); we applied scaling specifically to take that into account where scale normalised values of B-factors ranged from 0 to 1 allowing them to be conceptually integrated into the fingerprint score equation (Figure 3, supplementary material). Finally, dividing the cumulative sum by the number of residues we can measure an average value for the region or Fi-score.

### Protein characterisation and Fi-score performance testing

A diverse set of 3352 randomly selected proteins was used for the comparative studies (Figure 1 and Table 1, supplementary

material) which allowed us to contrast varied regions of target proteins based on their Fi-score values. After seeing that Fi-score differentiated between different structural motifs of 50,043 element test set (Figure 1, supplementary material), we further probed Fi-score, normalised B-factor as well as dihedral angle distributions and sequence alignment data of selected proteins (Figure 1, Table 1; Figure 3, supplementary material). In addition, the selected region was scored for hydrophobicity and a RMSD value was identified for two target sequences. Target sequences that share higher similarity have closer Fi-score values which also correspond to a more similar distribution profile (Figure 2), for example, a sequence from human GABA-A receptor, subunit beta-2 (PDB ID: 6D6U) sharing 78.1% similarity with human glycine receptor alpha-3 (PDB ID: 5CFB) differ by 0.07702 in their Fi-scores. When compared to a case of 100% similarity as is for the chain A and B of human heat shock protein 90,  $\Delta$ Fi-score value drops to 0.04451. However, the sequence similarity alone does not play a defining role as illustrated by catalytic antibody 21H3 with hapten (PDB ID: 1UM4) and therapeutical antibody fragment of canakinumab light chains (PDB ID: 4G5Z) that although share 100% sequence similarity have almost a half of Fi-score difference between them; this is because the Fi-score captures the 3D distribution of the amino acids, side chain orientation and the predicted atom movements. The slight shifts in the amino acid and their side change orientation (Figure 1) will have a noticeable effect on the Fi-score. Moreover, protein regions that have large structural differences as showcased by the interleukin-1 beta mutant F146Y (PDB ID: 1TWM) and therapeutical antibody fragment of canakinumab (PDB ID: 4G5Z) will have large corresponding differences between the Fi-score and RMSD values (RMSD = 11.407 Å) (Table 1). In many cases smaller  $\Delta$ Fi-score values will mean that protein regions have similar structural and physicochemical profiles but in more ambiguous cases hydrophobicity analysis should be included as it indirectly captures the nature of amino acid composition



**Figure 2.** Dihedral angle and Fi-score distribution, respectively left and right panels, where  $\phi$  dihedral angle is represented in red and  $\psi$ —blue. Blue region delineates the protein section that was used to calculate Fi-score and in [Table 1, supplementary material](#). Right panels of the corresponding protein structure are colour coded in arbitrary increments of 50 amino acids. 3D molecule images rendered with PyMol and tables created with R/RStudio.

in the selected region as illustrated by the representative cases ([Table 1](#)). This information can be especially useful as it allows to compare protein regions of similar mobility or amino acid composition that have a matching structural profile, for example, the conserved DNA binding site of p53 (PDB ID: 2AC0) is quite similar to a region of SWIB/MDM2 domain (PDB ID: 2AXI) and a similar profile can be seen for the heavy and light chains of therapeutic antibody fragment of canakinumab (PDB ID: 4G5Z) ([Figure 1, Table 1](#)). These examples illustrate that depending on the 3D organisation of a region of interest, conservative substitutions of amino acids, dihedral angle and B-factor values will have an impact on the individual Fi-score values for amino acids and the overall cumulative score. There are many studies that support these findings as it has been established that B-factors can be used to identify flexibility in proteins and can also be linked to hydrophilicity as well as absolute net charge (Kuczera et al., 1990; Liu et al., 2014; Radivojac et al., 2004; Schlessinger & Rost, 2005; Vihinen et al., 1994). B-factors can also aid in identifying biologically active small molecules for a site of interest (Bornot et al., 2011; Li et al., 2017; Liu et al., 2014; Smith et al., 2003).

Another important criterion is the region size selected for the analysis since the Fi-score encapsulates conformational information and does not rely on sequence values alone

(where sequence influence arises in a form of dihedral angle and B-factor distribution) (Pang, 2016; Radivojac et al., 2004; Weiss, 2007; Yang et al., 2016; Yuan et al., 2003).

Selected window size for the analysis will have an effect on what information is contained within the Fi-score. A smaller window size of approximately 20–50 amino acids can reflect the profile of an average motif in a protein ([Figure 2](#)); however, larger window sizes averaging 100 or more amino acids reveal the averaged physicochemical information of that larger window size ([Figure 1](#)). This is especially evident when looking at individual Fi-score values per amino acid ([Figure 2](#)) where the Fi-score distribution captures different protein regions not easily recognised by looking at the dihedral angle or B-factor distribution alone ([Figure 2; Figures 2 and 3, supplementary material](#)).

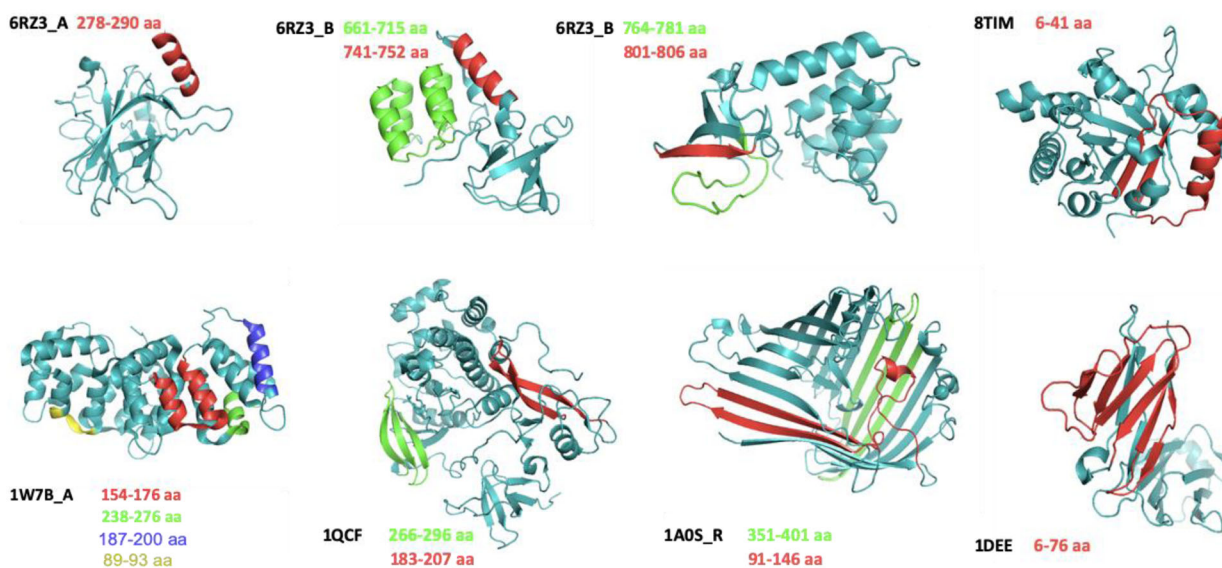
### **Verification of Fi-score's ability to capture structural topology features**

These observations prompted us to investigate a varied set of proteins with various structural elements ranging from  $\beta$ -sheets/strands to  $\alpha$ -helices as well as mixed or disordered regions ([Table 2, Figure 3](#)). By narrowing down to a unique structural motif we can capture not only its physicochemical



**Table 2.** Structural motif and domain physicochemical characterisation for selected protein structural elements and motifs.

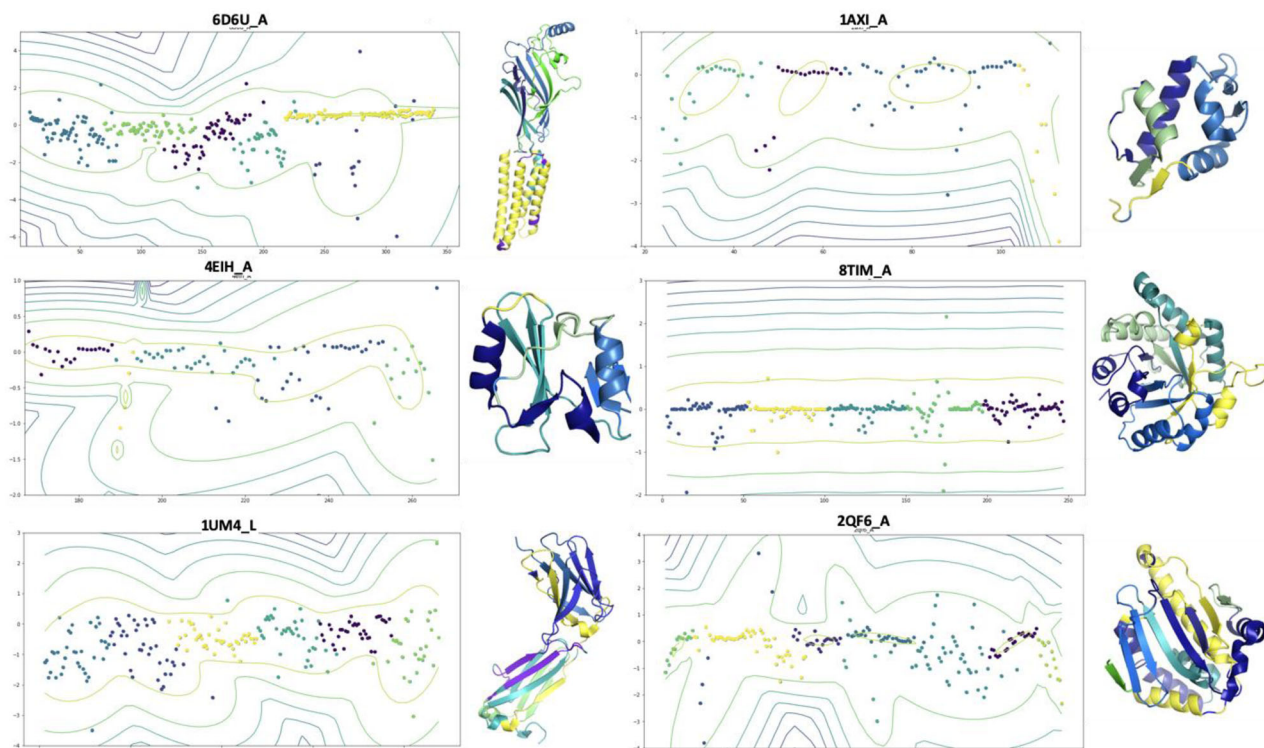
Protein PDB ID	Description	Amino acid number	Fi-score	Hydrophobicity score
6S34	$\alpha$ -helix with disordered linker	2–19	-0.1727632	0.3277778
6S34	Iregular C-end $\alpha$ -helix	11–17	0.6483593	-0.7
6RZ3, A	Single outer $\alpha$ -helix of cellular tumour antigen p53	278–290	0.06747168	-2.646154
6RZ3, B	Outer $\alpha$ -helix surrounded by smaller helices and unstructured regions of the carboxyl-terminal conserved region of inhibitor of apoptosis-stimulating protein of p53 (iASPP)	741–752	0.1493196	0.175
6RZ3, B	Four $\alpha$ -helices with long stretches of disordered linkers of the carboxyl-terminal conserved region of inhibitor of apoptosis-stimulating protein of p53 (iASPP)	661–715	-0.02906267	0.4854545
1W7B	Contorted $\alpha$ -helix joining two separate $\alpha$ -helices	154–176	-0.4146927	-0.1521739
1W7B	Contorted $\alpha$ -helix joining three separate $\alpha$ -helices	238–276	0.02634468	-0.4
1W7B	Single outer $\alpha$ -helix	187–200	0.0995633	-1.264286
1W7B	Single region of left handed $\alpha$ -helix like structure	89–93	-1.233356	0.72
1DEE, G	Protein A I mG binding domain, $\alpha$ -helical motif, chain G	1810–1852	0.122619	-0.7627907
1QCF	Antiparallel $\beta$ -strands	183–207	-0.8710873	-1.108
1QCF	Antiparallel $\beta$ -strands	266–296	-0.2612031	-0.4129032
1A0S, R	$\beta$ -barrel, with three random coil motifs	76–476	-0.3455974	-0.4897756
1A0S, R	Three antiparallel $\beta$ -sheets; outer pore region	351–401	-0.3455974	-0.4960784
1A0S, R	Three antiparallel $\beta$ -sheets connected via $\alpha$ -helix; inner pore region	91–146	-0.2178924	-0.5375
8TIM	TIM barrel	11–241	-0.04860643	-0.1311688
8TIM	TIM barrel motif of two $\beta$ -sheets and $\alpha$ -helix	6–41	-0.1594032	-0.09166667
1DEE, E	ImG Fab $\beta$ -sheet, chain E	2006–2076	-0.6376855	-0.2239437
1DEE, A	ImG Fab $\beta$ -sheet, chain A	6–76	-0.2326242	-0.2239437
6RZ3, B	A single stretch of $\beta$ -sheet strand	806–811	-1.700738	-1.016667
6RZ3, B	Disordered region of a protein	764–781	-0.7002029	-0.5944444

**Figure 3.** Representative protein structural motifs and regions from Table 2 are colour coded for specific amino acids. 3D molecule images rendered with PyMol.

properties but also reliably categorise it to either  $\alpha$ -helix or  $\beta$ -sheets/strands-like structures. However, structures that are a mixture of several components, e.g. PDB ID: 6S34 (Table 2), might have opposite sign values or values closer to 0 for Fi-scores because some less predominant structures of  $\alpha$ -helices and  $\beta$ -sheets occupy negative and positive basins, respectively.

The higher absolute Fi-score value, the more flexible the region is likely to be, for example, the outer  $\alpha$ -helix of DNA-binding domain of p53 (chain A, PDB ID: 6RZ34; Table 2, colour - red) might appear untethered and relatively flexible; however, based on the low absolute Fi-score value and a further inspection on the inter-chain H-bond formation (Figure

4, supplementary material), this structure is accurately predicted to be of a limited dynamic range. Other  $\alpha$ -helices complexes within the chain B of the inhibitor of apoptosis-stimulating protein of p53 (iASPP; chain B, PDB ID: 6RZ34; Table 2) are of varying flexibility because they are at the contact point between two chains, specifically: a shorter  $\alpha$ -helix (chain B) is less constrained by the polar contacts and interaction surface than the other participating elements of this contact site. Similar patterns can be observed in  $\alpha$ -helices and their complexes of annexin A2 (PDB ID: 1W7B, Table 2, Figure 3) where flexibility and the Fi-score value depends on the conformation. In the case of  $\beta$ -strands and  $\beta$ -sheets, these secondary structure elements have the same trend of



**Figure 4.** Representative proteins and their fi-score clustering based on density estimation GMM where colours of the clusters as well as density lines match structural element colours of a protein on the right. 3D molecule images rendered with PyMol and tables created with R/RStudio.

higher flexibility associated with higher Fi-score values (iASPP; PDB ID: 6RZ34, chain B; Table 2, Figure 3). More compact sites have minimal space of side chain and motif movement as can be seen in triose phosphate isomerase (PDB ID: 8TIM, chain A; Table 2, Figure 3). Finally, disordered regions or regions that combine several secondary structure elements might have a sign value depending on the dominating structural sub-motif. All of the above findings are supported by earlier studies showing that B-factors can act as indicators of the relative vibrational motion of atoms where low values belong to a well-ordered site, and the highest values come from the most flexible regions (Li et al., 2017; Obradovic et al., 2003; Pang, 2016; Siglioccolo et al., 2010; Tang et al., 2019; Yuan et al., 2003).

### **Fi-score classification to capture and predict topological features using machine learning**

Based on the findings that the Fi-score captures and allows to differentiate among varied protein regions, we wanted to check if applying clustering would allow us to categorise Fi-score values as we have already observed clear distribution patterns (Figure 1). However, some protein regions might be in transition states and thus, have similar or overlapping Fi-score values and in order to address that we selected Gaussian mixture models (GMM) (Dubey, 2004; Mann & Hermans, 2000; Parthasarathy & Murthy, 1997; Zhang et al., 2017). GMM is often categorised as a clustering algorithm, but it has much broader implications functioning as a density estimator. Since fundamentally GMM is a generative

probabilistic model, this algorithm was chosen to describe the distribution of the Fi-scores.

The covariance type for the fits of the majority of studied cases was left to be modelled as an ellipse of an arbitrary orientation for each cluster and the optimal number of components for a given dataset was determined using AIC and BIC approaches to avoid overfitting. Fi-score clustering revealed that GMM allows not only to capture different secondary structure elements (Figure 4) but at the same time group them into physicochemically similar units based on the dihedral angle determined side chain orientation and B-factor predicted amino acid oscillations amplitude. In the case study of catalytic antibody 21H3 with hapten (PDB ID: 1UM4, chains H and L; Figure 4; Figure 6, supplementary material) we can see that the Fi-score evaluation and clustering successfully determined complementarity determining region  $\beta$ -turns and different  $\beta$ -strands in the immunoglobulin fold. The heavy and light chain contact sites are also captured through different chain topology and relevant atomic movement. Another case study of triose phosphate isomerase (PDB ID: 8TIM) demonstrated how a Fi-score based method allows to differentiate secondary structure elements of  $\alpha$ -helices and loops at the C-terminal ends of the  $\beta$ -barrel which are known to be involved in catalytic activity (Reardon & Farber, 1995). Similarly, N-terminal loops performing a stabilising function were also distinguished from the surrounding structural elements. This ubiquitous enzyme fold can be further resolved into different motifs of interchanging  $\alpha$ -helices and  $\beta$ -strands forming the structure's core (Figure 4).

As illustrated, Fi-score centered analysis can be a powerful tool to gain insight into structural topology of a target of interest. Furthermore, by including density estimation

contours we can predict the changes of a protein region if, for example, the structure is not in a crystal but in a solution (Bryn Fenwick et al., 2014; Powers et al., 1993). This method could also be expanded to estimate effects of mutations and what changes in the Fi-score value are the most optimal. Finally, drug screening studies can benefit from classifying target sites and cross-referencing with known binders which could reduce off-target effects as well as allow to address and better understand cases of unspecific binding or dynamic instability.

## Discussion

*In silico* target evaluation and compound screening have become a focal point in drug discovery studies (Gangadharan et al., 2017); this paradigm shift from *in vitro* to computational setting during early stages of pilot studies represents a need to establish reliable approaches in selecting targets and evaluating pharmacological intervention strategies. The druggability of a protein of interest can be defined as the likelihood that the target will be amendable to functional modulation by a compound. This concept can be also extended to biologics and new therapeutic modalities where the main therapeutic requirement is that there is an active binding spot to be engaged by the said modulator (Dias & de Azevedo, 2008; Huang & Dixit, 2016). Thus, our research aim was to devise an effective way to capture structural and physicochemical features and use that to not only investigate sites of interest but also to classify the protein features providing a scalable way to compare proteins under investigation.

Protein conformation determination and capturing of the physicochemical properties remain one of the most important topics in drug discovery (Huang & Dixit, 2016; Yang et al., 2016). That is, defining protein regions that share similar dynamic range is a significant challenge and in order to address that we developed a method to capture the side chain as well as mean atomic displacement distribution to provide a value that can aid in comparing and characterising regions of interest which we call a fingerprint score or 'Fi-score'. We showed that the Fi-score can capture both local and distal information via dihedral angle and B-factor distribution which allows us to evaluate potential physicochemical properties and also extract information on structural motifs (Figures 1–3; Tables 1 and 2; Table 1, supplementary material). Numerous past reports (Hartmann et al., 1982; Kuczera et al., 1990; Liu et al., 2014; Radivojac et al., 2004; Schlessinger & Rost, 2005; Vihinen et al., 1994) have already established that B-factors can be employed to define protein region hydrophobicity, flexibility or can even be used for small molecule search against a target site (Bornot et al., 2011; Li et al., 2017; Liu et al., 2014; Smith et al., 2003); similarly, dihedral angles (De Juan et al., 2013; Faraggi et al., 2009; Heffernan, 2015; Schlessinger & Rost, 2005) are used for protein structure modelling and interaction predictions. However, despite these insights, to our knowledge, there have been no attempts to capture this information in a unified way. As a result, we, for the first time, demonstrate that

the information of both, B-factors and dihedral angles, can be successfully combined in a single equation.

One of the main challenges of successful therapeutics development is the establishment of the binding site profile that could be compared to other sites in a target or other proteins with similar features (Dias & de Azevedo, 2008; Fauman et al., 2011; Fuller et al., 2009; Gangadharan et al., 2017; Hartmann et al., 1982; Knapp, 2016; Li et al., 2017; Pérot et al., 2010; Santos et al., 2017). This is especially important when trying to minimise off-target effects or designing high-throughput virtual screenings with multiple hot spots in proteins (Bryn Fenwick et al., 2014; Powers et al., 1993). Our described method provides a solution by allowing to inspect the differences in dihedral angle and B-factor distributions as well as score individual motifs and compare them across all sites of interest. As illustrated, the Fi-score can provide valuable insights into structural profiles across different target groups (Figures 2–4). By applying machine learning approaches we can get density estimation contours which can then be used to predict the changes in a specific protein region (Figure 4).

The described methodology could aid in the identification of target families that are affected by drug treatment since fingerprinting does not rely on the scanning of sequence similarity but actually measures physicochemical properties of the binding site. Fi-score visualisation provides a way to capture amino acid interactions over a selected span of a protein sequence and the clustering of Fi-scores can reveal dynamically similar sites.

This strategy can become especially relevant in the future as the pharmaceutical industry is shifting toward more complex targets and protein complexes and this requires a systemic approach that could be easily applied to many different proteins and would not rely on just the sequence information but would also take into account multidimensional distributions (Brazhnik et al., 2002; Huang & Dixit, 2016).

All of this could help reduce costs and the time needed in computationally expensive screenings by helping to prioritise targets, their sites as well as estimate potential off-target effects. In addition, topological feature based evaluation could allow to predict compound action by juxtaposing similar sites based on the Fi-score when one site has known interactors and the other does not. Finally, our work sets the ground for protein feature scoring that could be used in a relational way across multiple targets and we aim with our future research to deliver a robust R package so that a scientific user could quickly test their Fi-score for a selected target.

## Conclusion

We provide a new method to characterise proteins prior to *in silico* screening by evaluating potentially dynamically active regions or predicting sites that share similar qualities in the side chain distribution and movement. Incorporating the Fi-score with other physicochemical parameters, such as hydrophobicity, could greatly improve detecting valuable

multiple sites within a target or capturing similar profiles across different targets which in turn could be subjected to docking studies. Moreover, we showed that by using machine learning approaches we can expand the analysis of multiple targets by extracting and defining structural elements and motifs of various proteins. Fi-score focused analysis can aid in not only primary target selection studies but also advance drug or biologics formulation methods by evaluating potential binding sites or interaction surfaces. In summary, this innovative biophysical analysis method could significantly improve target selection, pre-screening analysis and speed-up biologics engineering.

## Acknowledgements

The authors would like to thank Dr. Farhad Forouhar, Columbia University, for kindly offering technical advice on structure visualisation

## Authors' contributions

AK devised the methodology, performed the analysis and wrote the manuscript. CB and EV critically reviewed the manuscript and provided suggestions, MH critically reviewed and edited the manuscript. All authors read and approved the final manuscript.

## Disclosure statement

The authors declare having no competing interests.

## Ethical approval

The study did not require ethical approval and did not involve human participants or test animals.

## Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. CB is supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement #674983 (ITN-MIMIC).

## ORCID

Austė Kanapekaitė  <http://orcid.org/0000-0001-6829-4082>

## References

- BLAST: Basic Local Alignment Search Tool. (n.d.). <https://blast.ncbi.nlm.nih.gov/Blast.cgi>
- Bornot, A., Etchebest, C., & De Brevern, A. G. (2011). Predicting protein flexibility through the prediction of local structures. *Proteins*, 79(3), 839–852. <https://doi.org/10.1002/prot.22922>
- Brazhnik, P., De La Fuente, A., & Mendes, P. (2002). Gene networks: How to put the function in genomics. *Trends in Biotechnology*, 20(11), 467–472. [https://doi.org/10.1016/s0167-7799\(02\)02053-x](https://doi.org/10.1016/s0167-7799(02)02053-x)
- Bryn Fenwick, R., Van Den Bedem, H., Fraser, J. S., & Wright, P. E. (2014). Integrated description of protein dynamics from room-temperature X-ray crystallography and NMR. *Proceedings of the National Academy of Sciences of the United States of America*, 111(4), E445–E454. <https://doi.org/10.1073/pnas.1323440111>
- Buckle, A. M., Cramer, P., & Fersht, A. R. (1996). Structural and energetic responses to cavity-creating mutations in hydrophobic cores: Observation of a buried water molecule and the hydrophilic nature of such hydrophobic cavities. *Biochemistry*, 35(14), 4298–4305. <https://doi.org/10.1021/bi9524676>
- Carugo, O. (2018). How large B-factors can be in protein crystal structures. *BMC Bioinformatics*, 19(1), 61. <https://doi.org/10.1186/s12859-018-2083-8>
- Carugo, O., & Argos, P. (1998). Accessibility to internal cavities and ligand binding sites monitored by protein crystallographic thermal factors. *Proteins: Structure, Function, and Genetics*, 31(2), 201–213. [https://doi.org/10.1002/\(SICI\)1097-0134\(19980501\)31:2<201::AID-PROT9>3.0.CO;2-O](https://doi.org/10.1002/(SICI)1097-0134(19980501)31:2<201::AID-PROT9>3.0.CO;2-O)
- Carugo, O., & Argos, P. (1999). Reliability of atomic displacement parameters in protein crystal structures. *Acta Crystallographica Section D, Biological Crystallography*, 55(Pt 2), 473–478. <https://doi.org/10.1107/s0907444998011688>
- Carugo, O., & Eisenhaber, F. (1997). Probabilistic evaluation of similarity between pairs of three-dimensional protein structures utilizing temperature factors. *Journal of Applied Crystallography*, 30(5), 547–549. <https://doi.org/10.1107/S0021889897003427>
- De Juan, D., Pazos, F., & Valencia, A. (2013). Emerging methods in protein co-evolution. *Nature Reviews Genetics*, 14(4), 249–261. <https://doi.org/10.1038/nrg3414>
- Delano, W. L. (2002). *Pymol: An open-source molecular graphics tool*. CCP4 Newsletter On Protein Crystallography, 40, 82–92.
- Dias, R., & de Azevedo, W. F. (2008). Molecular docking algorithms. *Current Drug Targets*, 9(12), 1040–1047. <https://doi.org/10.2174/138945008786949432>
- Dubey, A. (2004). *Clustering protein sequence and structure space with infinite Gaussian mixture models*. Pacific Symposium on Biocomputing, vol. 9.
- EMBOSS programs, EMBL-EBI. (n.d.). <https://www.ebi.ac.uk/Tools/emboss/>
- Faraggi, E., Xue, B., & Zhou, Y. (2009). Improving the prediction accuracy of residue solvent accessibility and real-value backbone torsion angles of proteins by guided-learning through a two-layer neural network. *Proteins*, 74(4), 847–856. <https://doi.org/10.1002/prot.22193>
- Fauman, E. B., Rai, B. K., & Huang, E. S. (2011). Structure-based druggability assessment-identifying suitable targets for small molecule therapeutics. *Current Opinion in Chemical Biology*, 15(4), 463–468. <https://doi.org/10.1016/j.cbpa.2011.05.020>
- Fuller, J. C., Burgoyne, N. J., & Jackson, R. M. (2009). Predicting druggable binding sites at the protein-protein interface. *Drug Discovery Today*, 14(3–4), 155–161. <https://doi.org/10.1016/j.drudis.2008.10.009>
- Gangadharan, N. T., Venkatachalam, A. B., & Sugathan, S. (2017). High-throughput and in silico screening in drug discovery. *Bioresources and Bioprocess in Biotechnology*, 1, 247–273.
- Grant, B. J., Rodrigues, A. P. C., Elsayy, K. M., Mccammon, J. A., & Caves, L. S. D. (2006). Bio3d: An R package for the comparative analysis of protein structures. *Bioinformatics (Oxford, England)*, 22(21), 2695–2696. <https://doi.org/10.1093/bioinformatics/btl461>
- Hartmann, H., Parak, F., Steigemann, W., Petsko, G. A., Ponzi, D. R., & Frauenfelder, H. (1982). Conformational substates in a protein: Structure and dynamics of metmyoglobin at 80 K. *Proceedings of the National Academy of Sciences*, 79(16), 4967–4971. <https://doi.org/10.1073/pnas.79.16.4967>
- Heffernan, R. (2015). Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. *Scientific Reports.*, 5, 1–11.
- Huang, X., & Dixit, V. M. (2016). Drugging the undruggables: Exploring the ubiquitin system for drug development. *Cell Research*, 26(4), 484–498. <https://doi.org/10.1038/cr.2016.31>
- Knapp, S. (2016). Emerging target families: Intractable targets. In *Handbook of Experimental Pharmacology* (vol. 232, pp. 43–58). Springer.
- Kuczera, K., Kuriyan, J., & Karplus, M. (1990). Temperature dependence of the structure and dynamics of myoglobin. A simulation approach. *Journal of Molecular Biology*, 213(2), 351–373. [https://doi.org/10.1016/S0022-2836\(05\)80196-2](https://doi.org/10.1016/S0022-2836(05)80196-2)

- Kyte, J., & Doolittle, R. F. (1982). A simple method for displaying the hydrophobic character of a protein. *Journal of Molecular Biology*, 157(1), 105–132.
- Li, X., Anderson, M., Collin, D., Muegge, I., Wan, J., Brennan, D., Kugler, S., Terenzio, D., Kennedy, C., Lin, S., Labadia, M. E., Cook, B., Hughes, R., & Farrow, N. A. (2017). Structural studies unravel the active conformation of apo ROR $\gamma$ t nuclear receptor and a common inverse agonism of two diverse classes of ROR $\gamma$ t inhibitors. *The Journal of Biological Chemistry*, 292(28), 11618–11630. <https://doi.org/10.1074/jbc.M117.789024>
- Liu, Q., Li, Z., & Li, J. (2014). Use B-factor related features for accurate classification between protein binding interfaces and crystal packing contacts. *BMC Bioinformatics*, 15(S16), 1–16. <https://doi.org/10.1186/1471-2105-15-S16-S3>
- Mann, G., & Hermans, J. (2000). Modeling protein-small molecule interactions: Structure and thermodynamics of noble gases binding in a cavity in mutant phage T4 lysozyme L99A. *Journal of Molecular Biology*, 302(4), 979–989. <https://doi.org/10.1006/jmbi.2000.4064>
- Obradovic, Z., Peng, K., Vucetic, S., Radivojac, P., Brown, C. J., & Dunker, A. K. (2003). Predicting intrinsic disorder from amino acid sequence. *Proteins*, 53(S6), 566–572. <https://doi.org/10.1002/prot.10532>
- Okonechnikov, K., Golosova, O., Fursov, M., & UGENE Team (2012). Unipro UGENE: A unified bioinformatics toolkit. *Bioinformatics (Oxford, England)*, 28(8), 1166–1167. <https://doi.org/10.1093/bioinformatics/bts091>
- Osorio, D., Rondón-Villarreal, P., & Torres, R. (2015). Peptides: A package for data mining of antimicrobial peptides. *The R Journal*, 7(1), 4–14. <https://doi.org/10.32614/RJ-2015-001>
- Pang, Y. P. (2016). Use of multiple picosecond high-mass molecular dynamics simulations to predict crystallographic B-factors of folded globular proteins. *Heliyon*, 2(9), e00161. <https://doi.org/10.1016/j.heliyon.2016.e00161>
- Parthasarathy, S., & Murthy, M. R. N. (1997). Analysis of temperature factor distribution in high-resolution protein structures. *Protein Science*, 6(12), 2561–2567. <https://doi.org/10.1002/pro.5560061208>
- Pérot, S., Sperandio, O., Miteva, M. A., Camproux, A. C., & Villoutreix, B. O. (2010). Druggable pockets and binding site centric chemical space: A paradigm shift in drug discovery. *Drug Discovery Today*, 15(15–16), 656–667. <https://doi.org/10.1016/j.drudis.2010.05.015>
- Pfam. (n.d.). Home page. <https://pfam.xfam.org/>
- Powers, R., Clore, G. M., Garrett, D. S., & Gronenborn, A. M. (1993). Relationships between the precision of high-resolution protein NMR structures, solution-order parameters, and crystallographic B factors. *Journal of Magnetic Resonance, Series B*, 101(3), 325–327. <https://doi.org/10.1006/jmrb.1993.1051>
- Radivojac, P., Obradovic, Z., Smith, D. K., Zhu, G., Vucetic, S., Brown, C. J., Lawson, J. D., & Dunker, A. K. (2004). Protein flexibility and intrinsic disorder. *Protein Science*, 13(1), 71–80. <https://doi.org/10.1110/ps.03128904>
- RCSB PDB. (n.d.). Homepage. <https://www.rcsb.org/>
- Reardon, D., & Farber, G. K. (1995). The structure and evolution of alpha/beta barrel proteins. *FASEB Journal*, 9(7), 497–503. <https://doi.org/10.1096/fasebj.9.7.7737457>
- RStudio. (n.d.). Open source & professional software for data science teams. RStudio. <https://rstudio.com/>
- Santos, R., Ursu, O., Gaulton, A., Bento, A. P., Donadi, R. S., Bologa, C. G., Karlsson, A., Al-Lazikani, B., Hersey, A., Oprea, T. I., & Overington, J. P. (2017). A comprehensive map of molecular drug targets. *Nature Reviews Drug Discovery*, 16(1), 19–34. <https://doi.org/10.1038/nrd.2016.230>
- Schlessinger, A., & Rost, B. (2005). Protein flexibility and rigidity predicted from sequence. *Proteins*, 61(1), 115–126. <https://doi.org/10.1002/prot.20587>
- scikit-learn: Machine Learning in Python. (n.d.). scikit-learn 0.23.2 documentation. <https://scikit-learn.org/stable/>
- SCOP | Structural Classification of Proteins. (n.d.). <http://scop.mrc-lmb.cam.ac.uk/>
- Shen, Y., Roche, J., Grishaev, A., & Bax, A. (2018). Prediction of nearest neighbor effects on backbone torsion angles and NMR scalar coupling constants in disordered proteins. *Protein Science*, 27(1), 146–158. <https://doi.org/10.1002/pro.3292>
- Siglioccolo, A., Gerace, R., & Pascarella, S. (2010). "Cold spots" in protein cold adaptation: Insights from normalized atomic displacement parameters (B'-factors). *Biophysical Chemistry*, 153(1), 104–114. <https://doi.org/10.1016/j.bpc.2010.10.009>
- Smith, D. K., Radivojac, P., Obradovic, Z., Dunker, A. K., & Zhu, G. (2003). Improved amino acid flexibility parameters. *Protein Science*, 12(5), 1060–1072. <https://doi.org/10.1110/ps.0236203>
- Tang, H., Shi, K., Shi, C., Aihara, H., Zhang, J., & Du, G. (2019). Enhancing subtilisin thermostability through a modified normalized B-factor analysis and loop-grafting strategy. *The Journal of Biological Chemistry*, 294(48), 18398–18407. <https://doi.org/10.1074/jbc.RA119.010658>
- Tosatto, S. C. E., & Battistutta, R. (2007). TAP score: Torsion angle propensity normalization applied to local protein structure evaluation. *BMC Bioinformatics*, 8, 155. <https://doi.org/10.1186/1471-2105-8-155>
- Vihinen, M. (1987). Relationship of protein flexibility to thermostability. *Protein Engineering*, 1:477–480.
- Vihinen, M., Torkkila, E., & Riikonen, P. (1994). Accuracy of protein flexibility predictions. *Proteins*, 19(2), 141–149. <https://doi.org/10.1002/prot.340190207>
- Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods*, 17(2), 228–243. <https://doi.org/10.1037/a0027127>
- Weiss, M. S. (2007). On the interrelationship between atomic displacement parameters (ADPs) and coordinates in protein structures. *Acta Crystallographica Section D, Biological Crystallography*, 63(Pt 12), 1235–1242. <https://doi.org/10.1107/S0907444907052146>
- Yang, J., Wang, Y., & Zhang, Y. (2016). ResQ: An approach to unified estimation of B-factor and residue-specific error in protein structure prediction. *Journal of Molecular Biology*, 428(4), 693–701. <https://doi.org/10.1016/j.jmb.2015.09.024>
- Yin, H., Li, Y.-Z., & Li, M.-L. (2011). On the relation between residue flexibility and residue interactions in proteins. *Protein and Peptide Letters*, 18(5), 450–456. <https://doi.org/10.2174/092986611794927974>
- Yuan, Z., Zhao, J., & Wang, Z.-X. (2003). Flexibility analysis of enzyme active sites by crystallographic temperature factors. *Protein Engineering*, 16(2), 109–114. <https://doi.org/10.1093/proeng/gzg014>
- Zhang, H., Jiang, T., Shan, G., Xu, S., & Song, Y. (2017). Gaussian network model can be enhanced by combining solvent accessibility in proteins. *Scientific Reports*, 7, 1–13.
- Zhao, G., & London, E. (2006). An amino acid "transmembrane tendency" scale that approaches the theoretical limit to accuracy for prediction of transmembrane helices: Relationship to biological hydrophobicity. *Protein Science*, 15(8), 1987–2001. <https://doi.org/10.1110/ps.062286306>

## Integrative *omics* approaches for new target identification and therapeutics development

### 5. *Fiscore* package: effective protein structural data visualisation and exploration

The experimental chapter is based on the published software package and publication in preparation

1. [Kanapeckaitė A.](#) *Fiscore: Effective Protein Structural Data Visualisation and Exploration*. CRAN. 2021 Sep. 02. Version 0.1.3; <https://cran.r-project.org/web/packages/Fiscore/index.html>
2. [Kanapeckaitė A.](#) *Fiscore: effective protein structural data visualisation and exploration*. *Accepted and in preparation*.

#### Conclusion of this chapter

My goal when developing the *Fiscore* package was to allow a user-friendly exploration of PDB structural data and the integration of that information into various machine learning methods. The package was benchmarked through several analytical stages that involved a diverse set of proteins (3352) to assess scoring principles and package functionalities (1337 structures). With a number of helpful functions, including distribution analyses or hydrophobicity assessment in the context of structural elements, *Fiscore* enables the exploration of new target families and comprehensive data integration since the described fingerprinting captures protein sequence and physicochemical properties. Such analyses could be very helpful when exploring therapeutically relevant proteins. Similarly, *Fiscore* could aid in drug repurposing studies when a chemical compound needs to be juxtaposed to a number of potential targets. This was also demonstrated during a native ligand search for the Nur77 protein. Thus, the *Fiscore* package provides an extensive analytical environment where in-depth analyses are streamlined for non-experts.

#### Contribution to this chapter (100%)

- Methodology development, equation and scoring function derivation, as well as machine learning pipeline implementation.
- Developed new programmatic features expanding various structural analyses.
- Performed software package development and testing.
- Conceptualised and wrote the documentation files and vignettes, including the figure preparation.
- Conceptualised and wrote the manuscript, including the figure preparation.
- Corresponding author and maintainer.



## Fiscore package: Effective protein structural data visualisation and exploration

Auste Kanapeckaite

School of Pharmacy, University of Reading, Hopkins Building, Reading RG6 6UB, United Kingdom

### A B S T R A C T

The lack of bioinformatics tools to quickly assess protein conformational and topological features motivated to create an integrative and user-friendly R package. Moreover, the *Fiscore* package implements a pipeline for Gaussian mixture modelling making such machine learning methods readily accessible to non-experts. This is especially important since probabilistic machine learning techniques can help with a better interpretation of complex biological phenomena when it is necessary to elucidate various structural features that might play a role in protein function. Thus, *Fiscore* builds on the mathematical formulation of protein physicochemical properties that can aid in drug discovery, target evaluation, or relational database building. In addition, the package provides interactive environments to explore various features of interest. Finally, one of the goals of this package was to engage structural bioinformaticians and develop more robust and free R tools that could help researchers not necessarily specialising in this field. Package *Fiscore* (v.0.1.3) is distributed free of charge via CRAN and Github.

### 1. Introduction

*Fiscore* R package was developed to quickly take advantage of protein topology/conformational feature assessment and perform various analyses allowing a seamless integration into relational databases as well as machine learning pipelines [1]. The package builds on protein structure and topology studies which led to the derivation of the Fi-score equation capturing protein dihedral angle and B-factor influence on amino acid residues (Eqs. (1) and (2)) [1]. The introduced tools can be very beneficial in rational therapeutics development where successful engineering of biologics, such as antibodies, relies on the characterisation of potential binding or contact sites on target proteins [1,2]. Moreover, translating structural data into scores can help with target classification, target-ligand information storage, screening studies, or integration into machine learning pipelines [1,2]. As a result, Fi-score, a first-of-its-kind *in silico* protein fingerprinting approach, created a premise for the development of a specialised and freely distributed R package to assist with protein studies and new therapeutics development [1].

*Fiscore* package allows capturing dihedral angle and B-factor effects on protein topology and conformation. Since these physicochemical characteristics could help with the identification or characterisation of a binding pocket or any other therapeutically relevant site, it is important to extract and combine data from structural files to allow such information integration [1,3,4]. Protein dihedral angles were selected as they contain information on the local and global protein structural features where protein backbone conformation can be highly accurately recreated based on the associated dihedral angles [1,4]. Furthermore, since Ramachandran plot, which provides a visualisation for dihedral angle distributions, namely  $\phi$  (phi) and  $\psi$  (psi), allows only a holistic description of conformation and cannot be integrated with traditional para-

metric or non-parametric density estimation methods, a specific transformation was required to use this data. An additional parameter, specifically the oscillation amplitudes of the atoms around their equilibrium positions (B-factors) in the crystal structures, was also used. B-factors encompass a lot of information on the overall biomolecule structure; for example, these parameters depend on conformational disorder, thermal motion paths, and the rotameric state of amino acids side-chains. B-factors also show dependence on the three-dimensional structure as well as protein flexibility [1,4]. Normalised dihedral angles (standard deviation scaling to account for variability and distribution) and scaled B-factors (min-max scaling) (Eq. (1)) were integrated into the Fi-score equation (Eq. (2)). It is important to highlight that B-factors need to be scaled so that different structural files can be compared and that the dihedral angle normalisation transforms angular data into adjusted values based on the overall variability [1]. Thus, combining dihedral angle and B-factor values into a single parameter provides a way to extract information on individual residues, residue clusters, motifs, and structural features. This information can be efficiently transferred into machine learning to detect data characteristics not easily identifiable otherwise.

$$B_{i-norm} = \frac{B_i - B_{min}}{B_{max} - B_{min}}$$

**Equation 1.** Min-max normalisation and scaling of B-factors where  $B_{i-norm}$  is a scaled B-factor,  $B_i$  - B-factor for a selected  $C_\alpha$  atom in a chain,  $B_{max}$  - the largest B-factor value for all  $C_\alpha$  B-factors in a protein,  $B_{min}$  - the smallest B-factor value for all  $C_\alpha$  B-factors in a protein. B-factor normalisation is based on the full length protein.

$$Fiscore = \frac{1}{N} \sum_i \frac{\phi_i \psi_i}{\sigma_{\phi_i} \sigma_{\psi_i}} B_{i-norm}$$

**Equation 2.** Fi-score evaluation where N is the total number of atoms for which dihedral angle information is available,  $\phi$  and  $\psi$  values represent dihedral angles for a specific  $C_\alpha$  atom,  $\sigma_{\phi_i}$  and  $\sigma_{\psi_i}$  represent cor-

E-mail address: [auste.kan@algorithm379.com](mailto:auste.kan@algorithm379.com)

<https://doi.org/10.1016/j.ailsci.2021.100016>

Received 16 November 2021; Accepted 18 November 2021

Available online 26 November 2021

2667-3185/© 2021 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

responding standard deviations for the torsion angles and  $B_{i-norm}$  is a normalised B-factor value for the  $C_{\alpha}$  atom. B-factor,  $\sigma_{\phi_i}$  and  $\sigma_{\psi_i}$  normalisation is based on the full length protein.

In order to identify meaningful clusters based on the structural complexity, Gaussian mixture models (GMM) were selected as a primary machine learning classifier [1]. The strength of GMM lies in the probabilistic model nature since all data points are assumed to be derived from a mixture of a finite number of Gaussian distributions with unknown parameters [1,5]. Consequently, the soft classification of GMM where a data point has a probability of belonging to a cluster is much more suitable to assess biological parameters compared to other hard classification techniques in machine learning, such as k-means, which provide only a strict separation between classes. GMM pipeline offers a number of benefits to categorise protein structural features and the information can be used to explore amino acid grouping based on their physicochemical parameters. The designed GMM implementation takes care of the information criterion assessment to fine tune the number of clusters for modelling and predicts the best suited model for the expectation-maximisation (EM) algorithm to maximise the likelihood of data point assignments [1,5]. As a result, protein residues can be grouped based on their Fi-scores where this information can be used to identify emerging patterns in the protein conformation or topology.

Nur77 protein was used as a case example to demonstrate various package functionalities. Nuclear receptor subfamily 4 group A member 1 (NR4A1), also known as Nur77/TR3/NGFIB, is a member of the nuclear receptor superfamily and regulates the expression of multiple target genes [6]. This nuclear receptor is classified as an orphan receptor since there are no known endogenous ligands. Nur77 has the typical structure of a nuclear receptor which consists of N-terminal, DNA binding, and ligand-binding domains. This regulatory protein plays many potentially therapeutically relevant roles regulating cell proliferation and apoptosis [6]. Consequently, the Nur77 protein is an excellent example to highlight how in-depth structural analysis and classification could be used in better understanding protein functions and finding druggable binding sites or identifying ligands.

Based on the need to develop integratable and specialised tools for protein analyses, the *Fiscore* package was developed to assist with a wide spectrum of research questions ranging from exploratory analyses to therapeutic target assessment (Fig. 1). The introduced set of new tools provides an interactive exploration of targets with an easy integration into downstream analyses. Importantly, the package and associated tools are written to be easy to use and freely available facilitating analyses for non-specialists in structural biology or machine learning.

## 2. Methods

*Fiscore* package architecture is divided into exploratory and advanced functions (Fig. 1). Several key packages, such as ggplot2 [7], Bio3D [8], plotly [9], and mclust [10], are also employed to create an easy-to-use analytical environment where a user-friendly machine learning pipeline of GMM [1] allows for a robust structural analysis. GMM implementation is designed to include the optimal cluster number evaluation (Bayesian information criterion; BIC), automatic model fitting in the EM phase of clustering, model-based hierarchical clustering, density estimation, as well as discriminant analysis [1,11]. Researchers also have an option to perform advanced exploratory studies or integrate the package into their development pipelines. *Fiscore* also takes care of raw data pre-processing and evaluation with optional settings to adjust how the analyses are performed. The package was built using functional programming principles with several R S3 methods to create objects for PDB files [12]. *Fiscore* is accompanied by documentation and vignette files to help the users with their analyses [11]. Since PDB files are typically large, the documentation provides a compressed testing environment as well as a detailed tutorial. Additional visualisations were generated with PyMol [13]. Proteins were retrieved from the Protein Data Bank database [14]. Protein sequence alignments were performed with PSI-

BLAST using default parameters and a single iteration [15]. Hydrophobicity plots for Nur77 functional analysis were generated with the following parameters: window = 15, weight = 25, model="exponential". Student's *t*-test (two-sided, unpaired, sig. level=95%) was performed in the R programming environment.

## 3. Results

### 3.1. Data preparation

The workflow begins with the PDB file pre-processing and preparation. The user should also generally assess if the structure is suitable for the analysis; that is, the crystallographic data provides a good resolution and there are no or a minimal number of breakages within the reported structure. Function *PDB\_process* takes a PDB file name which can be expressed as 6KZ5.pdb or path/to/the/file/6KZ5.pdb. One of the function's dependencies is package Bio3D [8], this useful package provides several tools to begin any PDB file analysis. In addition, the *PDB\_process* function can take a path parameter which can point to a directory where to split PDB files into separate chain files (necessary for the downstream analysis). If this option is left empty, a folder in the working directory will be created automatically. If the user splits multiple PDB files in a loop, they will be continuously added to the same folder. After the processing, the function *PDB\_process* returns a list of split chain names. It is important to highlight that PDB files need to be split for the downstream processing so that separate chains can be analysed independently.

After a file or files are pre-processed the function *PDB\_prepare* can be used to prepare a PDB file to generate Fi-score and normalised B-factor values as well as secondary structure designations. The function takes a PDB file name that was split into separate chains, e.g., 6KZ5\_A.pdb, where a letter designates a split chain. The file is then cleaned and only the complete entries for amino acids are kept for the analysis, i.e., amino acids from the terminal residues that do not contain both dihedral angles are removed. The function returns a data frame with protein secondary structure information 'Type', Fi-score values per residue 'Fi\_score', as well as normalised B-factor values for each amino acid  $C_{\alpha}$  'B\_normalised' (Fig. 2). Extracting protein secondary structure information, i.e., 'Type', helps to prepare a data object so that the information about a target can be supplied into cheminformatics or other bioinformatics pipelines where structural features are important to assess protein sites and amino acid composition. These features are new and extend structural file exploration possibilities compared to, for example, other software packages, such as Bio3D [8].

Function calls are simple and user-friendly:

```
#General function for pre-processing raw PDB files
```

```
pdb_df<-PDB_process(pdb_path)
```

```
#Cleaning and preparation of PDB file
```

```
pdb_df<-PDB_prepare(pdb_path)
```

```
#Explore the output
```

```
head(pdb_df)
```

```
#The package allows to call test data directly  
for the Nur77 example file
```

```
pdb_path<- system.file("extdata", "6kz5.pdb", package="Fiscore")
```

### 3.2. Exploratory analyses

The scope of the exploratory analyses provides options to evaluate physicochemical parameters, such as dihedral angles, B-factors, or hydrophobicity scores, and visualise their distribution (Fig. 1).





Fig. 1. Schematic visualisation of the package features.

##		phi	psi	chi1	chi2	chi3	chi4	chi5	df_resno
##	31.A.ALA	54.94701	53.80667	NA	NA	NA	NA	NA	31
##	32.A.ASN	-60.91976	-18.01379	-157.93838	-76.10748	NA	NA	NA	32
##	33.A.LEU	-64.18792	-45.94048	176.74103	46.17107	NA	NA	NA	33
##	34.A.LEU	-65.44501	-38.46630	-88.61643	158.49518	NA	NA	NA	34
##	35.A.THR	-67.68541	-43.43184	75.18019	NA	NA	NA	NA	35
##	36.A.SER	-76.88914	-17.40880	157.45159	NA	NA	NA	NA	36
##		df_res	B_factor	B_normalised	Fi_score	Type			
##	31.A.ALA	ALA	30.07	0.35506724	0.37663226	Right-handed alpha helix			
##	32.A.ASN	ASN	15.68	0.18227666	0.07176640	Right-handed alpha helix			
##	33.A.LEU	LEU	7.79	0.08753602	0.09261096	Right-handed alpha helix			
##	34.A.LEU	LEU	9.85	0.11227185	0.10140389	Right-handed alpha helix			
##	35.A.THR	THR	20.79	0.24363593	0.25696344	Right-handed alpha helix			
##	36.A.SER	SER	23.69	0.27845821	0.13372748	Right-handed alpha helix			

Fig. 2. PDB file processing output.

Basic analyses are accessed through simple function calls to explore how dihedral angles and B-factors are distributed. These analyses offer interactive and easy visualisations of key parameters that are currently not offered in any other package. For example, while Bio3D [8] has many useful functionalities for the exploration of PDB files, 'Fi\_score' extends exploratory analyses by allowing a simplified and in-depth look into the key physicochemical parameters, such as B-factor value visualisation or generation of Ramachandran plots. Similarly, other freely available tools (distributed as an online service), such as ExPASy ProtScale [16,17], provide only one dimensional assessment without incorporating structural features and do not process PDB files. 'Fi\_score', however, combines sequence, structural, and physicochemical analyses in simple function calls to quickly explore the user data.

```
#Calling a Ramachandran plot function
phi_psi_plot(pdb_df)
#Visualisation of dihedral angle juxtaposed distributions
phi_psi_bar_plot(pdb_df)
#B plot value visualisation
B_plot_normalised(pdb_df)
#Interactive plot to map amino acids via 2D distribution
```

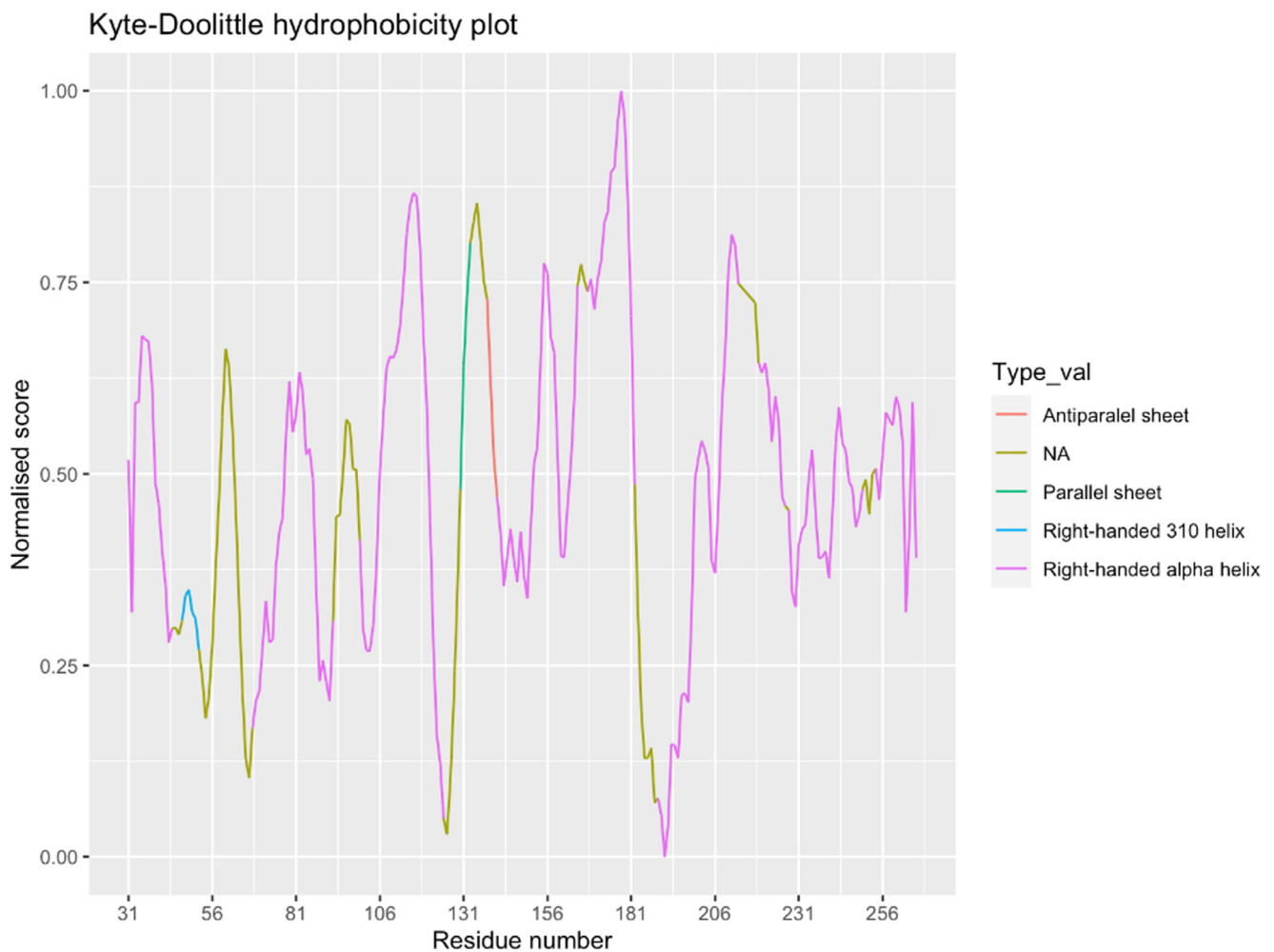


Fig. 3. Hydrophobicity plot with secondary structure superimposition.

#to precisely see what parameters an individual amino acid has

```
phi_psi_interactive(pdb_df)
```

#3D visualisation of dihedral angles and B-factor values

```
phi_psi_3D(pdb_df)
```

An especially useful functionality is the hydrophobicity visualisation with the superimposed secondary structure elements. To the author's knowledge, there are currently no tools implementing such a visualisation (Fig. 3). In contrast to ExPASy ProtScale [16,17], it is possible to visualise hydrophobicity values and their corresponding secondary structure elements as extracted from the PDB file. Such an assessment provides a direct way to compare structural features based on their affinity to water. This can be very helpful in evaluating or predicting potential binding sites as well as bioengineering new proteins.

The package provides an easy to use wrapper:

```
#Alternatively an exponential model can be selected
```

```
hydrophobicity_plot(pdb_df,window = 9,weight = 25,model = "exponential")
```

The nuclear receptor was assessed to provide a case example for the introduced hydrophobicity analysis. The evaluation revealed an overall dynamic profile for the protein. Moreover, Nur77 evidently contains a

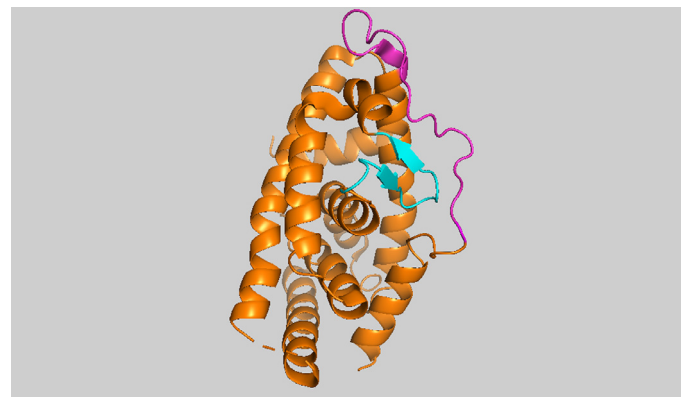


Fig. 4. Nur77 protein where magenta highlights are used to define a likely disordered region between 50 and 70 amino acids and the cyan color indicates a region between 127 and 140 amino acids.

relatively large number of right-handed alpha helices with the majority showing a hydrophobic profile, i.e., the larger the score, the more hydrophobic the region. Some likely disordered regions can be seen spanning 50–70 amino acids (Fig. 4). Another interesting region is around 126–136 amino acids since these amino acids undergo significant shifts in their hydrophilicity and hydrophobicity. Similarly, the region around 180–210 amino acids appears to be actively changing preferences from

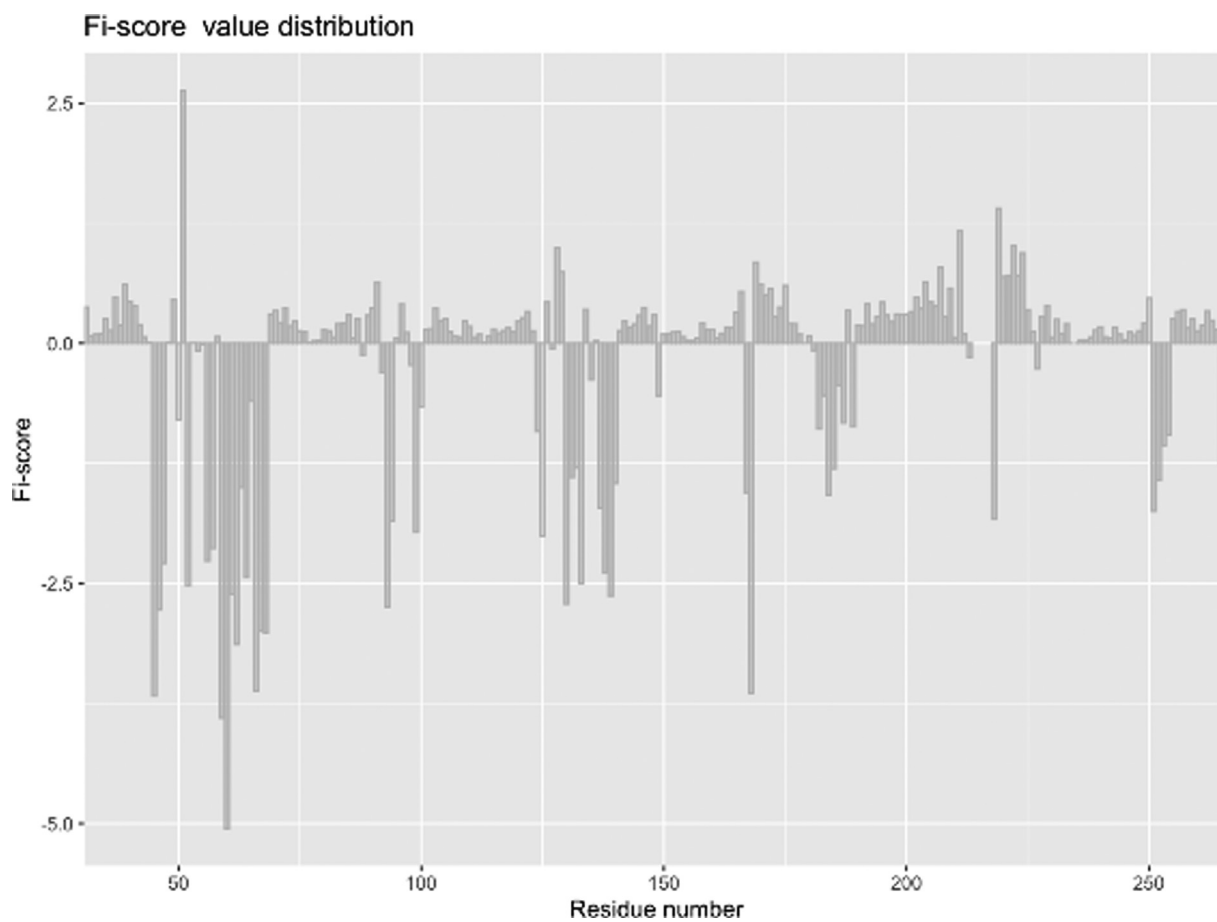


Fig. 5. Fi-score distribution for Nur77.

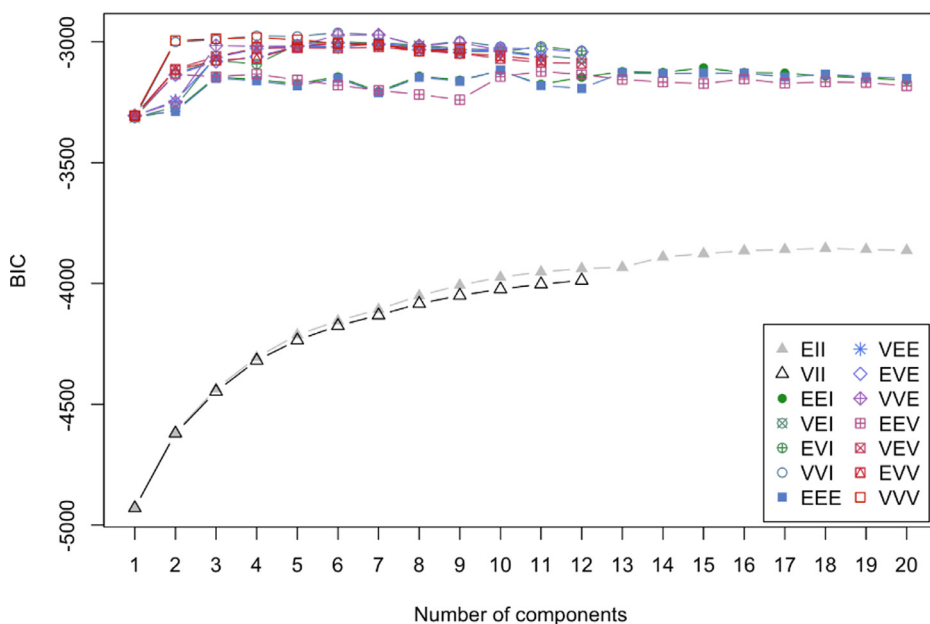


Fig. 6. Gaussian mixture modelling output showing Bayesian information criterion evaluation.

little solvent to being solvent exposed. This might suggest that the site undergoes considerable movements or actively engages other proteins or the DNA sequence. The disordered elements in this sequence stretch also imply that the region has to likely accommodate various rearrangements. Thus, studying these sites could provide hints at functionally important protein domains or subdomains (Figs. 3 and 4). Finally, evalu-

ating N and C terminal sites for the purpose of protein engineering, we can see that a histidine tag would not significantly disrupt the conformation of the molecule and the C-terminus is probably the best site for the tag.

It is worth commenting on the derivation of the hydrophobicity scoring since the algorithmic nature of the process provides several impor-

```
## Best BIC values:
##      VVI,6      VVE,7      VVE,6
## BIC      -2962 -2971.310677 -2971.756243
## BIC diff      0      -9.310508      -9.756073
## -----
## Dimension reduction for model-based clustering and classification
## -----
##
## Mixture model type: Mclust (VVI, 6)
##
## Clusters n
## 1 15
## 2 16
## 3 41
## 4 87
## 5 39
## 6 34
##
## Estimated basis vectors:
##      Dir1      Dir2
## Residue_number 0.00047634 0.056189
## Fi_score      0.99999989 -0.998420
##
##      Dir1      Dir2
## Eigenvalues 0.74861 0.55583
## Cum. %      57.38945 100.00000
```

Fig. 7. Output table for Gaussian mixture modelling evaluation.

tant analytical angles. The function builds on the Kyte-Doolittle hydrophobicity scale [1,18] to detect hydrophobic regions in proteins. Regions with a positive value are hydrophobic and those with negative values are hydrophilic. This scale can be used to identify both surface-exposed as well as transmembrane regions, depending on the

window size used. However, to make comparisons easier, the original scale is transformed from 0 to 1 (similar scaling is also implemented in Expsy ProtScale [16,17]). The function requires a PDB data frame generated by *PDB\_prepare* and the user needs to specify a window parameter to determine the size of the window for hydrophobicity calculations. The selection must be any odd number between 3 and 21 with the default being 21. Another parameter is weight that needs to be supplied to the function to establish a relative weight of the window edges compared to the window center (%); the default setting is 100%. Finally, a model parameter provides an option for weight calculation; that is, the selection determines whether weights are calculated linearly ( $y = k \cdot x + b$ ) or exponentially ( $y = a \cdot b^x$ ); the default model is 'linear'. The function evaluates each amino acid in a selected window where a hydrophobic influence from the surrounding amino acids is calculated in. While the terminal amino acids cannot be included into the window for centering and weighing, they are assigned unweighted values based on the Kyte-Doolittle scale [18]. The plot values are all scaled from 0 to 1 so that different proteins can be compared without the need to convert.

Thus, the hydrophobicity analysis can be especially useful when preparing to engineer proteins for various expression systems as the superimposition of structural features and hydrophobicity scores can help deciding if a protein region or domain is likely to be solvent exposed or prefer hydrophobic environments. For example, assessing the hydrophobicity and structural milieu of the N or C terminal amino acids can help selecting which terminal site should be tagged (as was demonstrated with Nur77). Moreover, this tool could be broadly applied in drug discovery studies involving the assessment of protein-protein interactions, protein-nucleic acid interactions, and membrane association events based on physicochemical characteristics.



Fig. 8. The Nur77 protein cluster identification with secondary structure elements.



Fig. 9. Nur77 cluster identification.

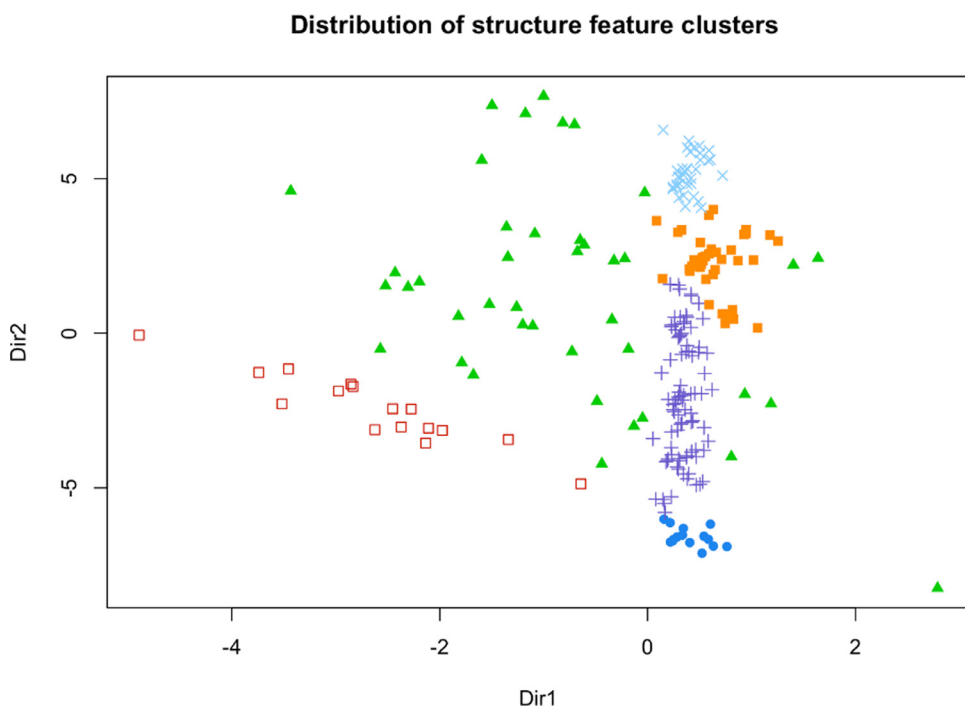


Fig. 10. Dimension reduction plot for the identified clusters.

### 3.3. Advanced analyses

Advanced analyses provide an opportunity to evaluate F1-score distributions and take advantage of a streamlined GMM pipeline (Fig. 1). The main impetus for the development of this pipeline was the need for functions and data modelling tools that could be made freely

accessible to non-experts. By contrast, commercial solutions, namely Schrödinger chemical simulation software [19], or non-commercial/semi-commercial solutions, including PSIPRED, AutoDock, MGLtools, and Expaty [16,20–22], lack a simple software platform to summarise and assess protein structural data that can be integrated into machine learning. While the mentioned software suites or online workbenches

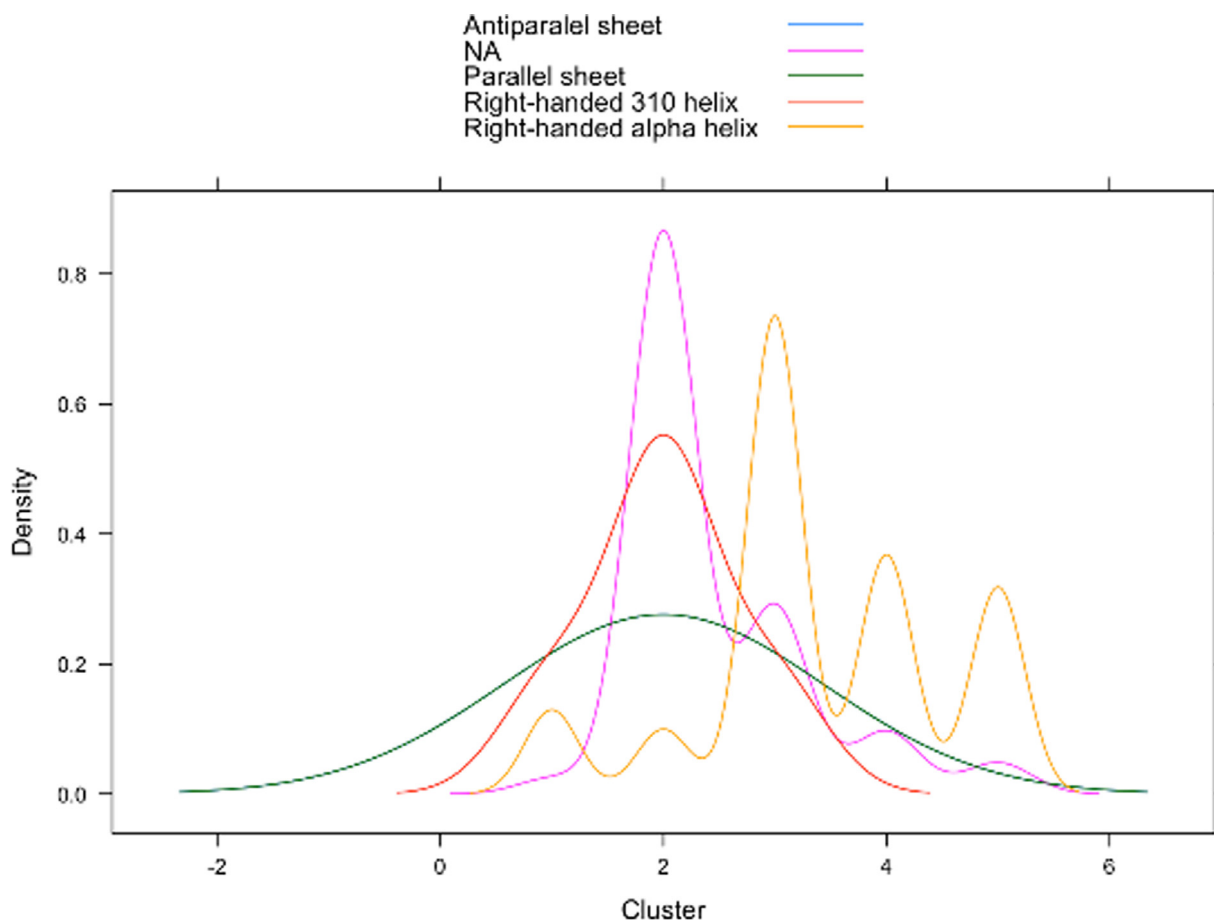


Fig. 11. Protein cluster density plots.

provide many useful functionalities, there is no one solution to use machine learning based inferences on the user's structural data. '*Fi\_score*' enables researchers to quickly extract, assess, and summarise key features of their data and incorporate that information into downstream analyses or custom pipelines. That is, more advanced users are also given opportunity to supply custom parameters for the GMM workflow and extract probabilities from the output to use scores in other analyses or integrate the values in their own discovery pipelines.

#Fi-score distribution plot to explore scores for corresponding amino acids

```
Fi_score_plot(pdb_df)
```

```
#Fi-score for a selected region
```

```
#this value for multiple sites can be stored in relational databases
```

```
Fi_score_region(pdb_df,50,70)
```

```
#Plot of Fi-score values with superimposed secondary structures
```

```
Fiscore_secondary(pdb_df)
```

For example, a Fi-score distribution plot captures several interesting regions in Nur77 around the 50, 130, and 180 amino acids (Fig. 5) that coincide with the Fi-score shifts and mirroring patterns. Some other regions are also picked up which should be studied in more detail based on the amino acid composition and 3D conformations. The uncovered characteristics can be juxtaposed to other similar sites to better understand

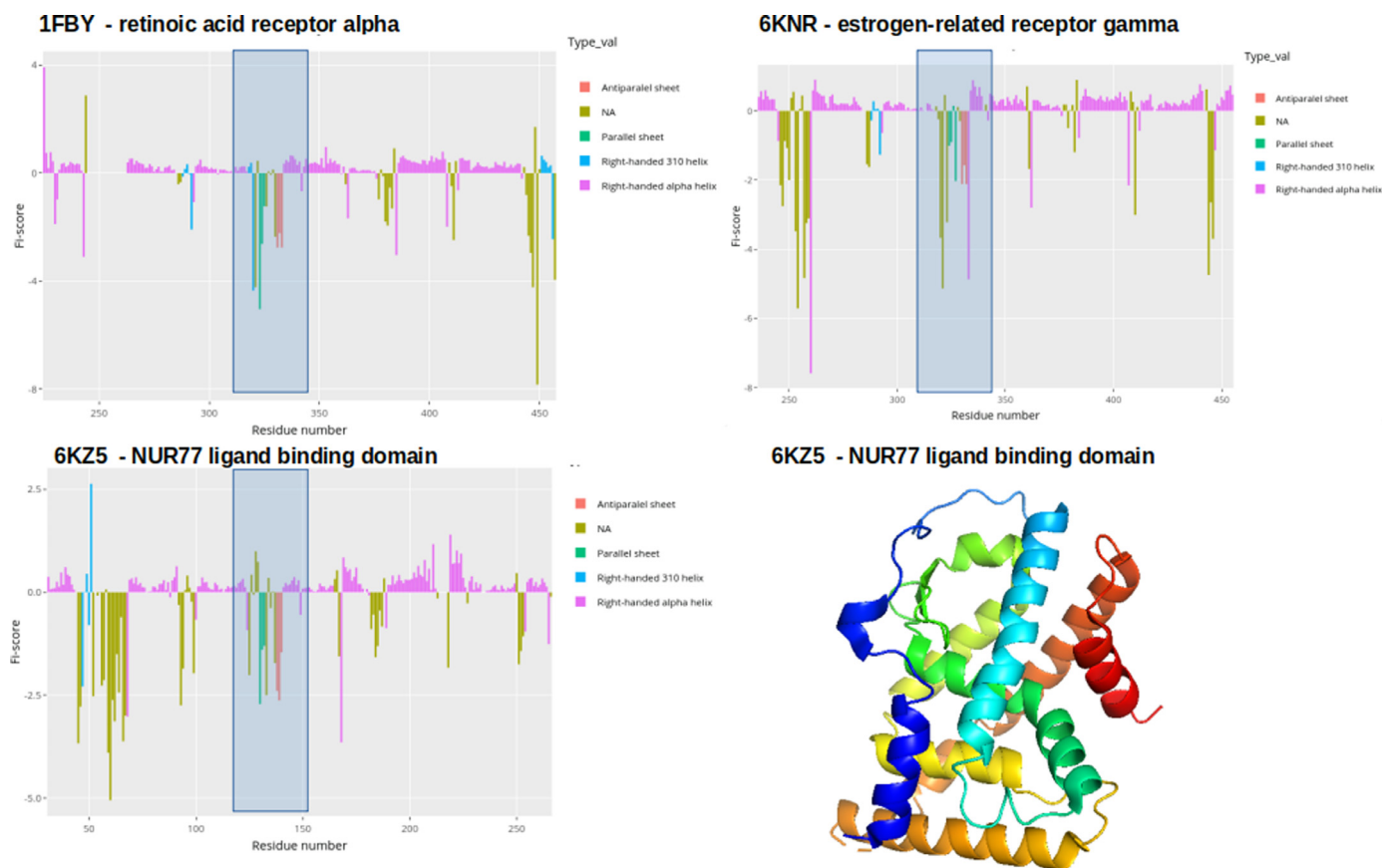
interaction mechanisms. Such approaches are especially useful when comparing known structures with the newly identified or investigating potential structural outliers.

Extracted Fi-score values can be used in machine learning modelling and this is enabled through the function *cluster\_ID*. This function groups structural features using the Fi-score and Gaussian mixture modelling where an optimal number of clusters and a model to be fitted during the EM phase of clustering for GMM are automatically selected (Fig. 6). The output of this analytical tool summarises cluster information and also provides plots to visualise the identified clusters based on the cluster number and BIC value (Fig. 7). These outputs can be used to better assess model performance for the select parameters if the users chose to customise their model building.

```
#User selected parameters
```

```
df<-cluster_ID(pdb_df,clusters = 5, modelNames = "VVI")
```

The users are advised to set seed for more reproducible results when initiating their projects. *cluster\_ID* takes a data frame containing a processed PDB file with Fi-score values as well as a number of clusters to consider during model selection; by default 20 clusters ('max\_range') are explored. In addition, a 'secondary\_structures' parameter is needed to define whether the information on secondary structure elements from the PDB file needs to be included when plotting; the default value is TRUE. Researchers also have an option to select a cluster number to test 'clusters' together with 'modelNames'. However, it is important to stress that both optional entries need to be selected and defined, if the



**Fig. 12.** Fi-score distribution plots with the Nur77 ligand binding domain (PDB ID:6KZ5), retinoic acid receptor alpha (PDB ID: 1FBY), and estrogen-related receptor gamma (PDB ID: 6KNR). Rainbow spectrum of the Nur77 structure allows to visualise the sequence from N-terminal (blue) to C-terminal (red). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

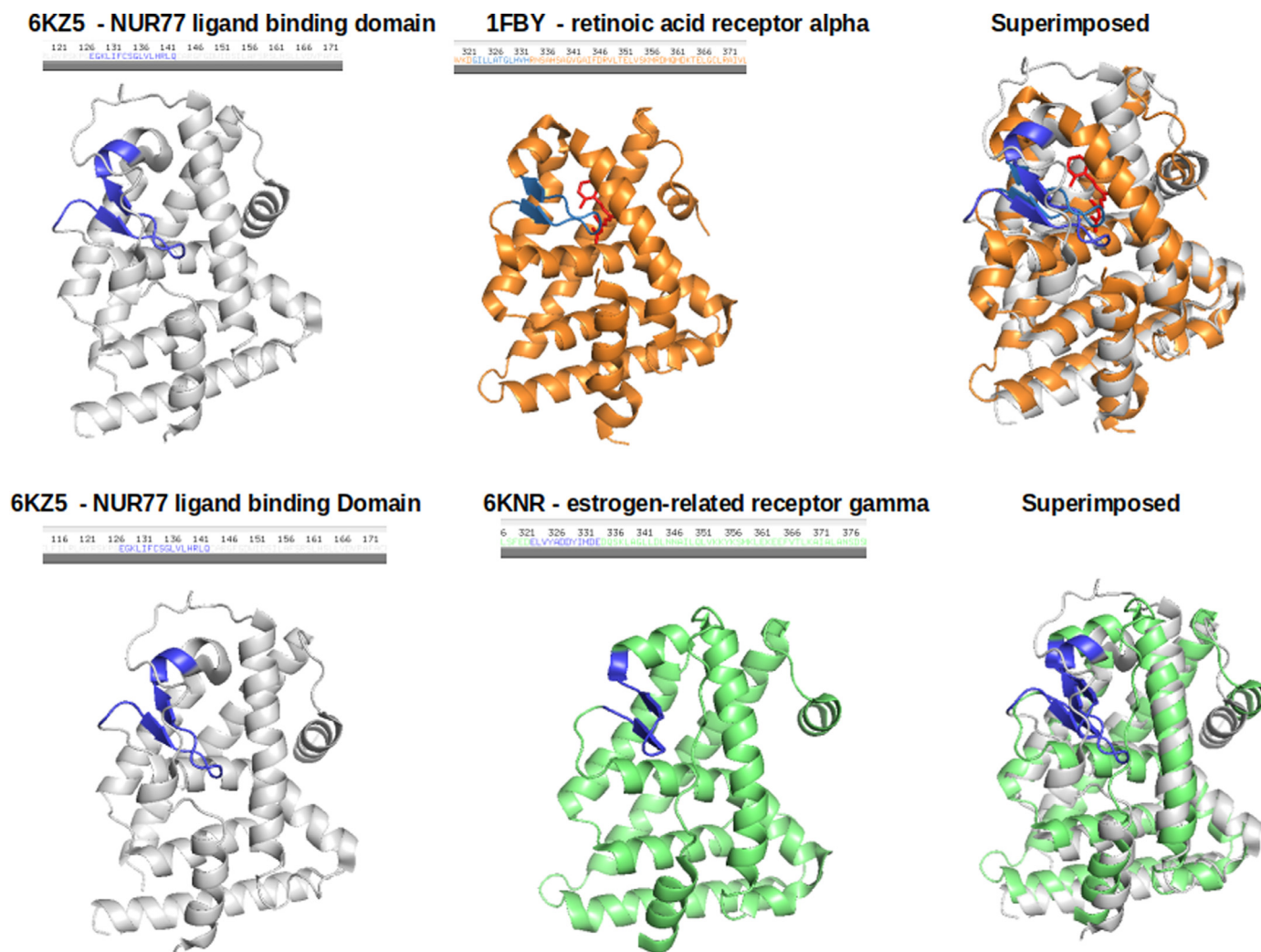
user wants to test other clustering options that were not provided by the automated BIC output. This is an advanced option and the user should assess the BIC output to decide which model and what cluster number he or she wants to try out. It is important to note that *cluster\_ID* offers a user-friendly implementation of GMM where most technical decisions are already incorporated automatically.

A dimension reduction method for the visualisation of clustering is also automatically provided (Fig. 10). Dimension reduction is a useful technique to explore multi-dimensional biological data through key eigenvalues that define the largest information content of the explored features [10]. In other words, one can infer how well the explored characteristics define the data and if the classification is sufficient for downstream analyses. For example, in the case of Nur77 Fi-score clustering, this analysis allowed assessing how well the number of clusters separates data points based on their distribution features. Nur77 has six clusters which might indicate functionally and structurally distinct regions in the target protein. It appears that the data points are well separated into groups accounting for the different variability. The dimension reduction approach could also help deciding if a different number of clusters might better classify Fi-scores. One of the goals of building this software package was not only to provide accessible and easy-to-use functions but also to generate additional plots allowing to assess model performance and data point distribution.

In addition, one of the most valuable features of this set of functions is to generate clusters with secondary structure information (Figs. 8 and 9). The produced interactive plots enable researchers to explore structural characteristics of a protein of interest (Figs. 8 and 9). Thus, the subdivision of a protein structure based on the physicochemical features

offers a new way to detect and explore functional sites or structural elements. Figs. 9 and 10 clearly indicate that some structural elements in Nur77 are likely similar in their function and physicochemical characteristics. For example, different types of helices as well as beta sheets in some cases overlap in their Fi-score characteristics and the assigned cluster type. This detailed capture of structural elements can help evaluate conformational outliers or infer similarities for different motifs. Moreover, it can be clearly seen that the region around the 50 amino acid is set to be distinct from the other two sites around 130 and 180 amino acids which could suggest overall different motion and interaction profiles. These findings also correspond with the earlier observations for the hydrophobicity features (Fig. 3). A similar trend can be seen for N and C terminus clusters which form distinct groups and might indicate sites where the receptor mediates specific functions [6]. GMM guided analyses offer a novel way to extract patterns that might not be observable using other methods dependent on sequence based analytics, e.g., ExPasy [16,17].

All previous analyses tie in with the function *density plots* which provides a density plot set for  $\phi/\psi$  angle distributions, Fi-scores, and normalised B-factors. 3D visualisation of dihedral angle distribution for every residue is also included. These plots can be used for a quick assessment of the overall parameters as well as to summarise the observations. Density plots are very useful when evaluating how well the selected features or scores separate protein structural elements and if a protein structure is of good quality (i.e., dihedral angles, B-factors, or Fi-scores provide reasonable separation between elements). The function also gives another reference point to establish if the selected number of clusters differentiates residues well based on the secondary structure



**Fig. 13.** PyMol generated plots to visualise protein structures where blue colors indicate the region identified through Fi-score patterns where cis-9 retinoic acid is highlighted in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

elements. In order to get this information, the user is only required to supply the output from the `cluster_ID` function (Fig. 11).

Data summary and evaluation

```
density_plots(pdb_df)
```

Data summary and evaluation including GMM outputs

```
cluster_IDs<-cluster_ID(pdb_df)
```

```
density_plots(cluster_IDs)
```

### 3.4. Case study: exploring potential ligands for the Nur77 orphan receptor

To demonstrate some of the *Fiscore* applications, potential ligands were searched for the Nur77 receptor which can be considered as a complex target since no endogenous ligands are known for this orphan receptor [6]. The first analysis step involved searching for other similar human proteins that did not belong to the Nuclear receptor subfamily 4. PSI-BLAST alignment analysis led to several candidate proteins, namely the retinoic acid receptor alpha (PDB ID: 1FBY) and estrogen-related receptor gamma (PDB ID: 6KNR) [15]. These proteins showed a good alignment to the Nur77 ligand binding domain sequence (average percent identity 31.68%; Suppl. Table 1) and were subsequently

used for the structural and functional exploration. Comparing Nur77 Fi-scores with the retinoic acid receptor alpha and estrogen-related receptor gamma Fi-score distributions revealed several interesting patterns (Fig. 12). The shaded blue region highlights a matching distribution pattern for all the proteins and the Student's *t*-test confirmed that none of the distributions differed significantly (Fig. 12; Suppl. Table 2). Intriguingly, this region is involved in mediating interactions with retinoic acid in the retinoic acid receptor alpha (Fig. 13). Similarly, the estrogen-related receptor gamma (PDB ID: 6KNR) has a known inverse agonist binding to the same cavity created by paired alpha-helices, an anti-parallel beta sheet, and disordered stretches [23]. The inverse agonist exhibits several structural features, such as the scaffold size/orientation and key aromatic groups, that are similar to retinoic acid. Moreover, despite the different amino acid composition, the key physicochemical features are preserved in this site across the investigated proteins as can be seen from the superimposition studies (Fig. 13). These observations point to the fact that this region might be essential in accommodating binding events. Importantly, machine learning exploration (Fig. 8 & 9) helped to classify Fi-scores around this region which revealed a repeating pattern very different from a surrounding N- and C-terminal regions. This further implies a site of special functional importance where data was grouped based on emerging probabilistic patterns in data point values. This case study suggests an interesting possibility that Nur77 with no known ligands might bind to chemical entities similar to retinoic



acid [6]. This is also supported by the alignment data and hydrophobicity plots (Suppl. Figs. 1–3) where Nur77 and the retinoic acid receptor alpha show substantial structural and physicochemical overlaps at this interactor site. Further molecular modelling and docking studies could aid in better understanding binding energetics and emerging interactions.

Overall, this example reveals that extracting patterns through scoring and machine learning could help identify proteins that have shared and functionally related features. Thus, *Fiscore* allows an easy implementation of protein structural data mining and classification without necessarily performing multiple visual inspections of the structures. These analytical principles can also be applied to explore other proteins of interest and their potential ligands.

#### 4. Discussion

*Fiscore* package was developed to address the need for a simple-to-use, freely available, and adaptable set of tools for protein physicochemical feature exploration via machine learning. By contrast, other commercial, semi-commercial, or free software tools lack machine learning pipeline implementation to explore structural features and in most cases users need special knowledge to employ these pieces of software [8,16,19–22].

*Fiscore* package (Fig. 1) allows a user-friendly exploration of PDB structural data and integration with various machine learning methods. The package was benchmarked through several analytical stages that involved a diverse set of proteins (3352) to assess scoring principles [1] and package functionalities (1337 structures) [11]. With a number of helpful functions, including distribution analyses or hydrophobicity assessment in the context of structural elements, *Fiscore* enables the exploration of new target families and comprehensive data integration since the described fingerprinting captures protein sequence and physicochemical properties. Such analyses could be very helpful when exploring therapeutically relevant proteins. In addition, provided tutorials and documentation should guide researchers through their analysis and allow adapting the package based on individual project needs [1]. *Fiscore* could also aid in drug repurposing studies when a chemical compound needs to be juxtaposed to a number of potential targets. This was demonstrated during a native ligand search for Nur77 where a case study of a nuclear receptor revealed the usefulness of the introduced scoring and physicochemical data capturing via GMM. Furthermore, the novel scoring system as well as machine learning applications can lead to interesting insights about sites of structural and functional importance. The retrieved information could be used in comparative studies to search for other proteins that share similar features. For example, some of the shifts in Fi-score values coincide or precede post-translational modifications in Nur77 (Fig. 5) [24]. This information could be included in the future studies together with fingerprinting to better understand structural characteristics of this receptor.

Another important aspect of the *Fiscore* package is the simplification of complex analytical steps so that the researchers without an extensive background in structural bioinformatics or machine learning could still use the tools for their analyses, such as protein engineering, protein assessment, and data storage based on specific target sites. Thus, the interactive analytical and visualisation tools could become especially relevant in the pharmaceutical research and drug discovery studies as more complex targets and protein-protein interactions need to be assessed in a streamlined fashion. In other words, ability to translate structural data into parameters could accelerate target classification, target-ligand studies, or machine learning integration. Since target evaluation is paramount for rational therapeutics development, there is an undeniable need for specialised analytical tools and techniques that can be used in R&D or academic research. Implementing these novel approaches could significantly improve our ability to assess new targets and develop better therapeutics. As a result, the *Fiscore* package was developed to aid with therapeutic target assessment and make machine

learning techniques free-to-use and more accessible to a wider scientific audience.

#### 5. Funding

The package development was not supported by outside funding.

#### Declaration of Competing Interest

The author reports no conflicts of interest.

#### Acknowledgments

The author would like to thank the anonymous reviewers and code testers for helping to improve the package with their valuable suggestions and advice.

#### Supplementary material

Supplementary material associated with this article can be found, in the online version, at [10.1016/j.aiisci.2021.100016](https://doi.org/10.1016/j.aiisci.2021.100016).

#### References

- [1] Kanapeckaitė A, Beurivage C, Hancock M, Verschueren E. Fi-score: a novel approach to characterise protein topology and aid in drug discovery studies. *J Biomol Struct Dyn* 2021. doi:[10.1080/07391102.2020.1854859](https://doi.org/10.1080/07391102.2020.1854859).
- [2] Du J, Guo J, Kand D, Li Z, Wang G, Wu J, et al. New techniques and strategies in drug discovery. *Chin Chem Lett* 2020;31. doi:[10.1016/j.ccl.2020.03.028](https://doi.org/10.1016/j.ccl.2020.03.028).
- [3] Fauman E, Rai B, Huang E. Structure-based druggability assessment—identifying suitable targets for small molecule therapeutics. *Curr Opin Chem Biol* 2011;15. doi:[10.1016/j.cbpa.2011.05.020](https://doi.org/10.1016/j.cbpa.2011.05.020).
- [4] Faraggi E, Xue B, Zhou Y. Improving the prediction accuracy of residue solvent accessibility and real-value backbone torsion angles of proteins by guided-learning through a two-layer neural network. *Proteins* 2009;74. doi:[10.1002/prot.22193](https://doi.org/10.1002/prot.22193).
- [5] Reynolds D. Gaussian mixture models2015; 10.1007/978-1-4899-7488-4
- [6] Wu L, Chen L. Characteristics of nur77 and its ligands as potential anticancer compounds. *Mol Med Rep* 2018. doi:[10.3892/mmr.2018.9515](https://doi.org/10.3892/mmr.2018.9515).
- [7] Wickham H. ggplot2: Elegant graphics for data analysis2016; <https://ggplot2.tidyverse.org>.
- [8] Grant BJ, Rodrigues APC, ElSawy KM, McCammon JA, Caves LSD. Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics* 2006;22:2695–6.
- [9] Sievert C. Interactive web-based data visualization with R, Plotly, and Shiny2020; <https://plotly-r.com>.
- [10] Scrucca L, Fop M, Murphy TB, Raftery AE. Mclust 5: clustering, classification and density estimation using gaussian finite mixture models. *R J* 2016;8(1):289–317. doi:[10.32614/RJ-2016-021](https://doi.org/10.32614/RJ-2016-021).
- [11] Kanapeckaite A. Fiscore: effective protein structural data visualisation and exploration2021;R package version 0.1.3; <https://github.com/AusteKan/Fiscore>.
- [12] Chambers J. Object-oriented programming, functional programming and R. *Stat Sci* 2014;29. doi:[10.1214/13-STS452](https://doi.org/10.1214/13-STS452).
- [13] DeLano W.. The PyMOL molecular graphics system, version 2.3.02021;.
- [14] Berman H, Westbrook J, Feng Z, Gilliland G, Bhat T, Weissig H, et al. The protein data bank. *Nucleic Acids Res* 2000. doi:[10.1093/nar/28.1.235](https://doi.org/10.1093/nar/28.1.235).
- [15] Altschul S, Madden T, Schäffer A, Zhang J, Zhang Z, Miller W, et al. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res* 1997. doi:[10.1093/nar/25.17.3389](https://doi.org/10.1093/nar/25.17.3389).
- [16] Gasteiger E., Hoogland C., Gattiker A., vaud S.D., ans Ron D. Appel M.R.W., Bairoch A.. Protscale2021;.
- [17] Gasteiger E., Hoogland C., ans S'everine Du vaud A.G., ans Ron D. Appel M.R.W., Bairoch A.. Protein identification and analysis tools on the expasy server2005; 10.1385/1592598900
- [18] Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. *J Mol Biol* 1982;157(1):105–32. doi:[10.1016/0022-2836\(82\)90515-0](https://doi.org/10.1016/0022-2836(82)90515-0).
- [19] Schrödinger. Schrödinger platform2021; <https://www.schrodinger.com/platform>.
- [20] UCL. Predict secondary structure (psipred)2021; <http://bioinf.cs.ucl.ac.uk/index.php?id=779>.
- [21] Institute T.S.R.. Autodock suite2021; <https://autodock.scripps.edu/>.
- [22] the Sanner lab at the Center for Computational Structural Biology (CCRB). Mgltools software suite2021; <https://ccsb.scripps.edu/mgltools/>.
- [23] Kimag J, Hwanga H, HeeseokYoona, Jae-EonLeed, Oh JM, An H, et al. An orally available inverse agonist of estrogen-related receptor gamma showed expanded efficacy for the radioiodine therapy of poorly differentiated thyroid cancer. *Eur J Med Chem* 2020. doi:[10.1016/j.ejmech.2020.112501](https://doi.org/10.1016/j.ejmech.2020.112501).
- [24] Hornbeck P, Zhang B, Murray B, Kornhauser J, Latham V, Skrzypek E. Phosphositeplus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res* 2015. doi:[10.1093/nar/gku1267](https://doi.org/10.1093/nar/gku1267).

## **Integrative *omics* approaches for new target identification and therapeutics development**

### **6. *In silico* drug discovery for a complex immunotherapeutic target - human c-Rel protein**

**The experimental chapter is based on the following publication**

Kanapeckaitė A, Beaurivage C, Jančorienė L, Mažeikienė A. *In silico* drug discovery for a complex immunotherapeutic target-human c-Rel protein. Biophysical Chemistry. 2021 Sep 1;276:106593; doi: 10.1016/j.bpc.2021.106593. Selected as the issue cover.

#### **Conclusion of this chapter**

In this chapter I introduce my work on the efficient development of therapeutic agents through an early *in silico* analysis which can reduce both the costs and time needed to discover promising lead-like compounds. My developed screening methodology is an efficient drug screening approach when no crystal structure exists for a target of interest. This variant of *in silico* screening can become central in drug discovery and can be used to better understand the molecular basis of target interactions prior to performing costly *in vitro* screens. By using computational methods and the 3D structural information of c-Rel, it was possible to investigate the differences in ligand-c-Rel binding and validate that with HTVS. Computational analysis resulted in the identification of 15 promising compounds that could be further tested *in vitro* for the c-Rel protein inhibition or modulation. Finally, my research helped to demonstrate that immunotherapies can be developed by relying more on discovering new drug candidates *in silico* which could be more quickly and cost-efficiently translated into *in vitro* screens.

#### **Contribution to this chapter (95%)**

- Devised the methodology and screening pipeline.
- Performed all the analytical, data mining, molecular modelling, screening, and experimental work as well as formulated conclusions.
- Performed benchmarking and comparative analyses.
- Conceptualised and wrote the manuscript, including the figure preparation.
- Corresponding author.



## *In silico* drug discovery for a complex immunotherapeutic target - human c-Rel protein

Austė Kanapekaitė<sup>a,\*</sup>, Claudia Beurivage<sup>c</sup>, Ligita Jančorienė<sup>b</sup>, Asta Mažeikienė<sup>d</sup>

<sup>a</sup> Algorithm379, Laisves g. 7, Vilnius LT-12007, Lithuania

<sup>b</sup> Vilnius University Medical Faculty Institute of Clinical Medicine, Clinic of Infectious Diseases and Dermatovenereology, Santariškių str. 14, 08406 Vilnius, Lithuania

<sup>c</sup> Galapagos BV, Zernikedreef 16, 2333CL Leiden, the Netherlands

<sup>d</sup> Department of Physiology, Biochemistry, Microbiology and Laboratory Medicine, Institute of Biomedical Sciences, Faculty of Medicine, Vilnius University, M. K. Čiurlionio g. 21, LT-03101, Vilnius, Lithuania

### ARTICLE INFO

#### Keywords:

Drug discovery  
Normal mode analysis  
Mode of action prediction  
NFκB  
Molecular dynamics  
Machine learning

### ABSTRACT

Target evaluation and rational drug design rely on identifying and characterising small-molecule binding sites on therapeutically relevant target proteins. Immunotherapeutics development is especially challenging because of complex disease etiology and heterogenous nature of targets. c-Rel protein, a promising target in many human inflammatory and cancer pathologies, was selected as a case study for an effective *in silico* screening platform development since this transcription factor currently has no successful therapeutic inhibitors or modulators. This study introduces a novel *in silico* screening approach to probe binding sites using structural validation sets, molecular modelling and describes a method of a computer-aided drug design when a crystal structure is not available for the target of interest. In addition, we showed that binding sites can be analysed with the machine learning as well as molecular simulation approaches to help assess and systematically analyse how drug candidates can exert their mode of action. Finally, this cutting-edge approach was subjected to a high through-put virtual screen of selected 34 M drug-like compounds filtered from a library of 659 M compounds by identifying the most promising structures and proposing potential action mechanisms for the future development of highly selective human c-Rel inhibitors and/or modulators.

### 1. Introduction

High-throughput screening (HTS) of large compound libraries against a therapeutically-relevant target allows to identify compounds showing pharmacological promise and select new hit compounds that could be further optimised in hit-to-lead phase. In order to reduce costs and delivery time many leading pharmaceutical companies have their HTS preceded by *in silico* screens. The main advantage of the preparatory computational analysis is increased resource savings and a better understanding of relevant biological activity. As a result, a varied assortment of computational platforms are used for early stage screening studies to successfully identify candidates and narrow down the chemical search space [1–5]. With the advance of new computational methods, it is possible to achieve a more accurate and robust optimisation of the pharmacological properties of selected compounds which leads to the overall improvement of *in vitro* HTS. This study aimed to combine the best existing practices of high through put virtual screening

(HTVS) and introduce additional analytical and control approaches to formulate an effective HTVS pipeline for complex immunotherapy targets.

The human genome contains about 25,000 genes, but only about a tenth of expressed proteins is amenable to small-molecule modulation [6–9]. In addition, less than 5% of proteins that can be pharmacologically targeted show any therapeutic potential and drug development is further complicated by a very low success rate since less than 2% of lead compounds successfully reach the marketing stage [6,10]. Immunotherapeutics development is made even more difficult by a complex network of interactors that might share a varying degree of similarity which can lead to off-target effects and limit disease-specific therapeutic approaches. Consequently, knowing the binding sites and physico-chemical properties of any complex target prior to the screening or optimisation of lead compounds would be extremely beneficial in terms of cost reduction and faster turnover. In addition, introducing molecular dynamics analyses for the target site characterisation permits

\* Corresponding author.

E-mail address: [info@algorithm379.com](mailto:info@algorithm379.com) (A. Kanapekaitė).

<https://doi.org/10.1016/j.bpc.2021.106593>

Received 18 December 2020; Received in revised form 28 March 2021; Accepted 12 April 2021

Available online 24 April 2021

0301-4622/© 2021 Elsevier B.V. All rights reserved.

researchers to better evaluate large-scale motions in molecules and predict compound binding effects. For example, conformational flexibility can have a significant influence on the structure–function relationship where the flexibility is the key structural determinant for binding partner interactions [11,12]. This type of analyses can shed new light on how binding events evolve, what potentially new sites for compound binding are formed and how introduced mutations can alter the site and interactions. Normal mode analysis (NMA) is one of the major simulation techniques that can be used to address the above questions as it takes advantage of the small oscillations physics to describe flexible states in a protein at an equilibrium [13–16]. The central idea of this method is that when a macromolecule is in an energy minimum conformation, even if it is slightly perturbed, the restoring forces return the system to its equilibrium. This biophysical concept allows the integration of other methods to further enhance the predictiveness of NMA. Moreover, machine learning techniques applied to probe structural characteristics can reveal additional features that could be used to understand protein topology [17]. In other words, combination of biophysical, cheminformatics and machine learning approaches can greatly enhance our ability to analyse potential therapeutic targets. As a result, we focused our efforts on NF- $\kappa$ B as an excellent candidate for such a study because of a significant unmet need for drugs that could effectively target this complex.

The transcription factor NF- $\kappa$ B plays a multitude of roles through the regulation of key genes in pro-survival and pro-apoptotic pathways. As a master regulator, NF- $\kappa$ B, consists of hetero- or homo-dimers formed by the Rel transcription factor family members: p50, p52, Rel A (p65), Rel B, and c-Rel - all of which share N-terminal homology with the v-Rel oncogene [18–20]. The regulation of NF- $\kappa$ B is achieved through the binding of  $\kappa$ B inhibitor proteins which can be proteolytically degraded after the engagement of I $\kappa$ B kinase complex (IKK). This results in the release of NF- $\kappa$ B allowing Rel dimers to translocate to the nucleus [20]. NF- $\kappa$ B can be activated through two pathways, the canonical and the non-canonical, where the former pathway is mainly involved in the immune system activation and cellular survival, while the non-canonical pathway is primarily functional in lymphoid organogenesis. The canonical pathway induction results in p50 and p65 or p50 and c-Rel heterodimer formation. In contrast, the non-canonical pathway signalling is achieved only through p52-RelB heterodimers [21,22]. The signalling has additional layers of complexity since NF- $\kappa$ B complexes containing either p65 or c-Rel are known to be involved in distinct biological roles; for example, NF- $\kappa$ B complexes formed with p65 play a role in cellular metabolism and inflammatory response regulation, such as glutamine homeostasis and cytokine production, respectively. However, NF- $\kappa$ B complexes containing c-Rel are involved in a more specialised immune response and lymphoid development. This is supported by deletion experiments where a germline deletion of p65 leads to embryonic lethality but there are no effects on viability and only limited immunological defects when c-Rel expression is eliminated [23,24].

Although NF- $\kappa$ B is at the nexus of multiple signals, so far no significant strides have been made in developing specific therapeutics that would have minimal side effects [20,24,25]. This difficulty can mainly be attributed to the wide expression of NF- $\kappa$ B in multiple tissues; yet, one aspect of the NF- $\kappa$ B signalling pathway can be exploited to achieve better therapeutic characteristics. That is, NF- $\kappa$ B is assembled from different dimers that vary between tissues and pathologies and thus, targeting a specific partner of the dimer pair could increase specificity and reduce off-target effects. For example, it has been recently demonstrated that the pharmacological inhibition of c-Rel function delayed melanoma growth by impairing effector Treg-mediated immunosuppression. This immunotherapy approach was even further potentiated when combined with anti-PD-1 treatment proving that the inhibition of NF- $\kappa$ B c-Rel is a viable therapeutic target [21,26,27]. Furthermore, targeting c-Rel to modulate Treg activity in cancer revealed that c-Rel, but not p65, was susceptible to pentoxifylline, an FDA-approved drug. Specifically, it was reported that pentoxifylline caused a selective

degradation of c-Rel without affecting p65 [27]. All of these earlier analyses provided an incentive to perform a first in-depth structural modelling analysis as well as a focused preclinical *in silico* screening of likely human c-Rel inhibitors and modulators as a way to control NF- $\kappa$ B in cancer pathologies when p65 and c-Rel play the driving role.

Most small-molecule drugs currently on the market were developed to target protein–ligand interactions [28]. These interaction sites are usually concave with complex topological features which contrasts with the sites normally found on protein surfaces [29–32]. Computational analysis of likely and/or unusual binding sites, especially on protein surfaces, helps to evaluate their physicochemical properties and determine if the protein of interest is druggable. We employed NMA [14–16], GROMACS [33] molecular dynamics as well as Gaussian mixture models [34] based algorithm [17] to characterise the target and address the common issue with the crystal structures of not being able to assess how a protein behaves *in vitro* and what natural conformational states these macromolecules can pose [35,36]. In order to evaluate, alternate conformations and capture the range of motions, it was necessary to go beyond a snapshot structure conformation. In addition to this, we used SiteMap, developed by Schrödinger, to identify potential binding sites since this tool makes use of linking together “site points” that are likely to contribute to tight protein–ligand or protein–protein binding [37–39]. However, it is necessary to note that there is not one universal algorithm that could suit all drug discovery scenarios; therefore, this study combined multiple levels of analysis, namely sequence, structure, dihedral angle distribution machine learning assessment as well as physicochemical characteristics, and incorporated those findings into Schrödinger Maestro suite for the screening of the drug-like library of 34 M compounds. This library was prepared by filtering 659 M chemical entities to generate a diverse compound set for the final rounds of docking. As a result, the characterisation of binding sites and potential drug-protein interaction mechanisms led to the discovery of 15 hit compounds specific for the human c-Rel protein. These hit compounds were additionally tested with a different docking program – Autodock Vina [40] and yielded similar results.

While the experimental validation was beyond the scope of this study and there are many other reports employing only the *in silico* strategy [5,41–43], the authors would like to highlight that the identified hit compounds should be further explored in an appropriate *in vitro* and biophysical assay set-up. For example, over-expression or knock-down phenotypic studies of c-Rel could potentially allow to better evaluate any pharmacological intervention effects. Furthermore, the mode of action, such as inhibition or promotion of degradation, should also be investigated. We would also like to advise to combine both transcriptome and proteome analysis to uncover the real kinetics of the pathway in the context of an active drug-like molecule.

With this research we aimed to set the computational groundwork for a focused analysis of complex targets. *In silico* techniques demonstrated here could greatly enhance the discovery process and ensure that only the most promising candidates reach the expensive wet-lab testing pipelines. Finally, we hoped to address the growing need of clearly defined bioinformatics and cheminformatics methods that could aid in the immune therapeutics development.

## 2. Methods

### 2.1. Target identification and characterisation

Structures of chicken c-Rel (PDB:1GJI, 2.85 Å, 281 amino acids) and mouse p65 (PDB: 5U01, 2.50 Å, 291 amino acids) as well as additional good quality PDB structures were downloaded directly from RCSB Protein Data Bank (PDB) [44] detailed information on all structures used for the analysis can be found in Sup. Table 1; bound DNA fragments were removed. Multiple sequence alignment (MSA) (MUSCLE algorithm, default parameters) [45] was used for structure superimposition studies using Bio3D package [46]. Protein Blast [47], T-coffee [48] MSA

**Table 1**

SiteMap analysis for the chicken c-Rel protein (PDB: 1GJI) when dividing the protein into 5 regions.

Name	SiteScore	Size, Å <sup>2</sup>	Dscore	Volume, Å <sup>3</sup>	Exposure	Enclosure	Phobic	Philic
Site 1	0.877	881	0.991	460.649	0.822	0.359	0.095	0.504
Site 2	0.887	445	1.003	268.569	0.815	0.37	0.131	0.49
Site 3	0.887	430	0.99	261.709	0.727	0.393	0.154	0.57
Site 4	0.887	242	1.013	122.794	0.809	0.352	0.167	0.425
Site 5	0.881	216	0.981	122.108	0.749	0.392	0.07	0.594

analyses (default parameters) as well as Gaussian mixture models based Fi-score [17] were employed to further assess sequence and corresponding structure characteristics to determine potential binding sites and sites of interest for *in silico* modelling. MSA analyses were performed in JalView environment using corresponding amino acid sequences of the selected PDB structures (Sup. Table 1) which approximate the first 300 amino acids capturing the RHD. Colombic electrostatic potential and hydrophobicity based on Kyte-Doolittle scale [49] were visualized using Chimera-X platform [50], the same platform was used for the structure superimposition via Needleman-Wunsch global alignment (matrix-Blosum 62) [51].

## 2.2. Homology modelling and structure validation

Selected human c-Rel sequence (capturing RHD, NCBI:NP\_002899.1) was subjected to homology-threading and *ab initio* structure modelling using Phyre2 intensive mode [52]. Generated models were validated by Phyre2 [52] and independently with Schrödinger Maestro protein structure validation tool (Schrödinger Release 2019–2: Protein Preparation Wizard, Schrödinger [53]). Modelled and crystal structure similarity was assessed with an independent *t*-test for phi ( $\Phi$ ) and psi ( $\Psi$ ) torsion angles. After confirming consistency between templates, the modelled structures were used for the protein docking preparation and subsequent grid generation followed by HTVS.

## 2.3. Molecular dynamics and normal mode analysis

Molecular dynamics analysis was performed using GROMACS [33] suite and the following parameters (using TIP3P water model, Amber99sb-ildn force field, dodecahedron 1 Å box, adding neutralizing Na and Cl ions concentration at 0.1 mM, simulation time 1 ns; all parameter files are provided with the supplementary materials). Normal mode analysis was done using Bio3D R suite [46] with *calpha* forcefield to employ a spring force constant allowing the differentiation between the nearest-neighbour pairs along the protein backbone and all other pairs [54]. The parametrization was achieved by fitting a local minimum of a crambin model using AMBER94 force field.

## 2.4. Compound selection

Compound structures (659 M;  $2 < \log P < 4$ ;  $300 < MW < 500$ ) were downloaded from ZINC5 database of commercially-available compounds for virtual screening [55]. The downloaded set of compounds (SDF format) was fingerprinted (type: FP2, a path-based fingerprint that indexes fragments of a small molecule using linear segments of up to 7 atoms; ChemmineR package [56–58]), an add-on package fmcsR was used to identify maximum common substructures (MCSs) for structure similarity search and cluster based analysis. Based on this evaluation, similar compounds were removed from the set to generate a diverse set of molecules. This was followed by a structure and activity based search against PubChem database [59] to select candidate compounds that were the most drug-like; this selection resulted in 34 M compounds ( $\log P \leq 3$ ;  $300 < MW \leq 375$ , standard reactivity, default pH = 7.4) that were used for the HTS *in silico* screen.

## 2.5. Virtual screening and molecular docking

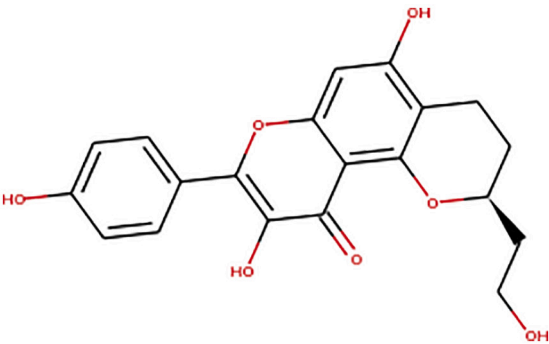
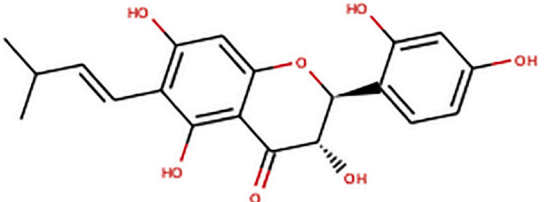
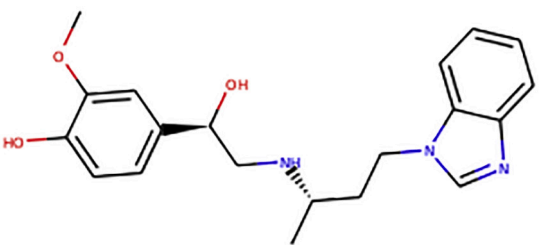
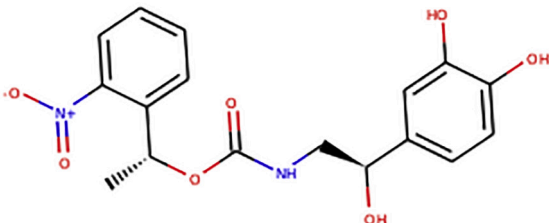
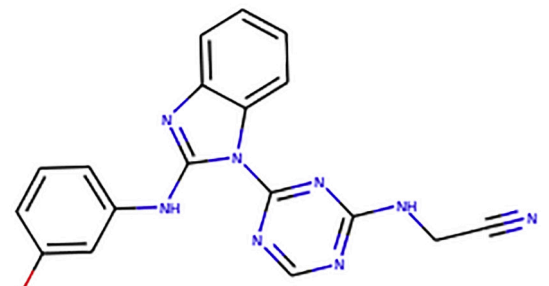
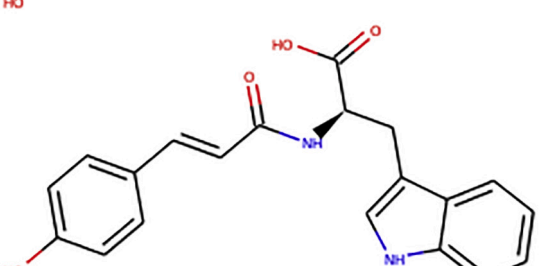

Both crystal and simulated structures of proteins were prepared using Protein Preparation Wizard and every protein was divided into five regions and their binding sites were predicted using SiteMap with shallow site identification allowed (Schrödinger Release 2019–2: SiteMap and Protein Preparation Wizard, Schrödinger [53]). These sites were ranked using SiteScore and Dscore, three highest ranking regions were selected for a further evaluation with Poisson-Boltzmann (APBS) method for their electrostatic surface (solute dielectric constant: 1.0, solvent dielectric constant: 80.0, solvent radius: 1.4 Å, temperature: 298.0 K, radius: 5.0 Å). Selected top scoring sites were used to generate a grid with the grid size chosen sufficiently large to include all residues potentially involved in ligand binding per site (Extra Precision (XP) mode) (Schrödinger Release 2019–2: Glide, Schrödinger [53]). A combination of site scores (SiteMap) as well as docking and glide scores (Glide) were used as an empirical scoring functions to evaluate and predict free energy for ligand binding to the selected site. In addition, the predicted area of a protein that could mediate binding was also evaluated in comparison to other respective sites on the same protein and homologous proteins. LigPrep module (Schrödinger Release 2019–2: Maestro, Schrödinger [53]) was used in preparation for virtual screening to evaluate the ligand library ( $\log P \leq 3$ ;  $300 < MW \leq 375$ , standard reactivity, default pH = 7.4, charges = default (all allowed), total screened compounds 34 million) and confirm its suitability for docking and screening. All ligand conformers were docked to each receptor grid using Glide (Schrödinger Release 2019–2: Glide, Schrödinger [53]). High-Throughput Virtual Screening (HTVS) mode with default settings was initially used to filter out unlikely ligands (cut-off for binding interactions  $\Delta G < -2$  kJ/mol); compounds that were hits in both mouse p65 and chicken c-Rel were removed from the next stages to increase the specificity of hits to c-Rel. This was followed by re-docking of the highest-ranking compounds from HTVS using Glide SP (Standard-Precision Glide with default settings, cut-off for binding interactions  $\Delta G < -2$  kJ/mol). Finally, the top ranking fragments from Glide SP were docked using Glide XP (Extra-Precision Glide, cut-off for binding interactions  $\Delta G < -3$  kJ/mol) for extra precision. All of the screening steps were of increasing stringency for binding interactions (Sup. Table 2). *In silico* ADME (absorption, distribution, metabolism, and excretion) of top hits was evaluated with QikProp (Schrödinger Release 2019–2: QikProp, Schrödinger [28]). Additional validation of the top scoring compounds per site were analysed with an alternative docking program AutoDock Vina [40].

## 2.6. Statistical analysis

Graphs and statistical analysis (unpaired *t*-test, two-tailed) performed with R studio (Version 1.1.463 [60]). Data filtering was performed with BASH (Ubuntu 18.04.2 LTS [61]).

**Table 2**

Top five compounds for each chicken c-Rel (PDB:1GJI) and modelled human c-Rel site. Each row provides the location of the compound followed by the interaction descriptions.\*

Site	Structure	Compound	$\Delta G$	MW	QPlogPw	QPlogPo/w	RuleOfFive
1GJI							
1		ZINC000085569496	-4.679	370.358	15.792	1.335	0
1		ZINC000095909670	-4.527	372.374	14.63	1.757	0
1		ZINC000003785475	-4.411	355.436	12.062	2.604	0
1		ZINC000098052562	-3.976	362.338	15.214	1.529	0
1		ZINC000064744186	-3.973	358.362	15.668	1.898	0
2		ZINC000014824074	-7.582	350.373	12.659	3.424	0
2		ZINC000064744186	-7.552	N/A	N/A	N/A	0

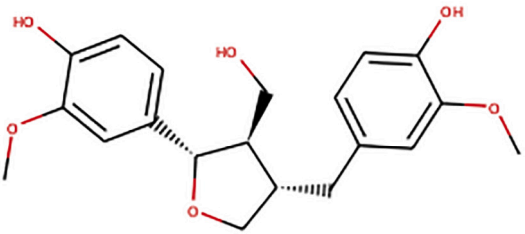
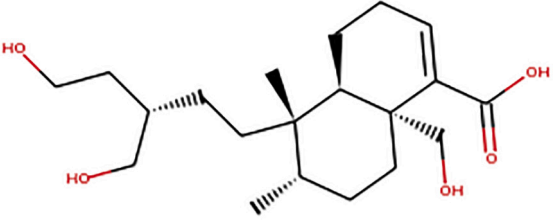
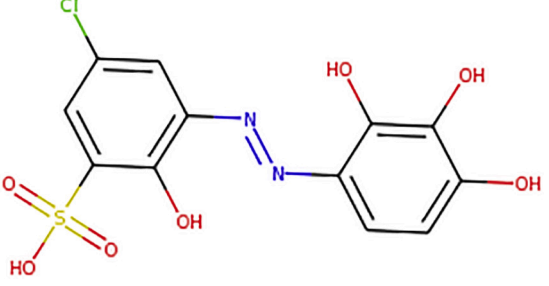
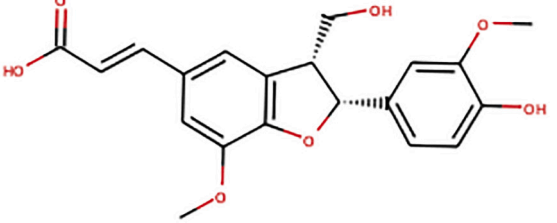
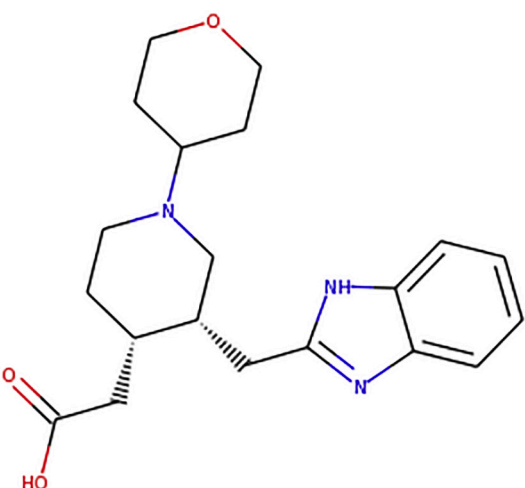
(continued on next page)

Table 2 (continued)

Site	Structure	Compound	$\Delta G$	MW	QPlogPw	QPlogPo/w	RuleOfFive
2		ZINC000095920801	-5.155	363.375	14.473	2.018	0
2		ZINC000012495519	-4.977	352.47	12.964	2.786	0
2		ZINC000043772464	-4.97	366.453	13.283	2.514	0
3		ZINC000029041971	-7.428	360.406	11.337	2.806	0
3		ZINC000038794072	-6.571	360.406	12.388	2.512	0

(continued on next page)

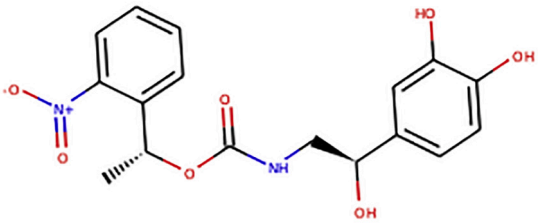
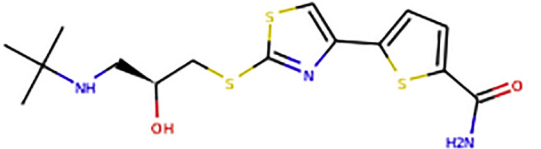
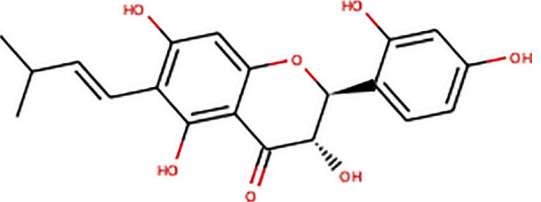
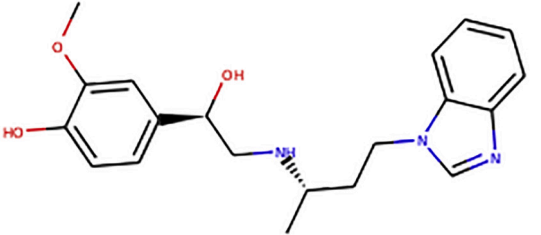
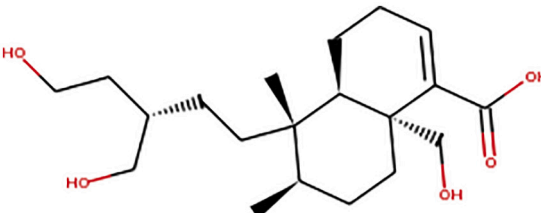
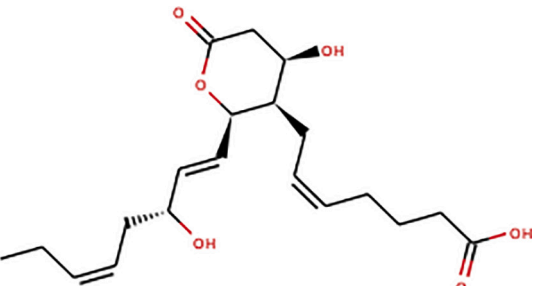
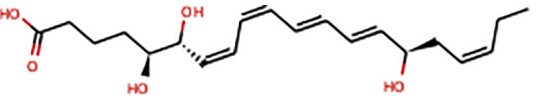
Table 2 (continued)

Site	Structure	Compound	$\Delta G$	MW	QPlogPw	QPlogPo/w	RuleOfFive
3		ZINC000072320355	-6.432	354.486	13.236	2.198	0
3		ZINC000065748825	-6.333	360.725	15.148	0.384	0
3		ZINC000031163554	-6.269	372.374	13.045	2.671	0
Human modelled c-Rel 1		ZINC000514288546	-5.651	357.452	12.569	0.769	0
1		ZINC000098052562	-4.083	362.338	15.043	1.418	0

(continued on next page)



Table 2 (continued)

Site	Structure	Compound	$\Delta G$	MW	QPlogPw	QPlogPo/w	RuleOfFive
1		ZINC000001542905	-3.777	371.53	14.057	2.234	0
1		ZINC000095909670	-3.418	372.374	14.538	1.749	0
1		ZINC000003785475	-3.359	355.436	12.539	2.814	0
2		ZINC000072320354	-7.209	354.486	12.876	1.953	0
2		ZINC000043772464	-4.869	366.453	12.861	2.103	0
2		ZINC000027647260	-4.252	350.454	12.244	3.124	0
2		ZINC000012495519	-3.936	352.47	12.864	2.511	0

(continued on next page)

Table 2 (continued)

Site	Structure	Compound	$\Delta G$	MW	QPlogPw	QPlogPo/w	RuleOfFive
2		ZINC000014824074	-3.898	350.373	12.652	3.447	0
3		ZINC000029041971	-4.851	360.406	11	2.619	0
3		ZINC000038794072	-4.264	360.406	12.413	2.568	0
3		ZINC000065748825	-4.201	360.725	15.198	0.398	0
3		ZINC000100388550	-3.717	351.404	15.954	0.477	0
3		ZINC000031163554	-3.427	372.374	13.032	2.611	0

\* QPlogPw- predicted water/gas partition coefficient; QPlogPo/w - predicted octanol/water partition coefficient; MW- molecular weight,  $\Delta G$ - Gibbs free energy, kJ/mol; RuleOfFive - number of violations of Lipinski's rule of five; N/A - need to be determined experimentally.

### 3. Results

#### 3.1. Rel family protein sequence and structure analysis revealed that the c-Rel protein is an excellent immunotherapeutic target with a potential for high specificity and selectivity

The main structural feature of c-Rel is the Rel homology domain (RHD) (300 amino acids) which is reported to form a contact surface with a single turn of the major groove of double-stranded DNA [62,63]. RHD consists of two immunoglobulin-like (Ig-like) domains, where the N-terminal domain contains the first Ig-like structure (approximately 160–210 amino acids) followed by a short flexible linker of 10 amino acids. C-terminal dimerisation domain of RHD spans about 100 amino acids adopting the second Ig-like fold and this sequence also contains a nuclear localisation signal element [62,64,65] (Fig. 1). Preliminary analysis of the whole c-Rel protein identified that the C-terminal region outside of the RHD is disordered and unstable; thus, from a drug developing perspective, it is unreliable to model the full structure and it may not be a good anchoring point for a small molecule. In contrast, the RHD, as an ordered structure, offers an easier target for crystallisation studies which could help to understand compound binding and could also complement computational screens. Furthermore, the Rel homology domain is the main domain mediating protein-protein as well as protein-DNA interactions making it the most likely target for drug-like compounds that are able to disrupt such interactions or affect c-Rel promoting its degradation [62,63,66,67]. Considering all of the above, a more defined region, such as the RHD (300 amino acids), with already existing homologous crystal structures was selected for the *in silico* analysis and subsequent HTVS.

In preparation for the modelling, an in-depth sequence and structure analysis was performed (Sup. Table 1). That is, structure superimposition and sequence studies allowed to answer several questions: how similar and/or dissimilar the RHD is among the selected species for the same protein, how the RHD amino acids distribute in MSA when aligning REL family proteins and how a structural analysis can help evaluate closely related proteins for drug discovery. Specifically, it was also necessary to explore how similar p65 would be to c-Rel since the aim of the screening was to select compounds with the highest selectivity for human c-Rel that would not otherwise affect p65. Previous crystallography studies of p65 and other Rel proteins [66,68] prompted us to pinpoint potential regions within RHD at around 30–75, 90–130 and 150–220 amino acids which we deemed to be likely involved in mediating protein-protein as well as DNA-protein contacts. We reasoned that such sites might be more susceptible to drug-like compound interference. This was an arbitrary selection in order to have some indexing within the sequence around which we could compare binding events, if

such happened.

To establish a better structural understanding of the selected Rel family sequences (Sup. Table 1), both sequence and structure based multiple alignments were performed (Sup. Fig. 1 & Fig. 2). T-Coffee sequence alignment using default settings revealed ordered regions of higher identity and the most prominent consensus stretches around the RHD (Sup. Fig. 1). Since full sequences were aligned, it was interesting to observe that p50 and p52 show some alignment to cRel and p65 toward the C-terminus as full length p50 (p105) and p52 (p100) undergo a C-terminus directed proteolytic processing [69]. The seen motifs that align when analysing full sequences might indicate that for p50 and p52 function analogous regions need to be removed when pairing occurs. This would be an interesting structural analysis avenue to explore further. Moreover, it was necessary to explore amino acid distributions for the RHD to identify both highly and less conserved regions that might hint toward unique function or binding activities. Selecting only regions that form mature structures (Fig. 2), the RHD alignment captured conserved motifs across different REL family members which also originated from different species. REL family sequences are homologous and highly-conserved across different organisms [70]; thus, while the majority of structures are non-human, they can still be used for the analyses. Furthermore, it can be seen that specific gaps in the alignment or lower identity/conservation sequences can be exploited for pharmacological targeting. These observations were followed by a more in-depth structural analysis.

#### 3.2. c-Rel structural analysis and molecular dynamics offered new insights into potential binding sites and interactions

Multiple sequence and structure alignment for specific regions revealed how c-Rel regions showing the most variability or situated in close proximity to mediate protein-DNA or protein-protein interactions could be good targets to not only increase specificity but also disrupt the dimer formation (Fig. 2 & Sup. Fig. 2). The most interesting sites for the pharmaceutical intervention are around the regions that ensure the DNA binding and contact points between the dimers (Fig. 3, A). This was confirmed when we mapped MSA results on the structure (Fig. 3, B) by highlighting some of the secondary structure elements. Moreover, superimposing multiple structures *via* alignment (Fig. 3, C) allowed us to show how REL family proteins have the most structural differences around C-terminus helixes and N-terminus beta-strands with some variation in the hinge region. The conventional structural superimposition relies on the root mean square difference superimposition for sets of residues and while we performed such an analysis (Sup. Fig. 3, C), we additionally used another method to get a fuller evaluation of the rearrangements of c-Rel domains [12]. The method developed by Romanowska et al. provides means to assess the true atomic displacement by anchoring the most invariant region; in other words, it provides a sub-structure superimposition to capture relevant domain rearrangement. Analysing c-Rel domain rearrangements maintaining restrained motion volume ( $\leq 1 \text{ \AA}^3$ ) for the anchored superimposition, we found that across all superimposed structures the most invariant region was the one mediating the dimer contact formation toward the C terminus (Fig. 3, D). This early observation hinted toward likely differences in functional dynamics for the protein sub-domains around the hinge region.

To supplement this analysis and better understand the topological landscape, we performed a hydrophobicity and electrostatic charge distribution evaluation for the c-Rel protein investigating these specific sites. Exploring Colombyic charge distributions as well as hydrophobic regions on the c-Rel surface (Sup. Fig. 2) revealed that contact sites responsible for the dimerisation have alternating positively and negatively charged regions with a hydrophobic patch. As expected (Sup. Fig. 2, A&B) the c-Rel site that clamps the negatively charged DNA also has mostly positive residues that could be targeted to disrupt this interaction.

While physicochemical and structure characterisation are incredibly

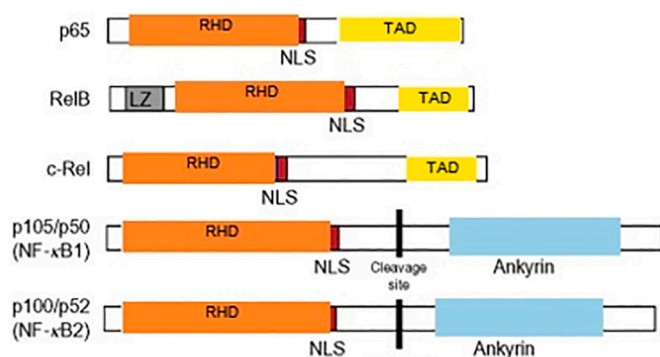
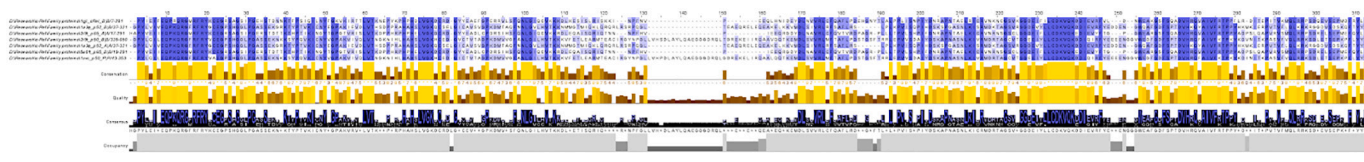
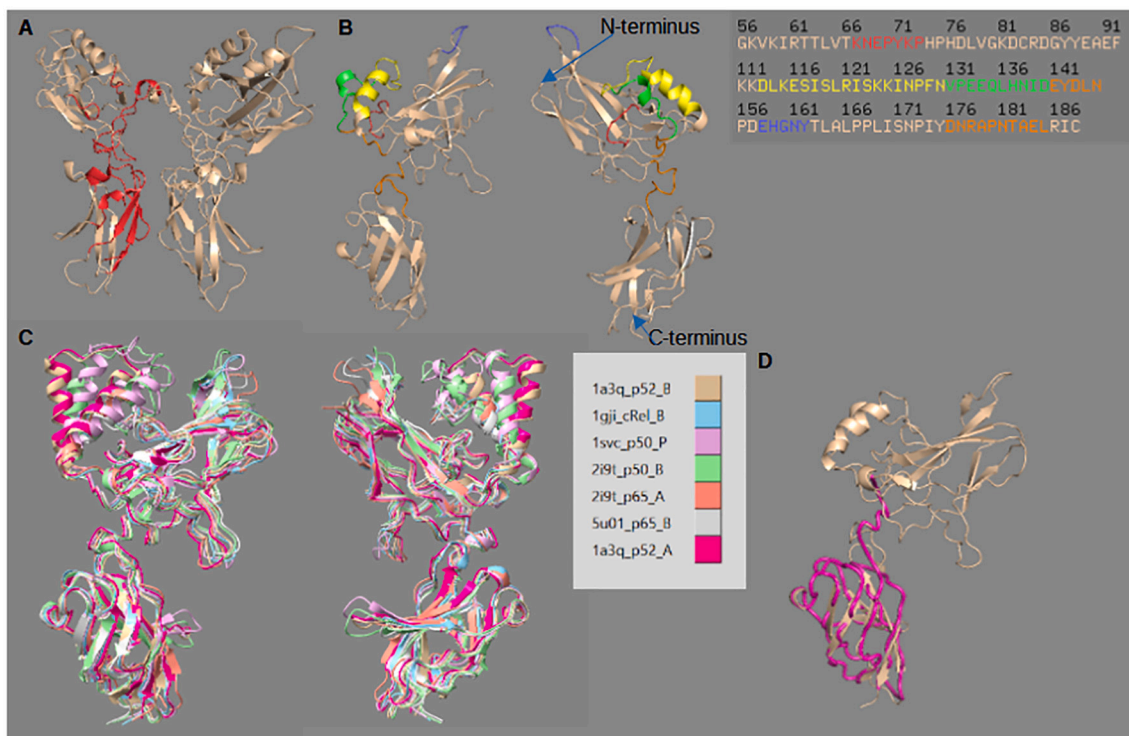


Fig. 1. The NF-κB family members. NF-κB family members are represented with their specific domains where TAD, transactivation domain; RHD, the Rel-homology domain; NLS, nuclear localisation signal; LZ, leucine zipper; Ankyrin, ankyrin repeats.



**Fig. 2.** T-Coffee structure sequence alignment using default settings where the higher identity percentage is represented with a more intense blue colour; additional parameters, such as the alignment quality score, conservation score, occupancy and consensus sequence are also provided with the alignment. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)



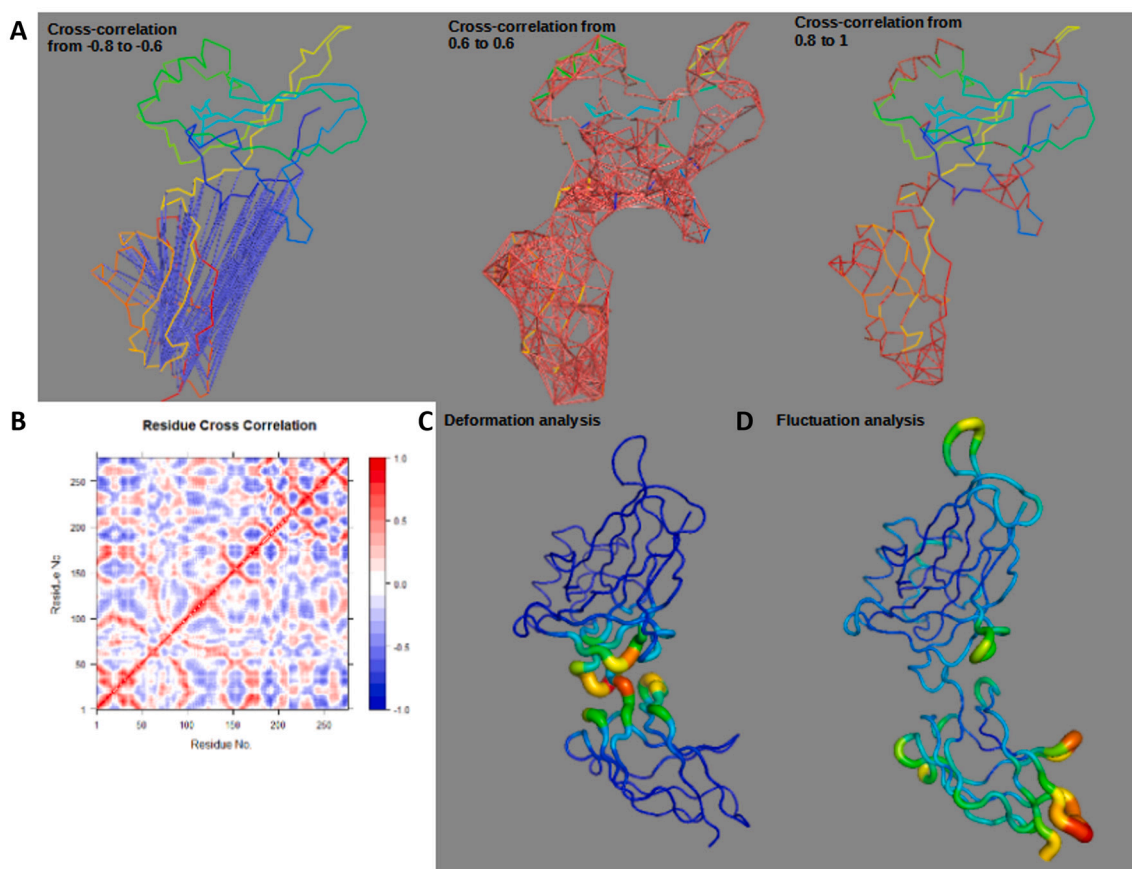
**Fig. 3.** c-Rel DNA-protein and protein-protein contact sites are highlighted in red (A). T-coffee multiple sequence alignment identified regions of interest are provided in a colour coded manner (B). Additional features are shown for the secondary structure alignment based on Needleman-Wunsch algorithm (C) and the invariant region based on anchored superimposition ( $\leq 1 \text{ \AA}^3$  volume motion) (D). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

valuable tools when selecting potential binding sites or deciding on a therapeutic compound mode of action, molecular dynamics and normal mode analyses become widely employed to understand atomic and coarse-grained domain motions. NMA was used to capture the nature of c-Rel motions prior to the screening with the aim to prioritize binding sites. NMA predicted fluctuations per residue matched the earlier identified sites around 30–75, 90–130 and 150–220 amino acids (Sup. Fig. 3, A). These sites also showed the most variability in MSA (Fig. 2 & Sup. Fig. 1). Visualisation of motions at different frequencies revealed how different movement modes undergo complex spatiotemporal movements within a protein as it forms contact sites with the DNA (Sup. Fig. 3, B). This observation also offers a new perspective on how dimers recognise DNA sequences (e.g., sweeping motions) and how binding partners initiate and maintain their contact which might not be a clamp but rather a dynamic gear-like rotation around DNA to maintain the thermodynamic binding equilibrium. Both movies (mode\_7.pdb – high frequency (0.004 s) and mode\_12.pdb-low frequency (0.0015 s)) for the structure movements are provided with the supplementary materials and can be visualized with PyMOL [71]. Such observations highlight the value of the molecular dynamics analyses as they provide new perspectives for the pharmaceutical design.

Cross-correlation analysis provided interesting insights into locally

and globally correlating and anti-correlated regions in the protein (Fig. 4, A&B). Hints of domain organized movement can be seen at C and N termini as well as around the hinge region (Fig. 4, B). That is, a globular-like domains connected through a hinge have an anti-correlating motion when referenced against each other (Fig. 4 A). However, sub-domain groups of interacting secondary structures have a closely coordinated positive and highly correlating movements. This provides additional clues that targeting the hinge region can potentially destabilize not only the dimers but also disrupt DNA binding as harmonized movements are likely necessary to ensure a proper function.

Domain and protein region correlation assessment was followed by a deformation analysis which allowed to measure the local flexibility of the structure. This analysis relies on measuring atomic motion relative to the surrounding atom groups. It is important to stress that this type of analysis differs from root mean square fluctuations (RMSF) which only provide amplitudes of the absolute atomic motion. We can see that the hinge region accommodated the largest deformation shifts in the protein structure (Fig. 4, C) and atomic fluctuations were the most significant for protein termini (Fig. 4, D). All of this points to the specific functions of these regions where inherently different motions can accommodate precise interactions to other co-binders when the transcription regulatory complex is formed.



**Fig. 4.** c-Rel protein (PDB ID: 1GJI) normal mode analysis (NMA) was used to determine correlated and anti-correlated residues which are depicted with red and blue lines, respectively (A). All cross-correlation graphic (B) provides full 2D view of the interactions. NMA mode deformation (C) and fluctuation (D) analyses provided as colour and size spectrum based on the value size ranging from low- blue to red-high. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Finally, domain analysis was used to identify regions of c-Rel moving as rigid parts (Sup. Fig. 3, C). This method relies on generating a conformational ensemble by using interpolation of the eigenvectors of the first 5 normal modes from NMA. We were able to show that while two clear domain sections stood out, especially around C-terminus which already had been shown to have a tendency of lower fluctuations (Fig. 3), they are dependent on the coherent movement of much smaller sub-domains.

### 3.3. Fi-score distribution analysis and dihedral angle analysis for modelled structures provided alternative means to assess the functional domains

Dihedral angle as well as associated B-factor distribution can be used to better understand both 3D conformation of the structure and local *Calpha* atom mobility. Capturing these parameters allows to establish a comparative measure of physicochemical characteristics of a protein of interest and we used our earlier devised Fi-score to comprehensively capture these parameters [17]. c-Rel (PDB ID: 1GJI, chain B) was assessed using dihedral angle and Fi-score distributions to uncover potentially interesting regions that show unique motion potential and map that information onto the structure (Fig. 5). This analysis can further supplement protein topology and molecular dynamics analysis by clustering protein sub-domains or regions that show similar Fi-score distributions. Moreover, the size of Fi-score change going from one region to another can indicate local mobility and dominating secondary structures [17].

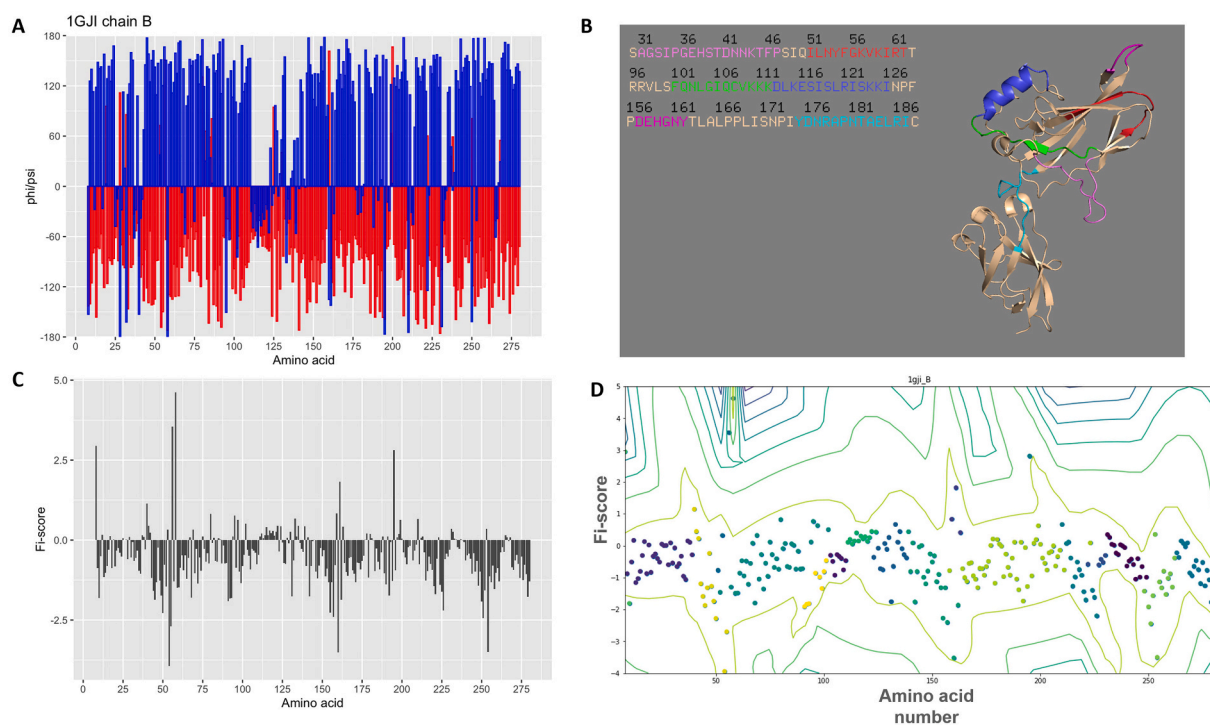
While a robust and precise method for protein-ligand complex

investigation is usually based on existing crystal structures, lack of available 3D structures for some targets has hindered efforts to design drug-like compounds. Homology modelling is rapidly becoming the method of choice for modelling protein structures [72] that can be exploited in HTVS even when the experimental data is lacking. This study aimed to build and investigate a model for human c-Rel using Phyre2 algorithm [52].

To find an alternative way to investigate binding pocket dynamics and quickly compare modelled and crystal structures, we looked into the combinations of  $\varphi/\psi$  angles in chicken c-Rel and compared that to the modelled human structure. Unpaired *t*-test revealed that the observed variation for  $\varphi/\psi$  angles between two types of structures was not significant (unpaired *t*-test for  $\varphi$ : two-tailed  $p > 0.8642$ ,  $t = 0.1711$ ,  $N = 271$ ; unpaired *t*-test for  $\psi$ : two-tailed  $p > 0.4792$ ,  $t = 0.7081$ ,  $N = 271$ ). As an additional control, no significant differences between 1GJI homodimer chains A and B were found and the observed deviations are similar to those of the crystal and modelled structures (unpaired *t*-test for  $\varphi$ : two-tailed  $p > 0.7467$ ,  $t = 0.3232$ ,  $N = 273$ ; unpaired *t*-test for  $\psi$ : two-tailed  $p > 0.7565$ ,  $t = 0.3102$ ,  $N = 273$ ). It is also important to highlight that even very similar structures might have dihedral angle shifts in critical regions and it is necessary to capture all small changes to understand the dynamics of the binding site. Thus, appropriate test steps, as discussed above, should always be implemented.

### 3.4. 34 M drug library screening reveals potential therapeutic targets for human c-Rel and hints toward possible interaction mechanisms

Molecular docking and the establishment of ligand and target



**Fig. 5.** Phi/Psi angle value distribution for c-Rel (PDB ID: 1GJI, chain B) where red bars represent phi angles and the blue - psi angles (A); structure elements of marked Fi-score shifts are colour code for amino acid clusters (B). Fi-Score (C) and Fi-score GMM distributions (D) are also provided revealing 17 distinct clusters based on Fi-Score values. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

interactions are at the centre of *in silico* drug discovery [73]. In this screening, selected two available high resolution crystal structures, namely chicken c-Rel (2.85 Å) and mouse p65 (2.50 Å), as well as a respective model of human c-Rel were used to screen against a 34 million drug-like compound library from ZINC15 [74].

Each protein was analysed with SiteMap allowing the detection of even shallow binding sites. Initially predicted five binding sites in each protein were ranked and three highest scoring regions were selected for drug screening. This selection was aided by previous structural and normal mode analyses. Grid generation for docking of the selected sites was achieved using Glide in XP mode. All binding sites for chicken c-Rel were ranked by SiteScore and Dscore to evaluate which sites are the most likely candidates for ligand-protein interactions (Table 1, Sup. Fig. 4), as can be seen all predicted sites for chicken c-Rel have SiteScore close to 0.9 and the druggability score (Dscore) of about 1 which suggest a high potential for ligand-protein interactions. While similar values were found for mouse p65 sites (Sup. Table 2, Sup. Fig. 5), this was not surprising given protein homology and the structural similarity. However, more subtle differences were observed based on the site size and amino acid composition in terms of hydrophobicity and hydrophilicity. For the purpose of the current analysis, three selected sites were analysed again with Poisson-Boltzmann (APBS) method to compare different binding pockets and their electrostatic surfaces (Sup. Fig. 6&7). Clear shifts could be observed accentuating the more electrostatically positive areas around the hinge regions for both c-Rel proteins; however, the outer side of the hinge regions that is exposed to the cytosol and does not interact with DNA was more electronegative. Interestingly, the N-terminal part of the c-Rel proteins was also highly electropositive. Mouse p65 showed a more dispersed profile with less distinct boundaries of electrostatically positive and negative regions (Sup. Fig. 7). These findings were in agreement with earlier sequence and structural analyses (Fig. 1-4) revealing that even small differences in the binding pocket amino acid composition can be fundamental in establishing target specificity and the mode of ligand-protein interaction.

Electrostatic distribution and site scoring allowed a rational selection

of three top sites based on the SiteMap parameters (Table 1). In addition, it was necessary to maintain diversity across the selected regions for chicken c-Rel (Sup. Fig. 4, Table 1) and mouse p65 (sup. Fig. 5, Sup. Table 2) so that it was possible to capture different binding modes and perform an in-depth screening. Again, it can be observed that these sites differed between the c-Rel and p65 proteins which was overall promising for the *in silico* drug screening as a way to increase compound binding specificity and minimise off-target effects.

Overall HTVS strategy was first to screen the compound library using HTVS mode for both chicken c-Rel and mouse p65 proteins, remove any hits that were identified to bind both proteins or were below the set threshold of  $-2$  kJ/mol,  $\Delta G$ . Resulting unique hits for each chicken c-Rel site were resubmitted for SP screening mode. SP screening top scoring compounds were subsequently submitted for a final refinement with XP screening mode to ensure a gradually increasing stringency in the screening (Sup. Table 3). This strategy allowed to determine five most promising compounds for each site of chicken c-Rel (Fig. 5, Table 2). As predicted in the earlier analyses, three regions of c-Rel around 30–75, 90–130 and 150–220 amino acids appear to be important in the binding of the majority of compounds; intriguingly these regions coincided with shifts in protein mobility as well as showed distinct features for the dihedral angle shifts (Fig. 4&5, Sup. Fig. 3).

One of the core aims of this study was to find a way to reduce a large compound library to an effective size for a focused screen and extract compound characteristics allowing to capture the most information. That is, by performing initial filtering to remove similar compounds, we reduced the set of the compounds to only the representative members for that group. This enabled us to explore a wider set of chemical groups and address any binding mechanisms without the need to screen a huge number of similar or very similar compounds. As predicted, all of the identified compounds had a diverse set of structural features; heterocyclic and/or aromatic groups, hydrogen donors/acceptors in combination with aliphatic groups allowed anchoring of the ligand through hydrophobic, dipole mediated or hydrogen bond interactions (Fig. 6, Table 2). For example, compounds with aliphatic and aromatic groups

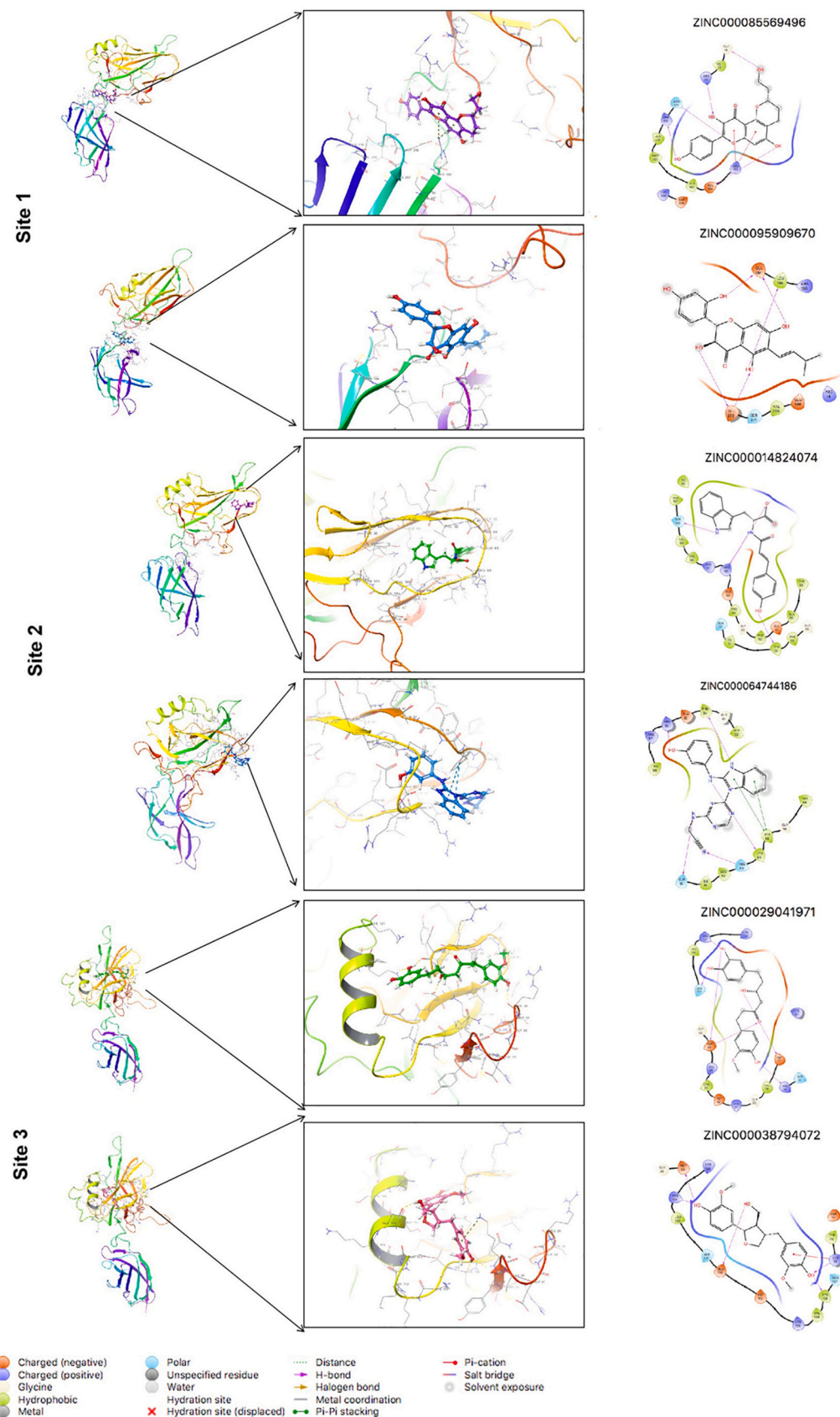
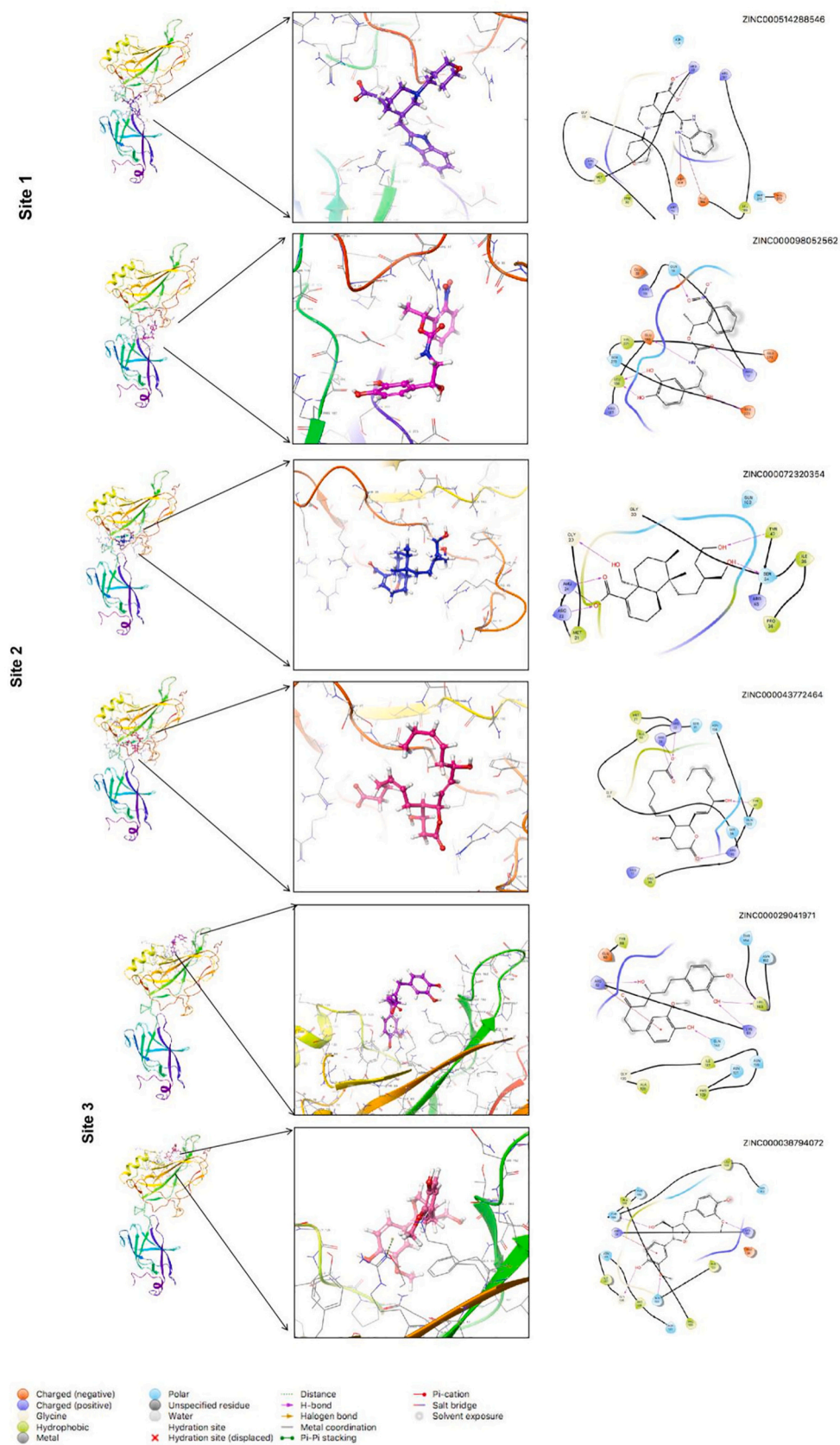


Fig. 6. Top two compounds for each chicken c-Rel (PDB ID:1GJI) site. Each row of panels provides an overall location of the compound followed by the binding region visualisation and schematic representation of the observed interactions.



**Fig. 7.** Top two compounds for each modelled human c-Rel site. Each row of panels provides an overall location of the compound followed by the binding region visualisation and schematic representation of the observed interactions.



such as, ZINC000095909670 (Fig. 6, site 1) and ZINC000038794072 (Fig. 6, site 3) provide anchoring points to orient other functional groups for better positively charged or hydrophobic interactions. This diverse set of compounds allows to associate specific amino acids with varied functional groups and establish privileged structures that could be used to build lead compounds (Fig. 6, Table 2). In addition, based on the described strategic sectioning of the c-Rel protein it is possible to propose potential mechanisms of action of these compounds. For example, conformational instability could be achieved by compound binding to Site 3 where compounds dock closely to the N-terminus of RHD, this could destabilize the Ig-like fold and expose hydrophobic amino acids to the overall highly electropositive environment which in turn could promote the destabilisation of the protein. Site 2 that engulfs the outer region of c-Rel (opposite to the hinge region) consists of a number of hydrophobic residues and displacing these amino acids could promote protein unfolding, aggregate formation and subsequent degradation. Finally, a compound binding to Site 1 embedded around the hinge region can dislodge the protein from interacting with DNA and/or its binding partner. As a result, the exposure of the electropositive core could increase solvation surface leading to conformational instability. The loss of structural stability is especially likely around Site 1 since the hinge region is loop-based without any other stabilising interactions within the protein and the majority of the stabilisation comes from protein-protein as well as DNA-protein interactions (Fig. 4).

The docking exercise of the top scoring compounds per site was subjected to a different docking approach using Autodock Vina [40]; however, the returned binding energies were very similar to the identified earlier (Sup. Fig. 8). To investigate if there were any exclusion volumes due to the side chain movements, we performed a 1-ns-long GROMACS dynamics analysis which did not reveal any restrictions locally (Sup. Fig. 9).

All of this further underlines that complex targets do not receive enough attention or dedicated computational solutions. c-Rel as well as other similar targets could benefit from our described analytical approach to establish varied sets of promising compounds that could be screened *in vitro*.

All of the chicken c-Rel XP screening compounds were then docked using Glide XP docking mode with the respective sites of human modelled c-Rel which again identified a new set of compounds that showed the highest specificity for human c-Rel (Fig. 7, Table 2). As predicted due to amino acid variation, top hits changed from chicken to human c-Rel; nevertheless,  $\Delta G$  remained consistent with a high binding capacity as observed for chicken c-Rel. Interestingly, there was a number of compounds that were matched between the sites of the crystal structure and modelled c-Rel protein. For example, three compounds were found to dock Site 1 in both proteins, namely ZINC000095909670, ZINC000003785475 and ZINC000098052562. The most shared compounds for both sites had Site 3 where 4 drug-like hits (ZINC000029041971, ZINC000038794072, ZINC000065748825, ZINC000031163554) were identified while Site 2 had only two shared compounds (ZINC000012495519, ZINC000014824074) (Table 2). These findings highlight the efficacy of this methodology where compounds can be tested for homologous structures and good quality models can be successfully used to uncover potential hit compounds.

Nearly 40% of drug candidates fail in clinical trials because of unfavourable ADME properties; thus the detection of problematic candidates is essential in early screening stages [75–79]. Computer-based methods are becoming more widely used as initial means to eliminate compounds that would likely present poor pharmacokinetic and toxicity profiles. This strategy accentuates how *in silico* approaches can reduce the high costs of drug discovery by evaluating candidates before submitting them to expensive *in vitro* testing. For the final validation of c-Rel screened compounds QikProp was used to predict the widest variety of pharmaceutically relevant properties and determine favourable ADME characteristics. At this stage the compounds were evaluated only using computational models and should be further assessed with ADME assays

to establish how hit compounds perform *in vitro*. For selected lead compounds, the partition coefficient (QPlogPo/w) and water solubility (QPlogS) was within the permissible range of  $-2.0$  to  $6.5$  and  $-6.5$  to  $0.5$ , respectively. Lipinski's rule of five for the physicochemical properties of drug likeness had no violations (Table 2). Thus, the *in silico* screened compounds showed not only a high binding capacity to the target but also *in silico* ADME profiling confirmed the 'drug-likeness' properties of the hits. This demonstrates that before embarking into costly wet-lab set-ups compounds should be filtered and analysed employing already existing knowledge bases and models. This approach could facilitate capturing any features linked to toxicity or poor pharmacokinetics and could further refine the hit compound set to be tested *in vitro*.

#### 4. Discussion

Despite the potential of c-Rel, as a therapeutic target, no potent and selective inhibitors for c-Rel have been developed at present [21,65,66,80,81]. As a first step toward discovering both potent and specific inhibitors and/or modulators of this transcription factor a detailed analysis of the sequences, structural features and relevant domains was initiated in preparation to an *in silico* high-throughput screening. This was done to address the common issue of not having a crystal structure of a target protein. By evaluating existing differences of closely related target proteins, this study demonstrated that capturing the biophysical properties of selected regions can translate into binding pocket identification, site characterisation and the improved detection of the most promising hits from the screen. With this study we showed how machine learning can be applied to assess protein topology features using dihedral angles and B-factors and we integrated this information with a molecular dynamics and biophysical parameter assessment, such as the electrostatic potential, hydrophobicity and predicted mobility. Thus, the described method of an *in silico* target analysis using structural controls as well as the biophysical characterisation of a protein of interest could be a helpful addition in building compound screening pipelines and prioritising compounds for the downstream analyses to reduce a large chemical space. Such practices have been increasingly employed where only *in silico* studies are the focal point of the analysis to refine the screening libraries or identify new therapeutic compounds [1,5,41–43]. That is, the central idea of the screening is not to identify binding affinities (as such might depend on other *in vitro* factors that are difficult to simulate) but provide a ranking of compounds to offer a directed approach for screening efforts that can help with the fast-tracking therapeutics development [1,5,40,43,77,82–86].

A virtual library of 34 million drug-like compounds was docked against the high-resolution chicken c-Rel protein and the highest scoring compounds after two rounds of refinement were docked to the corresponding sites of the high-resolution human modelled c-Rel. In total 15 hits with 6 overall being the highest scoring compounds were identified; all of which showed favourable ADME characteristics when analysed using *in silico* predictors for pharmacokinetics. Based on the analysis of binding regions it was possible to further predict potential mode of action of the identified hits. Moreover, this analysis provides a specifically selected diverse set of compounds which allows to both capture different molecules and use them as a guide to build hit-to-lead structures. That is, these results of the first extensive *in silico* analysis for the c-Rel and p65 proteins illustrate the key interactions between potential therapeutic compounds and c-Rel and can be used as a basis for a rational drug design when performing a focused large scale *in vitro* screen.

We also demonstrated how NMA based approaches can be used to explore the full scope of molecular movements [16] and connect this information with both the sequence and binding site characteristics. For example, the c-Rel protein was captured to have swinging motions and these observations invite further studies to investigate how DNA-protein and protein-protein interactions occur in this specific dimerisation event. That is, the transcription regulating complex might function *via*

alternating motions to expose DNA bases and alter torsional constraints [87]. Molecular dynamics and other analyses will likely become an inseparable part of the drug discovery and docking studies as they offer a spatiotemporal evaluation for the sites of interest [13,88].

Finally, the current report reveals that modelled structures in combination with NMA, machine learning and molecular dynamics can be used as a useful emulation when crystallographic data is not available. Our designed HTVS study identified promising compounds that may be considered as good candidate leads for further development of highly selective c-Rel inhibitors/modulators. We wanted to highlight the need to rethink current paradigms in drug discovery and especially in immunotherapeutics development because a computer-aided target validation and screening can facilitate the designing of better *in vitro* screens with minimal early investments. We would very much welcome the scientific community partaking in this analysis and exploring our discovered structures further.

## 5. Conclusion

Efficient development of therapeutic agents can be successfully achieved through an early *in silico* analysis which can reduce both costs and time needed to discover promising lead-like compounds. Our reported method is an efficient drug screening approach when no crystal structure exists for a target of interest. This variant of *in silico* screening can become central in drug discovery and can be used to better understand the molecular basis of target interactions prior to performing costly *in vitro* screens. By using computational methods and the 3D structural information of c-Rel, it was possible to investigate the differences in ligand-c-Rel binding and validate that with HTVS. Computational analysis resulted in the identification of 15 promising compounds that could be further tested *in vitro* for the c-Rel protein inhibition or modulation. Finally, this work shows that immunotherapies can be developed by relying more and more on discovering new drug candidates *in silico* which could be more quickly and cost-efficiently translated into *in vitro* screens.

## Author contributions

AK devised the methodology, performed the analysis and wrote the manuscript. CB critically reviewed the manuscript and provided suggestions, LJ critically reviewed the manuscript. All authors read and approved the final manuscript.

## Declaration of Competing Interest

The authors declare having no competing interests.

## Data availability

Data will be made available on request.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.bpc.2021.106593>.

## References

- C. Choudhury, Fragment tailoring strategy to design novel chemical entities as potential binders of novel corona virus main protease, *J. Biomol. Struct. Dyn.* 1 (2020), <https://doi.org/10.1080/07391102.2020.1771424>.
- E. Lionta, G. Spyrou, D. Vassilatis, Z. Cournia, Structure-based virtual screening for drug discovery: principles, applications and recent advances, *Curr. Top. Med. Chem.* 14 (2014) 1923–1938.
- Y. Chen, B.K. Shoichet, Molecular docking and ligand specificity in fragment-based inhibitor discovery, *Nat. Chem. Biol.* 5 (2009) 358–364.
- P.G. Jamkhande, M.H. Ghante, B.R. Ajgunde, Software based approaches for drug designing and development: A systematic review on commonly used software and its applications, *Bull. Fac. Pharm. Cairo. Univ.* 55 (2017) 203–210.
- S.J.Y. Macalino, J.B. Billones, V.G. Organo, M.C.O. Carrillo, *In silico* strategies in tuberculosis drug discovery, *Molecules* 25 (2020) 665.
- A.L. Hopkins, C.R. Groom, The druggable genome, *Nat. Rev. Drug Discov.* 1 (2002) 727–730.
- S. Knapp, Emerging target families: Intractable targets, in: *Handbook of Experimental Pharmacology* 232, Springer New York LLC, 2016, pp. 43–58.
- C. Finan, et al., The druggable genome and support for target identification and validation in drug development, *Sci. Transl. Med.* 9 (2017).
- P. Schmidtke, X. Barril, Understanding and predicting druggability. A high-throughput method for detection of drug binding sites, *J. Med. Chem.* 53 (2010) 5858–5867.
- Z. Guo, et al., Identification of protein-ligand binding sites by the level-set variational implicit-solvent approach, *J. Chem. Theory Comput.* 11 (2015) 753–765.
- W. Singh, T.G. Karabencheva-Christova, G.W. Black, O. Sparagano, C.Z. Christov, Conformational flexibility influences structure-function relationships in tyrosyl protein sulfotransferase-2, *RSC Adv.* 6 (2016) 11344–11352.
- J. Romanowska, K.S. Nowiński, J. Trylska, Determining geometrically stable domains in molecular conformation sets, *J. Chem. Theory Comput.* 8 (2012) 2588–2599.
- L. Skjaerven, X.Q. Yao, G. Scarabelli, B.J. Grant, Integrating protein structural dynamics and evolutionary analysis with Bio3D, *BMC Bioinform.* 15 (2014) 399.
- H. Wako, S. Endo, Normal mode analysis as a method to derive protein dynamics information from the Protein data Bank, *Biophys. Rev.* 9 (2017) 877–893.
- I. Bahar, T.R. Lezon, A. Bakan, I.H. Shrivastava, Normal mode analysis of biomolecular structures: functional mechanisms of membrane proteins, *Chem. Rev.* 110 (2010) 1463–1497.
- A.J. Bauer, V. Bauerová-Hlínková, Normal mode analysis: a tool for better understanding protein flexibility and dynamics with application to homology models, in: *Homology Molecular Modeling - Perspectives and Applications* [Working Title], 2020, <https://doi.org/10.5772/intechopen.94139>. IntechOpen.
- A. Kanapeckaitė, C. Beauvillage, M. Hancock, E. Verschuere, Fi-score: a novel approach to characterise protein topology and aid in drug discovery studies, *J. Biomol. Struct. Dyn.* (2020) 1–11, <https://doi.org/10.1080/07391102.2020.1854859>.
- M.S. Hayden, S. Ghosh, Signaling to NF- $\kappa$ B, *Genes Dev.* 18 (2004) 2195–2224.
- M.S. Hayden, S. Ghosh, Shared principles in NF- $\kappa$ B signaling, *Cell* 132 (2008) 344–362.
- A.Y. Ting, D. Endy, Decoding NF- $\kappa$ B signaling, *Science* 298 (2002) 1189–1190.
- Y. Grinberg-Bleyer, et al., NF- $\kappa$ B c-Rel is crucial for the regulatory T cell immune checkpoint in cancer, *Cell* 170 (2017) 1096–1108, e13.
- Q. Ruan, Y.H. Chen, Nuclear factor- $\kappa$ B in immunity and inflammation: the Treg and Th17 connection, *Adv. Exp. Med. Biol.* 946 (2012) 207–221.
- F. Köntgen, et al., Mice lacking the c-rel proto-oncogene exhibit defects in lymphocyte proliferation, humoral immunity, and interleukin-2 expression, *Genes Dev.* 9 (1995) 1965–1977.
- A. Hoffmann, G. Natoli, G. Ghosh, Transcriptional regulation via the NF- $\kappa$ B signaling module, *Oncogene* 25 (2006) 6706–6716.
- A. Hoffmann, A. Levchenko, M.L. Scott, D. Baltimore, The I $\kappa$ B-NF- $\kappa$ B signaling module: Temporal control and selective gene activation, *Science* (80) 298 (2002) 1241–1245.
- Y. Grinberg-Bleyer, et al., The alternative NF- $\kappa$ B pathway in regulatory T cell homeostasis and suppressive function, *J. Immunol.* 200 (2018) 2362–2371.
- A. Li, T. Jacks, Driving Rel-iant Tregs toward an identity crisis, *Immunity* 47 (2017) 391–393.
- J.C. Fuller, N.J. Burgoyne, R.M. Jackson, Predicting druggable binding sites at the protein-protein interface, *Drug Discov. Today* 14 (2009) 155–161.
- S.J. Campbell, N.D. Gold, R.M. Jackson, D.R. Westhead, Ligand binding: functional site location, similarity and docking, *Curr. Opin. Struct. Biol.* 13 (2003) 389–395.
- J.A. Wells, C.L. McClendon, Reaching for high-hanging fruit in drug discovery at protein-protein interfaces, *Nature* 450 (2007) 1001–1009.
- R.P. Bahadur, P. Chakrabarti, F. Rodier, J. Janin, A dissection of specific and non-specific Protein-Protein interfaces, *J. Mol. Biol.* 336 (2004) 943–955.
- I.M.A. Nooren, J.M. Thornton, Structural characterisation and functional significance of transient protein-protein interactions, *J. Mol. Biol.* 325 (2003) 991–1018.
- Molecular Dynamics Simulations - Gromacs. [https://www.gromacs.org/Documentation/outdated\\_versions/Terminology/Molecular\\_Dynamics\\_Simulations](https://www.gromacs.org/Documentation/outdated_versions/Terminology/Molecular_Dynamics_Simulations).
- D. Reynolds, Gaussian mixture models, in: *Encyclopedia of Biometrics*, Springer US, 2009, pp. 659–663, [https://doi.org/10.1007/978-0-387-73003-5\\_196](https://doi.org/10.1007/978-0-387-73003-5_196).
- C. Greenwell, G.J.O. Beran, Inaccurate conformational energies still hinder crystal structure prediction in flexible organic molecules, *Cryst. Growth Des.* 20 (2020) 4875–4881.
- M.L. Peach, R.E. Cachau, M.C. Nicklaus, Conformational energy range of ligands in protein crystal structures: The difficult quest for accurate understanding, *J. Mol. Recognit.* 30 (2017).
- K. Wang, J.A. Horst, G. Cheng, D.C. Nickle, R. Samudrala, Protein meta-functional signatures from combining sequence, structure, evolution, and amino acid property information, *PLoS Comput. Biol.* 4 (2008).
- E.B. Fauman, B.K. Rai, E.S. Huang, Structure-based druggability assessment-identifying suitable targets for small molecule therapeutics, *Curr. Opin. Chem. Biol.* 15 (2011) 463–468.

- [39] T.A. Halgren, Identifying and characterizing binding sites and assessing druggability, *J. Chem. Inf. Model.* 49 (2009) 377–389.
- [40] O. Trott, A.J. Olson, AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading, *J. Comput. Chem.* 31 (2009).
- [41] J.M. Planesas, R.M. Claramunt, J. Teixidó, J.I. Borrell, V.I. Pérez-Nueno, Improving VEGFR-2 docking-based screening by pharmacophore postfiltering and similarity search postprocessing, *J. Chem. Inf. Model.* 51 (2011) 777–787.
- [42] N. Ben Nasr, H. Guillemain, N. Lagarde, J.F. Zagury, M. Montes, Multiple structures for virtual ligand screening: defining binding site properties-based criteria to optimize the selection of the query, *J. Chem. Inf. Model.* 53 (2013) 293–311.
- [43] R. Yu, L. Chen, R. Lan, R. Shen, P. Li, Computational screening of antagonists against the SARS-CoV-2 (COVID-19) coronavirus by molecular docking, *Int. J. Antimicrob. Agents* 56 (2020) 106012.
- [44] RCSB PDB, [Homepage](https://www.rcsb.org/). <https://www.rcsb.org/>, 2020.
- [45] R.C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput, *Nucleic Acids Res.* 32 (2004) 1792–1797.
- [46] Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics*, Oxford Academic, 2020. <https://academic.oup.com/bioinformatics/article/22/21/2695/252414>.
- [47] Protein BLAST: Search Protein Databases Using a Protein Query. <https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins>, 2020.
- [48] T-COFFEE Multiple Sequence Alignment Server. <http://tcoffee.crg.cat/>.
- [49] J. Kyte, R.F. Doolittle, A simple method for displaying the hydropathic character of a protein, *J. Mol. Biol.* 157 (1982) 105–132.
- [50] UCSF, ChimeraX Home Page. <https://www.rbvi.ucsf.edu/chimerax/>.
- [51] S.B. Needleman, C.D. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins, *J. Mol. Biol.* 48 (1970) 443–453.
- [52] L.A. Kelley, S. Mezulis, C.M. Yates, M.N. Wass, M.J.E. Sternberg, The Phyre2 web portal for protein modeling, prediction and analysis, *Nat. Protoc.* 10 (2015) 845–858.
- [53] Drug Discovery, Schrödinger. <https://www.schrodinger.com/drug-discovery>, 2020.
- [54] K. Hinsen, A.J. Petrescu, S. Dellerue, M.C. Bellissent-Funel, G.R. Kneller, Harmonicity in slow protein dynamics, *Chem. Phys.* 261 (2000) 25–37.
- [55] T. Sterling, J.J. Irwin, ZINC 15 - ligand discovery for everyone, *J. Chem. Inf. Model.* 55 (2015) 2324–2337.
- [56] X. Chen, C.H. Reynolds, Performance of similarity measures in 2D fragment-based similarity searching: comparison of structural descriptors and similarity coefficients, *J. Chem. Inf. Comput. Sci.* 42 (2002) 1407–1414.
- [57] Y. Cao, A. Charisi, L.C. Cheng, T. Jiang, T. Girke, ChemmineR: a compound mining framework for R, *Bioinformatics* 24 (2008) 1733–1734.
- [58] ChemmineR, Cheminformatics Toolkit for R. <https://www.bioconductor.org/packages/devel/bioc/vignettes/ChemmineR/inst/doc/ChemmineR.html>, 2020.
- [59] PubChem. <https://pubchem.ncbi.nlm.nih.gov/>, 2020.
- [60] RStudio, Open source & Professional Software for Data Science Teams - RStudio. <https://rstudio.com/>, 2020.
- [61] Enterprise Open Source and Linux | Ubuntu. <https://ubuntu.com/>, 2020.
- [62] C.W. Müller, F.A. Rey, M. Sodeoka, G.L. Verdine, S.C. Harrison, Structure of the NF-kappa B p50 homodimer bound to DNA, *Nature* 373 (1995) 311–317.
- [63] C.W. Müller, F.A. Rey, S.C. Harrison, Comparison of two different DNA-binding modes of the NF-kappa B p50 homodimer, *Nat. Struct. Biol.* 3 (1996) 224–227.
- [64] T. Huxford, G. Ghosh, A structural guide to proteins of the NF-kappaB signaling module, in: *Cold Spring Harbor Perspectives in Biology* 1, 2009.
- [65] T.S. Fulford, D. Ellis, S. Gerondakis, Understanding the roles of the NF-kB pathway in regulatory T cell development, differentiation and function, in: *Progress in Molecular Biology and Translational Science* 136, Elsevier B.V., 2015, pp. 57–67.
- [66] B. Berkowitz, D. Huang, F.E. Bin Chen-Park, P.B. Sigler, G. Ghosh, The X-ray crystal structure of the NF-kB p50-p65 heterodimer bound to the interferon beta-kB site, *J. Biol. Chem.* 277 (2002) 24694–24700.
- [67] C.B. Phelps, L.L. Sengchanthalangsy, S. Malek, G. Ghosh, Mechanism of kB DNA binding by Rel/NF-kB dimers, *J. Biol. Chem.* 275 (2000) 24392–24399.
- [68] F.E. Chen, D.B. Huang, Y.Q. Chen, G. Ghosh, Crystal structure of p50/p65 heterodimer of transcription factor NF-kappaB bound to DNA, *Nature* 391 (1998) 410–413.
- [69] B. Miraghazadeh, M.C. Cook, Nuclear factor-kappaB in autoimmunity: man and mouse, *Front. Immunol.* 9 (2018) 613.
- [70] T.D. Gilmore, Nuclear factor Kappa B, in: *Encyclopedia of Biological Chemistry: Second Edition*, Elsevier Inc., 2013, pp. 302–305, <https://doi.org/10.1016/B978-0-12-378630-2.00335-2>.
- [71] PyMOL | pymol.org. <https://pymol.org/2/>, 2020.
- [72] A. Tramontano, V. Morea, Assessment of homology-based predictions in CASP5, in: *Proteins: Structure, Function and Genetics* 53, Proteins, 2003, pp. 352–368.
- [73] G. Sliwoski, S. Kothiwale, J. Meiler, E.W. Lowe, Computational methods in drug discovery, *Pharmacol. Rev.* 66 (2014) 334–395.
- [74] T. Sterling, J.J. Irwin, ZINC 15—ligand discovery for everyone, *J. Chem. Inf. Model.* 55 (2015) 2324–2337.
- [75] V. Mandlik, P.R. Bejugam, S. Singh, Application of artificial neural networks in modern drug discovery, in: *Artificial Neural Network for Drug Design, Delivery and Disposition*, Elsevier Inc., 2016, pp. 123–139, <https://doi.org/10.1016/B978-0-12-801559-9.00006-5>.
- [76] A.P. Li, Screening for human ADME/Tox drug properties in drug discovery, *Drug Discov. Today* 6 (2001) 357–366.
- [77] J. Jain, et al., In silico analysis of natural compounds targeting structural and nonstructural proteins of chikungunya virus, *F1000Research* 6 (2017) 1601.
- [78] L.R. de Souza Neto, et al., In silico strategies to support fragment-to-lead optimization in drug discovery, *Front. Chem.* 8 (2020).
- [79] R.S. Jaleel, U.C. A, Toxicity prediction of anti tuberculosis active molecules, *Nat. Preced.* (2011), <https://doi.org/10.1038/npre.2011.6236.1>.
- [80] Y. Shono, et al., Characterization of a c-Rel inhibitor that mediates anticancer properties in hematologic malignancies by blocking NF-kB-controlled oxidative stress responses, *Cancer Res.* 76 (2016) 377–389.
- [81] J.E. Hunter, J. Leslie, N.D. Perkins, c-Rel and its many roles in cancer: an old story with new twists, *Br. J. Cancer* 114 (2016) 1–6.
- [82] A. Lavecchia, C. Giovanni, Virtual screening strategies in drug discovery: a critical review, *Curr. Med. Chem.* 20 (2013) 2839–2860.
- [83] N.T. Gangadharan, A.B. Venkatachalam, S. Sugathan, High-throughput and In Silico screening in drug discovery, in: *Bioresources and Bioprocess in Biotechnology 1*, Springer Singapore, 2017, pp. 247–273.
- [84] A.B. Kinghorn, L.A. Fraser, S. Lang, S.C.C. Shiu, J.A. Tanner, Aptamer bioinformatics, *Int. J. Mol. Sci.* 18 (2017).
- [85] A. Lavecchia, C. Cerchia, In silico methods to address polypharmacology: current status, applications and future perspectives, *Drug Discov. Today* 21 (2016) 288–298.
- [86] C.H. Andrade, et al., In Silico Chemogenomics drug repositioning strategies for neglected tropical diseases, *Curr. Med. Chem.* 26 (2018) 4355–4379.
- [87] J. Hörberg, A. Reymer, Specifically bound BZIP transcription factors modulate DNA supercoiling transitions, *Sci. Rep.* 10 (2020).
- [88] Y.P. Pang, Use of multiple picosecond high-mass molecular dynamics simulations to predict crystallographic B-factors of folded globular proteins, *Heliyon* 2 (2016).

## 7. General discussion

### 7.1. Towards new R&D strategies: cardiomyopathies study revealed how to improve complex disease analyses and find new therapeutic avenues

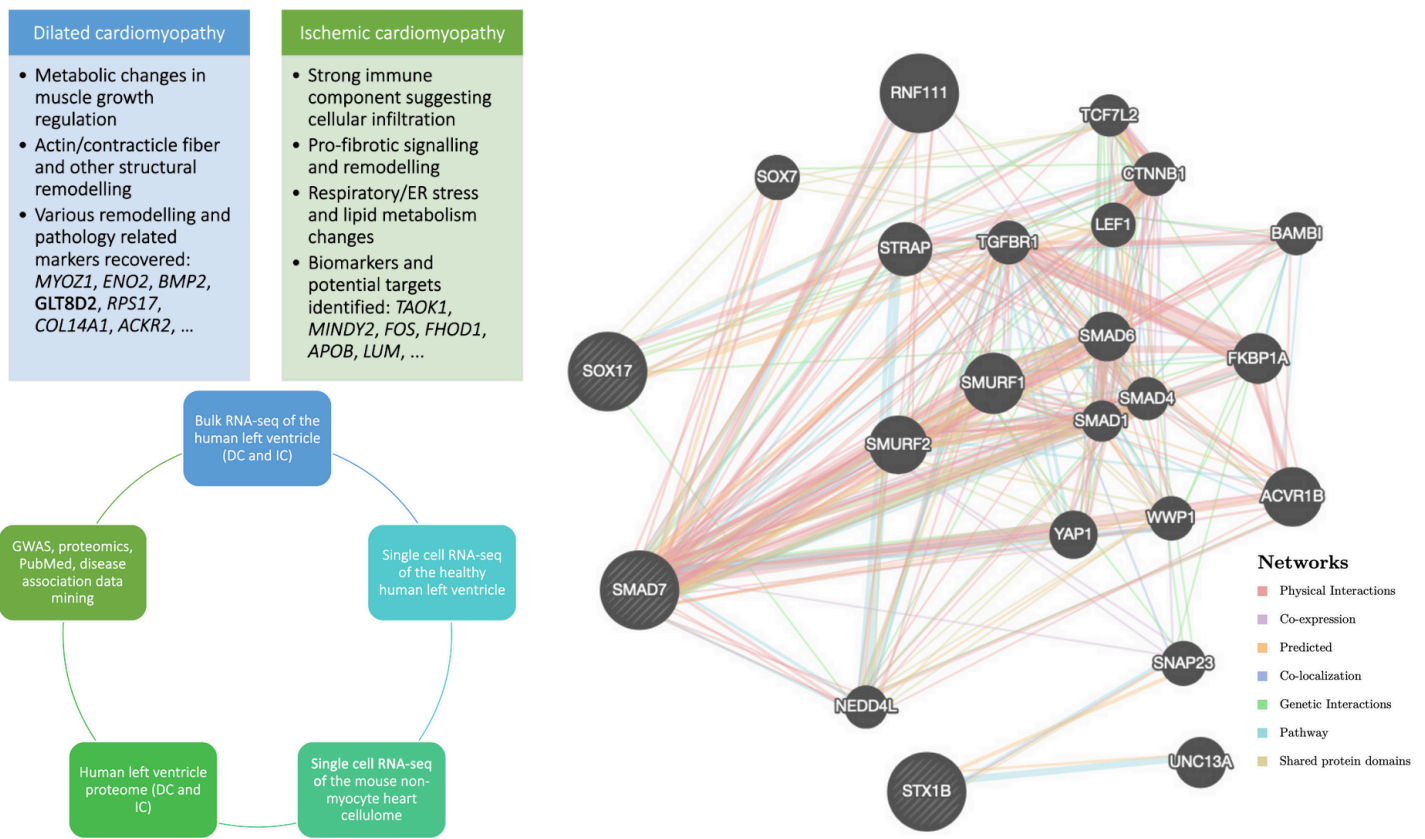
Mounting research and commercial pressures for novel therapeutics highlight why better strategies are needed for R&D and drug discovery<sup>2,13,17,18,22</sup>. This need becomes especially evident when considering how older target-centric or ‘one gene equals a disease’ approaches fail to explain multifactorial aspects of many pathologies and do not offer effective treatment options<sup>2,10,13,25</sup>. As a result, the first experimental chapter of the thesis (Chapter 2: Insights into therapeutic targets and biomarkers using integrated multi-‘*omics*’ approaches for dilated and ischemic cardiomyopathies) addressed the lack of integrative, *omics*-driven, and network-centric approaches in drug discovery<sup>2,10,13,45,46,132</sup>. The chapter focused on developing a highly integrative multi-*omics* and machine learning analytical system to probe relevant gene expression patterns and associate the identified changes with pathways and cellular processes. The main rationale for selecting cardiomyopathies as a case study was the fact that this multifactorial cardiac syndrome accurately illustrates complex diseases. This can be appreciated when considering the stochastic nature of this pathology with many predisposing elements, including genetic, epigenetic, familial, and environmental factors<sup>207</sup> (Fig. 3). Moreover, limited treatment options also provided a motivation to explore how better we could help patients<sup>88,95,208,209</sup>. Thus, CVD (specifically, HF with left ventricular dysfunction) served as a model to build a multi-*omics* analytical framework that could also be applied to other complex diseases and provide new insights to improve therapeutic outcomes<sup>88,208</sup>.

The introduced approaches for dilated and ischemic cardiomyopathies underscored how heterogeneous diseases can be explored to identify new targets or biomarkers by employing different datasets, encompassing bulk RNA-seq, single cell RNA-seq, and proteomics. A new scoring system was also developed that can be easily adapted based on individual researcher’s needs. Moreover, this scoring system can be integrated with machine learning pipelines to unveil novel links within the interactome. The introduced research revealed that dilated and ischemic cardiomyopathies are driven by a nexus of shared and diverging pathways, namely oxidative stress, metabolic perturbations, and immune system modulators. In addition, the cellome of cardiac tissue was found to maintain a complex heterogeneity of infiltrating immune and pro-fibrotic cells. This, in line with the expressome data, suggested that these cell populations and their proportions depend on the cardiac tissue state<sup>91-99,101-103,209</sup>. Another important aspect of this study was to address a common issue

in clinical studies when there are a limited number of samples that can be analysed<sup>207</sup>. HF analysis demonstrated that multi-*omics* based enrichment, multiple data points integration, and data mining can aid in uncovering disease associated pathways even with smaller sample sets. In addition, machine learning can be applied to further deconvolute complex expression features<sup>81,89,91-99,101-103</sup>. To complement this work, an R software package was created to make these analyses more readily available to researchers so that it is possible to quickly explore and integrate lab generated expression and *omics* data. This software tool set offers an expanded range of functionalities and scoring functions with a possibility to take advantage of machine learning.

Building this analytical architecture involved many different datasets and provided a wealth of interesting findings. However, several key therapeutically relevant highlights demonstrate why a holistic research approach is needed and why such strategies can lead to new clinical applications. The first part of the multi-*omics* analysis explored the bulk RNA-seq data of the human left ventricle tissue for two indications: DC and IC. The identified significantly changed genes allowed the uncovering of subtle differences between hypertrophic and ischemic heart conditions. Specifically, the studied DC samples revealed a number significantly upregulated genes (e.g., *BMP2*, *MYOZ1*, and *ENO2*) that have strong links to myocardial tissue remodelling and structural changes not seen in the healthy samples (Fig. 6)<sup>81,91,145,210-217</sup>. Other genes, such as *RPS17*, *SLITRK4*, and *GLT8D2*, represent a group of genes only recently implicated in DC, where these genes play a role in protein synthesis, post-translational modifications, and growth control<sup>145,202,218-221</sup>. Intriguingly, genes that were significantly downregulated (e.g., *C11*, *ICAM3*, and *ELOVL2*) are responsible for a wide spectrum of metabolic functions from cellular respiration to membrane integrity<sup>145,217,222-225</sup>. By contrast, ischemic heart conditions showed a marked upregulation of pro-inflammatory and pro-fibrotic genes where *CX3CL1* is an especially intriguing gene as it encodes an atypical chemokine. This chemokine can exist as either a membrane-bound or soluble protein and the membrane-associated form is largely expressed on endothelial cells in myocardial ischemia and HF<sup>91,145,226-228</sup>. Moreover, a number of identified chemokine ligands (e.g., *CXCL11*, *CXCL10*, and *CCL5*), some chemokine receptors (e.g., *CXCR3* and *CCR7*), as well as other markers, such as *CD2*, were also found to be significantly changed under myocardial ischemia. Subsequently, these findings led to hypothesise that the inflammatory gene upregulation might be precipitated by a significant infiltration of immune and immune system-linked cells<sup>81,83,85,91,145,210,229,230</sup>. This was supported by the finding that a significant proportion of fibroblasts and fibroblast-like cells populated a healthy human heart in addition to various immune cells (based on the single cell RNA-seq analysis). Under myocardial stress conditions, these cell types can be repopulated to promote a pro-inflammatory

and pro-fibrotic environment<sup>91,125,218,230-232</sup>. The presence of immune cell populations was also found in an additional/control analysis of the mouse left ventricle non-myocardial cellulome (single cell RNA-seq). As a result, normal subpopulations of immune cells identified in the heart, such as monocytes, macrophages, mast cells, eosinophils, neutrophils, B and T cells, have the potential to become activated and propagate the inflammatory state<sup>107,125,218,230-234</sup>. This study highlighted that analysing datasets without considering the full *omics* landscape can potentially lead to the misinterpretation of the results. Thus, bulk RNA-seq or proteomics data should always be weighed against the possibility of mixed cell populations in the tissue. These findings also underscored the shortcomings of some previous studies that attempted to explore the pathological milieu of heart disease. The limited statistical and technical exploration in the earlier studies resulted in missing the complexity of the cellular makeup which subsequently restricted new therapeutic target discovery<sup>81,90</sup>. Furthermore, the research discussed in this thesis also revealed that proteomics data juxtaposed with bulk RNA-seq does not always share the same expression patterns. One such example of contrasting expression patterns was the titin protein (encoded by the *TTN* gene) with the downregulated gene expression and upregulated protein expression levels. While *TTN* mutations are well-documented for DC and mutated or truncated *TTN* proteins lead to the disease parthenogenesis, the role of the higher protein expression is not clear<sup>235</sup>. In the case of IC, a similar expression profile could be found for *APOB* where this protein was also overexpressed despite the reduced mRNA levels. Thus, further investigation into the *APOB* expression variation across the transcriptome and expression could provide a glimpse into the perturbed energy metabolism and compensatory mechanisms<sup>236-240</sup>. These observations of different biological readouts and their cross-referencing could be exploited in building *omics* biomarker panels (Fig. 6 and 7).



**Figure 6.** Summary of the key findings for the cardiopathologies study listing the main observations. The circular graph captures the types of data used for the analyses and the network analysis provides an example of how the introduced methodology can help build pathways and interactor networks. The cluster identified by machine learning (*SOX17*, *SMAD7*, and *STX1B*) was searched against known interactors. Various functional connections were established (Networks - legend) demonstrating how specific clusters can be analysed through data mining when building regulatory networks. The network map was generated using the GeneMANIA software tool; Warde-Farley et al., 2010<sup>241</sup>.

This study also aimed to explore how different data points can be integrated into a single analytical framework to generate further pathology-related inferences. In order to address the main challenge of biological data integration and current limitations in combining different *omics* datasets<sup>45,131-136,138,140</sup>, a scoring system was devised with a focus on bulk RNA-seq enriched with the data mined from multiple resources that combined proteome, experimental, clinical, and predictive readouts. Specifically, the derivation of  $LFC_{score}$  allowed the evaluation of how a gene participates in the expressome network and to what extent it can cause perturbations if the gene function is disrupted. Such clustering is the first step to integrate LFC, differentially expressed genes, and protein-protein interaction-based networks when recreating a signalling interactome. Moreover, this strategy could be especially useful if researchers enriched the scoring with additional weights derived from their experimental work to add new information during clustering. To provide this specific option, a complimentary software package, *OmicInt*, was developed (Chapter 3: *OmicInt* pack-

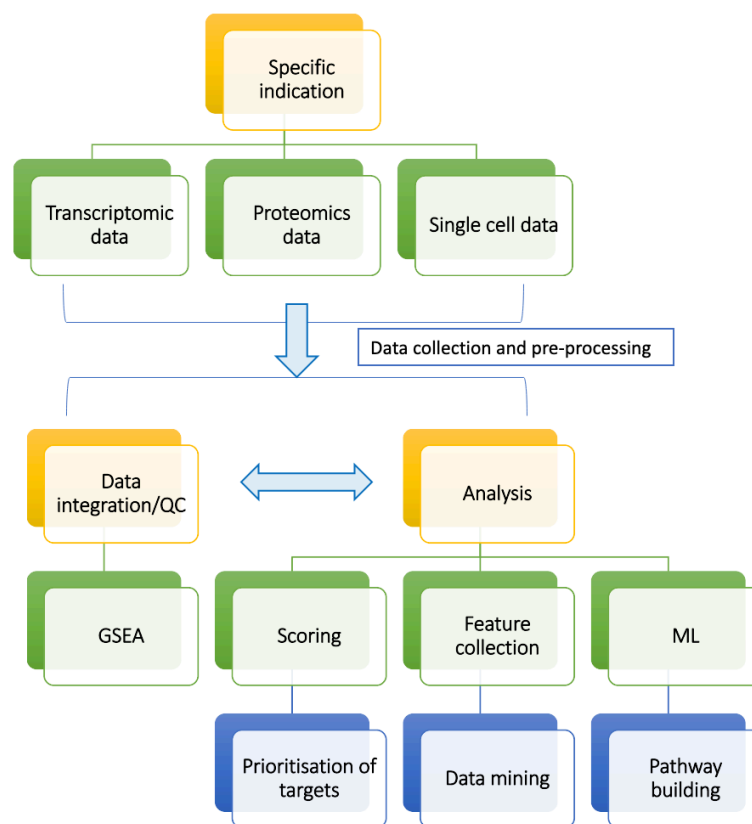
age: exploring *omics* data and regulatory networks using integrative analyses and machine learning). *OmicInt* facilitates further data integration and analysis so that researchers can take advantage of their experimental readouts, database mining, and machine learning in a single seamless application.  $LFC_{score}$  and the introduced machine learning approach were further tested with two additional cardiopathology datasets as well as cross-referenced with text or database mined resources for cardiovascular pathologies, such as the GWAS dataset of human heart disease genetic variants<sup>242</sup>, clinical/experimental evidence from Open Targets platform<sup>144,145</sup>, as well as complete PubMed records<sup>160</sup> (>30 M). This analysis allowed to verify that the proposed method juxtaposes rarer or newly discovered targets with more known genes linked to dilated and ischemic cardiomyopathies<sup>144,145,160,241,242</sup> (Fig. 6).

The discussed research highlighted the importance of capturing different levels of *omics* datasets and pursuing integrative analyses, such as data mining, in order to reconstruct complex networks. Currently, *omics* studies are only just beginning to make inroads into R&D space and often such research is still very narrow without exploring robust statistical, enrichment, and classification methods<sup>45,46,81,132,138,202</sup>. Moreover, ‘big picture’ strategies, for example, publication, experimental, or clinical evidence mining, are not incorporated into building new models to consolidate many different types of experimental readouts<sup>2,45,46,81,132</sup>. This is evident from the current pathway and network analyses which are most commonly employed to assess cellular perturbation events using high-level analytical approaches, involving over-representation, rank-based, and topology-based methods<sup>138–141,243</sup>. These strategies do not provide additional insights from other clinical and experimental resources and do not include probabilistic models for feature prediction, such as GMM<sup>148,150</sup>. While useful, these compartmentalised analyses provided the motivation to develop a more integrative approach that could evolve depending on the research needs (Fig. 7). Furthermore, to anticipate potential shortcomings of the proposed analytical strategy, a software package was developed to provide more freedom with data integration and scoring. The versatility and adaptability of the introduced methodology allow researchers to adjust the scoring system based on their in-house data and known disease associations. Performing this scoring-based analysis prior to choosing targets for downstream screens can help avoid selecting groups of genes that belong to the same effector network. Of course, the success of this method and other less integrative approaches always depends on the quality of the experimental data and available resources to perform data enrichment<sup>138–142</sup>.

The second chapter of the thesis not only provided an overview of the present challenges in treating HF, but also demonstrated the urgent need to rethink current therapeutic paradigms for the



treatment of left ventricle dysfunction. While systematic *omics* studies and in-depth analyses of heterogeneous disease mechanisms are lacking, this study, for the first time, demonstrated how bulk and single cell RNA-seq, as well as the proteomics analysis of the human heart tissue can be integrated to uncover heart failure specific networks and potential therapeutic targets or biomarkers for dilated and ischemic cardiomyopathies<sup>207</sup> (Fig. 7). Importantly, focusing on metabolic changes as well as inflammatory signatures could open new avenues for a targeted pharmacological intervention rather than the currently practised symptomatic management.

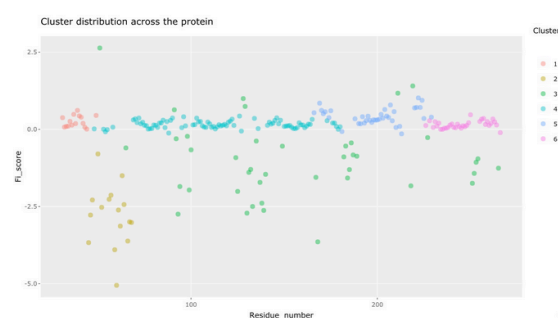
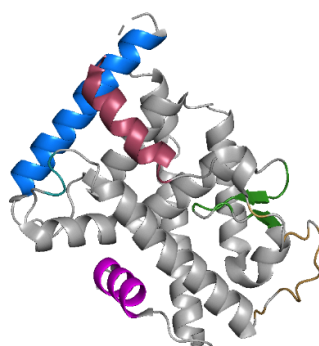
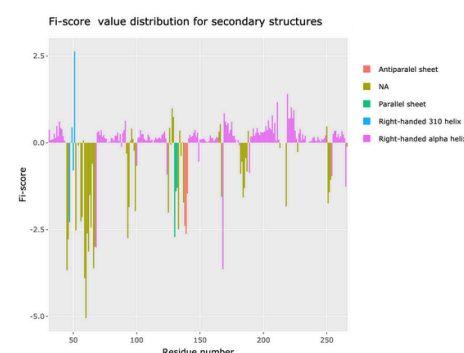
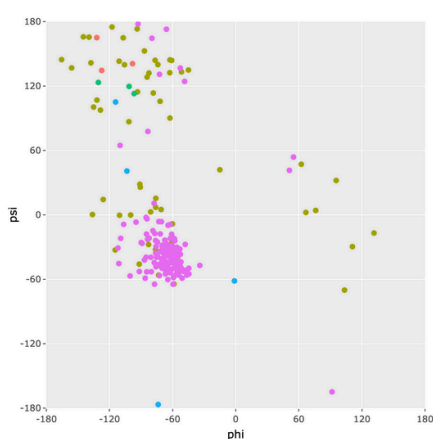


**Figure 7.** Multi-*omics* analytical framework depicting integration, analysis, and modelling principles. Several key steps are shown: *omics* data collection (top level), data integration/quality control (QC) (left bottom section), and the integrative analysis (right bottom section).

## 7.2. Implementing a streamlined target evaluation and classification: new solutions for discovery pipelines

The fourth chapter of the thesis (Fi-score: a novel approach to characterise protein topology and aid in drug discovery studies) provided a new method for the exploration of a protein topological and conformational organisation where an in-depth structural feature capture was achieved using structural bioinformatics and machine learning methods. The paradigm shift from *in vitro* to *in silico* in early pilot studies underscores the need to establish reliable approaches for target selection and the evaluation of pharmacological intervention options<sup>13,20,43,244</sup>. In other words, the critical steps in R&D are to assess the druggability of a protein of interest and to estimate the likelihood that the target will be amenable to pharmacological modulation<sup>50,54,244,245</sup>. Thus, establishing robust computational analysis principles is essential for the growth of this field and successful preclinical studies<sup>22,49,246,247</sup>. As demonstrated in the second chapter, the search for a therapeutic target or targets can generate a whole set of likely candidates which after filtering and classification still need to be evaluated from a structural perspective in order to establish their druggability<sup>153,154,169,246,247</sup>. Consequently, it became necessary to devise an effective way to capture structural and physico-chemical features of multiple targets so that it was possible to stratify these proteins prior to screening. Moreover, the introduced methodology allows not only the investigation of sites of interest but also the classification of protein features which could be implemented through relational databases. All these developments are integratable and scalable to support downstream analyses (Fig. 8).

$$Fi\text{-score} = \frac{1}{N} \sum_i \frac{\phi_i \psi_i}{\sigma_{\phi_i} \sigma_{\psi_i}} B_{i\text{-norm}}$$



**Figure 8.** Fi-score equation and associated analyses. Fi-score equation includes:  $N$  - the total number of atoms for which dihedral angle information is available,  $\varphi$  and  $\psi$  values - dihedral angles for the  $C\alpha$  atom,  $\sigma_\varphi$  and  $\sigma_\psi$  - corresponding standard deviations for the torsion angles and  $B_{i\text{-norm}}$  - a normalised B - factor value for the  $C\alpha$  atom. B-factor,  $\sigma_\varphi$ , and  $\sigma_\psi$  normalisation is based on the full-length protein. Plots provide information on dihedral angle and Fi-score distributions as well as Fi-score clustering using Gaussian mixture modelling. Provided structure represents the Nur77 protein (PDB ID: 6KZ5) highlighting some of the identified clusters (with matched colours, bottom right plot).

Protein conformation determination and capturing of physicochemical properties are some of the most important research aspects in drug discovery<sup>28,195,248,249</sup>. Protein features that have both structural and composition variability can present a significant challenge in identifying good binding pockets<sup>70,248</sup>. As a result, successful therapeutics development depends on the establishment of a binding site profile that can be contrasted with other known binding pockets in a protein or other targets<sup>169,250-252</sup>. This is especially relevant for homologous or highly similar targets where structural classification approaches are still lacking<sup>248,249,253,254</sup>. To address structural assessment and classification challenges, one of the central aims of this thesis was to develop a method to capture amino acid residue distributions providing a value that could be used to compare and characterise either regions of interest or entire structural elements. This topological fingerprinting technique or ‘Fi-score’ offers an integrative approach to capture both local and distal information via dihedral angle and B-factor distribution. Specifically, the Fi-score helps to evaluate residue physicochemical properties and extract information on structural motifs<sup>174</sup> (Fig. 8). There have been extensive studies showing that both dihedral angles and B-factors carry a lot of information that can be used to assess protein backbone orientation. B-factors can, for example, help quantify protein region flexibility or even hydrophobicity<sup>177-183,255-257</sup>. However, despite these insights, there have been no previous attempts to capture this information in a unified way so that a quantifiable parameter could be compared across different structures<sup>177-183,258,259</sup>. Furthermore, the Fi-score study demonstrated for the first time that the combination of dihedral angle values and B-factors into a single equation enables capturing structural and functional elements that might not be distinguished by analysing B-factors or dihedral angles alone<sup>174-183</sup>. Moreover, probabilistic density analysis, such as the implementation of Gaussian mixture modelling, permits a probability-based classification of features (or amino acids) to unveil conformational elements that depend on amino acid composition, flexibility, and other physicochemical parameters<sup>172-183,260</sup>. Thus, the described method offers a new way to inspect the differences in dihedral angle and B-factor distributions which can be, through scoring, linked to structural motifs or used to classify multiple targets<sup>174-176,189,256,261-266</sup> (Fig. 8). It is important to note, however, that the accuracy of the scoring is dependent on the quality of the available crystallographic data.

To enable access for such an assessment, the R software package was developed where researchers can explore their structures of interest in-depth (Chapter 5: *Fiscore* package: effective protein structural data visualisation and exploration). Importantly, *Fiscore* can be integrated into other analytical architectures and aid in studies where the expertise in machine learning is lacking since this package provides a user-friendly GMM exploration of any appropriate target<sup>267</sup>. Additional features, namely hydrophobicity-secondary structure plots or Fi-score-secondary structure plots as well as many other interactive graphs, build a highly integrative analytical framework that could help to quickly evaluate proteins prior to downstream analyses.

Thus, the introduced scoring system and machine learning applications could help to reduce not only costs but also the time needed to select targets and prioritise screening strategies. Specifically, developing more comprehensive R&D pipelines could improve drug discovery success rates and allow to target complex proteins as well as disease-causing networks. In addition, topological feature-based evaluation could advance drug repurposing efforts where known drug hotspots are searched against newly discovered targets<sup>133,170,175,176,180,181,268,269</sup>. Overall, the introduced research sets the ground for future studies to analyse structural characteristics in-depth and integrate this information with drug discovery pipelines. Implementation of these new strategies could greatly reduce the multi-dimensional complexity of therapeutics screening and target selection.

### **7.3. Highly integrative *in silico* screening pipeline: a better method to explore targets and capture potential hit compounds**

The final part of the thesis integrates the previous chapters' research by introducing a newly developed and highly integrative drug discovery pipeline focusing on a complex immunotherapeutic target (Chapter 6: *In silico* drug discovery for a complex immunotherapeutic target - human c-Rel protein). As discussed earlier, the growing R&D costs and decreasing new therapeutics outputs underline why it is imperative to rethink current discovery and development strategies<sup>17-24,51,52,151,174,270,271</sup>. Moreover, risk-averse approaches in the pharmaceutical industry limit novel therapies development and lead to increasing patient care costs<sup>70,105,170,270,271</sup>. This was also exemplified in a cardiomyopathies case study (Chapter 2) where current therapeutic options are only limited to symptomatic management with declining investments in the exploration of the alternatives<sup>11,13,17,22,244,270,271</sup>. Thus, these discovery and clinical challenges motivated the development of new HTVS strategies using holistic and integrative methods in computational biology and chemistry that could speed up the search of drug-like compounds and expand the screening space. Since

earlier *omics* analyses hinted at the potential involvement of the NF- $\kappa$ B pathway in cardiopathologies, a subunit of this transcription factor, the c-Rel protein, was selected as a complex immunotherapeutic target for the development of a HTVS pipeline<sup>2,123,125-129,244,272</sup> (Fig. 9).

Despite the potential of c-Rel, as a therapeutic target, there are no potent and selective inhibitors for this protein<sup>81,89,91,123,273</sup>. c-Rel has been implicated in many different diseases ranging from immunopathologies to cardiomyopathies<sup>115,116,121-129,274</sup>. Thus, studying this transcription factor subunit could help devise multi-modulatory strategies (e.g., homologous targets are engaged at a varying degree) as well as discover new mode of action highly specific therapeutics<sup>86,106,114,123</sup>. Therefore, the first step towards discovering both potent and specific inhibitors and/or modulators of NF- $\kappa$ B or any similar complex target relies on a detailed analysis of target sequences and structural features as well as the identification of relevant domains for protein-drug interactions<sup>121-129,274-276</sup>. Considering the above, a novel strategy of hierarchical analysis was developed in preparation for an *in silico* high-throughput screening that could serve as a blueprint for complex target HTVS (Fig. 9). Moreover, in-depth structural analysis and molecular modelling demonstrated how to overcome the common issue of not having a crystal structure for a target protein, as no X-Ray structures are currently resolved for human c-Rel<sup>115,116,121-130,273-278</sup>. The study focused on the evaluation of existing differences between closely related target proteins (e.g., c-Rel and p65)<sup>61,121-130,154,250-252,279-281</sup> and in combination with structural bioinformatics, molecular modelling, and machine learning it was possible to capture the biophysical properties of selected c-Rel regions. The identified sites were ranked based on the ligand binding probability and expected drug modes of action. Various studies have stressed the need to focus on binding pockets but little methodology has been developed to combine structural bioinformatics with computational chemistry<sup>168,195-197,278,282,283</sup>. Specifically, commercial platforms, such as Schrödinger<sup>201</sup>, or open-source tools, including Autodock Vina<sup>203</sup>, primarily focus on the computational chemistry without considering the structural bioinformatics component. Moreover, screening strategies still lack well-defined protocols and the developed methodologies are not made easily accessible for further development and testing. These limitations span library selection strategies, target selection and assessment, screening refinement, and result validation<sup>28,68,69,80,279,284</sup>. The current study expanded on the missing pieces in computational chemistry research by providing a detailed analysis strategy and showed how machine learning can be applied to assess protein topology and conformational features (Fig. 9). In addition, it was outlined how this information can be integrated with a molecular dynamics and biophysical parameter assessment, such as the electrostatic potential, hydrophobicity, and predicted mobility, to prepare for HTVS. This study also demonstrated how NMA and similar coarse grained

molecular modelling approaches can be used to explore the relevant scope of molecular movements<sup>53,131,170,197-199,285</sup> and how the incorporation of this information together with sequence and binding site characteristics can offer new insights into molecular dynamics. NMA revealed that c-Rel can potentially have winging motions which suggests that DNA-protein and protein-protein interactions are complex creating and opening new binding sites which go beyond a simple clamping seen via X-Ray crystallography studies<sup>113,118,119,187,275</sup>. The introduced strategies of an *in silico* target analysis could assist in selecting relevant targets and building compound screening pipelines<sup>28,286-292</sup>.

It is also important to highlight that the central idea of the *in silico* screening is not to identify exact binding affinities (i.e., to match experimental readouts) or to replace *in vitro* screens but to provide a ranking of potential hit compounds in order to fast-track therapeutics development<sup>28,68-70,278</sup>. Limitations of HTVS depend on the quality of target structures, selected libraries, and platform design<sup>68,69,255</sup>. The aim of the presented research was to create a framework that is adaptable and can evolve as better algorithms become available. Importantly, the developed pipeline combines both bioinformatics and cheminformatics tools to prepare for the screening which is expected to improve the detection of therapeutically promising compounds.

The results of the first extensive *in silico* analysis for the c-Rel protein outlined key protein-drug interactions and the developed analytical framework could be used as a basis for rational drug design in preparation for large-scale *in vitro* screens of complex targets<sup>278</sup>. Furthermore, this screening introduced a diverse set of compounds which can be used as a chemical guide to build improved hit-to-lead structures for c-Rel. Thus, this analytical map could be incorporated into therapeutic pipelines for the NF- $\kappa$ B pathway targets (Fig. 9). Considering the above, this study allowed to appreciate that drug discovery and complex target analysis can rely more on discovering new drug candidates through *in silico* methods. This type of search for new drugs could be more quickly and cost-efficiently translated into *in vitro* and *in vivo* screens.



**Figure 9.** Integrative *in silico* pipeline protocol for high-throughput virtual screening and drug discovery. The analytical schema highlights the synergy of structure- and ligand-based methods. In addition, the iterative screening provides many different opportunities to further optimise and adapt any compound library. The process concludes with hit compound selection, optimisation, and early validation studies. ADMET - absorption, distribution, metabolism, excretion, and toxicity.

#### 7.4. Programmatic approaches and data management: maintaining and designing robust workflows

Workflow development can become a key component in a successful *in silico* discovery pipeline enabling many different research aspects, including managing data flow and merging virtual analyses with *in vitro* or *in vivo* assessments<sup>292</sup>. Establishing good workflow building, maintenance, and use practices guarantees better research reproducibility and resource savings since data curation and retrieval can be automated or semi-automated processes<sup>293,294</sup>. Part of the research aims of this thesis was to develop software packages, namely *OmicInt* and *Fiscore*, and make the introduced analytical approaches more available to other researchers (Fig. 10). Moreover, it was necessary to create an interactive and user-friendly environment where machine learning analyses could be easily implemented by non-experts. In the case of cheminformatics, Schrödinger suite<sup>201</sup> provides a fully customisable set of tools and the user can easily adapt the protocol introduced in this thesis. To assist with early compound library preparation, a *Chemexpy* software package was also introduced for the Python programming environment. Together, these pieces of software and protocols create an adaptable set of research tools that can facilitate target evaluation and new drug discovery.

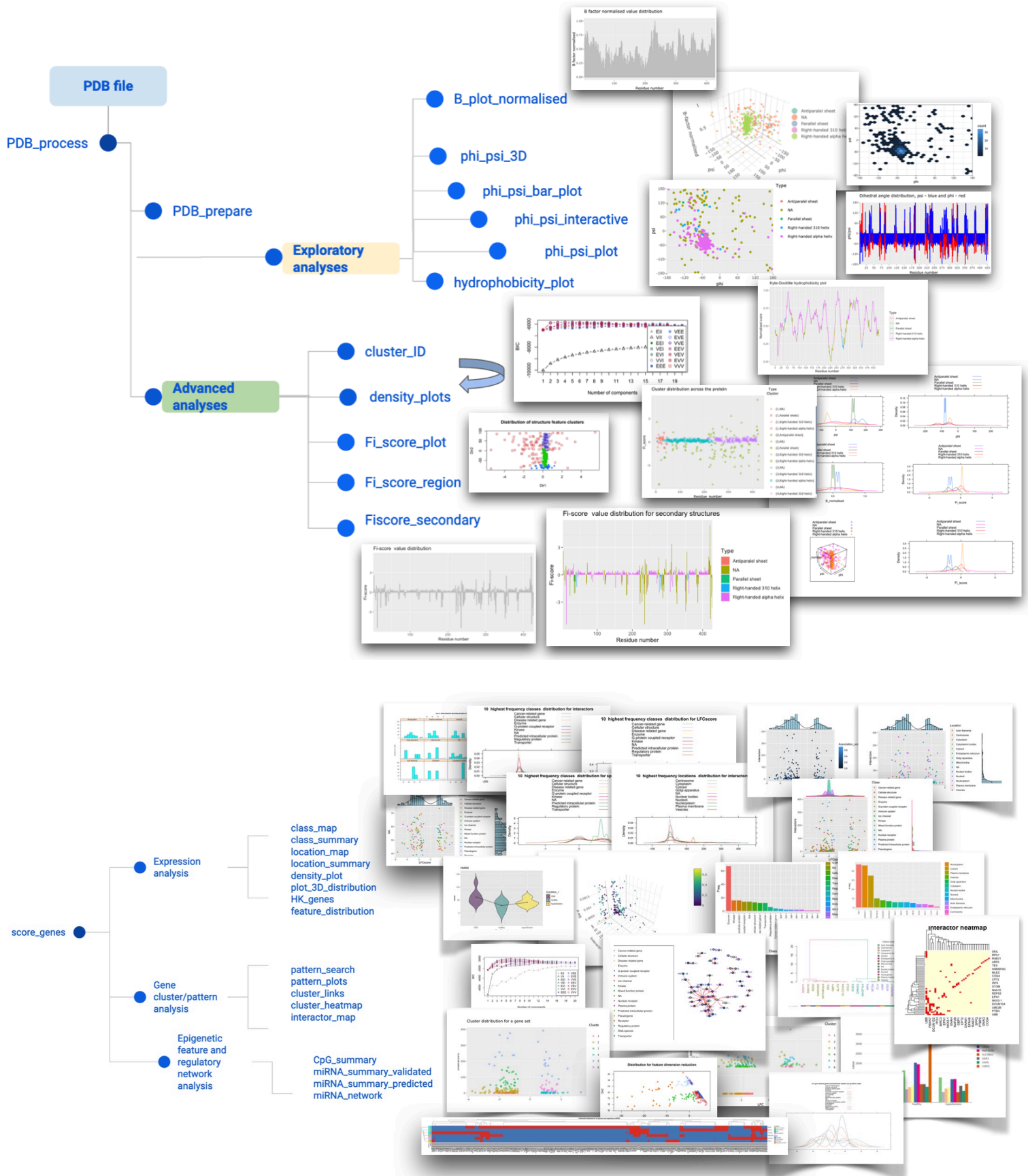
*OmicInt* is an R software package developed for an in-depth exploration of significantly changed genes, gene expression patterns, and the associated methylome as well as the miRNA environment. This piece of software accompanies the second chapter of the thesis focusing on *omics* analyses. The package helps to assess gene clusters based on their known or predicted interactors from several different resources, e.g., UniProt<sup>295</sup> and STRINGdb<sup>146,147</sup>. Moreover, *OmicInt* provides an easy Gaussian mixture modelling<sup>148,150,296</sup> pipeline for integrative analysis that can be used by a non-expert to explore gene expression datasets. Specifically, the package expands the  $LFC_{score}$  functionality by allowing single-cell and proteome experimental data integration. In addition, many other package functionalities can aid in studying specific gene networks, understanding cellular perturbation events, and exploring interactions that might not be easily detectable otherwise.

Lack of bioinformatics tools for a quick assessment of protein conformational and topological features motivated to create an integrative and user-friendly R software package - *Fiscore*<sup>246,297-300</sup>. This package complements the fourth chapter of the thesis. One of the key features of the *Fiscore* package is Gaussian mixture modelling to allow a probabilistic evaluation of complex structural features. The package builds on the mathematical formulation of protein physicochemical properties that can be easily visualised and explored with interactive plots.



All pieces of software are accompanied by vignettes and supporting documentation which are provided to guide the user through detailed tutorials and use cases<sup>267,301–304</sup>. In addition, Github<sup>305</sup> provides an opportunity to actively make suggestions for additional features.

R software packages are distributed as a part of the CRAN network<sup>306</sup> and the python cheminformatics package is on the PyPi platform<sup>307</sup>. All software tools promote open science practices and make research more accessible. With community inputs and suggestions, it is possible to expand the analytical scope and ensure the quality of programmatic solutions.



**Figure 10.** *Fiscore* (top) and *OmicInt* (bottom) package architecture visualisation with function organisation and sample outputs.

## 7.5. Thesis overview and conclusion

Drug discovery and development depends on our ability to identify therapeutically relevant targets and match that information with a complex chemical space that can lead to potential drug candidates<sup>15,23,152</sup>. However, in the last few decades the continued decline in new drug discovery and growing R&D expenditures prompted many companies to rethink their discovery strategies and begin focusing on computational methods<sup>17,18,22,244</sup>. As a result, computational biology, bioinformatics, and cheminformatics have made significant inroads into the pharmaceutical industry where *in silico* approaches now not only accelerate the exploration of new therapeutic candidates but also help to reduce R&D costs and the likelihood of missing relevant hits<sup>55,131,308</sup>. Despite many advancements in computational biology, bioinformatics, system biology, and computational chemistry, there is still a significant lack of end-to-end solutions for the right target identification and selection of the most promising compounds<sup>43,59,60,281</sup>. Considering the above challenges and the need for improved R&D strategies, this thesis aimed to introduce multi-*omics* and highly integrative analytical frameworks for a more streamlined target and therapeutics discovery approach.

The research began by developing an analytical strategy for studying complex diseases through multi-*omics* approaches to identify new therapeutically relevant targets. A case study of cardiomyopathies allowed to demonstrate how *omics* data integration, data enrichment, and machine learning can aid in better understanding multifaceted disease aetiologies. These insights provided an impetus to develop a protein structural and topological classification methodology so that multiple targets can be evaluated, grouped, and analysed based on structural and functional features. The final experimental chapter of the thesis combined the earlier analyses by introducing a novel hierarchical HTVS pipeline and comprehensive target analysis to reveal potential drug candidates for a challenging immunotherapeutic target. A case study target, the human c-Rel protein, was selected to build this analytical environment comprising structural bioinformatics, molecular modelling, cheminformatics, and machine learning. 15 new hit compounds were discovered after combining structure- and ligand-based approaches to parse an unprecedented size chemical library (659 M chemical entities)<sup>80,157,278</sup>. The devised screening blueprint can be applied to study complex targets and accelerate compound selection.

In summary, the outlined studies create a holistic research strategy for drug discovery in the computational space. Proposed solutions and novel insights in therapeutics development can significantly improve the current R&D strategies by reducing screening time, costs, and

turnaround<sup>2,13,17,18,28</sup>. Importantly, the introduced highly integrative and network-centric approaches offer a better understanding of pathological perturbations and can help deliver so much needed clinical solutions faster and with a safer profile.

## 8. Future work

New therapeutics development faces a number of challenges and, while some are the commercial pressures to maintain market dominance, the bigger issue is the constantly shrinking pool of easy-to-identify and viable targets<sup>2,4,15,140,169</sup>. The change in research strategy brought about by the use of *in silico*, ML/AI, and data mining is likely to continue to grow in popularity for preclinical research and development<sup>2,28,49,244,281,286</sup>. Thus, reducing associated R&D costs and the time needed to produce new pharmaceuticals will depend on how well we can take advantage of existing methodologies and continue evolving the *in silico* field.

The future work within the scope of drug discovery will continue to build on the findings and newly developed methodologies described in this thesis. The planned research trajectories could be divided into several themes. The first will focus on continuing to establish disease network and perturbation event exploratory analyses where significantly changed genes can be explored in a broader context of mined proteomics, single cell, and regulatory data. Specifically, creating mathematical and systems biology methods to better classify and prioritise the expressive patterns should enable a more sensitive and specific detection of causal gene networks. This work will tie in with the second major research theme of protein structural analysis where the main focus will be to improve conformational modelling and feature prediction to assist with drug discovery. For example, multiple homologous and non-homologous proteins could be screened using different scoring windows to predict which scoring approach is the best for various target site comparisons. Similarly, known-binding pocket survey could help gain additional insights into druggable proteome characteristics. The third theme will continue to be ligand- and structure-based drug discovery as well as HTVS protocol improvement juxtaposing *in silico* readouts with *in vitro* binding studies. Building such a screening library should be a valuable reference for both statistical analyses and machine learning based modelling. The exploration of therapeutic intervention options for the NF- $\kappa$ B pathway will remain an important future research aspect because understanding NF- $\kappa$ B targeting principles could greatly enhance our ability to engage other challenging proteins or complexes. Finally, any newly developed software packages or tools will be made freely available to other researchers so that the methods can evolve and improve.

## 9. References

1. Taylor, D. The Pharmaceutical Industry and the Future of Drug Development. *Issues Environ. Sci. Technol.* 1–33 (2015).
2. Earm, K. & Earm, Y. E. Integrative approach in the era of failing drug discovery and development. *Integr. Med. Res.* **3**, 211 (2014).
3. Pereira, D. A. & Williams, J. A. Origin and evolution of high throughput screening. *Br. J. Pharmacol.* **152**, 53 (2007).
4. Bikash, D., Laith, A. M. & Nouri, N. Are we living in the end of the blockbuster drug era? *Drug News Perspect.* **23**, 670–684 (2010).
5. Lorenz, M. & Peter, F. The future of high-throughput screening. *J. Biomol. Screen.* **13**, 443–448 (2008).
6. Christine, D. & Brian, M. The impact of genomics on drug discovery. *Annu. Rev. Pharmacol. Toxicol.* **40**, 193–208 (2000).
7. Roland, D. Historical overview of chemical library design. *Methods Mol. Biol.* **685**, 3–25 (2011).
8. Ohlstein, E. H., Johnson, A. G., Elliott, J. D. & Romanic, A. M. New Strategies in Drug Discovery. *Bioinforma. Drug Discov.* **316**, 1 (2006).
9. Skrepnek, G. H. & Sarnowski, J. J. Decision-making associated with drug candidates in the biotechnology research and development (R&D) pipeline. *J. Commer. Biotechnol.* **13**, 99–110 (2007).
10. Kola, I. & Landis, J. Can the pharmaceutical industry reduce attrition rates? *Nat. Rev. Drug Discov.* **3**, 711–716 (2004).
11. Khanna, I. Drug discovery in pharmaceutical industry: Productivity challenges and trends. *Drug Discovery Today.* **17**, 1088–1102 (2012).
12. Frank, S.-D. Is poor research the cause of the declining productivity of the pharmaceutical industry? An industry in need of a paradigm shift. *Drug Discov. Today.* **18**, 211–217 (2013).
13. Scannell, J. W. & Bosley, J. When Quality Beats Quantity: Decision Theory, Drug Discovery, and the Reproducibility Crisis. *PLoS One.* **11** (2016).
14. Takebe, T., Imai, R. & Ono, S. The Current Status of Drug Discovery and Development as Originated in United States Academia: The Influence of Industrial and Academic Collaboration on Drug Discovery and Development. *Clin. Transl. Sci.* **11**, 597 (2018).
15. Hopkins, A. L. & Groom, C. R. The druggable genome. *Nat. Rev. Drug Discov.* **1**, 727–730 (2002).
16. Mohs, R. C. & Greig, N. H. Drug discovery and development: Role of basic biological research. *Alzheimer's Dement. Transl. Res. Clin. Interv.* **3**, 651 (2017).

17. Herper, M. The Truly Staggering Cost Of Inventing New Drugs. *Forbes* <https://www.forbes.com/sites/matthewherper/2012/02/10/the-truly-staggering-cost-of-inventing-new-drugs/?sh=169c95294a94> (2012).
18. LaMattina, J. Should Pharma Companies Give Up Discovery Research? *Forbes* <https://www.forbes.com/sites/johnlamattina/2013/09/10/should-pharma-companies-give-up-discovery-research/> (2013).
19. Scannell, J. W., Blanckley, A., Boldon, H. & Warrington, B. Diagnosing the decline in pharmaceutical R&D efficiency. *Nature Reviews Drug Discovery*. **11**, 191–200 (2012).
20. Yildirim, O., Gottwald, M., Schüler, P. & Michel, M. C. Opportunities and Challenges for Drug Development: Public–Private Partnerships, Adaptive Designs and Big Data. *Front. Pharmacol.* **7**, 461 (2016).
21. Li, A. P. Screening for human ADME/Tox drug properties in drug discovery. *Drug Discovery Today*. **6**, 357–366 (2001).
22. Grainger, D. Why Too Many Clinical Trials Fail -- And A Simple Solution That Could Increase Returns On Pharma R&D. *Forbes* <https://www.forbes.com/sites/davidgrainger/2015/01/29/why-too-many-clinical-trials-fail-and-a-simple-solution-that-could-increase-returns-on-pharma-rd/?sh=75bf084db8b3> (2015).
23. Degoey, D. A., Chen, H. J., Cox, P. B. & Wendt, M. D. Beyond the Rule of 5: Lessons Learned from AbbVie’s Drugs and Compound Collection. *Journal of Medicinal Chemistry*. **61**, 2636–2651 (2018).
24. Rishton, G. Failure and Success in Modern Drug Discovery: Guiding Principles in the Establishment of High Probability of Success Drug Discovery Organizations. *Med. Chem. (Los. Angeles)*. **1**, 519–527 (2005).
25. Wong, C. H., Siah, K. W. & Lo, A. W. Estimation of clinical trial success rates and related parameters. *Biostatistics*. **20**, 273–286 (2019).
26. Fogel, D. B. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: A review. *Contemporary Clinical Trials Communications*. **11**, 156–164 (2018).
27. Cook, D. *et al.* Lessons learned from the fate of AstraZeneca’s drug pipeline: A five-dimensional framework. *Nature Reviews Drug Discovery*. **13**, 419–431 (2014).
28. Brogi, S., Ramalho, T. C., Kuca, K., Medina-Franco, J. L. & Valko, M. Editorial: In silico Methods for Drug Design and Discovery. *Front. Chem.* **8**, 612 (2020).
29. Hughes, J., Rees, S., Kalindjian, S. & Philpott, K. Principles of early drug discovery. *Br. J. Pharmacol.* **162**, 1239–1249 (2011).
30. Yun, T., Weiliang, Z., Kaixian, C. & Hualiang, J. New technologies in computer-aided drug design: Toward target identification and new chemical entity discovery. *Drug Discov. Today. Technol.* **3**, 307–313 (2006).

31. Ohlstein, E. H., Ruffolo, R. R. & Elliott, J. D. Drug discovery in the next millennium. *Annu. Rev. Pharmacol. Toxicol.* **40**, 177–91 (2000).
32. Li, C. I., Samuels, D. C., Zhao, Y. Y., Shyr, Y. & Guo, Y. Power and sample size calculations for high-throughput sequencing-based experiments. *Brief. Bioinform.* **19**, 1247–1255 (2017).
33. Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biol.* **17**, 13 (2016).
34. Pereira, M. A., Imada, E. L. & Guedes, R. L. M. RNA-seq: Applications and Best Practices. in *Applications of RNA-Seq and Omics Strategies - From Microorganisms to Human Health* (InTech, 2017). doi:10.5772/intechopen.69250.
35. Pederson, T. RNA interference and mRNA silencing, 2004: how far will they reach? *Mol. Biol. Cell.* **15**, 407–410 (2004).
36. Gygi, S., Rochon, Y., Franza, R. & Aebersold, R. Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol.* **19**, 1720–1730 (1999).
37. Service, R. F. Protein arrays step out of DNA's shadow. *Science.* **289**, 1673 (2000).
38. Gerhold, D., Jensen, R. & Gullans, S. Better therapeutics through microarrays. *Nat. Genet.* **32**, 547–552 (2002).
39. Watkins, S. & German, B. Toward the implementation of metabolomic assessments of human health and nutrition. *Curr. Opin. Biotechnol.* **13**, 512–516 (2002).
40. Doherty, A. *et al.* Discovery of a novel series of orally active non-peptide endothelin-A (ETA) receptor-selective antagonists. *J. Med. Chem.* **38**, 1259–1263 (1995).
41. Owens, J. Determining druggability. *Nat. Rev. Drug Discov.* **6**, 187–187 (2007).
42. Guiguemde, W. A. *et al.* Global phenotypic screening for antimalarials. *Chem. Biol.* **19**, 116–129 (2012).
43. Abi Hussein, H. *et al.* Global vision of druggability issues: applications and perspectives. *Drug Discovery Today.* **22**, 404–415 (2017).
44. Bowes, J. *et al.* Reducing safety-related drug attrition: The use of in vitro pharmacological profiling. *Nature Reviews Drug Discovery.* **11**, 909–922 (2012).
45. Noble, D. Why integration? *Integr. Med. Res.* **1**, 2–4 (2012).
46. Schneider, H. C. & Klabunde, T. Understanding drugs and diseases by systems biology? *Bioorg. Med. Chem. Lett.* **23**, 1168–1176 (2013).
47. Macalino, S. J. Y., Billones, J. B., Organo, V. G. & Carrillo, M. C. O. In silico strategies in tuberculosis drug discovery. *Molecules.* **25**, 665 (2020).
48. Andrade, C. H. *et al.* In Silico Chemogenomics Drug Repositioning Strategies for Neglected Tropical Diseases. *Curr. Med. Chem.* **26**, 4355–4379 (2018).
49. Lima, A. N. *et al.* Use of machine learning approaches for novel drug discovery. *Expert Opin Drug Discov.* **11**, 225–239 (2016).



50. Gawehn, E., Hiss, J. A. & Schneider, G. Deep Learning in Drug Discovery. *Mol. Inform.* **35**, 3–14 (2016).
51. Norman, R. A. *et al.* Computational approaches to therapeutic antibody design: Established methods and emerging trends. *Brief. Bioinform.* **21**, 1549–1567 (2020).
52. Wu, F. *et al.* Computational Approaches in Preclinical Studies on Drug Discovery and Development. *Front. Chem.* **8**, 726 (2020).
53. Ekins, S., Mestres, J. & Testa, B. In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling. *Br. J. Pharmacol.* **152**, 9 (2007).
54. Lavecchia, A. Machine-learning approaches in drug discovery: Methods and applications. *Drug Discovery Today*. **20**, 318–331 (2015).
55. Xia, X. Bioinformatics and Drug Discovery. *Curr. Top. Med. Chem.* **17**, 1709 (2017).
56. Drakeman, D. & Oraiopoulos, N. The Risk of De-Risking Innovation: Optimal R&D Strategies in Ambiguous Environments. *Calif. Manage. Rev.* **62**, 42–63 (2020).
57. Cosconati, S. *et al.* Virtual screening with AutoDock: Theory and practice. *Expert Opinion on Drug Discovery*. **5**, 597–607 (2010).
58. Cherkasov, A. *et al.* QSAR modeling: Where have you been? Where are you going to? *Journal of Medicinal Chemistry*. **57**, 4977–5010 (2014).
59. Braga, R. *et al.* Virtual Screening Strategies in Medicinal Chemistry: The State of the Art and Current Challenges. *Curr. Top. Med. Chem.* **14**, 1899–1912 (2014).
60. Santos, R. *et al.* A comprehensive map of molecular drug targets. *Nat. Rev. Drug Discov.* **16**, 19–34 (2016).
61. Knapp, S. Emerging target families: Intractable targets. in *Handbook of Experimental Pharmacology*. **232**, 43–58 (Springer New York LLC, 2016).
62. Finan, C. *et al.* The druggable genome and support for target identification and validation in drug development. *Sci. Transl. Med.* **9** (2017).
63. Alshalalfa, M. & Alhaji, R. Using context-specific effect of miRNAs to identify functional associations between miRNAs and gene signatures. *BMC Bioinformatics*. **14** (2013).
64. Yoon, S. *et al.* GScluster: Network-weighted gene-set clustering analysis. *BMC Genomics*. **20**, 352 (2019).
65. Brazhnik, P., De La Fuente, A. & Mendes, P. Gene networks: How to put the function in genomics. *Trends in Biotechnology*. **20**, 467–472 (2002).
66. Sumathipala, M. & Weiss, S. T. Predicting miRNA-based disease-disease relationships through network diffusion on multi-omics biological data. *Sci. Rep.* **10**, 1–12 (2020).
67. Bizzarri, M., Palombo, A. & Cucina, A. Theoretical aspects of Systems Biology. *Prog. Biophys. Mol. Biol.* **112**, 33–43 (2013).

68. Lavecchia, A. & Giovanni, C. Virtual Screening Strategies in Drug Discovery: A Critical Review. *Curr. Med. Chem.* **20**, 2839–2860 (2013).
69. Pinzi, L. & Rastelli, G. Molecular docking: Shifting paradigms in drug discovery. *International Journal of Molecular Sciences.* **20** (2019).
70. Lavecchia, A. & Cerchia, C. In silico methods to address polypharmacology: Current status, applications and future perspectives. *Drug Discovery Today.* **21**, 288–298 (2016).
71. Ji, B. Y. *et al.* Predicting miRNA-disease association from heterogeneous information network with GraRep embedding model. *Sci. Rep.* **10**, 1–12 (2020).
72. Dahlin, J. L., Inglese, J. & Walters, M. A. Mitigating risk in academic preclinical drug discovery. *Nature Reviews Drug Discovery.* **14**, 279–294 (2015).
73. Bharti, R. & Shukla, S. K. Molecules against Covid-19: An in silico approach for drug development. *J. Electron. Sci. Technol.* **19**, 100095 (2021).
74. Oselusi, S. O., Egieyeh, S. A. & Christoffels, A. Cheminformatic Profiling and Hit Prioritization of Natural Products with Activities against Methicillin-Resistant *Staphylococcus aureus* (MRSA). *Molecules* **26**, 3674 (2021).
75. Duncan, R. Drug development and regulation. *Medicine* **36**, 369–376 (2008).
76. Boran, A. D. & Iyengar, R. Systems approaches to polypharmacology and drug discovery. *Curr. Opin. Drug Discov. Devel.* **13**, 297 (2010).
77. Mitchell, K. J. What is complex about complex disorders? *Genome Biol.* **13**, 237 (2012).
78. Müller, B. *et al.* Improved prediction of complex diseases by common genetic markers: state of the art and further perspectives. *Hum. Genet.* **135**, 259 (2016).
79. Vasaikar, S., Bhatia, P., Bhatia, P. G. & Yaiw, K. C. Complementary Approaches to Existing Target Based Drug Discovery for Identifying Novel Drug Targets. *Biomedicines.* **4** (2016).
80. Gangadharan, N. T., Venkatachalam, A. B. & Sugathan, S. High-throughput and In Silico screening in drug discovery. in *Bioresources and Bioprocess in Biotechnology.* **1**, 247–273 (Springer Singapore, 2017).
81. Sweet, M. E. *et al.* Transcriptome analysis of human heart failure reveals dysregulated cell adhesion in dilated cardiomyopathy and activated immune pathways in ischemic heart failure. *BMC Genomics.* **19** (2018).
82. McCombe, P. A. & Henderson, R. D. The Role of immune and inflammatory mechanisms in ALS. *Curr. Mol. Med.* **11**, 246–54 (2011).
83. Swirski, F. K. & Nahrendorf, M. Cardioimmunology: the immune system in cardiac homeostasis and disease. *Nature Reviews Immunology.* **18**, 733–744 (2018).
84. Gracia-Hernandez, M., Sotomayor, E. M. & Villagra, A. Targeting Macrophages as a Therapeutic Option in Coronavirus Disease 2019. *Front. Pharmacol.* **11**, 1659 (2020).

85. Strassheim, D., Dempsey, E. C., Gerasimovskaya, E., Stenmark, K. & Karoor, V. Role of inflammatory cell subtypes in heart failure. *Journal of Immunology Research* (2019).
86. Liu, T., Zhang, L., Joo, D. & Sun, S.-C. NF- $\kappa$ B signaling in inflammation. *Signal Transduct. Target. Ther.* **2**, 1–9 (2017).
87. Packer, M. The Imminent Demise of Cardiovascular Drug Development. *JAMA Cardiol.* **2**, 1293 (2017).
88. Brito, D. & Cepeda, B. *Heart Failure, Congestive (CHF)*. *StatPearls* (StatPearls Publishing, 2018).
89. Metra, M. & Teerlink, J. R. Heart failure. *The Lancet.* **390**, 1981–1995 (2017).
90. Bowles, N. E., Bowles, K. R. & Towbin, J. A. The ‘final common pathway’ hypothesis and inherited cardiovascular disease: The role of cytoskeletal proteins in dilated cardiomyopathy. *Herz.* **25**, 168–175 (2000).
91. Yang, G., Chen, S., Ma, A., Lu, J. & Wang, T. Identification of the difference in the pathogenesis in heart failure arising from different etiologies using a microarray dataset. *Clinics.* **72**, 600–608 (2017).
92. Zrimec, J., Buric, F., Kokina, M., Garcia, V. & Zelezniak, A. Learning the Regulatory Code of Gene Expression. *Front. Mol. Biosci.* **8**, 530 (2021).
93. Sparber, P., Filatova, A., Khantemirova, M. & Skoblov, M. The role of long non-coding RNAs in the pathogenesis of hereditary diseases. *BMC Medical Genomics.* **12**, 63–78 (2019).
94. Meder, B. *et al.* Epigenome-Wide Association Study Identifies Cardiac Gene Patterning and a Novel Class of Biomarkers for Heart Failure. *Circulation.* **136**, 1528–1544 (2017).
95. Fordyce, C. B. *et al.* Cardiovascular drug development: Is it dead or just hibernating? *Journal of the American College of Cardiology.* **65**, 1567–1582 (2015).
96. Reed, B. N. & Sueta, C. A. A Practical Guide for the Treatment of Symptomatic Heart Failure with Reduced Ejection Fraction (HFrEF). *Curr. Cardiol. Rev.* **11**, 23 (2015).
97. Bhandari, B. & Masood, W. *Ischemic Cardiomyopathy*. *StatPearls* (StatPearls Publishing, 2019).
98. Mahmaljy, H., Yelamanchili, V. S. & Singhal, M. Dilated Cardiomyopathy. *Matern. Fetal Cardiovasc. Dis.* 97–106 (2020).
99. Savoji, H., Mohammadi, M.H., Rafatian, N., Toroghi, M.K., Wang, E.Y., Zhao, Y., Korolj, A., Ahadian, S., Radisic, M. Cardiovascular disease models: A game changing paradigm in drug discovery and screening. *Biomaterials.* **198**, 3–26 (2019).
100. Suhre, K. *et al.* Human metabolic individuality in biomedical and pharmaceutical research. *Nature.* **477**, 54–62 (2011).
101. Povsic, T. J. *et al.* Navigating the Future of Cardiovascular Drug Development—Leveraging Novel Approaches to Drive Innovation and Drug Discovery: Summary of Findings from the

- Novel Cardiovascular Therapeutics Conference. *Cardiovasc. Drugs Ther.* **31**, 445–458 (2017).
102. Alimadadi, A., Munroe, P. B., Joe, B. & Cheng, X. Meta-analysis of dilated cardiomyopathy using cardiac rna-seq transcriptomic datasets. *Genes (Basel)*. **11** (2020).
  103. Witt, E. *et al.* Correlation of gene expression and clinical parameters identifies a set of genes reflecting LV systolic dysfunction and morphological alterations. *Physiol Genomics* **51**, 356–367 (2019).
  104. Mandlik, V., Bejugam, P. R. & Singh, S. Application of Artificial Neural Networks in Modern Drug Discovery. in *Artificial Neural Network for Drug Design, Delivery and Disposition*. 123–139 (Elsevier Inc., 2016). doi:10.1016/B978-0-12-801559-9.00006-5.
  105. Iskar, M., Zeller, G., Zhao, X. M., van Noort, V. & Bork, P. Drug discovery in the age of systems biology: the rise of computational approaches for data integration. *Curr. Opin. Biotechnol.* **23**, 609–616 (2012).
  106. Peterson, J. M. *et al.* NF- $\kappa$ B inhibition rescues cardiac function by remodeling calcium genes in a Duchenne muscular dystrophy model. *Nat. Commun.* **9**, 1–14 (2018).
  107. Jing-Bo, X. *et al.* Hypoxia/ischemia promotes CXCL10 expression in cardiac microvascular endothelial cells by NF $\kappa$ B activation. *Cytokine* **81**, 63–70 (2016).
  108. Van Linthout, S., & Tschöpe, C. Inflammation - Cause or Consequence of Heart Failure or Both?. *Current heart failure reports*, **14**, 251–265 (2017). <https://doi.org/10.1007/s11897-017-0337-9>.
  109. Murciano-Goroff, Y. R., Warner, A. B. & Wolchok, J. D. The future of cancer immunotherapy: microenvironment-targeting combinations. *Cell Res.* **30**, 507–519 (2020).
  110. Scheller, J., Chalaris, A., Schmidt-Arras, D. & Rose-John, S. The pro- and anti-inflammatory properties of the cytokine interleukin-6. *Biochim. Biophys. Acta - Mol. Cell Res.* **1813**, 878–888 (2011).
  111. Lawrence, T. The nuclear factor NF-kappaB pathway in inflammation. *Cold Spring Harb. Perspect. Biol.* **1** (2009).
  112. Peng, C., Ouyang, Y., Lu, N. & Li, N. The NF- $\kappa$ B Signaling Pathway, the Microbiota, and Gastrointestinal Tumorigenesis: Recent Advances. *Front. Immunol.* **11**, 1387 (2020).
  113. Hayden, M. S. & Ghosh, S. Signaling to NF-kappaB. *Genes Dev.* **18**, 2195–224 (2004).
  114. Ting, A. Y. & Endy, D. Decoding NF-kappaB signaling. *Science*. **298**, 1189–90 (2002).
  115. Grinberg-Bleyer, Y. *et al.* The Alternative NF- $\kappa$ B Pathway in Regulatory T Cell Homeostasis and Suppressive Function. *J. Immunol.* **200**, 2362–2371 (2018).
  116. Ruan, Q. & Chen, Y. H. Nuclear factor- $\kappa$ B in immunity and inflammation: the Treg and Th17 connection. *Adv. Exp. Med. Biol.* **946**, 207–21 (2012).

117. Köntgen, F. *et al.* Mice lacking the c-rel proto-oncogene exhibit defects in lymphocyte proliferation, humoral immunity, and interleukin-2 expression. *Genes Dev.* **9**, 1965–1977 (1995).
118. Hoffmann, A., Natoli, G. & Ghosh, G. Transcriptional regulation via the NF- $\kappa$ B signaling module. *Oncogene.* **25**, 6706–6716 (2006).
119. Freitas, R., & Fraga, C. NF- $\kappa$ B-IKK $\beta$  Pathway as a Target for Drug Development: Realities, Challenges and Perspectives. *Current drug targets.* **19**, 1933–1942 (2018). <https://doi.org/10.2174/1389450119666180219120534>.
120. Hoffmann, A., Levchenko, A., Scott, M. L. & Baltimore, D. The I $\kappa$ B-NF- $\kappa$ B signaling module: Temporal control and selective gene activation. *Science.* **298**, 1241–1245 (2002).
121. Grinberg-Bleyer, Y. *et al.* NF- $\kappa$ B c-Rel Is Crucial for the Regulatory T Cell Immune Checkpoint in Cancer. *Cell.* **170**, 1096–1108.e13 (2017).
122. Fulford, T. S., Ellis, D. & Gerondakis, S. Understanding the Roles of the NF- $\kappa$ B Pathway in Regulatory T Cell Development, Differentiation and Function. in *Progress in Molecular Biology and Translational Science.* **136**, 57–67 (Elsevier B.V., 2015).
123. Gaspar-Pereira, S. *et al.* The NF- $\kappa$ B subunit c-Rel stimulates cardiac hypertrophy and fibrosis. *Am. J. Pathol.* **180**, 929–939 (2012).
124. Li, A. & Jacks, T. Driving Rel-iant Tregs toward an Identity Crisis. *Immunity.* **47**, 391–393 (2017).
125. Hamid, T. *et al.* Cardiomyocyte NF- $\kappa$ B p65 promotes adverse remodelling, apoptosis, and endoplasmic reticulum stress in heart failure. *Cardiovasc. Res.* **89**, 129–138 (2011).
126. Gordon, J. W., Shaw, J. A. & Kirshenbaum, L. A. Multiple facets of NF- $\kappa$ B in the heart: To be or not to NF- $\kappa$ B. *Circulation Research.* **108**, 1122–1132 (2011).
127. Kumar, R., Yong, Q. C. & Thomas, C. M. Do multiple nuclear factor kappa B activation mechanisms explain its varied effects in the heart? *Ochsner J.* **13**, 157–165 (2013).
128. Fiordelisi, A., Iaccarino, G., Morisco, C., Coscioni, E. & Sorriento, D. NfkappaB is a key player in the crosstalk between inflammation and cardiovascular diseases. *International Journal of Molecular Sciences.* **20** (2019).
129. Ayllón, B.T., Souilhol, C., Oakley, F. *et al.* A c-REL drives atherosclerosis at sites of disturbed blood flow. *Heart.* **107**:A176 (2021).
130. van der Heiden, K., Cuhlmann, S., Luong, L., Mustafa, Z. & Evans, P. Role of nuclear factor kappaB in cardiovascular health and disease. *Clin. Sci. (Lond).* **118**, 593–605 (2010).
131. Sliwoski, G., Kothiwale, S., Meiler, J., Lowe, E. W. & Jr. Computational methods in drug discovery. *Pharmacol. Rev.* **66**, 334–95 (2014).
132. Hopkins, L, A. Network Pharmacology. *Netw. Pharmacol.* **25**, 127–164 (2017).

133. Cheng, F. *et al.* Network-based approach to prediction and population-based validation of in silico drug repurposing. *Nat. Commun.* **9** (2018).
134. Yildirim, M. A., Goh, K. II, Cusick, M. E., Barabási, A. L. & Vidal, M. Drug-target network. *Nat. Biotechnol.* **25**, 1119–1126 (2007).
135. Albert-László, B. & Zoltán, O. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* **5**, 101–113 (2004).
136. Eric, S. Molecular networks as sensors and drivers of common human diseases. *Nature.* **461**, 218–223 (2009).
137. Mete, C. & Aldons, L. Systems genetics approaches to understand complex traits. *Nat. Rev. Genet.* **15**, 34–48 (2014).
138. Yan, J., Risacher, S. L., Shen, L. & Saykin, A. J. Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data. *Brief. Bioinform.* **19**, 1370–1381 (2018).
139. Nir, F., Michal, L., Iftach, N. & Dana, P. Using Bayesian networks to analyze expression data. *J. Comput. Biol.* **7**, 601–620 (2000).
140. Vijay, R., Li, S., Jason, M. & Andrew, S. Pathway analysis of genomic data: concepts, methods, and prospects for future development. *Trends Genet.* **28**, 323–332 (2012).
141. Enrico, G., Anaïs, B., Natalio, K., Reinhard, S. & Alfonso, V. EnrichNet: network-based gene set enrichment analysis. *Bioinformatics.* **28** (2012).
142. Hänzelmann, S., Castelo, R., & Guinney, J. GSEA: gene set variation analysis for microarray and RNA-seq data. *BMC bioinformatics.* **14**, 7 (2013). <https://doi.org/10.1186/1471-2105-14-7>.
143. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, (2014).
144. Koscielny, G. *et al.* Open Targets: A platform for therapeutic target identification and Validation. *Nucleic Acids Res.* **45**, D985–D994 (2017).
145. Home - Open Targets. <https://www.opentargets.org/>.
146. Szklarczyk, D. *et al.* STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).
147. STRING: functional protein association networks. <https://string-db.org/>.
148. Reynolds, D. Gaussian Mixture Models. *Encycl. Biometrics.* 659–663 (2009) doi:10.1007/978-0-387-73003-5\_196.
149. Gao, C., Zhu, Y., Shen, X., & Pan, W. Estimation of multiple networks in Gaussian mixture models. *Electronic journal of statistics.* **10**, 1133–1154 (2016). <https://doi.org/10.1214/16-EJS1135>.

150. Liu, Z., Song, Y., Xie, C. & Tang, Z. A new clustering method of gene expression data based on multivariate Gaussian mixture models. *Signal, Image Video Process.* **10**, 359–368 (2015).
151. Zhang, H., Jiang, T., Shan, G., Xu, S. & Song, Y. Gaussian network model can be enhanced by combining solvent accessibility in proteins. *Sci. Rep.* **7**, (2017).
152. Schmidtke, P. & Barril, X. Understanding and predicting druggability. A high-throughput method for detection of drug binding sites. *J. Med. Chem.* **53**, 5858–5867 (2010).
153. Cukuroglu, E., Engin, H. B., Gursoy, A. & Keskin, O. Hot spots in protein-protein interfaces: Towards drug discovery. *Prog. Biophys. Mol. Biol.* **116**, 165–173 (2014).
154. Fuller, J. C., Burgoyne, N. J. & Jackson, R. M. Predicting druggable binding sites at the protein-protein interface. *Drug Discovery Today.* **14**, 155–161 (2009).
155. Zhang, H., Yu, Z., He, J., Hua, B. & Zhang, G. Identification of the molecular mechanisms underlying dilated cardiomyopathy via bioinformatic analysis of gene expression profiles. *Exp. Ther. Med.* **13**, 273–279 (2017).
156. Vilar, S., Quezada, E., Uriarte, E., Costanzi, S., Borges, F., Viña, D., & Hripcsak, G. Computational Drug Target Screening through Protein Interaction Profiles. *Scientific reports.* **6**, 36969 (2016). <https://doi.org/10.1038/srep36969>.
157. Follmann, M. *et al.* An approach towards enhancement of a screening library: The Next Generation Library Initiative (NGLI) at Bayer — against all odds? *Drug Discovery Today.* **24**, 668–672 (2019).
158. Cao, Y., Charisi, A., Cheng, L. C., Jiang, T. & Girke, T. ChemmineR: A compound mining framework for R. *Bioinformatics.* **24**, 1733–1734 (2008).
159. Skjærven, L., Yao, X.-Q., Scarabelli, G. & Grant, B. J. Integrating protein structural dynamics and evolutionary analysis with Bio3D. *BMC Bioinformatics.* **15**, 399 (2014).
160. Aronson, S. J. & Rehm, H. L. Building the foundation for genomics in precision medicine. *Nature.* **526**, 336–342 (2015).
161. Han, Y., Cheng, L. & Sun, W. Analysis of Protein-Protein Interaction Networks through Computational Approaches. *Protein Pept. Lett.* **27**, 265–278 (2019).
162. Asarnow, D. & Singh, R. Automatic classification of protein structures using low-dimensional structure space mappings. *BMC Bioinformatics.* **15**, (2014).
163. Wong, W. C., Maurer-Stroh, S. & Eisenhaber, F. Not all transmembrane helices are born equal: Towards the extension of the sequence homology concept to membrane proteins. *Biol. Direct.* **6**, (2011).
164. Tramontano, A. & Morea, V. Assessment of Homology-Based Predictions in CASP5. in *Proteins: Structure, Function and Genetics.* **53**, 352–368 (Proteins, 2003).
165. Sylvestersen, K. B., Young, C. & Nielsen, M. L. Advances in characterizing ubiquitylation sites by mass spectrometry. *Curr. Opin. Chem. Biol.* **17**, 49–58 (2013).

166. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. E. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10**, 845–858 (2015).
167. Campbell, S. J., Gold, N. D., Jackson, R. M. & Westhead, D. R. Ligand binding: Functional site location, similarity and docking. *Current Opinion in Structural Biology.* **13**, 389–395 (2003).
168. Peach, M. L., Cachau, R. E. & Nicklaus, M. C. Conformational energy range of ligands in protein crystal structures: The difficult quest for accurate understanding. *Journal of Molecular Recognition.* **30** (2017).
169. Pérot, S., Sperandio, O., Miteva, M. A., Camproux, A. C. & Villoutreix, B. O. Druggable pockets and binding site centric chemical space: A paradigm shift in drug discovery. *Drug Discovery Today.* **15**, 656–667 (2010).
170. March-Vila, E. *et al.* On the Integration of In Silico Drug Design Methods for Drug Repurposing. *Front. Pharmacol.* **8**, 298 (2017).
171. Forli, S. *et al.* Computational protein-ligand docking and virtual drug screening with the AutoDock suite. *Nat. Protoc.* **11**, 905–919 (2016).
172. Batool, M., Ahmad, B. & Choi, S. A structure-based drug discovery paradigm. *International Journal of Molecular Sciences.* **20** (2019).
173. Zhou, A. Q., O’Hern, C. S. & Regan, L. Revisiting the Ramachandran plot from a new angle. *Protein Sci.* **20**, 1166–1171 (2011).
174. Kanapeckaitė, A., Beurivage, C., Hancock, M. & Verschueren, E. Fi-score: a novel approach to characterise protein topology and aid in drug discovery studies. *J. Biomol. Struct. Dyn.* 1–11 (2020) doi:10.1080/07391102.2020.1854859.
175. Faraggi, E., Xue, B. & Zhou, Y. Improving the prediction accuracy of residue solvent accessibility and real-value backbone torsion angles of proteins by guided-learning through a two-layer neural network. *Proteins Struct. Funct. Bioinforma.* **74**, 847–856 (2009).
176. Heffernan, R. *et al.* Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. *Sci. Rep.* **5**, 1–11 (2015).
177. Carugo, O. & Eisenhaber, F. Probabilistic evaluation of similarity between pairs of three-dimensional protein structures utilizing temperature factors. *J. Appl. Crystallogr.* **30**, 547–549 (1997).
178. Carugo, O. & Argos, P. Reliability of atomic displacement parameters in protein crystal structures. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **55**, 473–478 (1999).
179. Weiss, M. S. On the interrelationship between atomic displacement parameters (ADPs) and coordinates in protein structures. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **63**, 1235–1242 (2007).



180. Yin, H., Li, Y.-Z. & Li, M.-L. On the Relation Between Residue Flexibility and Residue Interactions in Proteins. *Protein Pept. Lett.* **18**, 450–456 (2011).
181. Parthasarathy, S. & Murthy, M. R. N. Analysis of temperature factor distribution in high-resolution protein structures. *Protein Sci.* **6**, 2561–2567 (2008).
182. Bornot, A., Etchebest, C. & De Brevern, A. G. Predicting protein flexibility through the prediction of local structures. *Proteins Struct. Funct. Bioinforma.* **79**, 839–852 (2011).
183. Mauno, V. Relationship of Protein Flexibility to Thermostability. *Protein Eng.* **1** (1987).
184. Lorenzo, O., Picatoste, B., Ares-Carrasco, S., Ramírez, E., Egido, J., & Tuñón, J. Potential role of nuclear factor  $\kappa$ B in diabetic cardiomyopathy. *Mediators of inflammation*. 652097 (2011). <https://doi.org/10.1155/2011/652097>.
185. Santos, D. G., Resende, M. F., Mill, J. G., Mansur, A. J., Krieger, J. E., & Pereira, A. C. Nuclear Factor (NF) kappaB polymorphism is associated with heart function in patients with heart failure. *BMC medical genetics.* **11**, 89 (2010). <https://doi.org/10.1186/1471-2350-11-89>.
186. Phelps, C. B., Sengchanthalangsy, L. L., Malek, S. & Ghosh, G. Mechanism of  $\kappa$ B DNA binding by Rel/NF- $\kappa$ b dimers. *J. Biol. Chem.* **275**, 24392–24399 (2000).
187. Müller, C. W. & Harrison, S. C. The structure of the NF- $\kappa$ B p50:DNA-complex a starting point for analyzing the Rel family. *FEBS Letters.* **369**, 113–117 (1995).
188. Taylor, W. R. & Orengo, C. A. Protein structure alignment. *J. Mol. Biol.* **208**, 1–22 (1989).
189. Radivojac, P. Protein flexibility and intrinsic disorder. *Protein Sci.* **13**, 71–80 (2004).
190. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453 (1970).
191. Wang, K., Horst, J. A., Cheng, G., Nickle, D. C. & Samudrala, R. Protein meta-functional signatures from combining sequence, structure, evolution, and amino acid property information. *PLoS Comput. Biol.* **4** (2008).
192. Osorio, D., Rondón-Villarreal, P. & Torres, R. Peptides: A package for data mining of antimicrobial peptides. *R J.* **7**, 4–14 (2015).
193. Molecular Dynamics Simulations - Gromacs. [https://www.gromacs.org/Documentation\\_of\\_outdated\\_versions/Terminology/Molecular\\_Dynamics\\_Simulations](https://www.gromacs.org/Documentation_of_outdated_versions/Terminology/Molecular_Dynamics_Simulations).
194. Ferreira, L. G., Santos, R. N. Dos, Oliva, G. & Andricopulo, A. D. Molecular Docking and Structure-Based Drug Design Strategies. *Mol.* **20**, 13384-13421 (2015).
195. Ferreira, L. L. G. & Andricopulo, A. D. Editorial: Chemoinformatics Approaches to Structure- and Ligand-Based Drug Design. *Front. Pharmacol.* **9**, 1416 (2018).
196. Greenwell, C. & Beran, G. J. O. Inaccurate Conformational Energies Still Hinder Crystal Structure Prediction in Flexible Organic Molecules. *Cryst. Growth Des.* **20**, 4875–4881 (2020).

197. Takeda-Shitaka, M., Takaya, D., Chiba, C., Tanaka, H. & Umeyama, H. Protein Structure Prediction in Structure Based Drug Design. *Curr. Med. Chem.* **11**, 551–558 (2005).
198. Bauer, J. & Bauerová-Hlinková, V. Normal Mode Analysis: A Tool for Better Understanding Protein Flexibility and Dynamics with Application to Homology Models. in *Homology Molecular Modeling - Perspectives and Applications* (IntechOpen, 2020). doi:10.5772/intechopen.94139.
199. Bahar, I., Lezon, T. R., Bakan, A. & Shrivastava, I. H. Normal mode analysis of biomolecular structures: Functional mechanisms of membrane proteins. *Chem. Rev.* **110**, 1463–1497 (2010).
200. ChemmineR: Cheminformatics Toolkit for R. <https://www.bioconductor.org/packages/devel/bioc/vignettes/ChemmineR/inst/doc/ChemmineR.html>.
201. Drug Discovery | Schrödinger. <https://www.schrodinger.com/drug-discovery>.
202. Galizzi, J. P., Lockhart, B. P. & Bril, A. Applying systems biology in drug discovery and development. *Drug Metabol. Drug Interact.* **28**, 67–78 (2013).
203. Trott, O. & Olson, A. J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **31** (2009).
204. Bhandari, B., Rodriguez, B. S. Q. & Masood, W. Ischemic Cardiomyopathy. *Congest. Hear. Fail. Card. Transplant. Clin. Pathol. Imaging Mol. Profiles.* 119–133 (2021).
205. Schuhmacher, A., Gassmann, O. & Hinder, M. Changing R&D models in research-based pharmaceutical companies. *Journal of Translational Medicine.* **14**, 105 (2016).
206. Chen, J. *et al.* Drug discovery and drug marketing with the critical roles of modern administration. *Am. J. Transl. Res.* **10**, 4302–4312 (2018).
207. Kanapeckaitė, A. & Burokienė, N. Insights into therapeutic targets and biomarkers using integrated multi-'omics' approaches for dilated and ischemic cardiomyopathies. *Integr. Biol. (Camb).* **13**, 121–137 (2021).
208. Sidney, S. *et al.* Recent trends in cardiovascular mortality in the United States and public health goals. *JAMA Cardiol.* **1**, 594–599 (2016).
209. Khakoo, A. Y., Yurgin, N. R., Eisenberg, P. R. & Fonarow, G. C. Overcoming Barriers to Development of Novel Therapies for Cardiovascular Disease. *JACC Basic to Transl. Sci.* **4**, 269–274 (2019).
210. Santos-Zas, I., Lemarié, J., Tedgui, A. & Ait-Oufella, H. Adaptive Immune Responses Contribute to Post-ischemic Cardiac Remodeling. *Frontiers in Cardiovascular Medicine.* **5** (2019).
211. Pierpont, M. E. *et al.* Genetic Basis for Congenital Heart Disease: Revisited: A Scientific Statement from the American Heart Association. *Circulation.* **138**, e653–e711 (2018).

212. Zhao, J., Lv, T., Quan, J., Zhao, W., Song, J., Li, Z., Lei, H., Huang, W., & Ran, L. Identification of target genes in cardiomyopathy with fibrosis and cardiac remodeling. *Journal of biomedical science*. **25**(1), 63 (2018). <https://doi.org/10.1186/s12929-018-0459-8>.
213. Pepin, M. E. *et al.* Chromatin and Epigenetics in Cardiovascular Disease: DNA methylation reprograms cardiac metabolic gene expression in end-stage human heart failure. *Am. J. Physiol. - Hear. Circ. Physiol.* **317**, H674 (2019).
214. Rivera-Feliciano, J. & Tabin, C. J. Bmp2 instructs cardiac progenitors to form the heart-valve-inducing field. *Dev. Biol.* **295**, 580–588 (2006).
215. Prados, B. *et al.* Myocardial Bmp2 gain causes ectopic EMT and promotes cardiomyocyte proliferation and immaturity. *Cell Death Dis.* **9**, 1–15 (2018).
216. Hsu, J. *et al.* Genetic Control of Left Atrial Gene Expression Yields Insights into the Genetic Susceptibility for Atrial Fibrillation. *Circ. Genomic Precis. Med.* **11**, e002107 (2018).
217. Arola, A. M. *et al.* Mutations in PDLIM3 and MYOZ1 encoding myocyte Z line proteins are infrequently found in idiopathic dilated cardiomyopathy. *Mol. Genet. Metab.* **90**, 435–440 (2007).
218. Zheng, X., Yang, Y., Huang Fu, C. & Huang, R. Identification and verification of promising diagnostic biomarkers in patients with hypertrophic cardiomyopathy associate with immune cell infiltration characteristics. *Life Sci.* **285** (2021).
219. Zhang, X. *et al.* Identification of New SRF Binding Sites in Genes Modulated by SRF Over-Expression in Mouse Hearts. *Gene Regul. Syst. Bio.* **5**, 41–59 (2011).
220. Bai, Z., Xu, L., Dai, Y., Yuan, Q. & Zhou, Z. ECM2 and GLT8D2 in human pulmonary artery hypertension: fruits from weighted gene co-expression network analysis. *J. Thorac. Dis.* **13**, 2242 (2021).
221. Pirillo, A., Svecla, M., Catapano, A. L., Holleboom, A. G. & Norata, G. D. Impact of protein glycosylation on lipoprotein metabolism and atherosclerosis. *Cardiovasc. Res.* **117**, 1033–1045 (2021).
222. Pletsch-Borba, L. *et al.* Vascular injury biomarkers and stroke risk: A population-based study. *Neurology.* **94**, e2337–e2345 (2020).
223. Lange, S. *et al.* MLP and CARP are linked to chronic PKC $\alpha$  signalling in dilated cardiomyopathy. *Nat. Commun.* **7**, 1–11 (2016).
224. Aspatwar, A., Tolvanen, M. E. E. & Parkkila, S. Phylogeny and expression of carbonic anhydrase-related proteins. *BMC Mol. Biol.* **11**, 1–19 (2010).
225. Alsaleh, A. *et al.* ELOVL2 gene polymorphisms are associated with increases in plasma eicosapentaenoic and docosahexaenoic acid proportions after fish oil supplement. *Genes Nutr.* **9**, (2014).

226. Wu, X. M., Liu, Y., Qian, Z. M., Luo, Q. Q. & Ke, Y. CX3CL1/CX3CR1 Axis Plays a Key Role in Ischemia-Induced Oligodendrocyte Injury via p38MAPK Signaling Pathway. *Mol. Neurobiol.* **53**, 4010–4018 (2016).
227. Elissa Altin, S. & Christian Schulze, P. Fractalkine: A novel cardiac chemokine? *Cardiovascular Research.* **92**, 361–362 (2011).
228. Ahn, J. H. *et al.* Expression changes of CX3CL1 and CX3CR1 proteins in the hippocampal CA1 field of the gerbil following transient global cerebral ischemia. *Int. J. Mol. Med.* **44**, 939–948 (2019).
229. Barth, A. S. *et al.* Identification of a Common Gene Expression Signature in Dilated Cardiomyopathy Across Independent Microarray Studies. *J. Am. Coll. Cardiol.* **48**, 1610–1617 (2006).
230. Katsuhito, F. & Ryozo, N. Contributions of cardiomyocyte-cardiac fibroblast-immune cell interactions in heart failure development. *Basic Res. Cardiol.* **108**, (2013).
231. Liu, H., Lin, Z. & Ma, Y. Suppression of Fpr2 expression protects against endotoxin-induced acute lung injury by interacting with Nrf2-regulated TAK1 activation. *Biomed. Pharmacother.* **125**, 109943 (2020).
232. Aran, D. *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol.* **20**, 163–172 (2019).
233. Kuo, P. T. *et al.* The role of CXCR3 and its chemokine ligands in skin disease and cancer. *Frontiers in Medicine.* **5**, 271 (2018).
234. Groom, J. R. & Luster, A. D. CXCR3 in T cell function. *Experimental Cell Research.* **317**, 620–631 (2011).
235. Tharp, C. A., Haywood, M. E., Sbaizero, O., Taylor, M. R. G. & Mestroni, L. The Giant Protein Titin's Role in Cardiomyopathy: Genetic, Transcriptional, and Post-translational Modifications of TTN and Their Contribution to Cardiac Disease. *Front. Physiol.* **10**, 1436 (2019).
236. Nielsen, L. B., Bartels, E. D. & Bollano, E. Overexpression of Apolipoprotein B in the Heart Impedes Cardiac Triglyceride Accumulation and Development of Cardiac Dysfunction in Diabetic Mice. *J. Biol. Chem.* **277**, 27014–27020 (2002).
237. Su, Q. *et al.* Apolipoprotein B100 acts as a molecular link between lipid-induced endoplasmic reticulum stress and hepatic insulin resistance. *Hepatology.* **50**, 77–84 (2009).
238. Bartels, E. D., Nielsen, J. M., Hellgren, L. I., Ploug, T. & Nielsen, L. B. Cardiac Expression of Microsomal Triglyceride Transfer Protein Is Increased in Obesity and Serves to Attenuate Cardiac Triglyceride Accumulation. *PLoS One.* **4**, e5300 (2009).
239. Wang, S. *et al.* IRE1 $\alpha$ -XBP1s Induces PDI Expression to Increase MTP Activity for Hepatic VLDL Assembly and Lipid Homeostasis. *Cell Metab.* **16**, 473–486 (2012).

240. Tsai, J. *et al.* Inflammatory NF-kappaB activation promotes hepatic apolipoprotein B100 secretion: evidence for a link between hepatic inflammation and lipoprotein production. *Am. J. Physiol. Gastrointest. Liver Physiol.* **296**, G1287-98 (2009).
241. Warde-Farley, D. *et al.* The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.* **38**, W214–W220 (2010).
242. Li, M. J. *et al.* GWASdb: A database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res.* **40** (2012).
243. Codreanu, S. G. & Liebler, D. C. Novel approaches to identify protein adducts produced by lipid peroxidation. *Free Radic. Res.* **49**, 881–7 (2015).
244. Chan, H. C. S., Shan, H., Dahoun, T., Vogel, H. & Yuan, S. Advancing Drug Discovery via Artificial Intelligence. *Trends in Pharmacological Sciences.* **40**, 592–604 (2019).
245. Zin, P. P. K., Williams, G. J. & Ekins, S. Cheminformatics Analysis and Modeling with MacrolactoneDB. *Sci. Rep.* **10** (2020).
246. de Souza Neto, L. R. *et al.* In silico Strategies to Support Fragment-to-Lead Optimization in Drug Discovery. *Frontiers in Chemistry.* **8** (2020).
247. Tsaïoun, K., Bottlaender, M. & Mabondzo, A. ADDME - Avoiding Drug Development Mistakes Early: Central nervous system drug discovery perspective. in *BMC Neurology.* **9** (BioMed Central, 2009).
248. Aminpour, M., Montemagno, C. & Tuszynski, J. A. An overview of molecular modeling for drug discovery with specific illustrative examples of applications. *Molecules.* **24** (2019).
249. Jamkhande, P. G., Ghante, M. H. & Ajgunde, B. R. Software based approaches for drug designing and development: A systematic review on commonly used software and its applications. *Bull. Fac. Pharmacy, Cairo Univ.* **55**, 203–210 (2017).
250. Weisel, M., Proschak, E., Kriegl, J. M. & Schneider, G. Form follows function: Shape analysis of protein cavities for receptor-based drug design. *Proteomics* **9**, 451–459 (2009).
251. Fauman, E. B., Rai, B. K. & Huang, E. S. Structure-based druggability assessment-identifying suitable targets for small molecule therapeutics. *Current Opinion in Chemical Biology.* **15**, 463–468 (2011).
252. Guo, Z. *et al.* Identification of protein-ligand binding sites by the level-set variational implicit-solvent approach. *J. Chem. Theory Comput.* **11**, 753–765 (2015).
253. Yang, J., Wang, Y. & Zhang, Y. ResQ: An Approach to Unified Estimation of B-Factor and Residue-Specific Error in Protein Structure Prediction. *J. Mol. Biol.* **428**, 693–701 (2016).
254. Huang, X. & Dixit, V. M. Drugging the undruggables: exploring the ubiquitin system for drug development. *Cell Res.* **26**, 484–498 (2016).
255. Dias, R. & de Azevedo Jr., W. Molecular Docking Algorithms. *Curr. Drug Targets.* **9**, 1040–1047 (2008).

256. Hartmann, H. *et al.* Conformational substates in a protein: structure and dynamics of metmyoglobin at 80 K. *Proc. Natl. Acad. Sci. U. S. A.* **79**, 4967–4971 (1982).
257. Li, X. *et al.* Structural studies unravel the active conformation of apo ROR $\gamma$ t nuclear receptor and a common inverse agonism of two diverse classes of ROR $\gamma$ t inhibitors. *J. Biol. Chem.* **292**, 11618–11630 (2017).
258. Powers, R., Clore, G. M., Garrett, D. S. & Gronenborn, A. M. Relationships Between the Precision of High-Resolution Protein NMR Structures, Solution-Order Parameters, and Crystallographic B Factors. *Journal of Magnetic Resonance, Series B.* **101**, 325–327 (1993).
259. Bryn Fenwick, R., Van Den Bedem, H., Fraser, J. S. & Wright, P. E. Integrated description of protein dynamics from room-temperature X-ray crystallography and NMR. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E445–E454 (2014).
260. Saravanan, K. M. & Selvaraj, S. Dihedral angle preferences of amino acid residues forming various non-local interactions in proteins. *J. Biol. Phys.* **43**, 265 (2017).
261. Schlessinger, A. & Rost, B. Protein flexibility and rigidity predicted from sequence. *Proteins Struct. Funct. Genet.* **61**, 115–126 (2005).
262. Liu, Q., Li, Z. & Li, J. Use B-factor related features for accurate classification between protein binding interfaces and crystal packing contacts. *BMC Bioinformatics.* **15** (2014).
263. Vihinen, M., Torkkila, E. & Riikonen, P. Accuracy of protein flexibility predictions. *Proteins Struct. Funct. Bioinforma.* **19**, 141–149 (1994).
264. Smith, D. K., Radivojac, P., Obradovic, Z., Dunker, A. K. & Zhu, G. Improved amino acid flexibility parameters. *Protein Sci.* **12**, 1060–1072 (2003).
265. Kuczera, K., Kuriyan, J. & Karplus, M. Temperature dependence of the structure and dynamics of myoglobin. A simulation approach. *J. Mol. Biol.* **213**, 351–373 (1990).
266. De Juan, D., Pazos, F. & Valencia, A. Emerging methods in protein co-evolution. *Nature Reviews Genetics.* **14**, 249–261 (2013).
267. Fscore: Effective Protein Structural Data Visualisation and Exploration version 0.1.3 from CRAN. <https://rdr.io/cran/Fscore/>.
268. Siglioccolo, A., Gerace, R. & Pascarella, S. Cold spots in protein cold adaptation: Insights from normalized atomic displacement parameters (B'-factors). *Biophys. Chem.* **153**, 104–114 (2010).
269. Tang, H. *et al.* Enhancing subtilisin thermostability through a modified normalized B-factor analysis and loop-grafting strategy. *J. Biol. Chem.* **294**, 18398–18407 (2019).
270. Simoens, S. & Huys, I. R&D Costs of New Medicines: A Landscape Analysis. *Front. Med.* **8**, 1891 (2021).
271. Pushpakom, S. *et al.* Drug repurposing: progress, challenges and recommendations. *Nat. Rev. Drug Discov.* **18**, 41–58 (2018).

272. Dolgos, H. *et al.* Translational Medicine Guide transforms drug development processes: The recent Merck experience. *Drug Discovery Today*. **21**, 517–526 (2016).
273. Pazos-López, P. *et al.* The causes, consequences, and treatment of left or right heart failure. *Vascular Health and Risk Management*. **7**, 237–254 (2011).
274. Hunter, J. E., Leslie, J. & Perkins, N. D. c-Rel and its many roles in cancer: an old story with new twists. *Br. J. Cancer*. **114**, 1–6 (2016).
275. Berkowitz, B., Huang, D. Bin, Chen-Park, F. E., Sigler, P. B. & Ghosh, G. The X-ray crystal structure of the NF- $\kappa$ B p50·p65 heterodimer bound to the interferon  $\beta$ - $\kappa$ B site. *J. Biol. Chem.* **277**, 24694–24700 (2002).
276. Shono, Y. *et al.* Characterization of a c-Rel inhibitor that mediates anticancer properties in hematologic malignancies by blocking NF- $\kappa$ B-controlled oxidative stress responses. *Cancer Res.* **76**, 377–389 (2016).
277. Chen, Y. & Shoichet, B. K. Molecular docking and ligand specificity in fragment-based inhibitor discovery. *Nat. Chem. Biol.* **5**, 358–364 (2009).
278. Kanapeckaitė, A., Beaurivage, C., Jančorienė, L. & Mažeikienė, A. In silico drug discovery for a complex immunotherapeutic target - human c-Rel protein. *Biophys. Chem.* **276**, (2021).
279. Choudhury, C. Fragment tailoring strategy to design novel chemical entities as potential binders of novel corona virus main protease. *J. Biomol. Struct. Dyn.* **1** (2020) doi:10.1080/07391102.2020.1771424.
280. Loving, K., Salam, N. K. & Sherman, W. Energetic analysis of fragment docking and application to structure-based pharmacophore hypothesis generation. *J. Comput. Aided. Mol. Des.* **23**, 541–554 (2009).
281. Hodos, R. A., Kidd, B. A., Shameer, K., Readhead, B. P. & Dudley, J. T. In silico methods for drug repurposing and pharmacology. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **8**, 186–210 (2016).
282. RCSB PDB: Homepage. <https://www.rcsb.org/>.
283. Dessau, M. A. & Modis, Y. Protein crystallization for X-ray crystallography. *J. Vis. Exp.* (2011) doi:10.3791/2285.
284. Du, J. *et al.* New techniques and strategies in drug discovery. *Chinese Chem. Lett.* **31**, 1695–1708 (2020).
285. Cirauqui Diaz, N., Frezza, E., & Martin, J. Using normal mode analysis on protein structural models. How far can we go on our predictions?. *Proteins*. **89**(5), 531–543 (2021). <https://doi.org/10.1002/prot.26037>.
286. Maia, E., Assis, L. C., de Oliveira, T. A., da Silva, A. M., & Taranto, A. G. Structure-Based Virtual Screening: From Classical to Artificial Intelligence. *Frontiers in chemistry*. **8**, 343 (2020). <https://doi.org/10.3389/fchem.2020.00343>.

287. Karin, M. NF-kappaB as a critical link between inflammation and cancer. *Cold Spring Harb. Perspect. Biol.* **1**, a000141 (2009).
288. Gilmore, T. D. Nuclear Factor Kappa B. in *Encyclopedia of Biological Chemistry: Second Edition* 302–305 (Elsevier Inc., 2013). doi:10.1016/B978-0-12-378630-2.00335-2.
289. Hinz, M., Arslan, S. Ç. & Scheidereit, C. It takes two to tango: I $\kappa$ Bs, the multifunctional partners of NF- $\kappa$ B. *Immunol. Rev.* **246**, 59–76 (2012).
290. Meriño-Cabrera, Y. *et al.* Rational design of mimetic peptides based on the interaction between Inga laurina inhibitor and trypsins for Spodoptera cosmioides pest control. *Insect Biochem. Mol. Biol.* **122**, (2020).
291. Natoli, G., Saccani, S., Bosisio, D. & Marazzi, I. Interactions of NF-kappaB with chromatin: the art of being at the right place at the right time. *Nat. Immunol.* **6**, 439–45 (2005).
292. Schneider, G. Automating drug discovery. *Nat. Rev. Drug Discov.* **17**, 97–113 (2017).
293. Subramanian, G., Mjalli, A. M., & Kutz, M. E. Integrated approaches to perform in silico drug discovery. *Current drug discovery technologies.* **3**(3), 189–197 (2006). <https://doi.org/10.2174/157016306780136790>.
294. Chichester, C. *et al.* Drug discovery FAQs: workflows for answering multidomain drug discovery questions. *Drug Discov. Today.* **20**, 399–405 (2015).
295. UniProt. <https://www.uniprot.org/>.
296. Scrucca, L., Fop, M., Murphy, T. B. & Raftery, A. E. mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *R J.* **8**, 289 (2016).
297. Kuhlman, B., & Bradley, P. Advances in protein structure prediction and design. *Nature reviews. Molecular cell biology.* **20**(11), 681–697 (2019). <https://doi.org/10.1038/s41580-019-0163-x>.
298. Alam, K. K., Chang, J. L. & Burke, D. H. FASTAptamer: A bioinformatic toolkit for high-throughput sequence analysis of combinatorial selections. *Mol. Ther. - Nucleic Acids.* **4**, e230 (2015).
299. Tanjo, T., Kawai, Y., Tokunaga, K., Ogasawara, O. & Nagasaki, M. Practical guide for managing large-scale human genome data in research. *Journal of Human Genetics.* **66**, 39–52 (2021).
300. Kooistra, A. J. *et al.* 3D-e-Chem: Structural Cheminformatics Workflows for Computer-Aided Drug Discovery. *ChemMedChem.* **13**, 614–626 (2018).
301. Kanapeckaitė, A. Fiscore Package: Effective Protein Structural Data Visualisation and Exploration. <https://github.com/AusteKan/Fiscore>. *GitHub repository* (2021).
302. Kanapeckaitė, A. Fiscore Package: Effective Protein Structural Data Visualisation and Exploration. *bioRxiv* 2021.08.25.457640 (2021) doi:10.1101/2021.08.25.457640.



303. AusteKan/OmicInt: OmicInt Package. <https://github.com/AusteKan/OmicInt>. *GitHub repository* (2021).
304. CRAN - Package OmicInt. <https://cran.r-project.org/web/packages/OmicInt/index.html>.
305. GitHub: Where the world builds software · GitHub. <https://github.com/>.
306. R: The R Project for Statistical Computing. <https://www.r-project.org/>.
307. PyPI · The Python Package Index. <https://pypi.org/>.
308. Vamathevan, J. *et al.* Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery*. **18**, 463–477 (2019).

## **Integrative *omics* approaches for new target identification and therapeutics development**

### **10. Supplementary materials**

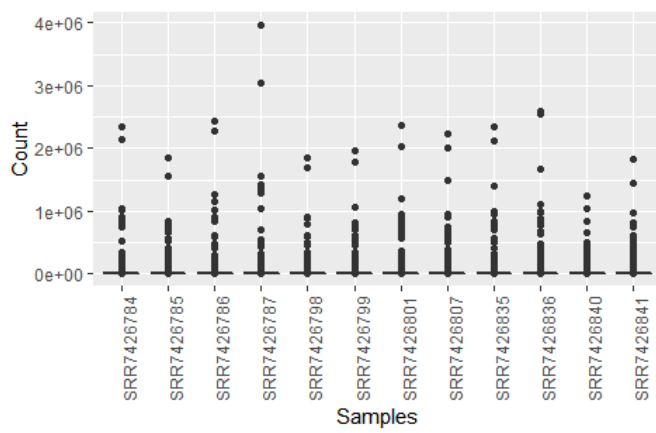
## **Integrative *omics* approaches for new target identification and therapeutics development**

### **10.1. Insights into therapeutic targets and biomarkers using integrated multi-‘*omics*’ approaches for dilated and ischemic cardiomyopathies**

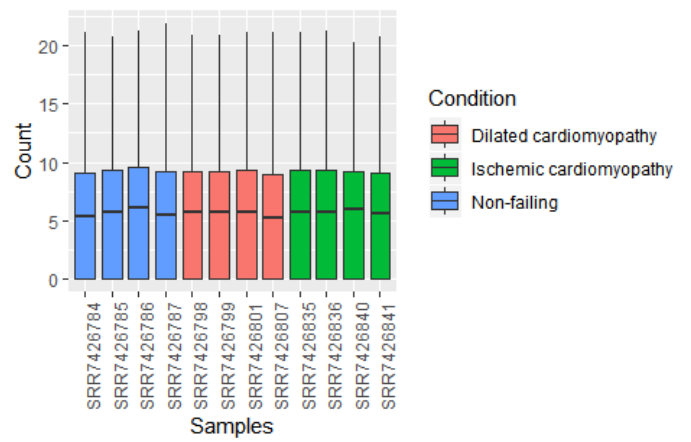
## Supplementary Figures and Data

$$\text{LFC}_{\text{Score}} = \text{LFC}(1 + \alpha)$$

**Equation 1.** Log2 Fold Change<sub>Score</sub> equation defines a scaled LFC (log2 fold change) value for a given contrast where  $\alpha$  is a value showing the strength of disease association for a given gene (the  $\alpha$  value is retrieved from Open Targets disease association scoring).

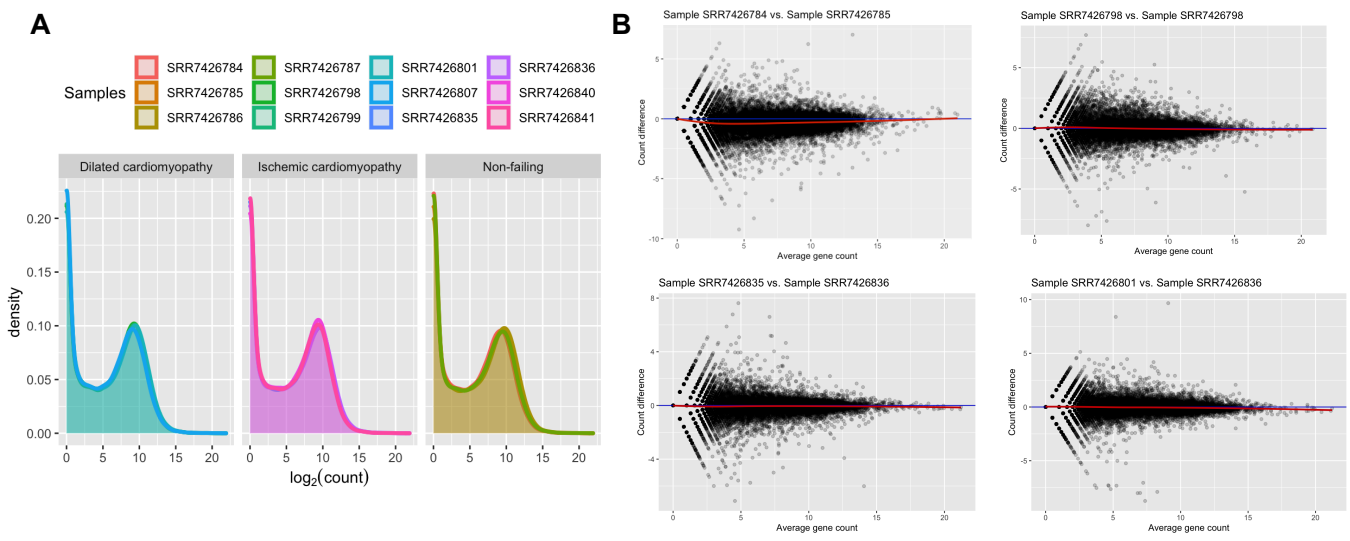


**A**

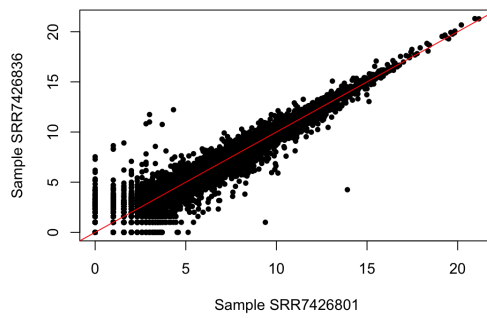
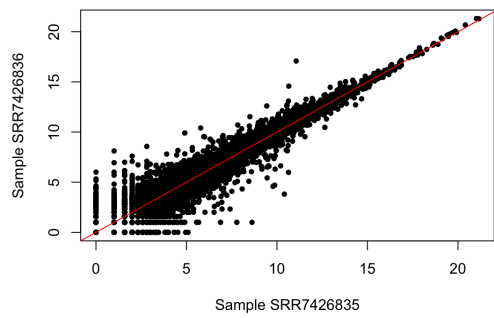
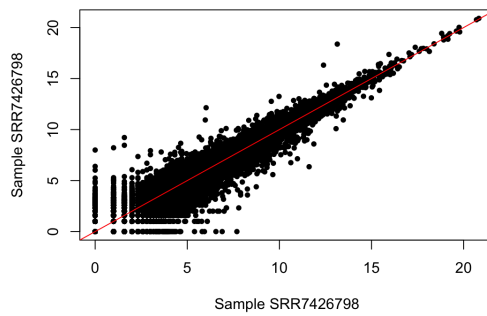
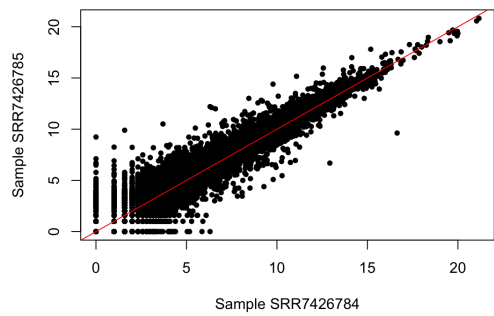


**B**

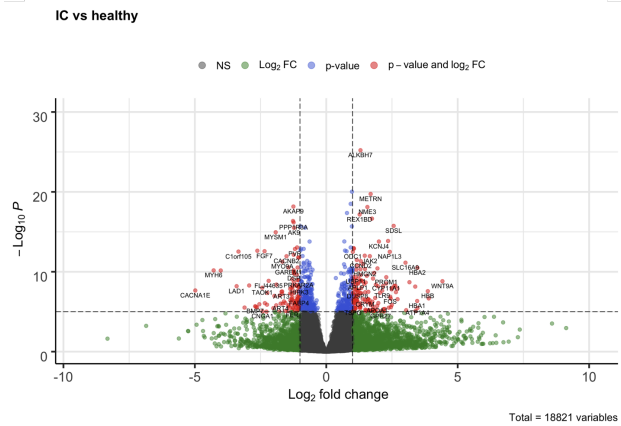
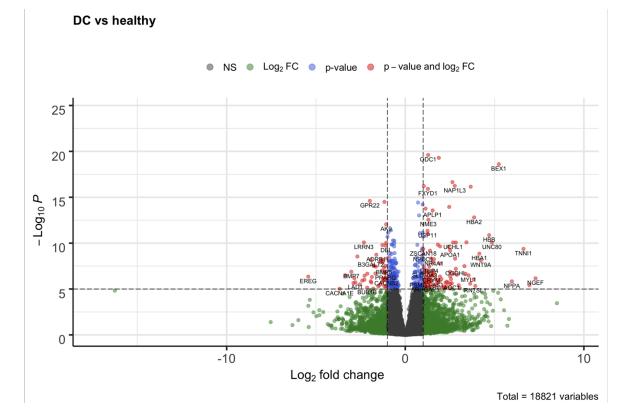
**Supplementary Figure 1.** Raw (A) and log<sub>2</sub>+1 normalised (B) sample count distributions for human left ventricle bulk RNA-seq (PRJNA477855).



**Supplementary Figure 2.** Density distribution of  $\log_2+1$  transformed sample counts (A) and MA plots for  $\log_2+1$  transformed counts (B) where the red line indicates average differences

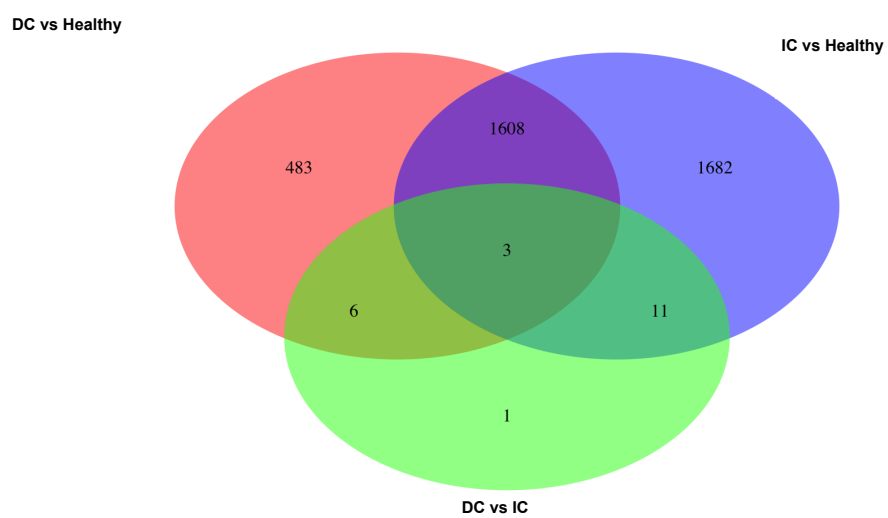


**Supplementary Figure 3.** Log<sub>2</sub>+1 transformed counts for samples plotted against each other to evaluate count distribution and sequencing depth. Representative combinations are shown.

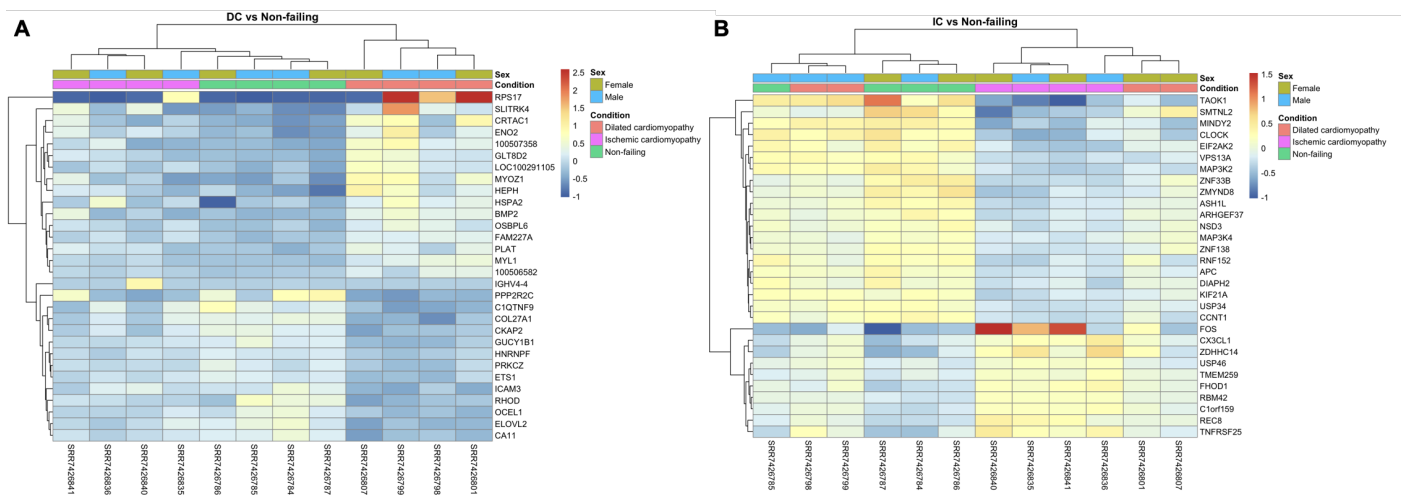


**Supplementary Figure 4.** Human left ventricle bulk RNA-seq (PRJNA477855) significantly changed gene count Volcano plots where FDR p-adjusted<0.00001 and LFC>|2|.



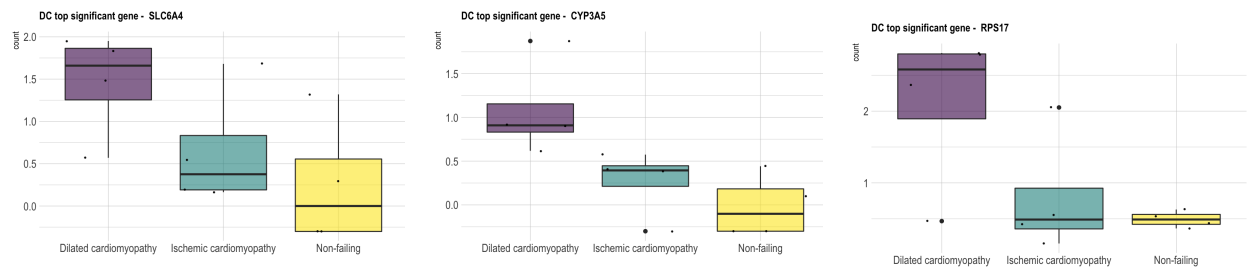


**Supplementary Figure 5.** Venn diagram for significantly changed genes when comparing changed genes across different contrast groups: DC vs Healthy, IC vs Healthy and DC vs IC.

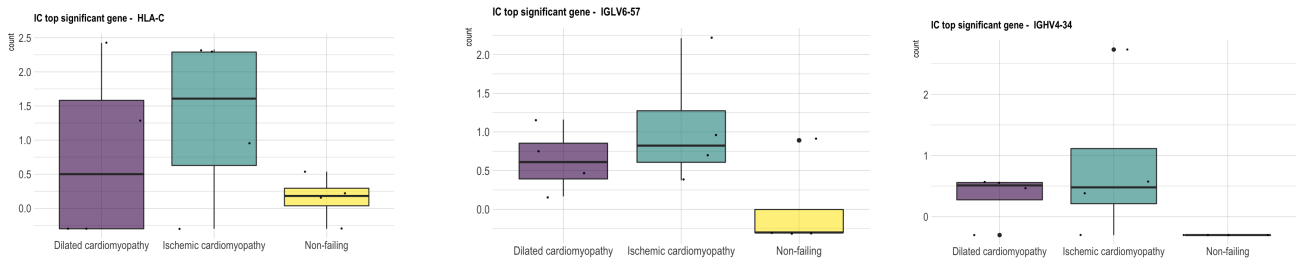


**Supplementary Figure 6.** Heatmap for significantly changed genes (ranking the top genes based on the p-adjusted value) that are unique for the contrasts: DC vs healthy samples (A) and IC vs healthy samples (B) where values are shown for all conditions. Reported counts are rlog transformed and mean standardised per gene.

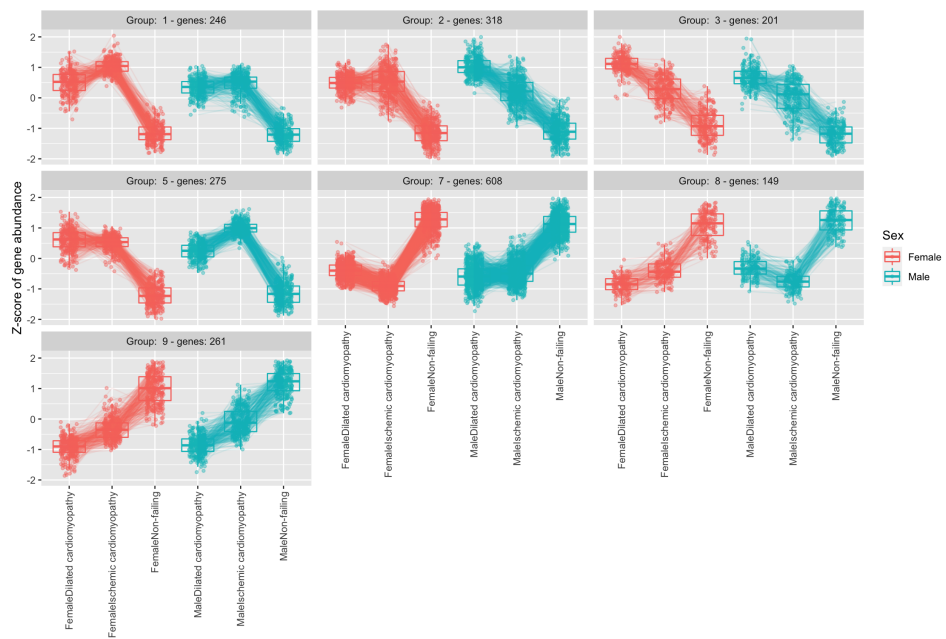
Genes that are uniquely and significantly changed in DC vs healthy



Genes that are uniquely and significantly changed in IC vs healthy



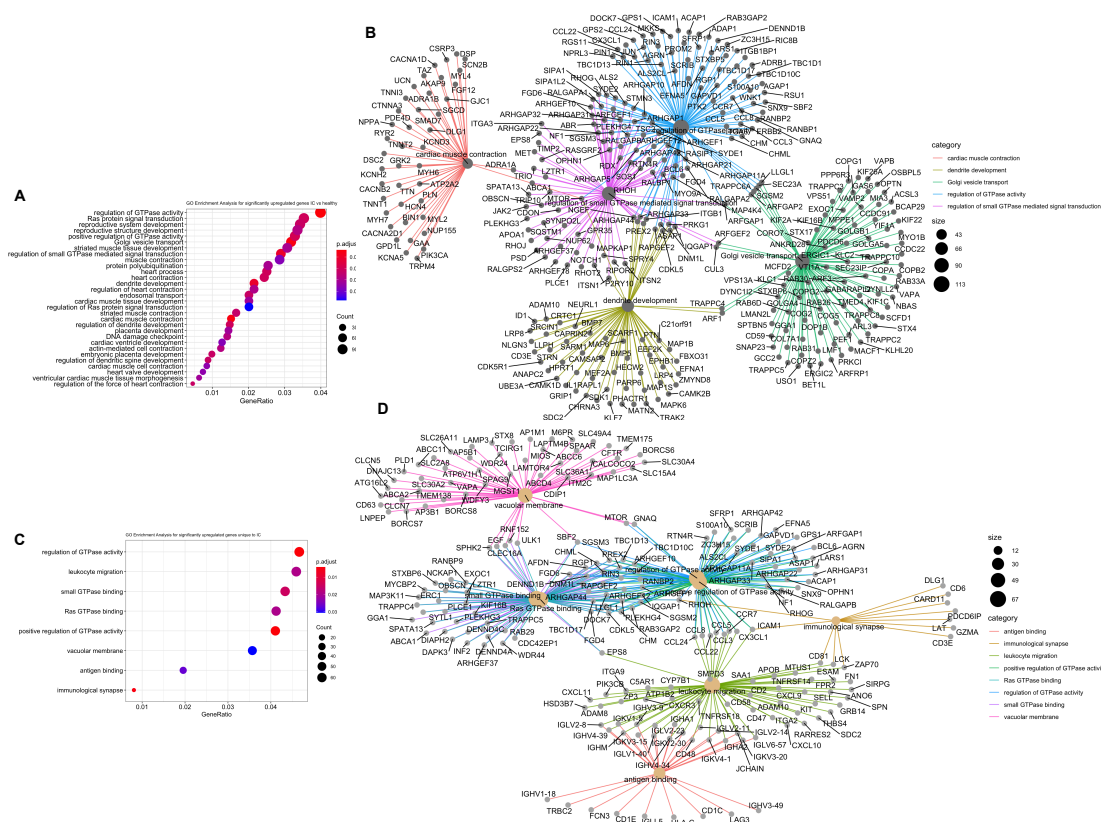
**Supplementary Figure 7.** Log10 scaled gene counts that changed significantly in a specific contrast groups: DC vs Healthy (top panels), IC vs Healthy (bottom panels) and belonged to the largest log fold change group.



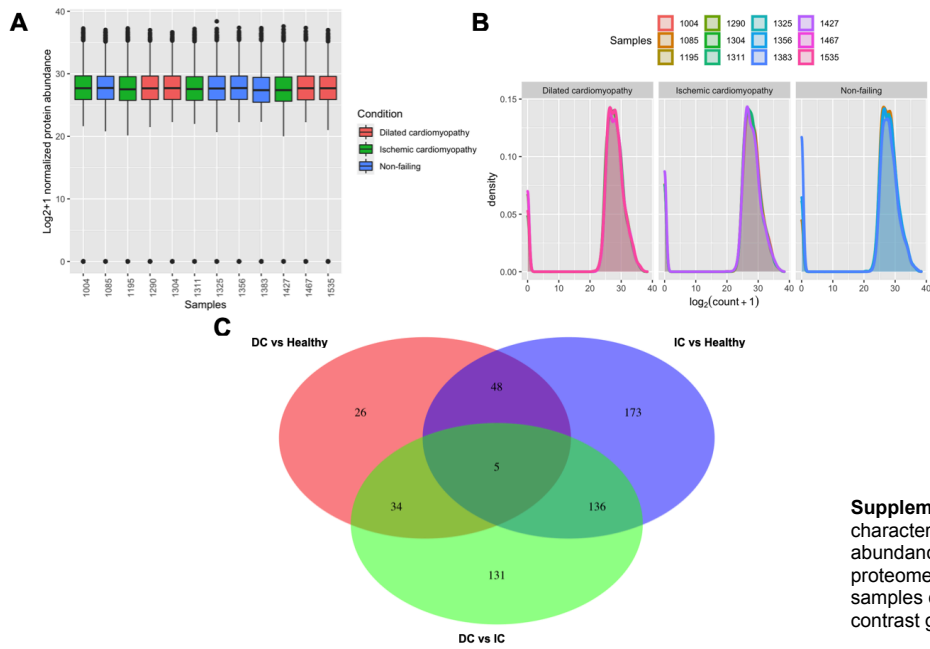
**Supplementary Figure 8.** Scaled gene counts that changed significantly in DC vs Healthy samples and where expression patterns formed distinct expression clusters across different human heart tissue states when comparing for both genders.



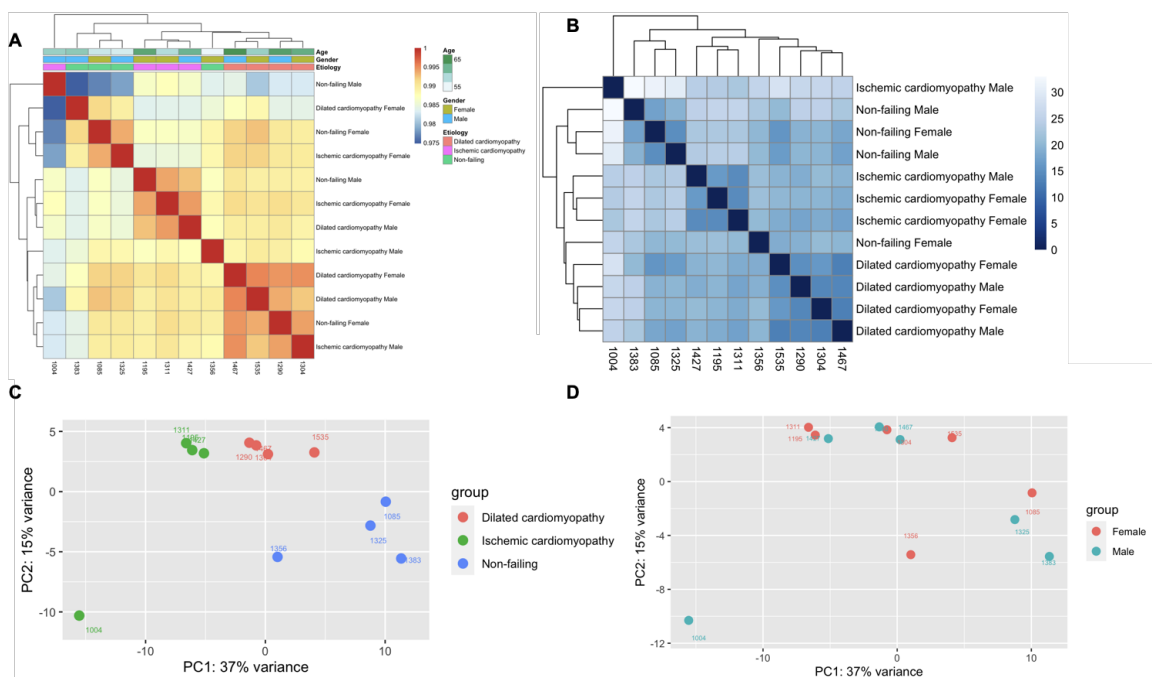
**Supplementary Figure 9.** Scaled gene counts that changed significantly in IC vs Healthy samples and where expression patterns formed distinct clusters across different human heart tissue states when comparing for both genders.



**Supplementary Figure 10.** Enrichment analysis for all significantly changed genes in the IC vs Healthy contrast group where enriched cellular processes (A) and the visualisation of the top highest ranking processes and corresponding genes (B) are provided in the distribution plots and network maps, respectively. Enrichment analysis for genes that changed significantly in IC vs Healthy but not in DC vs Healthy are plotted as cellular processes distribution (C) and the visualisation of the top highest ranking processes and corresponding genes are shown in network maps (D). Gene set size that was enriched and p-adjusted value provided with the plots.



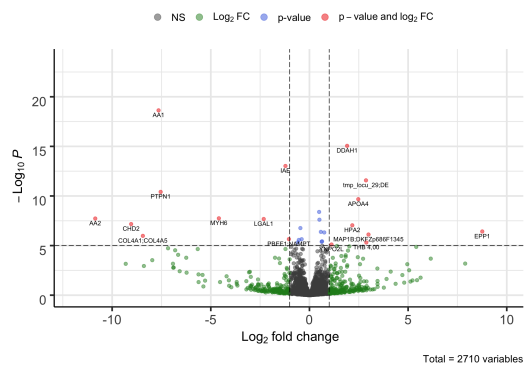
**Supplementary Figure 11.** Protein sample main characteristics: normalised ( $\log_2+1$  transformed) protein abundance/count (LFQ) values for human left ventricle proteome (PXD008934) (A), density plots of protein samples distribution (B) and shared proteins by different contrast groups are visualised via Venn diagram (C).



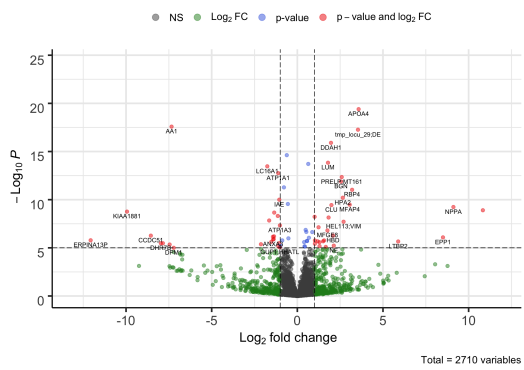
**Supplementary Fig. 12.** Human left ventricle bulk proteome (PXD008934) abundance clustering and distribution analysis showing Spearman correlation calculated distances (A) and euclidean distances (B) for  $\log$  transformed abundance values (LFQ) using complete-linkage hierarchical clustering method; sample distributions across top two principal components are shown in the PCA plot grouping by condition (C) and gender (D).



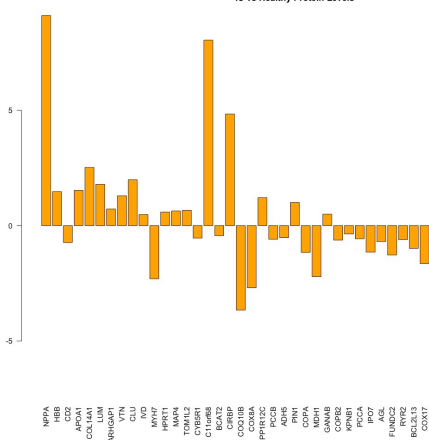
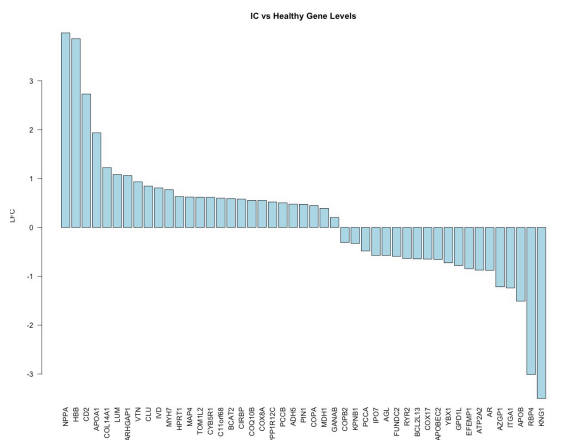
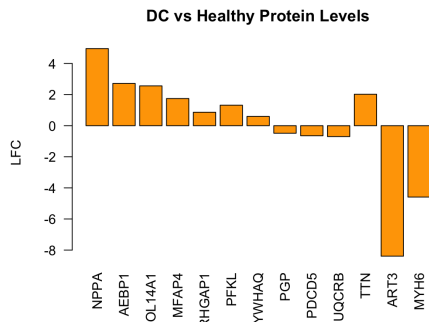
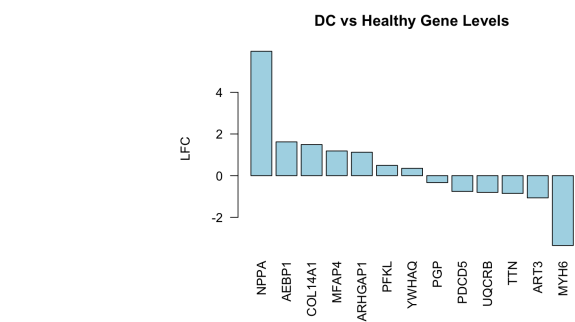
DC vs healthy



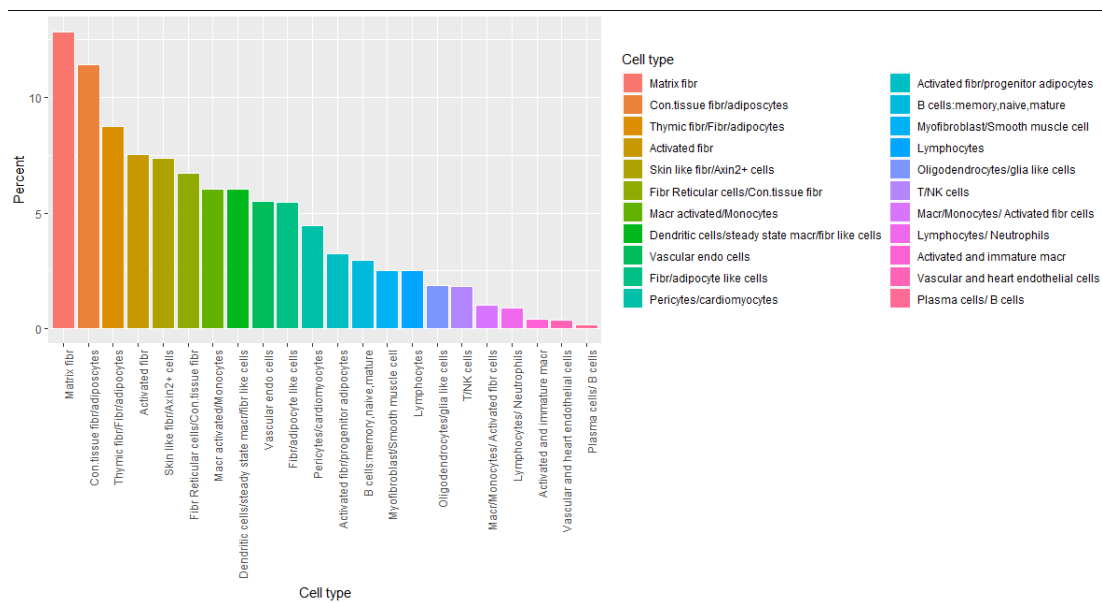
IC vs healthy



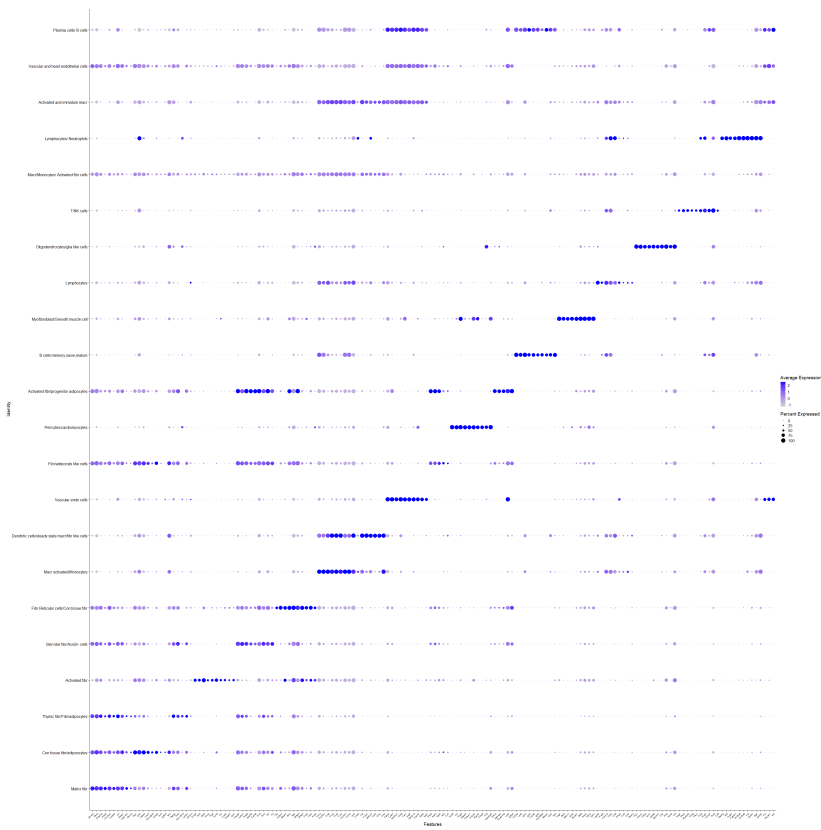
**Supplementary Figure 13.** Volcano plot for human left ventricle proteins (PXD008934) that changed significantly in specific contrasts, FDR p-adjusted < 0.00001 and Log fold change (LFC) > |2|.



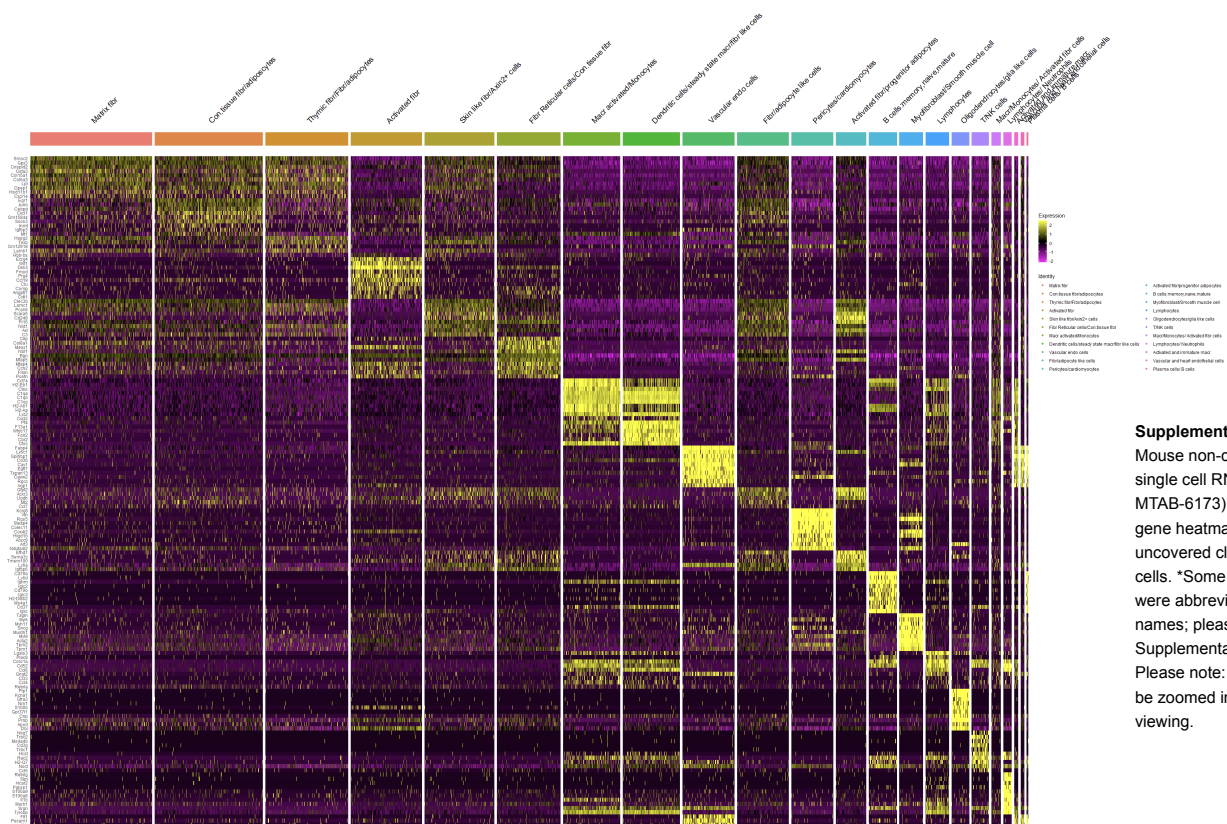
**Supplementary Figure 14.** Gene and protein LFC values for different contrasts where gene LFC values were generated from bulk RNA-seq (PRJNA477855) and protein LFCs were derived from proteome analysis (PXD008934).



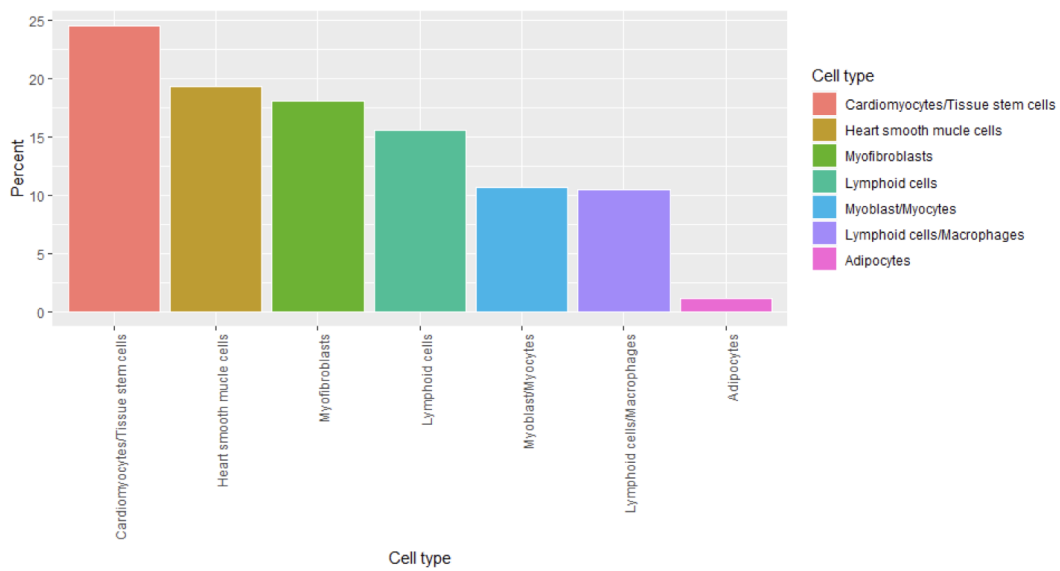
**Supplementary Figure 15.** Mouse non-cardiomyocyte single cell RNA-seq (E-MTAB-6173) cellome composition. \*Some longer names were abbreviated; for full names, please refer to Supplementary Table 9.



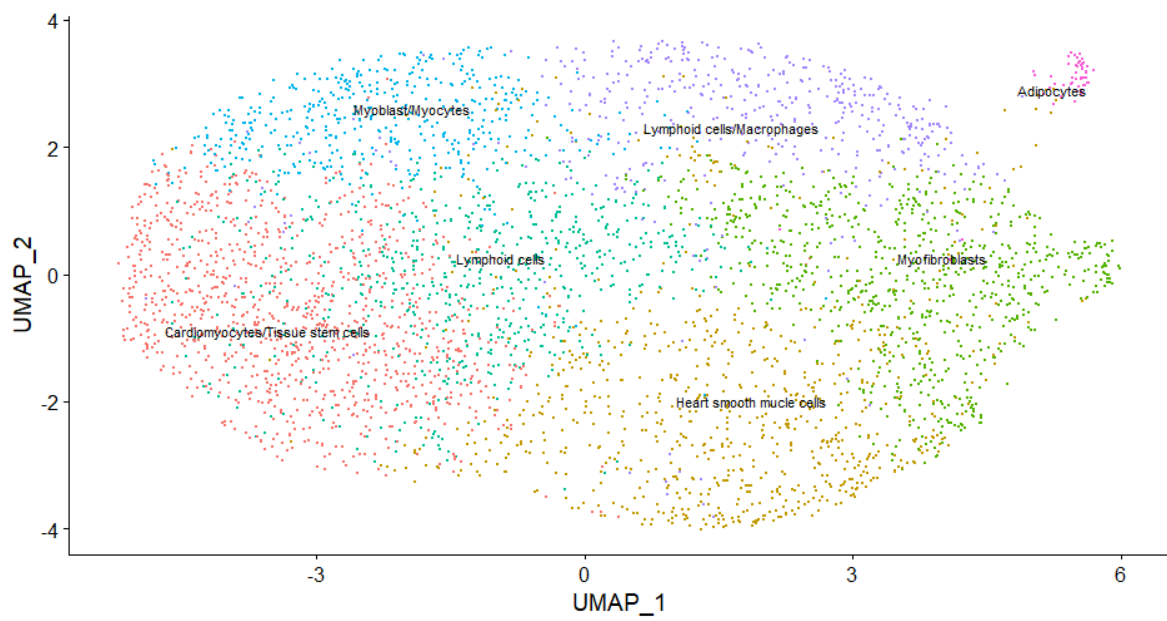
**Supplementary Figure 16.** Mouse non-cardiomyocyte single cell RNA-seq (E-MTAB-6173) cellome marker gene clusters for the uncovered cell types. Please note: image needs to be zoomed in for proper viewing.



**Supplementary Figure 17.**  
 Mouse non-cardiomyocyte single cell RNA-seq (E-MTAB-6173) celluome marker gene heatmap for the uncovered clusters of different cells. \*Some longer names were abbreviated, for full names; please refer to Supplementary Table 9. Please note: image needs to be zoomed in for proper viewing.



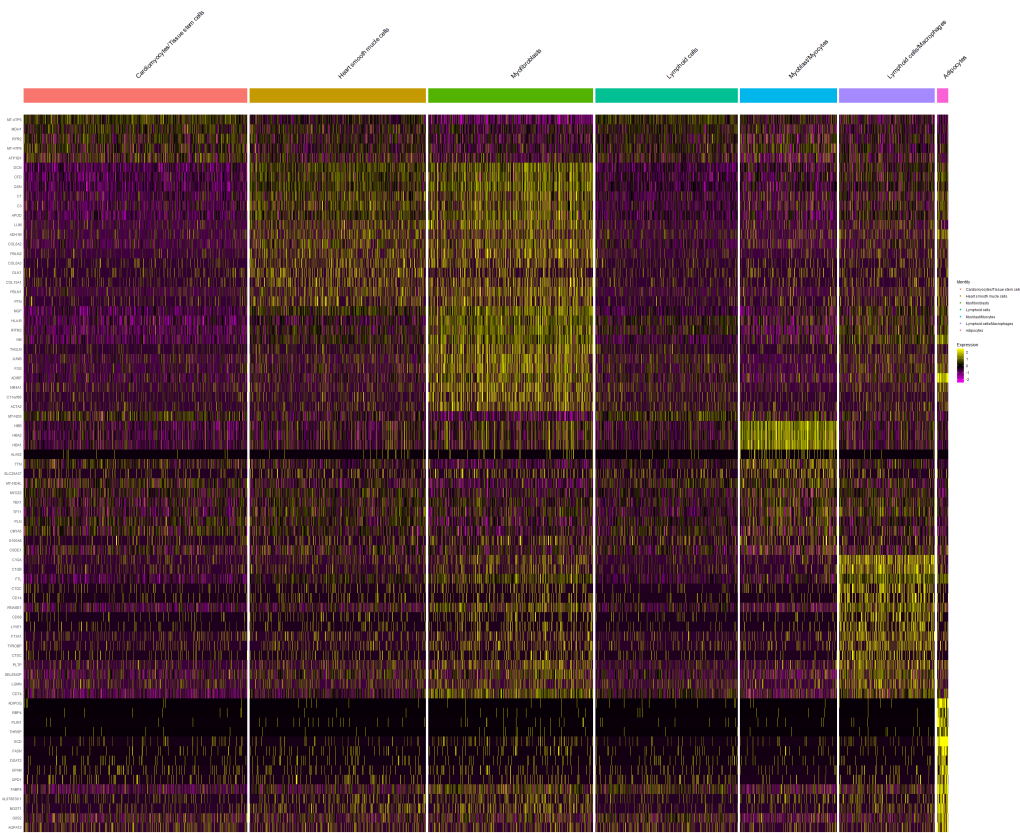
**Supplementary Figure 18.** Human heart left ventricle cellome composition.



**Supplementary Figure 19.** Human heart left ventricle cellulome UMAP decomposition showing relative distances and the uncovered clusters of different cells.

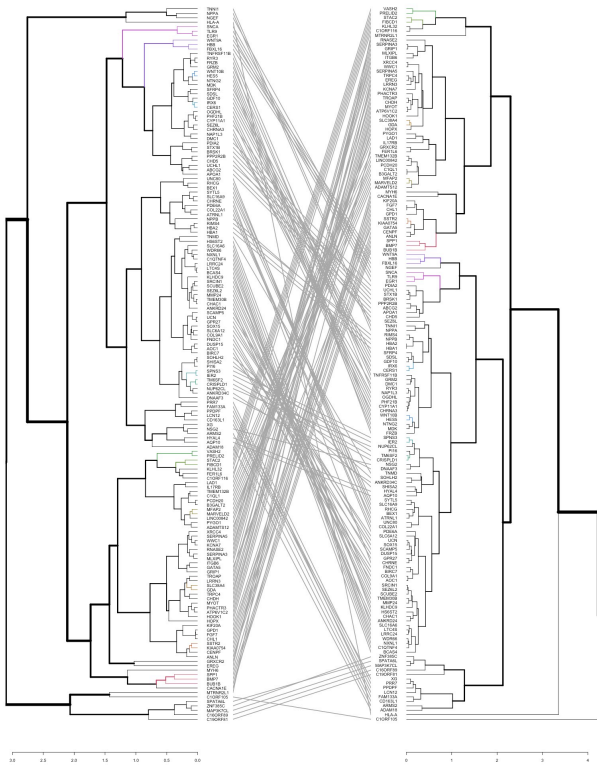




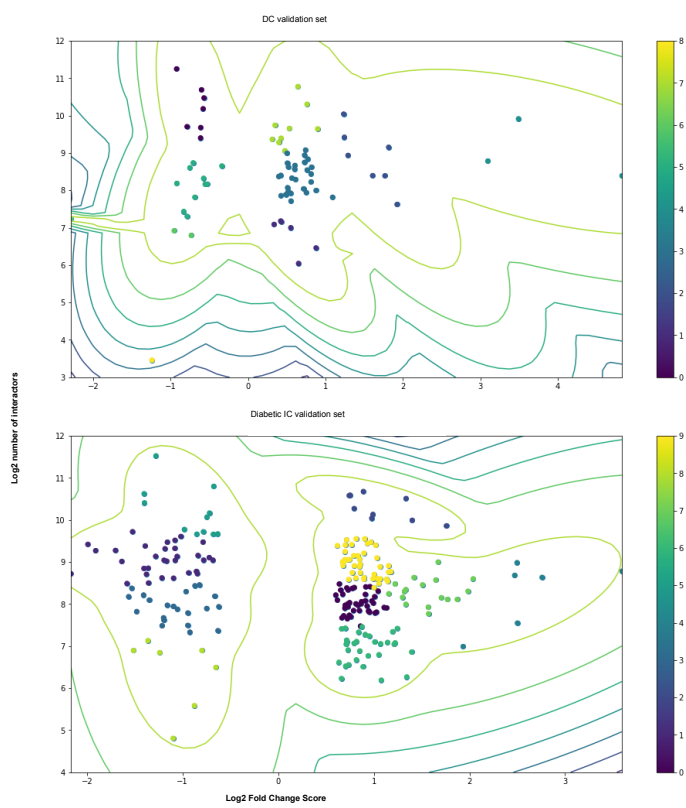


**Supplementary Figure 21.** Human heart left ventricle cellulome marker gene heatmap for the uncovered clusters of different cells. Please note: image needs to be zoomed in for proper viewing.

entanglement = 0.84



**Supplementary Figure 22.** Human heart left ventricle bulk RNA-seq shared significantly changed gene (n=160) set between dilated and ischemic cardiomyopathy contrast groups (disease vs healthy state) agglomerative hierarchical clustering based on Log2 Fold Change Score and known interactors returned the shared dendrogram. Coloured branches signify similar clustering patterns. Please note: image needs to be zoomed in for proper viewing.



**Figure 23.** GMM clustering showing specific grouping based on Log2 Fold Change Score against known or predicted number of interactions for that gene where significantly changed genes in biopsies of dilated heart (GEO: GSE3585) (DC validation) as well as diabetic heart failure samples (GEO: GSE26887) (Diabetic IC validation) were used for clustering. DC (dilated cardiomyopathy); IC (ischemic cardiomyopathy; ischemic tissue pathology).

**Supplementary Table 1. Randomly selected samples of PRJNA477855 RNA-seq for the heart failure in human left ventricles**

<b>Sample ID</b>	<b>Sex</b>	<b>Age</b>	<b>Condition</b>
SRR7426784	Male	43	Non-failing
SRR7426785	Male	54	Non-failing
SRR7426786	Female	41	Non-failing
SRR7426787	Female	56	Non-failing
SRR7426798	Male	38	Dilated cardiomyopathy
SRR7426799	Male	66	Dilated cardiomyopathy
SRR7426801	Female	66	Dilated cardiomyopathy
SRR7426807	Female	51	Dilated cardiomyopathy
SRR7426835	Male	63	Ischemic cardiomyopathy
SRR7426836	Male	56	Ischemic cardiomyopathy
SRR7426840	Female	57	Ischemic cardiomyopathy
SRR7426841	Female	60	Ischemic cardiomyopathy

**Supplementary Table 2. Randomly selected proteome samples (PXD008934 ) of the heart failure in human left ventricles**

Sample ID	Sex	Age	Condition
1085	Female	54	Non-failing
1356	Female	51	Non-failing
1383	Male	59	Non-failing
1325	Male	53	Non-failing
1535	Female	58	Dilated cardiomyopathy
1304	Female	63	Dilated cardiomyopathy
1467	Male	67	Dilated cardiomyopathy
1290	Male	65	Dilated cardiomyopathy
1311	Female	56	Ischemic cardiomyopathy
1195	Female	64	Ischemic cardiomyopathy
1427	Male	62	Ischemic cardiomyopathy
1004	Male	58	Ischemic cardiomyopathy

**Supplementary Table 3. DC vs Healthy significantly and uniquely expressed genes that cluster into functional processes**

Gene	Function	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
MYL1	myosin light chain 1	14.364641	3.5442291	0.6937477	5.108816	3.241839e-07	3.309117e-05
DNAH6	dynein axonemal heavy chain 6	8.166608	2.6515123	0.8483073	3.125651	1.774122e-03	2.042312e-02
MYOZ1	myozenin 1	67.628563	2.5242075	0.5377578	4.693949	2.679814e-06	1.651131e-04
ACKR2	atypical chemokine receptor 2	23.264701	2.1723989	0.6590295	3.298361	9.794617e-04	1.366374e-02
CPEB1	cytoplasmic polyadenylation element binding protein 1	20.604908	1.6807667	0.4578147	3.671282	2.413373e-04	4.872401e-03
SPOCK1	SPARC (osteonectin), cwcv and kazal like domains proteoglycan 1	2038.957418	1.4382561	0.4926295	2.919549	3.505379e-03	3.280854e-02
CCDC181	coiled-coil domain containing 181	41.082447	1.0713172	0.3905919	2.742804	6.091696e-03	4.805408e-02
CLSTN2	catsynenin 2	181.705133	1.0150829	0.3411624	2.975366	2.926394e-03	2.918652e-02
SCN3B	sodium voltage-gated channel beta subunit 3	121.951613	0.9871886	0.3396407	2.906567	3.654184e-03	3.382980e-02
MTCL1	microtubule crosslinking factor 1	399.153170	0.9570641	0.3243318	2.950879	3.168705e-03	3.080823e-02
SPACA9	sperm acrosome associated 9	92.324392	0.9510539	0.2954235	3.219290	1.285087e-03	1.653543e-02
STIM1	stromal interaction molecule 1	1423.623159	0.9285644	0.2951181	3.146416	1.652845e-03	1.943749e-02
DBN1	drebrin 1	1080.662131	0.7704495	0.2669485	2.886135	3.900043e-03	3.528259e-02
DPYSL2	dihydropyrimidinase like 2	3633.965755	0.7582409	0.2235979	3.391091	6.961502e-04	1.069271e-02
DPYSL3	dihydropyrimidinase like 3	2359.704732	0.7366823	0.2315744	3.181190	1.466713e-03	1.802526e-02
DYNC2H1	dynein cytoplasmic 2 heavy chain 1	180.805018	0.7031502	0.2276818	3.088302	2.013039e-03	2.232452e-02
DYNC2LI1	dynein cytoplasmic 2 light intermediate chain 1	266.016147	0.6649107	0.1777337	3.741050	1.832531e-04	4.011862e-03
STMN1	stathmin 1	1394.640523	0.6131505	0.1939413	3.161527	1.569443e-03	1.880988e-02
KIF13B	kinesin family member 13B	748.045214	0.5869961	0.1823980	3.218216	1.289905e-03	1.656006e-02
ARHGEF9	Cdc42 guanine nucleotide exchange factor 9	2062.960977	0.4829104	0.1541420	3.132894	1.730918e-03	2.006620e-02
BEX4	brain expressed X-linked 4	1429.322911	0.4552700	0.1443430	3.154085	1.610021e-03	1.910182e-02
EMD	emerin	1216.220897	0.4175931	0.1428686	2.922918	3.467680e-03	3.265488e-02
GRIN2A	glutamate ionotropic receptor NMDA type subunit 2A	573.703544	-2.7942514	0.7787652	-3.588054	3.331555e-04	6.236027e-03
CKAP2L	cytoskeleton associated protein 2 like	11.961873	-2.6685788	0.8531485	-3.127918	1.760491e-03	2.030932e-02
BIRC5	baculoviral IAP repeat containing 5	21.241373	-2.2754492	0.7494786	-3.036043	2.397053e-03	2.525290e-02
MYL7	myosin light chain 7	18087.277809	-2.2153761	0.7113916	-3.114144	1.844793e-03	2.089653e-02
GRIN3A	glutamate ionotropic receptor NMDA type subunit 3A	64.115114	-1.9192671	0.5051655	-3.799284	1.451149e-04	3.428558e-03
WDR62	WD repeat domain 62	892.755862	-1.5955712	0.4224322	-3.777106	1.586614e-04	3.599424e-03
TPX2	TPX2 microtubule nucleation factor	62.949467	-1.5698211	0.5693498	-2.757217	5.829562e-03	4.696565e-02
DAB1	DAB adaptor protein 1	150.512063	-1.4456068	0.5159792	-2.801677	5.083781e-03	4.238482e-02
SLC6A9	solute carrier family 6 member 9	46.797095	-1.3703820	0.4704159	-2.913128	3.578276e-03	3.331668e-02
XIRP2	xin actin binding repeat containing 2	24918.002954	-1.3193753	0.3904855	-3.378807	7.280109e-04	1.105679e-02
RAPGEF4	Rap guanine nucleotide exchange factor 4	431.263667	-1.1737832	0.2944133	-3.986856	6.695474e-05	1.962003e-03
NAV3	neuron navigator 3	149.225690	-1.1374313	0.3527139	-3.224798	1.260614e-03	1.629980e-02
SYT7	synaptotagmin 7	758.842852	-1.0877793	0.2804863	-3.878191	1.052361e-04	2.704219e-03
CASQ1	calsequestrin 1	1625.708040	-0.9884939	0.3025740	-3.266950	1.087130e-03	1.470147e-02
CKAP2	cytoskeleton associated protein 2	263.132918	-0.9700421	0.2016049	-4.811599	1.497275e-06	1.049771e-04
CSRFP2	cysteine and glycine rich protein 2	246.791278	-0.9263227	0.3090644	-2.997183	2.724870e-03	2.785318e-02
CAVIN4	caveolae associated protein 4	2065.005630	-0.9151393	0.2894771	-3.161353	1.570381e-03	1.880988e-02
RACGAP1	Rac GTPase activating protein 1	140.665811	-0.9059803	0.3151308	-2.874934	4.041121e-03	3.621779e-02
HAUS8	HAUS augmin like complex subunit 8	76.137586	-0.8579109	0.2293639	-3.740391	1.837339e-04	4.016956e-03
ADGRL1	adhesion G protein-coupled receptor L1	598.092664	-0.8145279	0.2114153	-3.852739	1.168038e-04	2.936524e-03
LMOD3	leiomodin 3	7566.814823	-0.7681449	0.2451851	-3.132918	1.730779e-03	2.006620e-02
HOMER1	homer scaffold protein 1	1455.505991	-0.7475483	0.2050636	-3.645447	2.669277e-04	5.269231e-03
TPM2	tropomyosin 2	17977.301514	-0.7375223	0.1857629	-3.970235	7.180183e-05	2.063478e-03
TUBG1	tubulin gamma 1	778.972242	-0.7280100	0.2135228	-3.409518	6.507779e-04	1.015888e-02
PLS1	plastin 1	62.528767	-0.7257692	0.2465286	-2.943955	3.240467e-03	3.122597e-02
TWF1	twirfilin actin binding protein 1	221.217600	-0.7146674	0.2287339	-3.124450	1.781379e-03	2.049216e-02
DVL1	dishevelled segment polarity protein 1	2161.904668	-0.5868644	0.1575715	-3.724431	1.957560e-04	4.213721e-03
PRK CZ	protein kinase C zeta	354.010016	-0.5829840	0.1401274	-4.160386	3.177105e-05	1.122829e-03
TBCE	tubulin folding cofactor E	653.400540	-0.5807267	0.2080410	-2.791405	5.247982e-03	4.333182e-02
LZTS3	leucine zipper tumor suppressor family member 3	634.659755	-0.5485539	0.1959767	-2.799077	5.124887e-03	4.260661e-02
ARHGEF25	Rho guanine nucleotide exchange factor 25	637.517291	-0.5294155	0.1675045	-3.160605	1.574418e-03	1.882395e-02
PSEN2	presenilin 2	445.233639	-0.4988028	0.1707856	-2.920638	3.493160e-03	3.280585e-02
CALM3	calmodulin 3	8389.100991	-0.4961853	0.1478749	-3.355439	7.923901e-04	1.164319e-02

Gene	Function	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
WDR1	WD repeat domain 1	6316.830591	-0.4824212	0.1769203	-2.726771	6.395744e-03	4.963415e-02
NEDD1	NEDD1 gamma-tubulin ring complex targeting factor	493.649478	-0.4821382	0.1636056	-2.946953	3.209216e-03	3.101170e-02
HAUS6	HAUS augmin like complex subunit 6	284.843278	-0.4344779	0.1511268	-2.874922	4.041270e-03	3.621779e-02
YWHAZ	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein zeta	3386.395776	-0.4324969	0.1315951	-3.286573	1.014145e-03	1.401537e-02
MYO9B	myosin IXB	1680.830140	-0.4021050	0.1393136	-2.886329	3.897646e-03	3.528053e-02
NPTN	neuroplastin	4830.323592	-0.3672508	0.1320114	-2.781962	5.403129e-03	4.434208e-02
WASL	WASP like actin nucleation promoting factor	1388.656537	-0.3639156	0.1255122	-2.899444	3.738247e-03	3.435385e-02
MYO1C	myosin IC	6923.937386	-0.2748624	0.0937564	-2.931666	3.371490e-03	3.197706e-02

Supplementary Table 4. IC vs Healthy significantly and uniquely expressed genes that cluster into functional processes

Gene	Function	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
IGHV4-34	immunoglobulin heavy variable4-34	46.175900	9.1273257	2.79946948	3.260377	1.112642e-03	1.028591e-02
IGHV3-9	immunoglobulin heavy variable3-9	33.303123	7.3619659	2.34733350	3.136310	1.710882e-03	1.406509e-02
IGKV2-30	immunoglobulin kappa variable2-30	12.548717	6.7281601	1.81938084	3.698049	2.172627e-04	2.978809e-03
HLA-C	majorhistocompatibilitycomplex,classII,C	58.412783	6.4450762	2.46816217	2.611286	9.020254e-03	4.673391e-02
IGHM	immunoglobulin heavy constantmu	1109.346750	6.3814377	1.56024785	4.090015	4.313446e-05	8.611875e-04
IGKV3-15	immunoglobulin kappa variable3-15	34.046892	5.8891336	2.26006724	2.605734	9.167768e-03	4.720071e-02
IgLVL6-57	immunoglobulin lamda variable6-57	17.218947	5.7905202	1.59281575	3.635399	2.775511e-04	3.594990e-03
IGLV2-14	immunoglobulin lamda variable2-14	42.539114	5.7834533	1.76037108	3.285360	1.018520e-03	9.629910e-03
IGHV4-39	immunoglobulin heavy variable4-39	23.994505	5.3459247	1.80178927	2.967009	3.007123e-03	2.129960e-02
LAMP3	lysosomal associated membraneprotein3	16.302660	5.1466230	1.36461187	3.771492	1.622743e-04	2.392538e-03
IGHA1	immunoglobulin heavy constantalpha1	2443.532712	5.1407396	1.32273578	3.886445	1.017230e-04	1.685016e-03
IGKV1-5	immunoglobulin kappa variable1-5	112.983246	4.6876531	1.61221023	2.913797	3.570622e-03	2.391583e-02
IGKV3-20	immunoglobulin kappa variable3-20	62.827803	4.6014836	1.48521977	3.098184	1.947107e-03	1.542016e-02
CXCL9	C-X-Cmotifchemokineligand9	913.960195	4.5990793	1.47405748	3.120014	1.808427e-03	1.461932e-02
IGHV3-49	immunoglobulin heavy variable3-49	5.965469	4.5368830	1.74620542	2.598138	9.373076e-03	4.796419e-02
IGHV4-1	immunoglobulin kappa variable4-1	110.768233	4.5232171	1.56503104	2.906894	3.650370e-03	2.425527e-02
IGHV1-18	immunoglobulin heavy variable1-18	20.877473	4.4452646	1.50479891	2.954059	3.136241e-03	2.190526e-02
CCL22	C-Cmotifchemokineligand22	7.957771	4.3738857	1.23279685	3.547937	3.882609e-04	4.655491e-03
IGLV2-8	immunoglobulin lamda variable2-8	41.688621	4.2705233	1.42892513	2.988626	2.802346e-03	2.018442e-02
CCR7	C-Cmotifchemokinerceptor7	16.270529	4.2070196	1.15634803	3.638195	2.745555e-04	3.575672e-03
IGLV1-40	immunoglobulin lamda variable1-40	23.419365	4.1822933	1.23504016	3.386362	7.082584e-04	7.257142e-03
IGLV2-11	immunoglobulin lamda variable2-11	19.128619	4.1416331	1.38691320	2.986224	2.824459e-03	2.029096e-02
CCL24	C-Cmotifchemokineligand24	8.324075	4.1391190	1.17488714	3.522993	4.267028e-04	4.969534e-03
SIRPG	signalregulatoryproteingamma	9.277391	4.1372090	1.50304550	2.752551	5.913298e-03	3.457949e-02
TNFRSF18	TNFRceptorsuperfamilymember18	6.137487	4.1211102	1.49236900	2.761455	5.754440e-03	3.391286e-02
IGLV2-23	immunoglobulin lamda variable2-23	29.117633	3.9435233	1.21315995	3.250621	1.151532e-03	1.056906e-02
JCHAIN	joiningchainofmultimericIgAandIgM	777.145189	3.8890048	1.41118545	2.755842	5.854119e-03	3.436634e-02
CD1E	CD1emolecule	13.516190	3.8875102	1.33380078	2.914611	3.561323e-03	2.390440e-02
IgL5	immunoglobulin lambda like polypeptide 5	78.973895	3.8517529	1.41422220	2.723584	4.657781e-03	3.681843e-02
CXCL11	C-X-Cmotifchemokineligand11	62.101707	3.7301417	1.40623421	2.652575	7.988037e-03	4.266187e-02
CXCL10	C-X-Cmotifchemokineligand10	215.444156	3.6992948	1.37792875	2.684678	7.259974e-03	4.008053e-02
CXCR3	C-X-Cmotifchemokinerceptor3	17.585516	3.5336855	1.09658404	3.222448	1.271001e-03	1.137675e-02
IGHA2	immunoglobulin heavy constantalpha2(A2mmarker)	169.686771	3.4444847	1.25399405	2.746811	6.017779e-03	3.504261e-02
SYTL1	synaptotagminlike1	37.139142	3.0402394	0.67412463	4.509907	6.485606e-06	1.879131e-04
CD1C	CD1cmolecule	36.442241	2.9629920	0.85679532	3.458226	5.437447e-04	6.004660e-03
CCL5	C-Cmotifchemokineligand5	120.600192	2.8050933	0.73455924	3.818744	1.341331e-04	2.082883e-03
CD2	CD2molecule	65.872598	2.7316791	0.89080190	3.066539	2.165523e-03	1.675108e-02
ABCC11	ATP binding cassette subfamily C member11	5.339165	2.7308018	0.92742095	2.944512	3.234647e-03	2.236549e-02
TRBC2	Tcell receptor beta constant2	120.393934	2.7055916	0.78756045	3.435408	5.916620e-04	6.389565e-03
RHOH	ras homolog family member H	18.654429	2.5831203	0.93939776	2.608779	9.086596e-03	4.697528e-02
SMPD3	sphingomyelinphosphodiesterase3	14.827849	2.5648910	0.99179402	2.586113	9.706521e-03	4.920451e-02
LCK	LCKproto-oncogene,Srcfamilytyrosinekinase	53.963402	2.5544172	0.87630383	2.914990	3.556995e-03	2.390440e-02
CCL8	C-Cmotifchemokineligand8	35.291840	2.5517122	0.88916665	2.869779	4.107582e-03	2.650881e-02
TBC1D10C	TBC1domainfamilymember10C	59.102377	2.5039043	0.69261306	3.615156	3.001668e-04	3.813655e-03
CD3E	CD3emolecule	77.928499	2.4005368	0.8804373	2.726508	6.400841e-03	3.660674e-02
ZAP70	zeta chain of Tcell receptor associated proteinkinase 70	61.210586	2.3199826	0.72600368	3.195552	1.395636e-03	1.218416e-02
CD48	CD48molecule	86.079633	2.3082654	0.78227812	2.950697	3.170582e-03	2.208122e-02
LAG3	lymphocyteactivating3	21.310969	2.2867201	0.80174834	2.852167	4.342230e-03	2.756322e-02
CD6	CD6molecule	35.444640	2.2724584	0.74097799	3.066837	2.163371e-03	1.674222e-02
CCL3	C-Cmotifchemokineligand3	21.127129	2.1933377	0.81856241	2.679500	7.373228e-03	4.046710e-02
ABCC6	ATP binding cassette subfamily C member6	18.061060	2.1686285	0.61750385	3.511927	4.448701e-04	5.120162e-03
CARD11	caspase recruitment domain family member11	38.304206	2.0987357	0.73959088	2.837698	4.544012e-03	2.837832e-02
ZP3	zona pellucida glycoprotein3	32.803384	2.0761913	0.76308278	2.720794	6.512527e-03	3.700370e-02
GZMA	granzymeA	42.765519	1.9748701	0.70833314	2.788053	5.302593e-03	3.191727e-02
FPR2	formylpeptid receptor2	15.548507	1.8722902	0.63758011	2.936557	3.318780e-03	2.276508e-02
RARRES2	retinoic acid receptor responder2	212.750256	1.8125928	0.60737989	2.984134	2.843825e-03	2.041246e-02
ACAP1	ArGAPwithcoiled-coil,ankyrinrepeatandPHdomains1	108.506544	1.7419962	0.54888261	3.173714	1.505021e-03	1.288878e-02
ADAM8	ADAM metallopeptidase domain8	98.869433	1.7036468	0.57298825	2.973266	2.946485e-03	2.099528e-02
SELL	selectin L	119.443633	1.6517917	0.51594712	3.201475	1.367260e-03	1.199942e-02
RTN4R	reticulin 4 receptor	34.829720	1.5979149	0.47073653	3.394499	6.875419e-04	7.133194e-03
SPN	sialophrin	114.624826	1.4235531	0.43250105	3.291444	9.967435e-04	9.461667e-03
THBS4	thrombospondin4	11145.458291	1.4034856	0.35709402	3.930297	8.484094e-05	1.459174e-03
SFRP1	secreted frizzled related protein1	2353.776080	1.3669175	0.47295760	2.890148	3.850605e-03	2.522230e-02
LAT	linkerforactivationofTcells	89.104364	1.3538948	0.41915336	3.230070	1.237598e-03	1.115583e-02
KIT	KITproto-oncogene,receptortyrosinekinase	114.866777	1.2394858	0.39910218	3.105685	1.898385e-03	1.516429e-02
AGRN	agrin	1199.474245	1.2294127	0.31404098	3.914816	9.047316e-05	1.536935e-03
PLCE1	phospholipaseCepsilon1	1338.872595	1.1128473	0.41459473	2.684181	7.270772e-03	4.009053e-02
PLEKHG4	pleckstrin homology and RhoGEF domain containingG4	44.592923	1.0963012	0.34716566	3.157862	1.589308e-03	1.336288e-02



Gene	Function	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
SLC30A2	solute carrier family30member2	374.964976	1.0776259	0.37753688	2.854359	4.312375e-03	2.746864e-02
DAPK3	death associated protein kinase3	2335.661715	1.0478415	0.25986152	4.032307	5.523191e-05	1.042312e-03
ARHGAP22	RhoGTPaseactivatingprotein22	130.296969	1.0351433	0.39050095	2.650809	8.029931e-03	4.282979e-02
ICAM1	intercellular adhesion molecule1	713.444287	0.8942768	0.33776778	2.647608	8.106335e-03	4.308612e-02
TCIRG1	Tcell immuneregulator1.ATPaseH+transportingV0subunit3	815.555957	0.8889976	0.26126382	3.402682	6.672794e-04	7.010439e-03
CX3CL1	C-X3-Cmotifchemokineligand1	975.457095	0.8857654	0.17435314	5.080295	3.768491e-07	1.882095e-05
TNFRSF14	TNFreceptorsuperfamilymember14	558.697451	0.8715396	0.27090885	3.217095	1.294956e-03	1.154149e-02
ARHGAP33	RhoGTPaseactivatingprotein33	266.575908	0.8554598	0.28821651	2.968115	2.996322e-03	2.125025e-02
ATG16L2	autophagyrelated16like2	471.761412	0.7881138	0.28642815	2.751523	5.931879e-03	3.463942e-02
SLC26A11	solute carrier family26member11	179.839122	0.7772683	0.24638687	3.154666	1.606820e-03	1.343684e-02
ATP1B2	ATPaseNa+/K+transportingsubunitbeta2	256.866844	0.7678534	0.28310518	2.712255	6.682719e-03	3.776262e-02
EFNA5	ephrinA5	627.289860	0.7426469	0.17187356	4.320891	1.554007e-05	3.837880e-04
CSAR1	complementCSareceptor1	191.001370	0.7159740	0.26896233	2.661986	7.768104e-03	4.192730e-02
MAP3K11	mitogen-activatedproteinkinasekinase11	1200.883967	0.7011078	0.14632121	4.791566	1.654843e-06	6.297871e-05
OBSCN	obscurin.cytoskeletalcalmodulinandtitin-interactingRhoGEF	22219.858775	0.6689645	0.18069928	3.702087	2.138331e-04	2.946362e-03
SIPA1	signal-inducedproliferation-associated1	765.861455	0.6598632	0.16729570	3.944293	8.003585e-05	1.398352e-03
WDR24	WDRepeatdomain24	257.388518	0.6269852	0.16503574	3.799088	1.452298e-04	2.209805e-03
MAP1LC3A	microtubule associated protein 1light chain 3alpha	935.008140	0.6118128	0.19667237	3.110822	1.865673e-03	1.496047e-02
ARHGEF1	Rho guanine nucleotide exchange factor1	1157.662608	0.5977079	0.20110861	2.972065	2.958041e-03	2.104157e-02
SDC2	syndecan2	2026.476299	0.5915831	0.18514513	3.195240	1.397148e-03	1.218456e-02
ITM2C	integral membrane protein 2C	968.913915	0.5684845	0.18187985	3.125605	1.774399e-03	1.442688e-02
CDC42EP1	CDC42effectorprotein1	692.111050	0.5682077	0.20925602	2.715371	6.620157e-03	3.748717e-02
RIN3	RasandRabinteractor3	494.967490	0.5616817	0.19863837	2.827660	4.688961e-03	2.905583e-02
RHOG	ras homolog family memberG	558.357588	0.5481124	0.20339103	2.894870	7.041612e-03	3.918001e-02
LAMTOR4	lateendosomal/lysosomaladaptor.MAPKandMTOReactivator4	779.714674	0.5446591	0.14505332	3.754889	1.734183e-04	2.514247e-03
AP5B1	adapto related protein complex 5 subunit beta1	308.921609	0.5397804	0.15987982	3.376163	7.350424e-04	7.463059e-03
LLGL1	LLGL scribble cell polarity complex component1	463.398907	0.5385923	0.18048487	2.984141	2.843755e-03	2.041246e-02
TRAPPC4	trafficking protein particle complex4	594.237028	0.5364898	0.14186903	3.781585	1.558330e-04	2.332726e-03
SPHK2	sphingosinekinase2	206.692143	0.5251151	0.17332988	3.029571	2.449016e-03	1.831366e-02
INF2	invertedformin.FH2andWH2domaincontaining	1408.986050	0.5228602	0.17525174	2.983481	2.849901e-03	2.044647e-02
ESAM	endothelial cell adhesion molecule	1854.382839	0.5061595	0.19303821	2.622069	8.739776e-03	4.575192e-02
TBC1D17	TBC1 domainfamilymember17	1397.410589	0.5016635	0.14206912	3.531123	4.137994e-04	4.856221e-03
SGSM3	small Gprotein signaling modulator3	1217.047279	0.4972390	0.10830284	4.591191	4.407223e-06	1.372596e-04
CDIP1	cell death inducing p53target1	1557.631401	0.4904338	0.14176772	3.459418	5.413437e-04	5.983994e-03
TMEM175	transmembraneprotein175	483.882475	0.4902202	0.11872998	4.128866	3.645563e-05	6.843300e-04
SCRIB	scribble planar cell polarity protein	1145.041278	0.4875877	0.16550830	2.946002	3.219107e-03	2.230423e-02
SLC15A4	solute carrier family15member4	530.775194	0.4867267	0.13938940	3.491849	4.796900e-04	5.419649e-03
CLCN7	chloridevoltage-gatedchannel7	1314.817865	0.4848339	0.10409008	4.657830	3.195600e-06	1.048245e-04
ATP6V1H	ATPaseH+transportingV1subunitH	1142.814507	0.4637430	0.16298028	2.845393	4.435659e-03	2.797476e-02
HSD3B7	hydroxy-delta-5-steroiddehydrogenase,3beta-andsteroiddelta-isomerase7	229.861506	0.4631949	0.17078468	2.712157	6.684694e-03	3.776262e-02
ULK1	unc-51likeautophagyactivatingkinase1	2336.237042	0.4546957	0.15884349	2.862539	4.202612e-03	2.692359e-02
BORCS7	BLOC-1relatedcomplexsubunit7	733.629837	0.4481911	0.14951078	2.997717	2.720097e-03	1.977290e-02
BORCS6	BLOC-1relatedcomplexsubunit6	287.931465	0.4457790	0.17077191	2.610376	9.044263e-03	4.682912e-02
SLC2A8	solute carrier family2member8	294.535985	0.4370628	0.16434909	2.659356	7.829010e-03	4.205923e-02
LAPTM4B	lysosomalproteintransmembrane4beta	7994.693640	0.4364742	0.13917967	3.136049	1.712407e-03	1.406761e-02
SYDE1	synapsedefectiveRhoGTPasehomolog1	602.363179	0.4247276	0.16037455	2.648348	8.088629e-03	4.301950e-02
BORCS8	BLOC-1relatedcomplexsubunit8	190.478550	0.4218507	0.14682046	2.873594	4.058302e-03	2.629280e-02
CD81	CD81molecule	11597.009572	0.4183444	0.15704346	2.663877	7.724590e-03	4.178533e-02
ABCA2	ATP binding cassette subfamily A member2	1754.426524	0.4114990	0.14537293	2.830644	4.645442e-03	2.884821e-02
TBC1D13	TBC1 domainfamilymember13	834.053744	0.4100646	0.09897961	4.142920	3.429115e-05	7.302126e-04
TRAPPC5	traffickingproteinparticlecomplex5	529.558791	0.4023173	0.15440179	2.605652	9.169953e-03	4.720071e-02
AP1M1	adaptor related protein complex1subunit mu1	1130.811690	0.3965961	0.11650630	3.404074	6.638865e-04	6.988035e-03
ABCD4	ATP binding cassette subfamily D member4	687.530857	0.3815700	0.12593828	3.029817	2.447018e-03	1.830695e-02
PLEKHG3	pleckstrin homology and RhoGEF domain containingG3	540.113877	0.3768760	0.14063074	2.679898	7.364459e-03	4.043523e-02
CD47	CD47molecule	1813.099249	0.3739061	0.09996234	3.740470	1.836766e-04	2.626592e-03
SGSM2	smallGprotein signalingmodulator2	1437.659314	0.3610841	0.12970495	2.783889	5.371146e-03	3.220171e-02
TMEM138	transmembraneprotein138	247.067246	0.3400481	0.12808123	2.654941	7.932231e-03	4.239788e-02
ARFGAP1	ADPribosylationfactorGTPaseactivatingprotein1	1012.452064	0.3344690	0.10713441	3.121957	1.796533e-03	1.456050e-02
GGA1	golgiassociated,gammaadaptingcontaining,ARFBindingprotein1	1005.115325	0.3159003	0.10951263	2.884602	3.919092e-03	2.560032e-02
VAPA	VAMP associated proteinA	3670.660390	0.3090737	0.11386404	2.714410	6.639385e-03	3.757047e-02
LZTR1	leucine zipper like transcription regulator1	1142.755338	0.3014826	0.08607600	3.502517	4.608848e-04	5.264406e-03
GPS1	Gprotein pathway suppressor1	1800.451045	0.2705263	0.10171599	2.659624	7.822787e-03	4.205923e-02
MTOR	mechanistic target of rapamycinkinase	2386.202543	-0.2403381	0.09182295	-2.617408	8.860047e-03	4.619168e-02
RANBP9	RANbindingprotein9	1389.777831	-0.2740819	0.08577810	-3.195243	1.397129e-03	1.218456e-02
EXOC1	exocyst complex component1	1174.187133	-0.2826886	0.09862890	-2.866185	4.154519e-03	2.667096e-02
PDCD6IP	programmed cell death 6 interacting protein	3471.602008	-0.3168946	0.10555600	-3.002147	2.680829e-03	1.956335e-02
SNX9	sortingnexin9	1985.538134	-0.3202141	0.10209180	-3.136531	1.709593e-03	1.406145e-02

Gene	Function	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
WDR44	WD repeat domain44	663.541154	-0.3318733	0.10515439	-3.156057	1.599175e-03	1.340909e-02
ZC3H15	zincfingerCCH-typecontaining15	1244.736696	-0.3350394	0.10852343	-3.087254	2.020150e-03	1.584770e-02
CLEC16A	C-typelectindomaincontaining16A	972.411399	-0.3376316	0.11838414	-2.852001	4.344502e-03	2.756712e-02
M6PR	mannose-6-phosphatereceptor,cationdependent	1959.760078	-0.3435663	0.12889634	-2.665446	7.688620e-03	4.167843e-02
ADAM10	ADAM metalloproteinase domain10	1288.069375	-0.3475807	0.13186984	-2.635786	8.394255e-03	4.430493e-02
SPAG9	sperm associated antigen9	2750.499732	-0.3552854	0.10231630	-3.472422	5.157850e-04	5.757061e-03
GAPVD1	GTPaseactivatingproteinandVPS9domains1	1221.676997	-0.3627798	0.12122054	-2.992725	2.764986e-03	2.001854e-02
LARS1	leucyl-tRNA synthetase1	3110.365778	-0.3650385	0.11414568	-3.198006	1.383815e-03	1.211275e-02
NF1	neurofibromin1	2316.006017	-0.3667890	0.13310729	-2.755589	5.858659e-03	3.438086e-02
ERC1	ELKS/RAB6-interacting/CASTfamilymember1	2207.649441	-0.3737131	0.11727010	-3.186772	1.438701e-03	1.246736e-02
SBF2	SETbindingfactor2	1164.923561	-0.3759591	0.10757202	-3.494952	4.741466e-04	5.371616e-03
RALGAPB	RalGTPase activating protein non-catalytic betasubunit	1947.492914	-0.3845592	0.10014837	-3.839895	1.230872e-04	1.947728e-03
RANBP2	RANbindingprotein2	2647.912204	-0.3854343	0.12328510	-3.126366	1.769814e-03	1.441419e-02
ARHGAP44	RhoGTPase activating protein44	297.393044	-0.3870483	0.11746065	-3.295132	9.837549e-04	9.365099e-03
CHM	CHMR abescortprotein	562.143574	-0.3974689	0.13172382	-3.017441	2.549183e-03	1.888671e-02
NCKAP1	NCK associated protein1	6686.567394	-0.3987882	0.11668330	-3.417697	6.315339e-04	6.715499e-03
RAB3GAP2	RAB3 GTP aseactivatingnon-catalytic protein subunit2	1099.547705	-0.4053384	0.12436687	-3.259215	1.117210e-03	1.032240e-02
MIOS	meiosis regulator for oocyte development	708.243866	-0.4121173	0.14181654	-2.905989	3.660946e-03	2.429577e-02

**Supplementary Table 5. Significantly changed genes DC vs Healthy that also matched significantly changed proteins in the same comparison**

Gene	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
<b>NPPA</b>	64750.650	5.9694452	1.2395293	4.815897	1.465403e-06	1.034987e-04
<b>AEBP1</b>	3010.071	1.6207095	0.4774258	3.394684	6.870794e-04	1.059340e-02
<b>COL14A1</b>	1428.224	1.4920775	0.4377406	3.408589	6.529987e-04	1.016981e-02
<b>MFAP4</b>	3545.324	1.1847313	0.3634084	3.260055	1.113907e-03	1.496186e-02
<b>ARHGAP1</b>	4615.860	1.1231660	0.2429495	4.623043	3.781511e-06	2.228625e-04
<b>PFKL</b>	1919.688	0.4922763	0.1252458	3.930481	8.477615e-05	2.329339e-03
<b>YWHAQ</b>	4124.617	0.3505814	0.1240680	2.825720	4.717449e-03	4.013505e-02
<b>MYH6</b>	98127.708	-3.3541422	0.6566385	-5.108050	3.255003e-07	3.309117e-05
<b>ART3</b>	2301.617	-1.0665936	0.3073363	-3.470444	5.195979e-04	8.695246e-03
<b>TTN</b>	260372.843	-0.8533805	0.1489343	-5.729914	1.004818e-08	2.068908e-06
<b>UQCRB</b>	17654.412	-0.8025549	0.2124250	-3.778062	1.580537e-04	3.595666e-03
<b>PDCD5</b>	1259.482	-0.7584922	0.1943986	-3.901736	9.550525e-05	2.533361e-03
<b>PGP</b>	1117.789	-0.3333979	0.0871767	-3.824392	1.310951e-04	3.196992e-03

**Supplementary Table 6. Significantly changed proteins in DC vs Healthy that had matching significantly changed genes in the same comparison**

Gene	baseMean (LFQs)	log2FoldChange	lfcSE	stat	pvalue	padj
<b>NPPA</b>	914.25858	4.9582743	1.4719756	3.368449	7.559246e-04	2.226691e-02
<b>AEBP1</b>	308.52135	2.7175458	0.7334202	3.705305	2.111365e-04	8.414409e-03
<b>COL14A1</b>	850.48807	2.5608028	0.7600312	3.369339	7.534869e-04	2.226691e-02
<b>TTN</b>	6750.06204	2.0204337	0.4615007	4.377964	1.197929e-05	1.082129e-03
<b>MFAP4</b>	600.37964	1.7467693	0.4873160	3.584470	3.377640e-04	1.207024e-02
<b>PFKL</b>	638.20791	1.3175970	0.4050131	3.253221	1.141047e-03	3.064628e-02
<b>ARHGAP1</b>	1621.36699	0.8639123	0.2052113	4.209867	2.555214e-05	1.753841e-03
<b>YWHAQ</b>	2490.18602	0.5970173	0.1543780	3.867243	1.100725e-04	5.326724e-03
<b>ART3</b>	17.04353	-8.3910270	2.5795193	-3.252942	1.142168e-03	3.064628e-02
<b>MYH6</b>	757.81447	-4.5902735	0.8151147	-5.631445	1.787061e-08	5.619916e-06
<b>UQCRB</b>	7093.72840	-0.6990583	0.1739609	-4.018480	5.857480e-05	3.239545e-03
<b>PDCD5</b>	386.59121	-0.6471055	0.1712485	-3.778750	1.576174e-04	7.002346e-03
<b>PGP</b>	542.65220	-0.4856027	0.1332880	-3.643260	2.692066e-04	1.027535e-02

**Supplementary Table 7. Significantly changed genes IC vs Healthy that also matched significantly changed proteins in the same comparison**

Gene	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
NPPA	6.475065e+04	3.9810476	1.23953548	3.211725	1.319404e-03	1.171543e-02
HBB	1.783859e+03	3.8629190	0.69460549	5.561314	2.677515e-08	2.137861e-06
CD2	6.587260e+01	2.7316791	0.89080190	3.066539	2.165523e-03	1.675108e-02
APOA1	1.178621e+03	1.9361251	0.40586255	4.770396	1.838640e-06	6.918194e-05
COL14A1	1.428224e+03	1.2246810	0.43777999	2.797480	5.150288e-03	3.123794e-02
LUM	8.659295e+03	1.0829886	0.36984100	2.928255	3.408707e-03	2.320516e-02
ARHGAP1	4.615860e+03	1.0633638	0.24294435	4.376985	1.203321e-05	3.097899e-04
VTN	1.746666e+03	0.9346608	0.32306629	2.893093	3.814683e-03	2.506350e-02
CLU	4.735043e+03	0.8458807	0.23042274	3.670995	2.416081e-04	3.245706e-03
IVD	3.218196e+03	0.8084392	0.15739534	5.136360	2.801102e-07	1.469563e-05
MYH7	7.525063e+05	0.7733415	0.21174226	3.652278	2.599246e-04	3.425361e-03
HPRT1	5.044695e+02	0.6328641	0.15518421	4.078147	4.539599e-05	8.977178e-04
MAP4	3.104889e+04	0.6234824	0.16175430	3.854503	1.159651e-04	1.863396e-03
TOM1L2	1.069734e+04	0.6189644	0.16401201	3.773897	1.607173e-04	2.382254e-03
CYB5R1	4.584060e+03	0.6171857	0.19505292	3.164196	1.555120e-03	1.318878e-02
C11orf68	1.018369e+03	0.5994157	0.14194986	4.222729	2.413623e-05	5.516179e-04
BCAT2	1.219605e+03	0.5912492	0.12562024	4.706640	2.518330e-06	8.780368e-05
CIRBP	7.594636e+03	0.5801297	0.15453285	3.754087	1.739747e-04	2.515976e-03
COQ10B	1.087776e+03	0.5566569	0.11665225	4.771935	1.824648e-06	6.881115e-05
COX8A	6.443199e+03	0.5551769	0.16765060	3.311512	9.279342e-04	8.925665e-03
PPP1R12C	6.316688e+03	0.5216105	0.16114937	3.236813	1.208724e-03	1.099687e-02
PCCB	2.797731e+03	0.5040773	0.18752495	2.688055	7.186961e-03	3.982884e-02
ADH5	3.836706e+03	0.4782080	0.15132423	3.160155	1.576852e-03	1.330036e-02
PIN1	1.222749e+03	0.4749581	0.14016051	3.388673	7.023181e-04	7.218944e-03
COPA	4.892705e+03	0.4479092	0.11160183	4.013458	5.983574e-05	1.111875e-03
MDH1	3.480311e+04	0.3933002	0.15007853	2.620630	8.776757e-03	4.582928e-02
GANAB	5.108525e+03	0.2048065	0.07563721	2.707748	6.774143e-03	3.815129e-02
KNG1	8.708153e+00	-3.5015224	1.08620148	-3.223640	1.265725e-03	1.134650e-02
RBP4	5.150649e+01	-3.0112881	0.80596479	-3.736253	1.867829e-04	2.658098e-03
APOB	1.457905e+03	-1.5096292	0.47412967	-3.184001	1.452547e-03	1.254273e-02
ITGA1	1.115450e+03	-1.2382714	0.18964214	-6.529516	6.598244e-11	1.463139e-08
AZGP1	1.877860e+03	-1.2162029	0.39720311	-3.061917	2.199246e-03	1.694431e-02
AR	2.921452e+02	-0.8772186	0.21534517	-4.073547	4.630254e-05	9.102335e-04
ATP2A2	8.817315e+04	-0.8740325	0.25647264	-3.407897	6.546551e-04	6.903975e-03
EFEMP1	2.269356e+03	-0.8453065	0.29247421	-2.890191	3.850073e-03	2.522230e-02
GPD1L	1.258384e+04	-0.7786781	0.24681414	-3.154917	1.605438e-03	1.343684e-02
YBX1	1.054430e+04	-0.7242123	0.17099695	-4.235235	2.283125e-05	5.289279e-04
APOBEC2	3.872554e+03	-0.6568878	0.17472567	-3.759538	1.702276e-04	2.479032e-03
COX17	1.173502e+03	-0.6481779	0.21244572	-3.051028	2.280591e-03	1.732604e-02
BCL2L13	2.965080e+03	-0.6436776	0.21650708	-2.973009	2.948953e-03	2.099776e-02
RYR2	6.533920e+04	-0.6349861	0.23220520	-2.734590	6.245794e-03	3.604226e-02
FUNDC2	4.125040e+03	-0.5920042	0.17621592	-3.359539	7.807266e-04	7.836007e-03
AGL	3.089582e+03	-0.5735931	0.20218325	-2.836996	4.554016e-03	2.841945e-02
IPO7	6.043559e+03	-0.5725417	0.08874020	-6.451887	1.104660e-10	2.161364e-08
PCCA	1.412112e+03	-0.4819731	0.14928290	-3.228589	1.244025e-03	1.119880e-02
KPNB1	4.499068e+03	-0.3279287	0.09265001	-3.539435	4.009848e-04	4.749842e-03
COPB2	2.886423e+03	-0.3072504	0.09912805	-3.099530	1.938276e-03	1.537186e-02

**Supplementary Table 8. Significantly changed proteins in IC vs Healthy that had matching significantly changed genes in the same comparison**

Gene	baseMean (LFQs)	log2FoldChange	lfcSE	stat	pvalue	padj
NPPA	914.25858	9.1124837	1.47053427	6.196716	5.765328e-10	8.223178e-08
C11orf68	30.10522	8.0450829	2.31548897	3.474464	5.118752e-04	7.498280e-03
CIRBP	19.18828	4.8415669	1.71840704	2.817474	4.840306e-03	3.892353e-02
RBP4	642.66213	3.1977299	0.46911446	6.816524	9.326971e-12	1.944315e-09
COL14A1	850.48807	2.5239521	0.76005018	3.320770	8.976937e-04	1.131512e-02
APOB	294.86064	2.4233619	0.58952286	4.110717	3.944316e-05	1.047951e-03
CLU	953.82664	1.9906723	0.31753242	6.269194	3.629220e-10	5.463992e-08
EFEMP1	232.92261	1.9321612	0.53140885	3.635922	2.769879e-04	4.633563e-03
LUM	3848.36226	1.7901360	0.23272311	7.692128	1.447073e-14	6.535946e-12
APOA1	23146.90214	1.5286333	0.40335221	3.789823	1.507549e-04	2.877083e-03
KNG1	1152.33142	1.4994651	0.34613139	4.332069	1.477145e-05	5.375421e-04
HBB	33276.88959	1.4745615	0.37755310	3.905574	9.400207e-05	2.071103e-03
VTN	1499.20573	1.2952025	0.28740491	4.506543	6.589243e-06	2.834420e-04
PPP1R12C	392.50775	1.2166470	0.41922320	2.902146	3.706155e-03	3.236944e-02
AZGP1	1692.67977	1.1895091	0.33963208	3.502346	4.611803e-04	6.904964e-03
PIN1	230.29457	1.0057528	0.31692409	3.173482	1.506224e-03	1.697622e-02
ARHGAP1	1621.36699	0.7214863	0.20529862	3.514326	4.408712e-04	6.712141e-03
TOM1L2	302.42573	0.6670240	0.15862253	4.205102	2.609640e-05	8.036504e-04
MAP4	3892.63803	0.6374412	0.08327922	7.654266	1.944198e-14	7.526824e-12
HPRT1	326.56634	0.5910349	0.18565923	3.183439	1.455366e-03	1.664153e-02
GANAB	3187.22455	0.4980595	0.13148731	3.787890	1.519318e-04	2.879268e-03
IVD	2814.55937	0.4809352	0.16882698	2.848687	4.390004e-03	3.629784e-02
ITGA1	37.93977	-7.4551057	1.62593307	-4.585125	4.537157e-06	2.049283e-04
COQ10B	25.76960	-3.6614749	1.29380219	-2.830011	4.654635e-03	3.788006e-02
COX8A	64.50582	-2.6965948	0.90325909	-2.985406	2.832025e-03	2.631100e-02
APOBEC2	538.25566	-2.3421796	0.68625970	-3.412964	6.426044e-04	8.979876e-03
MYH7	88222.32260	-2.3050101	0.69301991	-3.326037	8.809010e-04	1.117543e-02
MDH1	407.38258	-2.2105751	0.54704881	-4.040910	5.324416e-05	1.299925e-03
COX17	1153.03912	-1.6439617	0.29010619	-5.666758	1.455244e-08	1.516812e-06
FUNDC2	513.78221	-1.2744822	0.44303215	-2.876726	4.018240e-03	3.413614e-02
AR	166.10200	-1.2114963	0.29385106	-4.122824	3.742555e-05	1.038568e-03
COPA	250.12756	-1.1591848	0.40092386	-2.891284	3.836712e-03	3.300791e-02
IPO7	218.44644	-1.1471134	0.30505643	-3.760332	1.696882e-04	3.107128e-03
BCL2L13	719.29730	-0.9852841	0.21774942	-4.524853	6.043751e-06	2.641704e-04
YBX1	322.24257	-0.8710506	0.31688830	-2.748762	5.982075e-03	4.605518e-02
ATP2A2	7610.93019	-0.7945678	0.20407720	-3.893467	9.882178e-05	2.108717e-03
CD2	350.17987	-0.7332133	0.20579326	-3.562864	3.668314e-04	5.813527e-03
AGL	4127.01725	-0.6915786	0.22552103	-3.066581	2.165218e-03	2.165218e-02
COPB2	255.90592	-0.6275670	0.21901525	-2.865403	4.164786e-03	3.494294e-02
RYR2	4712.06417	-0.6041580	0.16349346	-3.695304	2.196242e-04	3.839881e-03
PCCB	2030.25586	-0.5911669	0.17026574	-3.472025	5.165485e-04	7.526056e-03
PCCA	1974.73516	-0.5708371	0.14609801	-3.907220	9.336412e-05	2.071103e-03
CYB5R1	4097.81450	-0.5375065	0.18367777	-2.926356	3.429582e-03	3.067382e-02
ADH5	4526.81439	-0.5213994	0.10674114	-4.884709	1.035818e-06	7.017665e-05
GPD1L	4495.10495	-0.5011858	0.17553792	-2.855142	4.301752e-03	3.575997e-02
BCAT2	2208.84041	-0.4337632	0.12991373	-3.338856	8.412426e-04	1.090798e-02
KPNB1	1752.28069	-0.3584892	0.12872289	-2.784968	5.353294e-03	4.205051e-02

**Supplementary table 9. Cell cluster full names**

Full names	Abbreviations
Matrix like fibroblasts	Matrix fibr
Connective tissue like fibroblasts/ adipose tissue like cells	Con.tissue fibr/adipocytes
Thymic fibroblasts/Fibroblasts/adipose tissue like cells	Thymic fibr/Fibr/adipocytes
Activated fibroblasts	Activated fibr
Skin like fibroblasts/Axin2+ like cells	Skin like fibr/Axin2+ cells
Mesenteric Lymph Node Fibroblastic Reticular cells/ Connective tissue like fibroblasts	Fibr Reticular cells/Con.tissue fibr
Macrophages/Macrophages activated/Monocytes	Macr activated/Monocytes
Dendritic cells/steady state macrophages/fibroblast like cells	Dendritic cells/steady state macr/fibr like cells
Vascular endothelial cells	Vascular endo cells
Fibroblast like cells/adipocyte like cells	Fibr/adipocyte like cells
Pericytes/cardiomyocytes	Pericytes/cardiomyocytes
Activated fibroblasts/progenitor like adipocytes	Activated fibr/progenitor adipocytes
B cells/B cells memory/B cells naive	B cells:memory,naive,mature
Myofibroblast/Smooth muscle cell	Myofibroblast/Smooth muscle cell
Lymphocytes	Lymphocytes
Oligodendrocytes/glia like cells	Oligodendrocytes/glia like cells
T cells/T memory cells/NK cells	T/NK cells
Macrophages/Monocytes/ Activated fibroblast like cells	Macr/Monocytes/ Activated fibr cells
Lymphocytes/ Neutrophils	Lymphocytes/ Neutrophils
Activated macrophages/immature macrophages	Activated and immature macr
Vascular endothelial cell/Heart endothelial cells	Vascular and heart endothelial cells
Plasma cells/ B cells	Plasma cells/ B cells

Supplemental table 10. DC vs healthy cluster cross-referencing with disease association databases and datasets

	Log2 Fold Change Score	Scaled interactor number	Cluster labels	Gene symbol	Ensembl	Association overall score - OT	Disease association - OT	PubMed Report number	GWAS standardised association score	Description
12	1.9947836466176	7.11894107272351		ANKRD34C	ENSG00000235711			0		ankyrin repeat domain 34C
45	1.53232136042079	5.90689059560852		C9ORF24	ENSG00000164972			1		
49	1.58694124507355	6.06608919045777		CCDC168	ENSG00000175820			0		coiled-coil domain containing 168
86	1.61142971071955	8.59245703726808		EDA2R	ENSG00000131080			2		ectodysplasin A2 receptor
92	1.65761219697529	8.37937836707126		F2RL2	ENSG00000164220			2		coagulation factor II thrombin receptor like 2
116	1.73235007216511	7.5622424242107		GPR85	ENSG00000164604			0		G protein-coupled receptor 85
149	1.60576050583933	7.78135971352466		KLHDC9	ENSG00000162755			0		kelch domain containing 9
153	1.69973389666383	6.37503943134692		LCN12	ENSG00000184925			0		lipocalin 12
156	1.68590224332899	7.49185309632967		LRRC24	ENSG00000254402			0		leucine rich repeat containing 24
191	1.8481036595783	7.18982455888002		NUP82CL	ENSG00000198088			0		nucleoporin 62 C-terminal like
239	1.82647897553963	8.21431912080077		SLC6A12	ENSG00000111811			4		solute carrier family 6 member 12
250	1.57879018657951	7.27612440527424		SPNS3	ENSG00000182557			1		sphingolipid transporter 3 (putative)
267	1.78393398468577	6.4594316186373		TMEM54	ENSG00000121900			0		transmembrane protein 54
189	1.83699438551428	8.76155123244448		NTNG2	ENSG00000196358	0.00620275403100288	Abnormality of the cardiovascular system	2		netrin G2
164	1.56451197201855	8.70390357344466		MDK	ENSG00000110492	0.334581090262546	arterial disorder	69		midkine
58	1.97239361439505	7.800898999203		CHAC1	ENSG00000128965	0.0165	cardiovascular disease	7		ChaC glutathione specific gamma-glutamylcyclotransferase 1
66	1.89302147069627	9.85174904141606		CNR1	ENSG00000118432		1 cardiovascular disease	73		cannabinoid receptor 1
71	1.68076668444065	8.47167521439204		CPEB1	ENSG00000214575	0.206964779293235	cardiovascular disease	5		cytoplasmic polyadenylation element binding protein 1
88	1.50411782671912	10.3106127816595		ENO2	ENSG00000111674	0.0722530203601724	cardiovascular disease	11		enolase 2
138	1.93955070368889	7.62935662007961		IL17D	ENSG00000172458	0.132140069059852	cardiovascular disease	3		interleukin 17D
183	1.50577624877796	9.280770701306		NR0B2	ENSG00000131910	0.190233333333333	cardiovascular disease	5		nuclear receptor subfamily 0 group B member 2
184	1.60739277448007	9.29691626687929		NR4A1	ENSG00000123358	0.0918987337952424	cardiovascular disease	153		nuclear receptor subfamily 4 group A member 1
235	1.70256448962541	7.78818432477693		SIK1	ENSG00000142178	0.971413312942282	cardiovascular disease	31		salt inducible kinase 1
244	1.76305284200499	9.57364718749332		SMAD7	ENSG00000101665		1 cardiovascular disease	290		SMAD family member 7
263	1.69896226113721	7.03342300153745		TM6SF2	ENSG00000213996	0.358231148247113	cardiovascular disease	73		transmembrane 6 superfamily member 2
269	1.84680590375758	9.13442632022093		TNFRSF11B	ENSG00000164761	0.341438608836088	cardiovascular disease	366		TNF receptor superfamily member 11b
278	1.58114171030336	8.0443941935845		UCN	ENSG00000163794	0.0860926975362822	cardiovascular disease	192		urocortin
283	1.83916521322178	8.84235034341381		WNT10B	ENSG00000169884	0.0068	congenital heart disease	20		Wnt family member 10B
105	1.8748706873852	8.98299357469431		FRZB	ENSG00000162998	0.056973109948679	dilated cardiomyopathy	15		frizzled related protein
207	1.90562977070203	6.4594316186373		PPDF	ENSG00000125534	0.0144197024176782	dilated cardiomyopathy	0		pancreatic progenitor cell differentiation and proliferation factor
82	1.55673898054257	6.84549005094437		DNAAF3	ENSG00000167646	0.408333333333333	Familial isolated dilated cardiomyopathy	4		dynein axonemal assembly factor 3
220	1.68793785794021	8.34429590791582		RUNDC3A	ENSG00000108309	0.19824	Familial progressive cardiac conduction defect	0		RUN domain containing 3A
15	1.53327011453769	9.44294349584873		APLP1	ENSG00000105290	0.02772	gastric cardia carcinoma	4		amyloid beta precursor like protein 1
260	1.77613185639517	8.280770701306		TCEAL2	ENSG00000184905	0.0125626104053031	gastric cardia carcinoma	0		transcription elongation factor A like 2
75	1.77534852029847	7.05528243550119		CRISPLD1	ENSG00000212005	0.0252	heart failure	2		cysteine rich secretory protein LCCL domain containing 1
107	1.70087992664471	8.0389189892923		GADD45G	ENSG00000130222	0.0244	heart failure	10		growth arrest and DNA damage inducible gamma
55	1.56302025572538	9.32867492732795		CDKN1C	ENSG00000129757	0.3	Heart murmur	74		cyclin dependent kinase inhibitor 1C
119	1.86957990174048	9.24555270625568		GRIA3	ENSG00000125675	0.7027	hypertension	3		glutamate ionotropic receptor AMPA type subunit 3
232	1.81225534180803	7.71424551766612		SEZ6L2	ENSG00000174938	0.000733126835909663	hypertensive heart disease	1		seizure related 6 homolog like 2
245	1.7841393659965	10.7846348455575		SNCA	ENSG00000145335		1 intrinsic cardiomyopathy	54		synuclein alpha
115	1.60729060356563	8.01680828768855		GPR27	ENSG00000170837	0.00761771373267391	ischemic cardiomyopathy	1		G protein-coupled receptor 27
222	1.82420947931998	7.99435343685886		SCAMP5	ENSG00000198794	0.090288415923715	mean arterial pressure	3		secretory carrier membrane protein 5



	Log2 Fold Change Score	Scaled interactor number	Cluster labels	Gene symbol	Ensembl	Association overall score - OT	Disease association - OT	PubMed Report number	GWAS standardised association score	Description
130	1.96294664806162	8.876516946565		0 HES5	ENSG00000197921	0.04085455555555556	pulmonary arterial hypertension	38		hes family bHLH transcription factor 5
248	1.66032903700229	9.2455270625568		0 SOX17	ENSG00000164736		1 pulmonary arterial hypertension	125		SRY-box transcription factor 17
159	1.59988565205825	7.52356195605701		0 LTC4S	ENSG00000213316	0.141941610723734	resting heart rate	16		leukotriene C4 synthase
29	1.56273663451547	10.2419831496943		0 BMP2	ENSG00000125845		1 short stature, facial dysmorphism, and skeletal anomalies with or without cardiac anomalies	678		bone morphogenetic protein 2
137	1.5027625796532	7.2667865406949		0 IER2	ENSG00000160888	0.507533520460129	venous thromboembolism	2		immediate early response 2
185	1.75250441317526	8.40541146313634		0 NR4A3	ENSG00000119508	0.19224	Glycogen storage disease due to muscle and heart glycogen synthase deficiency	77	0.049508	nuclear receptor subfamily 4 group A member 3
199	1.57076151056769	9.35974956032233		0 PDIA2	ENSG00000185615	0.29202651232481	cardiovascular disease	8	0.049702	protein disulfide isomerase family A member 2
160	1.54629170035045	8.1548181090521		0 LY6E	ENSG00000160932	0.330502962304304	familial cardiomyopathy	4	0.056197	lymphocyte antigen 6 family member E
256	1.93282627668144	8.52356195605701		0 STC1	ENSG00000159167	0.0532	cardiotoxicity	35	0.057459	stanniocalcin 1
290	1.9631556962664	7.65821148275179		0 ZNF365	ENSG00000138311	0.0919308837137536	cardiovascular disease	4	0.058627	zinc finger protein 365
53	1.59645473634056	6.16992500144231		0 CD163L1	ENSG00000177675			1	0.062029	CD163 molecule like 1
223	1.85608019769944	8.2240016741981		0 SCN11A	ENSG00000168356		1 cardiovascular disease	12	0.068964	sodium voltage-gated channel alpha subunit 11
44	1.76968025471781	7.65105169117893		0 C6	ENSG00000203937			2024	0.0699	complement C6
69	1.8171397161506	8.12928301694497		0 COL9A1	ENSG00000112280	0.0199	dilated cardiomyopathy	5	0.085871	collagen type IX alpha 1 chain
146	1.56553436797438	7.73470962022584		0 KCNT2	ENSG00000162687	0.000137947219061734	hypertensive retinopathy	15	0.102913	potassium sodium-activated channel subfamily T member 2
163	1.60416423336695	7.60733031374961		0 MCF2L2	ENSG00000053524			0	0.158409	MCF2 cell line derived transforming sequence-like 2
257	1.67407017460917	9.61654884377899		0 STX1B	ENSG00000099935	0.0640979185700417	hypertension	3	0.30863	syntaxin 1B
282	1.95101798871725	7.4594316186373		0 WDR66	ENSG00000158023	0.000484411379518669	cardiac arrhythmia	0	0.49309	WD repeat domain 66
24	1.63553938521082	7.53915881110803		0 BCAS4	ENSG00000124243			0	0.575039	breast carcinoma amplified sequence 4
251	1.8152214086476	8.38801728534514		0 SPOCK1	ENSG00000152377	0.02352	gastric non-cardia carcinoma	5	1.24288	SPARC (osteonectin), cwcv and kazal like domains proteoglycan 1
31	1.9266385270553	6.95999589242998		0 BRSK1	ENSG00000160469	0.00142081870334878	systolic heart failure	4	1.98052	BR serine/threonine kinase 1
8	1.62070948373321	7.82017896241519		0 AEBP1	ENSG00000106624	0.18258	Familial progressive cardiac conduction defect	30		AE binding protein 1
37	-1.5445275742128	7.32192809488736		1 C1QL1	ENSG00000131094			1		complement C1q like 1
54	-1.8574001880059	8.43462822763673		1 CDCA3	ENSG00000111665			1		cell division cycle associated 3
117	-1.5418882581767	9.03066713624694		1 GPM2	ENSG00000121957			4		G protein signaling modulator 2
125	-1.7415211970995	8.67948009950545		1 GSTT2	ENSG00000099984			9		glutathione S-transferase theta 2 (gene/pseudogene)
179	-1.5596652536575	9.55266909751427		1 NCAPH	ENSG00000121152			0		non-SMC condensin I complex subunit H
205	-1.7840646287159	8.61470984411521		1 POLR2J2	ENSG00000267645			0		RNA polymerase II subunit J2
213	-1.7979100668267	7.6724253419715		1 PYGO1	ENSG00000171016			3		pygopus family PHD finger 1
242	-1.5220892822109	7.94251450533924		1 SLC9C1	ENSG00000172139			1		solute carrier family 9 member C1
264	-1.6548081776658	7.46760550083		1 TMEM132B	ENSG00000139364			0		transmembrane protein 132B
133	-1.6561356842739	7.98868468677217		1 HOOK1	ENSG00000134709	0.183	Aicardi-Goutières syndrome	4		hook microtubule tethering protein 1
94	-1.5384011866924	9.86108690599539		1 FAIM2	ENSG00000135472	0.453020304441452	arterial stiffness measurement	16		Fas apoptotic inhibitory molecule 2
273	-1.5698211048197	9.4093909361377		1 TPX2	ENSG00000088325	0.0244	Arteritis	13		TPX2 microtubule nucleation factor
176	-1.5668621018310	8.15987133677839		1 MYOT	ENSG00000120729		1 cardiomyopathy	32		myotilin
6	-1.8636156397350	9.19967234483636		1 ADRA1B	ENSG00000170214		1 cardiovascular disease	104		adrenoreceptor alpha 1B
84	-1.8625379707147	9.32418054661874		1 DSP	ENSG00000096696		1 cardiovascular disease	522		desmoplakin
93	-1.5248099313577	9.35535109642481		1 FABP4	ENSG00000170323	0.312313303725802	cardiovascular disease	366		fatty acid binding protein 4
113	-1.8372791995641	9.4858293087019		1 GPD1	ENSG00000167588	0.254230226448427	cardiovascular disease	19		glycerol-3-phosphate dehydrogenase 1
204	-1.5386854548307	7.67948009950545		1 PLN	ENSG00000198523		1 cardiovascular disease	558		phospholamban
187	-1.6328402466491	7.443462822763672		1 NSG1	ENSG00000168824	0.050427857786417	congenital heart disease	0		neuronal vesicle trafficking associated 1
131	-1.5572025895760	8.5077946401987		1 HEY2	ENSG00000135547	0.338408265187396	Genetic cardiac anomaly	163		hes related family bHLH transcription factor with YRPW motif 2

	Log2 Fold Change Score	Scaled interactor number	Cluster labels	Gene symbol	Ensembl	Association overall score - OT	Disease association - OT	PubMed Report number	GWAS standardised association score	Description
281	-1.5955711845298 3	8.3037807481771		WDR62	ENSG00000075702	0.293382355555555 6	Genetic cardiac anomaly	5		WD repeat domain 62
208	-1.6438270827103 4	7.79441586635011		PPP1R1A	ENSG00000135447	0.0266	heart failure	2		protein phosphatase 1 regulatory inhibitor subunit 1A
241	-1.6183388708806 4	8.12928301694497		SLC6A9	ENSG00000196517	0.0405	hypertension	9		solute carrier family 6 member 9
165	-1.5090800613611 4	7.05528243550119		MEGF9	ENSG00000106780	0.00385124278181 082	ischemic cardiomyopathy	1		multiple EGF like domains 9
147	-1.6721299701192 4	9.09803208296053		KIAA0754	ENSG00000127603	0.21645329892635 3	peripheral arterial disease	1		KIAA0754
215	-1.5322118890134 2	9.32192809488736		RAPGEF4	ENSG00000091428	0.33226283756803 9	cardiovascular disease	34	0.047135	Rap guanine nucleotide exchange factor 4
56	-1.6542631778053 1	9.21916852046216		CENPF	ENSG00000117724	0.03948535130566 5	heart disease	7	0.047372	centromere protein F
99	-1.7742459974491 1	9.51569983828404		FGF7	ENSG00000140285	0.04447222222222 22	pulmonary arterial hypertension	54	0.048529	fibroblast growth factor 7
287	-1.7530206870971 3	7.69348695749933		XIRP2	ENSG00000163092	0.32604057977100 3	cardiomyopathy	23	0.049661	xin actin binding repeat containing 2
142	-1.5511467630484 1	10.270295326472		ITGB1	ENSG00000150093	0.29604338437263 8	cardiomyopathy	46	0.058312	integrin subunit beta 1
200	-1.8039602204495 6	8.19475685442225		PHACTR3	ENSG000000087495			2	0.087486	phosphatase and actin regulator 3
61	-1.8730668069956 7	9.52552080909507		CHL1	ENSG00000134121	9.23637951231764 e-05	congenital anomaly of cardiovascular system	13	0.191441	cell adhesion molecule L1 like
89	-1.5965866279621 2	9.48179943166575		EPHB1	ENSG00000154928	0.01455555555555 56	heart disease	16	0.505894	EPH receptor B1
20	-1.8650575816900 5	8.13442632022093		ATP6V1C2	ENSG00000143882			0	0.520142	ATPase H <sup>+</sup> -transporting V1 subunit C2
5	-1.5365126493018 1	7.73470962022584		ADAMTS12	ENSG00000151388	0.10959168753079 4	cardiovascular disease	8	0.657565	ADAM metalloproteinase with thrombospondin type 1 motif 12
288	-1.8745808736105 1	8.69348695749933		XRCC4	ENSG00000152422	0.0104	hypertension	16	0.696042	X-ray repair cross complementing 4
210	-1.7599819709811 3	9.64565843240871		PPP2R2C	ENSG00000074211			3	0.827715	protein phosphatase 2 regulatory subunit 2 gamma
276	-1.7067610214954 5	10.0042204663182		TTN	ENSG00000155657		1 cardiovascular disease	612	1.33249	titin
219	6.94802064808872	9.53915881110803		RPS17	ENSG00000182774			6		ribosomal protein S17
271	6.61799049327818	8.54303182025524		TNNI1	ENSG00000159173			29		troponin I1, slow skeletal type
132	10.3077145444924	9.40087943628218		HLA-A	ENSG00000206503	0.57063135542674 9	cardiovascular disease	501		major histocompatibility complex, class I, A
240	5.86144980575564	9.38370429247405		SLC6A4	ENSG00000108576		1 cardiovascular disease	232		solute carrier family 6 member 4
182	5.06890170624402	8.38801728534514		NPPB	ENSG00000120937	0.73362927273845 1	cardiovascular disease biomarker measurement	301		natriuretic peptide B
70	5.89925569841038	8.37068740680722		COMP	ENSG00000105664	0.31294441496913 6	dilated cardiomyopathy	9109		cartilage oligomeric matrix protein
25	5.32107294422745	8		BEX1	ENSG00000133169	0.03491944444444 44	heart disease	5		brain expressed X-linked 1
216	5.54937873653375	7.83289001416474		RHCG	ENSG00000140519	0.26201388888888 9	Infantile hypertrophic cardiomyopathy due to MRPL44 deficiency	30		Rh family C glycoprotein
180	7.29293557064031	9.05528243550119		NGEF	ENSG00000066248			3	0.150671	neuronal guanine nucleotide exchange factor
181	8.04397728581946	8.78135971352466		NPPA	ENSG00000175206		1 cardiovascular disease	437	0.170989	natriuretic peptide A
9	3.65838110652292	8.85174904141606		ALAS2	ENSG00000158578	0.25561338888888 9	cardiomyopathy	22		S-aminolevulinate synthase 2
19	3.9904844305577	7.65821148275179		ATP1B4	ENSG00000101892			1		ATPase Na <sup>+</sup> /K <sup>+</sup> -transporting family member beta 4
59	3.05220716699847	9.8008998999203		CHD5	ENSG00000116254			9		chromodomain helicase DNA binding protein 5
96	4.47462568183603	9.64024493622235		FBXL16	ENSG00000127585			2		F-box and leucine rich repeat protein 16
136	4.27481065712663	7.34872815423108		HYAL4	ENSG00000106302			1		hyaluronidase 4
174	3.54422913321816	9.4093909361377		MYL1	ENSG00000168530			13		myosin light chain 1
188	3.37289023262111	6.83289001416474		NSG2	ENSG00000170091			0		neuronal vesicle trafficking associated 2
217	3.91409704429358	8.47167521439204		RIMS4	ENSG00000101098			0		regulating synaptic membrane exocytosis 4
259	3.82373483620928	7.90689059560852		SYTL5	ENSG00000147041			0		synaptotagmin like 5
270	3.13563868733179	7.49185309632967		TNMD	ENSG00000000005			11		tenomodulin
279	4.85085048889719	8.0389189892923		UNC90	ENSG00000144408			1		unc-80 homolog, NALCN channel complex subunit
286	3.6360341169916	6.5077946401987		XG	ENSG00000124343			105		Xg glycoprotein (Xg blood group)
128	3.92488314747514	8.40087943628218		HBA2	ENSG00000188536	0.2	Aicardi-Goutières syndrome	33		hemoglobin subunit alpha 2
48	3.60375108954229	9.61470984411521		CALCA	ENSG00000110680		1 cardiovascular disease	360		calcitonin related polypeptide alpha
78	4.74954598736633	8.54303182025524		CYP3A5	ENSG00000106258	0.31213193995270 3	cardiovascular disease	820		cytochrome P450 family 3 subfamily A member 5
198	3.25962475538462	8.29462074889163		PDE6A	ENSG00000132915		1 cardiovascular disease	2		phosphodiesterase 6A

	Log2 Fold Change Score	Scaled interactor number	Cluster labels	Gene symbol	Ensembl	Association overall score - OT	Disease association - OT	PubMed Report number	GWAS standardised association score	Description
284	4.23206856642375	9.13955135239879		WNT9A	ENSG00000143816	0.761991770811379	cardiovascular disease		9	Wnt family member 9A
177	3.01403702527681	7.97727992349992		MYOZ1	ENSG00000177791	0.192951666666667	dilated cardiomyopathy		13	myozenin 1
129	4.81030163762285	9.53138146051631		HBB	ENSG00000244734	0.0270190913981237	gastric cardia carcinoma		79	hemoglobin subunit beta
77	3.0347905629289	8.96000193206808		CYP11A1	ENSG00000140459	0.202988212018141	hypertension		103	cytochrome P450 family 11 subfamily A member 1
63	3.56971747678817	8.08214904135387		CHRNE	ENSG00000108556		1 intracranial hypertension		3	cholinergic receptor nicotinic epsilon subunit
4	5.19892875253399	6.3037807481771		ADAM18	ENSG00000168619	5.9576943378641e-05	Polyarteritis Nodosa		0	ADAM metalloproteinase domain 18
21	3.18051500085422	8.01122725542325		ATRNL1	ENSG00000107518	0.00231788439106461	Paroxysmal supraventricular tachycardia		2	attractin like 1
237	3.66291878761574	8.00562454919388		SLC16A9	ENSG00000165449	0.00209062141407026	Arterial stenosis		8	solute carrier family 16 member 9
17	3.87318554341421	7.14974711950468		AOP10	ENSG00000143595				4	aquaporin 10
127	4.22203194733031	8.40514146313634		HBA1	ENSG00000206172	0.2	Aicardi-Goutières syndrome		9105	hemoglobin subunit alpha 1
209	3.31598118062034	9.8073549220576		PPP2R2B	ENSG00000156475	0.303587001569456	heart disease		5	protein phosphatase 2 regulatory subunit Bbeta
186	3.49986748217821	9.43879185257826		NRG1	ENSG00000157168	0.838868237015398	cardiovascular disease		202	neuregulin 1
201	2.98945841800273	8.92184093707449		PHF21B	ENSG00000056487				0	PHD finger protein 21B
67	2.97181911688648	8.11374216604919		COL22A1	ENSG00000169436				6	collagen type XXII alpha 1 chain
231	3.42460515913005	8.89784545600551		SEZL	ENSG00000100095	0.204944580793381	arterial stiffness measurement		3	seizure related 6 homolog like
16	3.74908488063969	6.71424551766612		ARMS2	ENSG00000254636	0.161172108916687	cardiovascular disease		37	age-related maculopathy susceptibility 2
62	3.34234394901801	8.97441458980553		CHRNA3	ENSG00000080644		1 cardiovascular disease		36	cholinergic receptor nicotinic alpha 3 subunit
22	-1.9229040243932	7.33091687811462		B3GALT2	ENSG00000162630				0	beta-1,3-galactosyltransferase 2
64	-2.66857878450178	8.25266543245025		CKAP2L	ENSG00000169607				1	cytoskeleton associated protein 2 like
80	-2.03196590170062	7.15987133677839		DISP2	ENSG00000140323				0	dispatched RND transporter family member 2
106	-1.86364800039181	7.79441586635011		GOS2	ENSG00000123689				26	GO/G1 switch 2
118	-2.54656789271846	8.51569983828404		GRB7	ENSG00000141738				8	growth factor receptor bound protein 7
144	-2.5990517633141	8.01680828768655		KANK4	ENSG00000132854				2	KN motif and ankyrin repeat domains 4
150	-2.12376787091781	6.78135971352466		KLHL32	ENSG00000186231				1	kelch like family member 32
154	-2.17489552183044	7.48381577726426		LINC00842	ENSG00000285294				0	long intergenic non-protein coding RNA 842
157	-2.31558667317575	8.3264294871223		LRRN3	ENSG00000173114				6	leucine rich repeat neuronal 3
202	-2.3214367225808	7.24792751344359		PH15	ENSG00000137558				86	peptidase inhibitor 15
226	-1.89766442224553	7.56985560833095		SEC14L5	ENSG00000103184				0	SEC14 like lipid binding 5
230	-2.00366439686964	6.8073549220576		SERTM1	ENSG00000180440				1	serine rich and transmembrane domain containing 1
258	-2.28760231530231	8.6582114827518		SYT13	ENSG0000019505				0	synaptotagmin 13
274	-1.9525022480234	8.29462074889163		TROAP	ENSG00000135451				0	trophinin associated protein
152	-1.89972852855174	8.61102479730735		LAMB3	ENSG00000196878	0.0696446821093559	arterial stiffness measurement		6	laminin subunit beta 3
167	-2.44009141242053	7.98953364497036		MKI67	ENSG00000148773	0.845973253250122	arterial stiffness measurement		958	marker of proliferation Ki-67
266	-1.6770088230723	6.39231742277876		TMEM40	ENSG00000088726	0.0814459696412086	cardiac edema		1	transmembrane protein 40
151	-2.78951596912917	7.79441586635011		LAD1	ENSG00000159166	0.283405423164368	cardiac troponin T measurement		8	ladinin 1
74	-2.60082059890161	8.49185309632967		CRHR2	ENSG00000106113	0.31575403368998	cardiovascular disease		27	corticotropin releasing hormone receptor 2
108	-2.808233069868	9.0389189892923		GATA5	ENSG00000130700		1 cardiovascular disease		120	GATA binding protein 5
218	-2.57098962677655	8.74146698640115		RNASE2	ENSG00000169385	0.031424216014505	cardiovascular disease		2	ribonuclease A family member 2
226	-2.23762020329757	8.73470962022584		SERPINA3	ENSG00000196136	0.214303720238637	cardiovascular disease		29	serpin family A member 3
254	-1.93324735234705	9.05528243550119		SSTR2	ENSG00000180616		1 cardiovascular disease		52	somatostatin receptor 2
114	-1.98316150997236	7.79441586635011		GPR22	ENSG00000172209	0.263732373714447	coronary artery disease		6	G protein-coupled receptor 22
227	-2.04073010887368	6.53915881110803		SERF1B	ENSG00000205572	0.1986	coronary artery disease, autosomal dominant 2		0	small EDRK-rich factor 1B
145	-2.18525684782395	8.60362634498619		KCNA7	ENSG00000104848	0.0334	Familial progressive cardiac conduction defect		6	potassium voltage-gated channel subfamily A member 7
272	-2.22496454592126	10.7047682393626		TOP2A	ENSG00000131747	0.03168	gastric cardia carcinoma		153	DNA topoisomerase II alpha

	Log2 Fold Change Score	Scaled interactor number	Cluster labels	Gene symbol	Ensembl	Association overall score - OT	Disease association - OT	PubMed Report number	GWAS standardised association score	Description	
97	-2.38015102562284	8.18487534290828		FCGBP	ENSG00000275395	0.0303797936422462	gastric non-cardia carcinoma		5	Fc fragment of IgG binding protein	
175	-2.2935788301116	9.23840473932508		MYL7	ENSG00000106631	0.938724436958744	heart failure		46	myosin light chain 7	
148	-2.29905103303797	9.72451385311995		KIF20A	ENSG00000112984	0.220411950722337	heart rate		9	kinesin family member 20A	
229	-2.24190675587965	8.51569983828404		SERPINA5	ENSG00000188488	0.032291	hypertension		7	serpin family A member 5	
195	-1.97481342706637	7.29462074889163		PCDH20	ENSG00000280165	0.0136	portal hypertension		2	protocadherin 20	
288	-2.12930974330921	7.34872815423108		TMEM63C	ENSG00000165548				2	0.046791	transmembrane protein 63C
166	-2.27392595497609	7.66533591718518		MFAP2	ENSG00000117122	0.1864	Aicardi-Goutières syndrome		6	0.048463	microfibril associated protein 2
211	-1.94091022315343		6	PRELID2	ENSG00000186314				2	0.048756	PREL domain containing 2
32	-2.64421120974753	10.1774195379892		BUB1B	ENSG00000156970	0.18596	Aicardi syndrome		13	0.050256	BUB1 mitotic checkpoint serine/threonine kinase B
280	-1.88751569143968	6.10852445677817		VASH2	ENSG00000143494	0.0308	pulmonary arterial hypertension		12	0.052454	vasohibin 2
275	-2.75168652690408	8.40514146313634		TRPC4	ENSG00000133107	0.0710127488599104	cardiovascular disease		155	0.053665	transient receptor potential cation channel subfamily C member 4
13	-2.05861995776965	9.27612440527424		ANLN	ENSG00000011426	0.0142	cardiovascular disease		6	0.056363	anillin actin binding protein
162	-2.08092916173716	7.59245703726908		MARVELD2	ENSG00000152939	0.0173777777777778	heart disease		6	0.057293	MARVEL domain containing 2
98	-2.59893307703872	7.10852445677817		FER1L6	ENSG00000214814				1	0.060283	fer-1 like family member 6
60	-2.70078934377649	8.24317398947295		CHDH	ENSG00000016391		1 cardiovascular disease		8	0.063677	choline dehydrogenase
190	-2.49764189851206	9.38154295118458		NUF2	ENSG00000143228				0	0.079543	NUF2 component of NDC80 kinetochore complex
238	-2.49021829905967	8.09803208296053		SLC38A4	ENSG00000139209				9	0.109267	solute carrier family 38 member 4
291	-2.57558296004614	7.89481776330794		ZNF385B	ENSG00000144331	0.873897848086887	cardiovascular disease		2	0.179581	zinc finger protein 385B
109	-2.27486252428201	8.18487534290828		GDA	ENSG00000119125	0.0637290136054422	myocardial infarction		562	0.501181	guanine deaminase
143	-2.45296876557447	8.9915218460757		ITGB6	ENSG00000115221	0.772906363010406	arterial stiffness measurement		11	0.503822	integrin subunit beta 6
120	-2.79425142739309	9.99859042974533		GRIN2A	ENSG00000183454		1 heart disease		8	0.528284	glutamate ionotropic receptor NMDA type subunit 2A
197	-2.41017980902592	8.1548181090521		PDE11A	ENSG00000128655	0.0104	pulmonary arterial hypertension		16	0.556685	phosphodiesterase 11A
122	-2.90845328901975	8.8548883826024		GRIP1	ENSG00000155974	0.2	Congenital vertebral-cardiac-renal anomalies syndrome		24	0.621951	glutamate receptor interacting protein 1
27	-2.30912583234393	10.1459321458205		BIRC5	ENSG00000089685	0.0139	hypertension		232	0.634612	baculoviral IAP repeat containing 5
285	-2.14266878260193	8.65105169117893		WWC1	ENSG00000113645				13	0.663947	WW and C2 domain containing 1
91	-2.00474207730438	8.85798099512757		EYA4	ENSG00000112319		1 heart disease		20	0.677699	EYA transcriptional coactivator and phosphatase 4
121	-1.91926708873315	8.61838550225861		GRIN3A	ENSG00000198785		1 Abnormality of cardiovascular system morphology		7	1.38272	glutamate ionotropic receptor NMDA type subunit 3A
168	-2.18556700814648	8.81057163474115		MLXIPL	ENSG00000099950	0.519043938998994	cardiovascular disease		66	1.74508	MLX interacting protein like
34	2.45333728185443	4.16992500144231		C19ORF81	ENSG00000235034				0		
43	1.61832055802933	5.39231742277876		C2ORF27A	ENSG00000197927				0		
249	1.51789304332459	5.08746284125034		SPATL8	ENSG00000106686				0		spermatogenesis associated 6 like
289	1.52813344556675	5.12928301694497		ZMYND12	ENSG00000066185				0		zinc finger MYND-type containing 12
292	2.12860078365148	5.08746284125034		ZNF385C	ENSG00000187595				0		zinc finger protein 385C
161	1.91129556374842		5	MAP3K7CL	ENSG00000156265	0.0104	coronary artery disease		3	0.673055	MAP3K7 C-terminal like
100	-2.63735489265179	6.61470984411521		FIBCD1	ENSG00000130720				0		fibrinogen C domain containing 1
255	-3.11590797652312	6.4757334309664		STAC2	ENSG00000141750				2		SH3 and cysteine rich domain 2
261	-2.7262271714906	5.28540221896225		TCF24	ENSG00000261787				0		transcription factor 24
30	-3.63044006169649	9.84862294042934		BMP7	ENSG00000101144	0.327475610575549	cardiovascular disease		188		bone morphogenetic protein 7
50	-3.2339108293811	9.04439411935845		CCL11	ENSG00000172156	0.0958047192835306	cardiovascular disease		186		C-C motif chemokine ligand 11
155	-3.3933052613818	8.04984854945056		LIPG	ENSG00000101670	0.310414855335407	cardiovascular disease		158		lipase G, endothelial type
252	-4.36447923961643	10.1910592145317		SPP1	ENSG00000118785	0.333324605482535	cardiovascular disease		603		secreted phosphoprotein 1
134	-3.44914367205226	8.20945336562895		HOPX	ENSG00000171476	0.295285022222222	familial cardiomyopathy		29		HOP homeobox
36	-2.90303076300225	6.84549005094437		C1ORF116	ENSG00000182795	0.02744	gastric non-cardia carcinoma		0		
39	-2.42777400564184	5.70043971814109		C1QTNF9	ENSG00000240854	0.2	Infantile hypertrophic cardiomyopathy due to MIRPL44 deficiency		4		C1q and TNF related 9

	Log2 Fold Change Score	Scaled interactor number	Cluster labels	Gene symbol	Ensembl	Association overall score - OT	Disease association - OT	PubMed Report number	GWAS standardised association score	Description
40	-2.42777400564184	5.70043971814109		C1QTNF9	ENSG00000240654	0.2	Infantile hypertrophic cardiomyopathy due to MRPL44 deficiency	4		C1q and TNF related 9
41	-2.42777400564184	5.70043971814109		C1QTNF9	ENSG00000240654	0.2	Infantile hypertrophic cardiomyopathy due to MRPL44 deficiency	4		C1q and TNF related 9
72	-3.0524926269486	8.31288295528435		CPLX3	ENSG00000213578	0.281049307901412	mean arterial pressure	5		complexin 3
126	-3.19574736169273	8.67948009950545		GSTT2B	ENSG00000133433	0.0116	primary hypertension	2		glutathione S-transferase theta 2B (gene/ pseudogene)
104	-3.48989011041154	10.1510165388922		FOXM1	ENSG00000111205	0.224023888888889	pulmonary arterial hypertension	74		forkhead box M1
139	-2.84441655394604	7.55458885167764		IL17RB	ENSG00000056736	0.29761689739624	cardiovascular disease	7	0.09034	interleukin 17 receptor B
47	-3.73091097283578	9.3037807481771		CACNA1E	ENSG00000198216		1 cardiovascular disease	18	0.096871	calcium voltage-gated channel subunit alpha 1 E
7	-3.24302583198899	9.24317398347295		ADRB1	ENSG00000043591		1 cardiovascular disease	3218	1.46277	adrenoceptor beta 1
3	-2.1723988037922	8.66533591718518		ACKR2	ENSG00000144648			3		atypical chemokine receptor 2
11	2.33159459960859	7.73470962022584		ANKRD24	ENSG00000089847			0		ankyrin repeat domain 24
23	2.8390579645635	7.85798099512757		BAAT	ENSG00000136881			11		bile acid-CoA:amino acid N-acyltransferase
26	2.83410764057654	7.82654848729092		BEX2	ENSG00000133134			2		brain expressed X-linked 2
28	2.25819152816438	8.13442632022093		BIRC7	ENSG00000101197			2		baculoviral IAP repeat containing 7
33	2.7882117058921	5.28540221886225		C16ORF89	ENSG00000153446			0		
68	2.04370806561401	8.10328780841202		COL8A2	ENSG00000171812			9		collagen type VIII alpha 2 chain
85	2.45912929047215	8.15987133677839		DUSP15	ENSG00000149599			1		dual specificity phosphatase 15
95	2.69672465800296	6.06608919045777		FAM133A	ENSG00000179083			1		family with sequence similarity 133 member A
112	2.06856921948313	9.35974956032233		GNG8	ENSG00000167414			1		G protein subunit gamma 8
123	2.18658728720147	9.08480838780436		GRM2	ENSG00000164082			3		glutamate metabotropic receptor 2
135	2.70496590387546	7.800899999203		HS6ST2	ENSG00000171004			3		heparan sulfate 6-O-sulfotransferase 2
140	2.04211140038593	8.3619437733524		IRX2	ENSG00000170561			8		iroquois homeobox 2
141	2.32368296222286	8.61102479730735		IRX6	ENSG00000159387			3		iroquois homeobox 6
192	2.15994536225414	7.43462822763672		NXNL1	ENSG00000171773			0		nucleoredoxin like 1
196	2.07145169757637	6.85798099512757		PCDHAC2	ENSG00000243232			0		protocadherin alpha subfamily C, 2
214	2.08843489809424	6.44294349584873		RADX	ENSG00000147231			0		RPA1 related single stranded DNA binding protein, X-linked
225	2.64071738403249	8.49984588708321		SDSL	ENSG00000139410			8		serine dehydratase like
234	2.45340214703444	7.11894107272351		SHISA2	ENSG00000180730			1		shisa family member 2
243	2.77407803349657	7.14974711950468		SLITRK4	ENSG00000179542			1		SLIT and NTRK like family member 4
247	1.99842293442822	8.08748284125034		SOX15	ENSG00000129194			3		SRV-box transcription factor 15
65	2.74213431922183	8.85798099512757		CLEC12A	ENSG00000172322	0.0172	Cardiofaciocutaneous syndrome	2		C-type lectin domain family 12 member A
158	2.14758188141638	7.49183309632967		LTPB4	ENSG00000090006	0.59443280361179	cardiomyopathy	28		latent transforming growth factor beta binding protein 4
14	2.60318351020564	8.16992500144231		AOC1	ENSG00000002726	0.0569439735818656	cardiovascular disease	1		amine oxidase copper containing 1
38	2.23835982235031	7.5077946401987		C1QTNF4	ENSG00000172247	0.0915705003619194	cardiovascular disease	0		C1q and TNF related 4
51	2.61231268356798	9.91288933622996		CCN2	ENSG00000118523	0.911169257585898	cardiovascular disease	512		cellular communication network factor 2
87	2.32732201943461	10.4356702609366		EGR1	ENSG00000120738	0.326228649622181	cardiovascular disease	473		early growth response 1
111	2.52619359939856	9.04984854945056		GDF15	ENSG00000130513	0.31018284691744	cardiovascular disease	357		growth differentiation factor 15
233	2.78001529946862	8.63299519714296		SFRP4	ENSG00000106483	0.0643642029123344	cardiovascular disease	40		secreted frizzled related protein 4
262	2.07884254987272	10.4136279290242		TLR9	ENSG00000239732	0.732929979867855	cardiovascular disease	262		toll like receptor 9
224	2.02490336151523	7.74819284958946		SCUBE2	ENSG00000175356		1 Cerebral arteriovenous malformation	12		signal peptide, CUB domain and EGF like domain containing 2
46	2.26253496058751	8.4262947547021		CA3	ENSG00000164879	0.0537994830033932	dilated cardiomyopathy	555		carbonic anhydrase 3
171	2.43136020592352	7.23840473932508		MSS51	ENSG00000166343	0.249386413342927	dilated cardiomyopathy	1		MSS51 mitochondrial translational activator
178	2.79780422637274	9.06608919045777		NAP1L3	ENSG00000186310	0.00912202178593987	dilated cardiomyopathy	1		nucleosome assembly protein 1 like 3
265	2.01088731554769	7.82017896241519		TMEM30B	ENSG00000182107	0.02548	gastric non-cardia carcinoma	1		transmembrane protein 30B

	Log2 Fold Change Score	Scaled interactor number	Cluster labels	Gene symbol	Ensembl	Association overall score - OT	Disease association - OT	PubMed Report number	GWAS standardised association score	Description
57	2.33247392487416	8.5468945988764		CERS1	ENSG00000223802	1	Genetic cardiac anomaly		5	ceramide synthase 1
103	2.26346939129182	9.17741953798924		FOSB	ENSG00000125740	0.19224	Glycogen storage disease due to muscle and heart glycogen synthase deficiency		87	FosB proto-oncogene, AP-1 transcription factor subunit
203	2.3893360816303	7.06608919045777		PI16	ENSG00000164530	0.0235721295175251	heart disease		113	peptidase inhibitor 16
277	2.73795505232141	9.55458885167764		UCHL1	ENSG00000154277	0.0541336727727073	heart disease		111	ubiquitin C-terminal hydrolase L1
193	2.83256720412549	8.91587937883577		OGDHL	ENSG00000197444	0.018	heart failure		6	oxoglutarate dehydrogenase like
81	2.68794201573077	9.19475685442225		DMC1	ENSG00000100206	0.0720431581139565	heart rate		8	DNA meiotic recombinase 1
79	2.15056200073272	7.65105169117893		DACT2	ENSG00000164488	0.0152	heart valve disease		2	dishevelled binding antagonist of beta catenin 2
283	2.02242252664875	7.68650052718322		SRCIN1	ENSG00000277363	0.115897536277771	hypertension		3	SRC kinase signaling inhibitor 1
110	2.28628746205277	8.48784003382305		GDF10	ENSG0000026524	0.486249446868896	mean arterial pressure		6	growth differentiation factor 10
52	2.14900226103209	9.94544383637791		CCNA1	ENSG00000133101	0.0104	pulmonary arterial hypertension		8	cyclin A1
10	2.63243069266142	8.37068740680722		AMHR2	ENSG00000135409	0.227225	X-linked intellectual disability - cardiomegaly - congestive heart failure		8	anti-Mullerian hormone receptor type 2
169	2.0977751395693	7.8703647195834		MMP24	ENSG00000125966	0.0856845263235322	venous thromboembolism		6	0.046945 matrix metalloproteinase 24
101	2.2208162398181	8.9971794803762		FMOD	ENSG00000122176	0.048	heart failure		20	0.047068 fibromodulin
236	2.5068639956886	7.5622424242107		SLC16A6	ENSG00000108932				1	0.053712 solute carrier family 16 member 6
246	2.43023639207744	7.12928301694497		SOHLH2	ENSG00000120669				0	0.056217 spermatogenesis and oogenesis specific basic helix-loop-helix 2
76	2.51277615136713	8.25266543245025		CRTAC1	ENSG00000095713	5.55463753426618e-05	systolic heart failure		1	0.067116 cartilage acidic protein 1
170	2.56701185551708	7.25738784269265		MRAP2	ENSG00000135324	0.2	Infantile hypertrophic cardiomyopathy due to MRPL44 deficiency		3	0.078202 melanocortin 2 receptor accessory protein 2
102	2.03193300733518	8.24317398347295		FNDC1	ENSG00000164894	0.00365269911261096	ischemic cardiomyopathy		5	0.078697 fibronectin type III domain containing 1
194	2.71824192123292	7.71424551766612		P3H2	ENSG00000090530	0.0101561530771077	pulmonary arterial hypertension		2	0.093045 prolyl 3-hydroxylase 2
212	2.5594447978589	6.39231742277876		PRR7	ENSG00000131188				0	0.095411 proline rich 7, synaptic
206	2.668409880871	7.92481250360578		POU6F2	ENSG00000106536				0	0.147318 POU class 6 homeobox 2
73	2.4984340890349	9.04712391211403		CRB1	ENSG00000134376	1	Pigmented paravenous retinochoroidal atrophy		15	0.152504 crumbs cell polarity complex component 1
2	2.49452675179688	9.68999797141945		ABCG2	ENSG00000118777	0.314731394121144	cardiovascular disease		501	0.33998 ATP binding cassette subfamily G member 2 (Junior blood group)
83	2.65151226952674	8.57364718749332		DNAH6	ENSG00000115423				3	0.669849 dynein axonemal heavy chain 6
1	2.07462783317426	9.64745842645492		ABCC8	ENSG00000006071	1	cardiovascular disease		170	0.820721 ATP binding cassette subfamily C member 8
221	1.98246051844236	9.01402047031493		RYR3	ENSG00000198838	0.717292547225952	arterial stiffness measurement		3292	1.19919 ryanodine receptor 3
16	2.5112833068297	9.90989308377004		APOA1	ENSG00000118137	1	cardiovascular disease		1162	1.67707 apolipoprotein A1
42	-1.801558816913	1.58496250072116		C20ORF202	ENSG00000215595				0	
124	-5.34677772735692	7.10852445677817		GRXCR2	ENSG00000204928				0	glutaredoxin and cysteine rich domain containing 2
172	-3.61592793709147	5.16992500144231		MTRNR2L1	ENSG00000256618				1	MT-RNR2 like 1
90	-5.43015727543702	8.37068740680722		EREG	ENSG00000124882	0.615417242050171	coronary artery calcification		23	epiregulin
35	-2.89595079904387	3.4594316186373		C10RF105	ENSG00000180999	0.00491658424315499	dilated cardiomyopathy		1	
173	-6.70828434364734	10.0953970227926		MYH6	ENSG00000197616	1	heart disease		329	0.386648 myosin heavy chain 6

Supplementary table 11. IC vs healthy cluster cross-referencing with disease association databases and datasets

Log2 Fold Change Score	Scaled interactor number	Cluster labels	Gene symbol	Ensembl	Association overall score - OT	Disease association - OT	PubMed Report number for a gene in the context of any cardiovascular indication	GWAS standardised association score for cardiovascular indication	Description
8.58916673230264	9.40087943628218	0	HLA-A	ENSG00000206503	0.570631355426749	cardiovascular disease	501		major histocompatibility complex, class I, A
6.44507620464629	9.20945336562895	0	HLA-C	ENSG00000204525			162	0.047144	major histocompatibility complex, class I, C
6.02101846831308	9.05528243550119	0	NGEF	ENSG00000066248			3	0.150671	neuronal guanine nucleotide exchange factor
-3.4115257065987	7.79441586635011	1	LAD1	ENSG00000159166	0.283405423164368	cardiac troponin T measurement	8		ladinin 1
-3.83530026412176	8.74146698640115	1	RNASE2	ENSG00000169385	0.031424216014505	cardiovascular disease	2		ribonuclease A family member 2
-2.39586391839252	6.84549005094437	1	C1ORF116	ENSG00000182795	0.02744	gastric non-cardia carcinoma	0		Chromosome 1 Open Reading Frame 116
-3.06294123793731	6.61470984411521	1	FIBCD1	ENSG00000130720			0		fibrinogen C domain containing 1
-2.29768583723687	6.61470984411521	1	GMNC	ENSG00000205835			1		geminin coiled-coil domain containing
-3.65179296086502	7.10852445677817	1	GRXCR2	ENSG00000204928			0		glutaredoxin and cysteine rich domain containing 2
-2.24353913401625	6.78135971352466	1	KLHL32	ENSG00000186231			1		kelch like family member 32
-2.44161744942345	5.90689059560852	1	PRR32	ENSG00000183631			0		proline rich 32
-3.02538408812299	6.4757334309864	1	STAC2	ENSG00000141750			2		SH3 and cysteine rich domain 2
-3.40486264488417	7.10852445677817	1	FER1L6	ENSG00000214814			1	0.060283	fer-1 like family member 6
-4.01495983096413	7.55458885167764	1	IL17RB	ENSG00000056736	0.29761689739624	cardiovascular disease	7	0.09034	interleukin 17 receptor B
1.77825191279389	7.2667865406949	2	IER2	ENSG00000160888	0.507533520460129	venous thromboembolism	2		immediate early response 2
1.98073116652793	6.84549005094437	2	DNAAF3	ENSG00000167646	0.408333333333333	Familial isolated dilated cardiomyopathy	4		dynein axonemal assembly factor 3
1.78389349940921	7.03342300153745	2	TM6SF2	ENSG00000213996	0.358231148247113	cardiovascular disease	73		transmembrane 6 superfamily member 2
1.87633203648745	9.13442632022093	2	TNFRSF11B	ENSG00000164761	0.341438608836088	cardiovascular disease	366		TNF receptor superfamily member 11b
1.67178546711004	8.70390357344466	2	MDK	ENSG00000110492	0.334581090262546	arterial disorder	69		midkine
1.84864059104871	7.68650052718322	2	SRIN1	ENSG00000277363	0.115897536277771	hypertension	3		SRC kinase signaling inhibitor 1
1.82447862050096	7.99435343685886	2	SCAMP5	ENSG00000198794	0.0990288415923715	mean arterial pressure	3		secretory carrier membrane protein 5
1.7580858151407	8.04439411935845	2	UCN	ENSG00000163794	0.0860926975362822	cardiovascular disease	192		urocortin
1.65649608205296	8.98299357469431	2	FRZB	ENSG00000162998	0.056973109948679	dilated cardiomyopathy	15		frizzled related protein
1.63540918814715	9.55458885167764	2	UCHL1	ENSG00000154277	0.054133672727073	heart disease	111		ubiquitin C-terminal hydrolase L1
1.82954960414151	8.876516946565	2	HESS	ENSG00000197921	0.040654555555555	pulmonary arterial hypertension	38		hes family bHLH transcription factor 5
1.8775345761873	7.82017896241519	2	TMEM30B	ENSG00000182107	0.02548	gastric non-cardia carcinoma	1		transmembrane protein 30B
1.870710784994	7.05528243550119	2	CRISPLD1	ENSG00000121005	0.0252	heart failure	2		cysteine rich secretory protein LCCL domain containing 1
1.63701588887323	7.06608919045777	2	PI16	ENSG00000164530	0.0235721295175251	heart disease	113		peptidase inhibitor 16
1.74734220905837	6.4594316186373	2	PPDPF	ENSG00000125534	0.0144197024176782	dilated cardiomyopathy	0		pancreatic progenitor cell differentiation and proliferation factor
1.77564128003123	8.76155123244448	2	NTNG2	ENSG00000196358	0.00620275403100288	Abnormality of the cardiovascular system	2		netrin G2
1.83006456780087	7.71424551766612	2	SEZ6L2	ENSG00000174938	0.000733126835909663	hypertensive heart disease	1		seizure related 6 homolog like 2
1.7036467612626	8.74483383749955	2	ADAM8	ENSG00000151651			19		ADAM metallopeptidase domain 8
1.51151884766014	8.85486838326024	2	CCR10	ENSG00000184451			16		C-C motif chemokine receptor 10
1.70481177683683	7.17990909001493	2	C2CD4B	ENSG00000205502			5		C2 calcium dependent domain containing 4B
1.57341687739654	8.08214904135387	2	CTRL	ENSG00000141086			507		chymotrypsin like
1.51445670155185	6.49185309632967	2	CSKMT	ENSG00000214756			0		citrate synthase lysine methyltransferase
1.60166024187191	8.67595703294175	2	CLDN5	ENSG00000184113			335		claudin 5

Log2 Fold Change Score	Scaled interactor number	Cluster labels	Gene symbol	Ensembl	Association overall score - OT	Disease association - OT	PubMed Report number for a gene in the context of any cardiovascular indication	GWAS standardised association score for cardiovascular indication	Description
1.61966432520463	9.64745842645492		CISH	ENSG00000114737			20		cytokine inducible SH2 containing protein
1.88669814815387	8.15987133677839		DUSP15	ENSG00000149599			1		dual specificity phosphatase 15
1.83107874393234	8.57742882803575		DUSP2	ENSG00000158050			30		dual specificity phosphatase 2
1.66283492831171	9.71596199025514		EGR2	ENSG00000122877			43		early growth response 2
1.61180157396318	7.98299357469431		ESM1	ENSG00000164283			134		endothelial cell specific molecule 1
1.93987798162356	6.06608919045777		FAM133A	ENSG00000179083			1		family with sequence similarity 133 member A
1.77496392805853	7.83289001416474		GJC2	ENSG00000198835			6		gap junction protein gamma 2
1.90318060375825	7.20945336562895		HSH2D	ENSG00000196684			6		hematopoietic SH2 domain containing
1.65385281506948	7.27612440527424		HAPLN3	ENSG00000140511			4		hyaluronan and proteoglycan link protein 3
1.75563330268562	8.49585502688717		HAPLN4	ENSG00000187664			0		hyaluronan and proteoglycan link protein 4
1.57265175225769	7.78135971352466		KLHDC9	ENSG00000162755			0		kelch domain containing 9
1.8598995760642	6.61470984411521		LIME1	ENSG00000203896			1		Lck interacting transmembrane adaptor 1
1.957991175897	7.49185309632967		LRRC24	ENSG00000254402			0		leucine rich repeat containing 24
1.6882853556259	7.3037807481771		METRN	ENSG00000103260			2		metastasin, glial cell differentiation regulator
1.59882727818822	7.34872815423108		NXPH4	ENSG00000182379			1		neurexophilin 4
1.56449154233014	8.6724253419715		NME3	ENSG00000103024			1		NME/NM23 nucleoside diphosphate kinase 3
1.57889743289431	7.10852445677817		NUAK2	ENSG00000163545			6		NUAK family kinase 2
1.56100690338488	7.18982455888002		NUP62CL	ENSG00000198088			0		nucleoporin 62 C-terminal like
1.9472723532423	7.14974711950468		ODF3B	ENSG00000177989			0		outer dense fiber of sperm tails 3B
1.92195119273719	8.96289600533726		PIM2	ENSG00000102096			16		Pim-2 proto-oncogene, serine/threonine kinase
1.7049732246303	9.07681559705083		PSD	ENSG00000059915			736		pleckstrin and Sec7 domain containing
1.94091146411242	7.64385618977472		PTPRCAP	ENSG00000213402			1		protein tyrosine phosphatase receptor type C associated protein
1.65752923832695	7.62935662007961		RAB26	ENSG00000167964			2		RAB26, member RAS oncogene family
1.81668934398203	8.96866679319521		RAB39B	ENSG00000155961			5		RAB39B, member RAS oncogene family
1.59791485332862	9.06069593168755		RTN4R	ENSG00000040608			25		reticulon 4 receptor
1.81250275355728	7.83289001416474		RARRES2	ENSG00000106538			29		retinoic acid receptor responder 2
1.64741023238842	7.94836723158468		RIPOR2	ENSG00000111913			0		RHO family interacting cell polarization regulator 2
1.66211684567822	8.67948009950545		SEMA4A	ENSG00000196189			11		semaphorin 4A
1.53943855669945	6.04439411935845		SAP25	ENSG00000205307			0		Sh3A associated protein 25
1.9584767133935	6.85798099512757		SLC44A5	ENSG00000137968			1		solute carrier family 44 member 5
1.56111051002513	8.21431912080077		SLC6A12	ENSG00000111181			4		solute carrier family 6 member 12
1.94405411985434	7.27612440527424		SPNS3	ENSG00000182557			1		sphingolipid transporter 3 (putative)
1.73929790537712	8.08746284125034		SOX15	ENSG00000129194			3		SRF-box transcription factor 15
1.51780876824547	9.56605403817109		SOD3	ENSG00000109610			188		superoxide dismutase 3
1.59158568931887	7.94251450533924		SNAP47	ENSG00000143740			1		synaptosome associated protein 47
1.59908215942757	8.8703647195834		TAS1R3	ENSG00000169962			6		taste 1 receptor member 3
1.75252826177825	8.96000193206806		TNFRSF4	ENSG00000186827			55		TNF receptor superfamily member 4
1.60440909460327	7.56985560833095		TLL2	ENSG00000095587			3		tolloid like 2
1.50000332104457	7.62205181945638		TMC8	ENSG00000167895			0		transmembrane channel like 8



Log2 Fold Change Score	Scaled interactor number	Cluster labels	Gene symbol	Ensembl	Association overall score - OT	Disease association - OT	PubMed Report number for a gene in the context of any cardiovascular indication	GWAS standardised association score for cardiovascular indication	Description
1.69140609297117	7.70043971814109		2 TMEM160	ENSG00000130748			0		transmembrane protein 160
1.57543206088795	8.30833903013941		2 TFF3	ENSG00000160180			38		trefoil factor 3
1.60057601039814	7.5698560833095		2 YJEFN3	ENSG00000250067			1		YjeF N-terminal domain containing 3
1.63967837037471	8.21916852046216		2 ZMYND15	ENSG00000141497			0		zinc finger MYND-type containing 15
1.51580485440068	6.71424551766612		2 ZNF467	ENSG00000181444			1		zinc finger protein 467
1.84132132816538	7.74819284958946		2 SCUBE2	ENSG00000175356		1 Cerebral arteriovenous malformation	12		signal peptide, CUB domain and EGF like domain containing 2
1.96129035869631	7.8703647195834		2 MMP24	ENSG00000125966	0.0856845263235322	venous thromboembolism	6	0.046945	matrix metalloproteinase 24
1.90232738951891	7.56224242422107		2 SLC16A6	ENSG00000108932			1	0.053712	solute carrier family 16 member 6
1.68608250252302	7.74146698640115		2 MED12L	ENSG00000144893			1	0.057872	mediator complex subunit 12L
1.59556390963892	6.16992500144231		2 CD163L1	ENSG00000177675			1	0.062029	CD163 molecule like 1
1.89986442667647	8.49984588708321		2 FCHO1	ENSG00000130475			1	0.064527	FCH and mu domain containing endocytic adaptor 1
1.61323056601836	9.47370574961942		2 ACAN	ENSG00000157766			33	0.103645	aggrecan
1.59990636957686	8.51569983828404		2 ST8SIA2	ENSG00000140557			7	0.187574	ST8 alpha-N-acetylneuraminidase alpha-2,8-sialyltransferase 2
1.67865865957693	9.61654884377899		2 STX1B	ENSG00000099365	0.0640979185700417	hypertension	3	0.30863	syntaxin 1B
1.59097015341644	7.53915881110803		2 BCAS4	ENSG00000124243			0	0.575039	breast carcinoma amplified sequence 4
1.828273505362	7.51569983828404		2 YIPF7	ENSG00000177752			0	1.41518	Yip1 domain family member 7
1.64504223507478	7.24792751344359		2 BEGAIN	ENSG00000183092			1	1.42467	brain enriched guanylate kinase associated
1.74199618941906	8.43879185257826		2 ACAP1	ENSG00000072818			1	1.48732	ArfGAP with coiled-coil, ankyrin repeat and PH domains 1
4.66346172781133	6.3037807481771		3 ADAM18	ENSG00000168619	5.95769433378641e-05	Polyarteritis Nodosa	0		ADAM metalloproteinase domain 18
4.41985828826816	9.13955135239879		3 WNT9A	ENSG00000143816	0.761991770811379	cardiovascular disease	9		Wnt family member 9A
3.3662454949879	7.83289001416474		3 RHCG	ENSG00000140519	0.26201388888889	Infantile hypertrophic cardiomyopathy due to MYRPL44 deficiency	30		Rh family C glycoprotein
3.46797527489724	8.40087943628218		3 HBA2	ENSG00000188536	0.2	Aicardi-Goutières syndrome	33		hemoglobin subunit alpha 2
3.38413790258724	8		3 BEX1	ENSG00000133169	0.0349194444444444	heart disease	5		brain expressed X-linked 1
3.86291902068675	9.53138146051631		3 HBB	ENSG00000244734	0.0270190913981237	gastric cardia carcinoma	79		hemoglobin subunit beta
4.66893434553611	8.74146698640115		3 ANKRD22	ENSG00000152766			1		ankyrin repeat domain 22
3.29841457303366	8.04984854945056		3 ALOX15	ENSG00000161905			290		arachidonate 15-lipoxygenase
4.37388573441346	8.64024493622235		3 CCL22	ENSG00000102962			61		C-C motif chemokine ligand 22
4.13911903907623	8.29462074889163		3 CCL24	ENSG00000106178			39		C-C motif chemokine ligand 24
2.83538834414389	10.3106127816595		3 CCL5	ENSG00000271503			849		C-C motif chemokine ligand 5
4.20701962044626	10.0279059965699		3 CCR7	ENSG00000126353			235		C-C motif chemokine receptor 7
3.69929482744647	10.2033480029798		3 CXCL10	ENSG00000169245			604		C-X-C motif chemokine ligand 10
3.73014173837633	9.3151495622563		3 CXCL11	ENSG00000169248			138		C-X-C motif chemokine ligand 11
3.53388547181941	9.78953364497036		3 CXCR3	ENSG00000186810			329		C-X-C motif chemokine receptor 3
3.6143805573298	8.84549005094438		3 CXCR6	ENSG00000172215			50		C-X-C motif chemokine receptor 6
2.98299198694532	9.2807707701306		3 CD1C	ENSG00000158481			25		CD1c molecule
4.00758613260115	9.32418054661874		3 CD27	ENSG00000139193			159		CD27 molecule
3.3494223929254	9.5018371849023		3 CD5	ENSG00000110448			210		CD5 molecule
4.59545403131854	9.00281501560705		3 CD79A	ENSG00000105369			86		CD79a molecule
2.82001066780688	10.0714623625566		3 CENPA	ENSG00000115163			9		centromere protein A
5.45850527668581	7.33985000288462		3 CLC	ENSG00000105205			353		Charcot-Leyden crystal galectin

Log2 Fold Change Score	Scaled interactor number	Cluster labels	Gene symbol	Ensembl	Association overall score - OT	Disease association - OT	PubMed Report number for a gene in the context of any cardiovascular indication	GWAS standardised association score for cardiovascular indication	Description
2.96541301513508	9.800898999203		CHD5	ENSG00000116254			9		chromodomain helicase DNA binding protein 5
4.6286593032555	9.64024493622235		FBXL16	ENSG000001127585			2		F-box and leucine rich repeat protein 16
4.49114925854412	10.4304525516655		FOXP3	ENSG00000049768			1157		forkhead box P3
3.38404700177082	8.43462822763673		GBP5	ENSG00000154451			4		guanylate binding protein 5
3.54225315258536	7.34872815423108		HYAL4	ENSG00000106302			1		hyaluronidase 4
3.8517528547932	9.13955135239879		IGLL5	ENSG00000254709			2		immunoglobulin lambda like polypeptide 5
3.88900475747716	8.16992500144231		JCHAIN	ENSG00000132465			0		joining chain of multimeric IgA and IgM
3.25929839493768	10.8462739113499		MMP9	ENSG00000100985			1863		matrix metalloproteinase 9
3.16404681887002	8.71424551766612		P2RY10	ENSG00000078589			1		P2Y receptor family member 10
4.34564662728636	8.21431912080077		PLD4	ENSG00000166428			0		phospholipase D family member 4
3.11105059227519	8.47167521439204		RIMS4	ENSG00000101098			0		regulating synaptic membrane exocytosis 4
3.3990947758927	7.11894107272351		SHISA2	ENSG00000180730			1		shisa family member 2
4.13720898139726	7.59245703726808		SIRPG	ENSG00000089012			1		signal regulatory protein gamma
3.24858513810853	8.21431912080077		SLAMF7	ENSG00000026751			5		SLAM family member 7
3.56462654658549	7.62205181945638		TIGIT	ENSG00000181847			16		T cell immunoreceptor with Ig and ITIM domains
5.00031965421217	7.33985000288462		TIFAB	ENSG00000255833			15		TIFA inhibitor
4.1211102189943	8.4178525148859		TNFRSF18	ENSG00000186891			14		TNF receptor superfamily member 18
4.09232165722855	8.54303182025524		TNNI1	ENSG00000159173			29		troponin I1, slow skeletal type
3.89731476566711	8.0389189892923		UNC90	ENSG00000144406			1		unc-90 homolog, NALCN channel complex subunit
3.68174924277724	8.51175265376738		MS4A1	ENSG00000156738			6	0.049694	membrane spanning 4-domains A1
3.98882416759078	10.0265234425198		KCNQ2	ENSG00000075043			77	0.050996	potassium voltage-gated channel subfamily Q member 2
3.16428626125551	8.93663793900257		IL4I1	ENSG00000104951			2	0.062012	interleukin 4 induced 1
3.23391662011405	8.01122725542325		ATRN1	ENSG00000107518	0.00231788439106461	Paroxysmal supraventricular tachycardia	2	0.063339	attractin like 1
4.59907931228781	9.7279204545632		CXCL9	ENSG00000138755			232	0.064711	C-X-C motif chemokine ligand 9
3.96379615514885	7.14974711950468		AQP10	ENSG00000143595			4	0.097956	aquaporin 10
3.88751021323304	8.4178525148859		CD1E	ENSG00000158488			1	0.10505	CD1e molecule
3.45875021625014	8.40514146313634		HBA1	ENSG00000206172	0.2	Aicardi-Goutières syndrome	9105	0.105576	hemoglobin subunit alpha 1
5.14662295313744	7.70043971814109		LAMP3	ENSG00000078081			5	0.113373	lysosomal associated membrane protein 3
3.47028864029895	8.83920378809694		ATP1A4	ENSG00000132681			4	0.160226	ATPase Na <sup>+</sup> /K <sup>+</sup> -transporting subunit alpha 4
4.031017177612	8.78135971352466		NPPA	ENSG00000175206		1 cardiovascular disease	437	0.170989	natriuretic peptide A
3.65426159155283	8.11374216604919		COL22A1	ENSG00000169436			6	0.546152	collagen type XXII alpha 1 chain
3.16400178024719	8.89784545600551		SEZ6L	ENSG00000100095	0.204944580793381	arterial stiffness measurement	3	0.858397	seizure related 6 homolog like
4.44972762183788	6.71424551766612		ARMS2	ENSG00000254636	0.161172108916687	cardiovascular disease	37	0.95488	age-related maculopathy susceptibility 2
2.12859741368208	10.4136279290242		TLR9	ENSG00000239732	0.73292979867855	cardiovascular disease	262		toll like receptor 9
2.58952386769066	10.4356702609366		EGR1	ENSG00000120738	0.326228649622181	cardiovascular disease	473		early growth response 1
2.27933629363019	8.96000193206808		CYP11A1	ENSG00000140459	0.202988212018141	hypertension	103		cytochrome P450 family 11 subfamily A member 1
2.1243635558976	9.19475685442225		DMC1	ENSG00000100206	0.0720431581139565	heart rate	8		DNA meiotic recombination 1
2.42313166343161	9.06608919045777		NAP1L3	ENSG00000186310	0.00912202178593987	dilated cardiomyopathy	1		nucleosome assembly protein 1 like 3
1.99197993602313	8.84235034341381		WNT10B	ENSG00000169884	0.0068	congenital heart disease	20		Wnt family member 10B

Log2 Fold Change Score	Scaled interactor number	Cluster labels	Gene symbol	Ensembl	Association overall score - OT	Disease association - OT	PubMed Report number for a gene in the context of any cardiovascular indication	GWAS standardised association score for cardiovascular indication	Description
2.43252828749429	9.09539702279256		4 BCL11B	ENSG00000127152			26		BAF chromatin remodeling complex subunit BCL11B
2.19333767492198	9.71596199025514		4 CCL3	ENSG00000277632			320		C-C motif chemokine ligand 3
2.39797392158601	9.54882190845875		4 EOMES	ENSG00000163508			42		eomesodermin
1.8722902241498	9.52552080909507		4 FPR2	ENSG00000171049			97		formyl peptide receptor 2
2.04596655162998	9.65642486327778		4 GABRD	ENSG00000187730			7		gamma-aminobutyric acid type A receptor delta subunit
1.61472493343347	10.0098286173681		4 GLI1	ENSG00000111087			133		GLI family zinc finger 1
2.16895096221027	9.08480838780436		4 GRM2	ENSG00000164082			3		glutamate metabotropic receptor 2
1.97487007888453	9.15987133677839		4 GZMA	ENSG00000145649			21		granzyme A
2.3632607166297	9.36194377373524		4 ITK	ENSG00000113263			55		IL2 inducible T cell kinase
1.65680103277328	10.7481928495885		4 JAK2	ENSG00000096968			1308		Janus kinase 2
2.55441722403153	10.4008794362822		4 LCK	ENSG00000182866			92		LCK proto-oncogene, Src family tyrosine kinase
2.07090406658569	9.93957921431469		4 MYCN	ENSG00000134323			69		MYCN proto-oncogene, bHLH transcription factor
2.00205286950728	9.39016895620018		4 KCNJ4	ENSG00000168135			30		potassium inwardly rectifying channel subfamily J member 4
2.01772710125384	9.51766938813381		4 PENK	ENSG00000181195			47		proenkephalin
2.18472670005788	8.876516946565		4 RAB33A	ENSG00000134594			0		RAB33A, member RAS oncogene family
2.59312029528846	9.93663793900257		4 RHOH	ENSG00000168421			7		ras homolog family member H
1.65179174325754	10.0927571409199		4 SELL	ENSG00000188404			372		selectin L
1.60675773129877	9.94104760634058		4 SH3GL2	ENSG00000107295			5		SH3 domain containing GRB2 like 2, endophilin A1
1.52964785010498	10.4125698468052		4 SNORD10	ENSG00000238917			0		small nucleolar RNA, C/D box 10
1.79211481293725	10.4125698468052		4 SNORA48	ENSG00000209582			0		small nucleolar RNA, H/ACA box 48
2.31330510199597	9.18239435340453		4 S1PR4	ENSG00000125910			11		sphingosine-1-phosphate receptor 4
2.35555813379308	9.96866679319521		4 SOCS1	ENSG00000185338			129		suppressor of cytokine signaling 1
2.31998259919929	9.9901039638575		4 ZAP70	ENSG00000115085			50		zeta chain of T cell receptor associated protein kinase 70
1.78324278068274	10.7846348455575		4 SNCA	ENSG00000145335		1 intrinsic cardiomyopathy	54		synuclein alpha
1.84453344921156	9.35974956032233		4 PDIA2	ENSG00000185615	0.29202651232481	cardiovascular disease	8	0.049702	protein disulfide isomerase family A member 2
1.92449488199414	9.53138146051631		4 CD74	ENSG00000019582			73	0.053395	CD74 molecule
2.00928683354396	8.99435343685886		4 FGF17	ENSG00000158815			3	0.055413	fibroblast growth factor 17
2.30826538071208	9.23122118071119		4 CD48	ENSG00000117091			35	0.056212	CD48 molecule
2.09873574959031	9.44086916761087		4 CARD11	ENSG00000198286			10	0.058737	caspase recruitment domain family member 11
2.39521253524188	9.8073549220576		4 PPP2R2B	ENSG00000156475	0.303587001569456	heart disease	5	0.12928	protein phosphatase 2 regulatory subunit Bbeta
2.50581286771842	9.82813648419411		4 STAT4	ENSG00000138378			98	0.186037	signal transducer and activator of transcription 4
2.06110718380973	9.65284497300198		4 NFXN2	ENSG00000110076			3	0.326205	neurexin 2
2.3966740284543	9.68999797141945		4 ABCG2	ENSG00000118777	0.314731394121144	cardiovascular disease	501	0.33998	ATP binding cassette subfamily G member 2 (Junior blood group)
2.28231349841461	9.66533591718518		4 PROM1	ENSG00000007062			375	0.532281	prominin 1
2.40053684125791	9.11113567023471		4 CD3E	ENSG00000198851			17	0.611353	CD3e molecule
2.21165821415797	9.56985560833095		4 DLGAP1	ENSG00000170579			3	0.835216	DLG associated protein 1
2.38680312197609	9.71596199025514		4 CD69	ENSG00000110848			252	1.08731	CD69 molecule
2.39229448702134	9.01402047031493		4 RYR3	ENSG00000198838	0.717292547225952	arterial stiffness measurement	3292	1.19919	ryanodine receptor 3
2.33129483975196	8.97441458980553		4 CHRNA3	ENSG00000080644		1 cardiovascular disease	36	1.23034	cholinergic receptor nicotinic alpha 3 subunit
1.93612514117864	9.90989308377004		4 APOA1	ENSG00000118137		1 cardiovascular disease	1162	1.67707	apolipoprotein A1

Log2 Fold Change Score	Scaled interactor number	Cluster labels	Gene symbol	Ensembl	Association overall score - OT	Disease association - OT	PubMed Report number for a gene in the context of any cardiovascular indication	GWAS standardised association score for cardiovascular indication	Description
1.78769934778717	9.6599589242998		4 BRSK1	ENSG00000160469	0.00142081870334878	systolic heart failure	4	1.98052	BR serine/threonine kinase 1
-3.34072223459275	3.4594316186373		5 C1ORF105	ENSG00000180999	0.00491658424315499	dilated cardiomyopathy	1		Chromosome 1 Open Reading Frame 105
-2.13247797041721	4.16992500144231		5 C11ORF91	ENSG00000205177			0		Chromosome 1 Open Reading Frame 91
-4.28286104178369	4.16992500144231		5 FAM9C	ENSG00000187268			0		family with sequence similarity 9 member C
-5.25065411873302	5.55458885167764		5 IGSF23	ENSG00000216588			0		immunoglobulin superfamily member 23
-4.2269421470993	5.16992500144231		5 MTRNR2L1	ENSG00000256618			1		MT-RNR2 like 1
-6.85337892128049	7.68650052718322		6 CALCB	ENSG00000175868			6		calcitonin related polypeptide beta
-3.45794193770006	9.78790255939143		6 MRAP	ENSG00000170262			86		melanocortin 2 receptor accessory protein
-4.08574850978445	9.50977500432694		6 PCK1	ENSG00000124253			32		phosphoenolpyruvate carboxykinase 1
-4.9833551942381	9.3037807481771		6 CACNA1E	ENSG00000198216		1 cardiovascular disease	18	0.096871	calcium voltage-gated channel subunit alpha1 E
-4.9352681994321	9.49785183695112		6 SAA1	ENSG00000173432			82	0.187328	serum amyloid A1
-4.30301522327876	10.0953970227926		6 MYH6	ENSG00000197616		1 heart disease	329	0.386648	myosin heavy chain 6
-3.37722329273652	10.429406741514		6 CFTR	ENSG00000001626			649	0.944108	CF transmembrane conductance regulator
-5.25901411089316	8.58496250072116		6 CSMD1	ENSG00000183117			17	1.11779	CUB and Sushi multiple domains 1
-3.50152237220186	10.140829770773		6 KNG1	ENSG00000113889			25	1.48714	kininogen 1
3.07823523365618	8.38801728534514		7 NPPB	ENSG00000120937	0.733629272738451	cardiovascular disease biomarker measurement	301		natriuretic peptide B
2.44735451122472	8.48784003382305		7 GDF10	ENSG00000266524	0.486249448688896	mean arterial pressure	6		growth differentiation factor 10
2.0432196840163	7.52356195605701		7 LTC4S	ENSG00000213316	0.141941610723734	resting heart rate	16		leukotriene C4 synthase
2.22785548237289	7.5077946401987		7 C1QTNF4	ENSG00000172247	0.0915705003619194	cardiovascular disease	0		C1q and TNF related 4
2.59750111489656	8.63299519714296		7 SFRP4	ENSG00000106483	0.0643642029123344	cardiovascular disease	40		secreted frizzled related protein 4
2.31201995863113	8.16992500144231		7 AOC1	ENSG00000002726	0.0569439735818656	cardiovascular disease	1		amine oxidase copper containing 1
2.45908415282001	8.91587937883577		7 OGDHL	ENSG00000197444	0.018	heart failure	6		oxoglutarate dehydrogenase like
2.19083910125356	7.800898999203		7 CHAC1	ENSG00000128965	0.0165	cardiovascular disease	7		ChaC glutathione specific gamma-glutamylcyclotransferase 1
2.08601590327918	8.01680828768655		7 GPR27	ENSG00000170837	0.00761771373267391	ischemic cardiomyopathy	1		G protein-coupled receptor 27
1.99977230020041	8.20945336562895		7 A1BG	ENSG00000121410			9		alpha-1-B glycoprotein
2.37825033900611	7.73470962022584		7 ANKRD24	ENSG00000089847			0		ankyrin repeat domain 24
2.64995001552377	7.11894107272351		7 ANKRD34C	ENSG00000235711			0		ankyrin repeat domain 34C
2.73080184353521	7.49984588708321		7 ABCC11	ENSG00000121270			6		ATP binding cassette subfamily C member 11
2.16862846288355	8.36194377373524		7 ABCC6	ENSG00000091262			241		ATP binding cassette subfamily C member 6
2.66537790842335	8.13442632022093		7 BIRC7	ENSG00000101197			2		baculoviral IAP repeat containing 7
2.55171219020732	8.33539035469392		7 CCL8	ENSG00000108700			56		C-C motif chemokine ligand 8
2.73167905780424	9.6599589242998		7 CD2	ENSG00000116824			617		CD2 molecule
2.68158107901429	9.38586240064146		7 CD3D	ENSG00000167286			4		CD3d molecule
2.67965066853657	8.12928301694497		7 CD8B	ENSG00000172116			31		CD8b molecule
2.93130193432858	8.33985000288462		7 CMA1	ENSG00000092009			3		chymase 1
2.25838495303674	6.8073549220576		7 DNASE1L2	ENSG00000167968			1		deoxyribonuclease 1 like 2
2.56560233530115	7.49984588708321		7 DNAJC22	ENSG00000178401			0		DnaJ heat shock protein family (Hsp40) member C22
2.93351746219208	8.63662462054365		7 GZMM	ENSG00000197540			2		granzyme M
1.98033877931149	8.73131903102506		7 HCST	ENSG00000126264			11		hematopoietic cell signal transducer
2.64995984666322	7.800898999203		7 HS6ST2	ENSG00000171004			3		heparan sulfate 6-O-sulfotransferase 2
2.22312815576303	8.61102479730735		7 IRX6	ENSG00000159387			3		iroquois homeobox 6
2.612273191131	8.76818432477693		7 KLRB1	ENSG00000111796			15		killer cell lectin like receptor B1

Log2 Fold Change Score	Scaled interactor number	Cluster labels	Gene symbol	Ensembl	Association overall score - OT	Disease association - OT	PubMed Report number for a gene in the context of any cardiovascular indication	GWAS standardised association score for cardiovascular indication	Description
2.67003597903346	7.96578428466209		LYPD1	ENSG00000150551			2		LY6/PLAUR domain containing 1
2.28672005195914	8.40514146313634		LAG3	ENSG00000089692			24		lymphocyte activating 3
2.60042807443691	8.17990909001493		LY9	ENSG00000122224			3		lymphocyte antigen 9
2.16200370996885	7.49185309632967		MT1G	ENSG00000125144			5		metallothionein 1G
2.18907040029136	8.70043971814109		NKG7	ENSG00000105374			5		natural killer cell granule protein 7
2.21952311686212	6.83289001416474		NSG2	ENSG00000170091			0		neuronal vesicle trafficking associated 2
2.25824566322869	7.4346222763672		NXNL1	ENSG00000171773			0		nucleoredoxin like 1
2.75548969313356	8.72451385311995		PLCH2	ENSG00000149527			0		phospholipase C eta 2
2.56389951825518	8.49984588708321		SDSL	ENSG00000139410			8		serine dehydratase like
2.34968484331985	7.32192809488736		SPTSSB	ENSG00000196542			0		serine palmitoyltransferase small subunit B
2.22310254848053	6.75488750216347		SUSD3	ENSG00000157303			0		sushi domain containing 3
3.04023939928333	7.876516946565		SYTL1	ENSG00000142765			1		synaptotagmin like 1
2.82788546999612	7.90689059560852		SYTL5	ENSG00000147041			0		synaptotagmin like 5
2.50390426134064	8.2045711442492		TBC1D10C	ENSG00000175463			6		TBC1 domain family member 10C
3.05666300604057	7.49185309632967		TNMD	ENSG00000000005			11		tenomodulin
2.40916087210493	7.93663793900257		VSX1	ENSG00000100987			5		visual system homeobox 1
2.2935954176009	6.85798099512757		WFIKKN1	ENSG00000127578			0		WAP, follistatin/kazal, immunoglobulin, kunitz and netrin domain containing 1
2.07619129212165	7.876516946565		ZP3	ENSG00000188372			11		zona pellucida glycoprotein 3
2.14796638558075	8.54689445988764		CERS1	ENSG00000223802		1 Genetic cardiac anomaly	5		ceramide synthase 1
2.04090525022728	8.08214904135387		CHRNE	ENSG00000108556		1 intracranial hypertension	3		cholinergic receptor nicotinic epsilon subunit
2.0240899183422	8.29462074889163		PDE6A	ENSG00000132915		1 cardiovascular disease	2		phosphodiesterase 6A
3.04263727921997	8.55842071326866		NELL2	ENSG00000184613			2	0.049539	neural EGFL like 2
2.63756499283572	8.90388184573618		LHCGR	ENSG00000138039			16	0.049653	lutetizing hormone/choriogonadotropin receptor
3.06950375830231	8.41362792902417		GZMK	ENSG00000113088			4	0.053505	granzyme K
2.95156742360353	7.12928301694497		SOHLH2	ENSG00000120869			0	0.056217	spermatogenesis and oogenesis specific basic helix-loop-helix 2
2.75597082192817	8.88569637333939		IL21R	ENSG00000103522			78	0.062337	interleukin 21 receptor
2.2724583876066	8.24317398347295		CD6	ENSG00000013725			23	0.063987	CD6 molecule
2.72197205469988	8.67595703294175		SLC4A1	ENSG00000004939			26	0.067153	solute carrier family 4 member 1 (Diego blood group)
3.01332661491172	8.00562454919388		SLC16A9	ENSG00000165449	0.00209062141407026	Arterial stenosis	8	0.067287	solute carrier family 16 member 9
2.38660226019745	8.43462822763673		MMP25	ENSG00000008516			4	0.076349	matrix metalloproteinase 25
2.63889469798875	8.24317398347295		FNDC1	ENSG00000164694	0.00365269911261096	ischemic cardiomyopathy	5	0.078697	fibronectin type III domain containing 1
2.3590838856802	8.96289600533726		NEURL1	ENSG00000107954			7	0.085472	neutralized E3 ubiquitin protein ligase 1
2.33253171672164	8.12928301694497		COL9A1	ENSG00000112280	0.0199	dilated cardiomyopathy	5	0.085871	collagen type IX alpha 1 chain
2.91112461231459	8.4512111183233		SCG5	ENSG00000166922			4	0.102619	secretogranin V
2.17387463883333	7.4594316186373		WDR6	ENSG00000158023	0.000484411379518669	cardiac arrhythmia	0	0.49309	WD repeat domain 66
2.3344924836178	8.92184093707449		PHF21B	ENSG00000056487			0	0.536388	PHD finger protein 21B
2.56489104266801	7.467605550083		SMPD3	ENSG00000103056			23	0.799531	sphingomyelin phosphodiesterase 3
2.68506839447976	7.74819284958946		CD96	ENSG00000153283			3	1.06777	CD96 molecule
3.00103776412044	8.41362792902417		UBASH3A	ENSG00000160185			3	1.14885	ubiquitin associated and SH3 domain containing A
2.3408208648566	7.4178525148859		MALRD1	ENSG00000204740			1	1.25609	MAM and LDL receptor class A domain containing 1
2.70990324049768	5.28540221886225		C16ORF89	ENSG00000153446			0		Chromosome 16 Open Reading Frame 89

Log2 Fold Change Score	Scaled interactor number	Cluster labels	Gene symbol	Ensembl	Association overall score - OT	Disease association - OT	PubMed Report number for a gene in the context of any cardiovascular indication	GWAS standardised association score for cardiovascular indication	Description
2.64138640019292	4.16992500144231		8 C19ORF81	ENSG00000235034			0	Chromosome 19 Open Reading Frame 81	
2.69394412190213	6.32192809488736		8 ANO9	ENSG00000185101			0	anoctamin 9	
1.8886180395277	4.4594316186373		8 CCDC154	ENSG00000197599			0	coiled-coil domain containing 154	
3.81217232635013	4.64385618977472		8 DCANP1	ENSG00000251380			0	dendritic cell associated nuclear protein	
2.29323689155923	5.55458885167764		8 FAM180B	ENSG00000196666			0	family with sequence similarity 180 member B	
1.84728904210943	0		8 FAM229A	ENSG00000225828			0	family with sequence similarity 229 member A	
2.43329398752359	11.0821490413539		8 FOS	ENSG00000170345			2928	Fos proto-oncogene, AP-1 transcription factor subunit	
2.49471525478438	6.37503943134692		8 LCN12	ENSG00000184925			0	lipocalin 12	
2.9607658928758	6.10852445677817		8 PVRIG	ENSG00000213413			1	PVR related immunoglobulin domain containing	
1.68452668286034	5.08746284125034		8 SPATA6L	ENSG00000106686			0	spermatogenesis associated 6 like	
2.10620412217147	6.5077946401987		8 XG	ENSG00000124343			105	Xg glycoprotein (Xg blood group)	
2.02854724940606	5.08746284125034		8 ZNF385C	ENSG00000187595			0	zinc finger protein 385C	
2.85382787755512	6.4757334309664		8 ZNF683	ENSG00000176083			1	zinc finger protein 683	
1.98277004045982	4.70043971814109		8 ANKRD33B	ENSG00000164236			0.061642	ankyrin repeat domain 33B	
2.13192409303079	6.39231742277876		8 PRR7	ENSG00000131188			0.095411	proline rich 7, synaptic	
2.62859428543533	6.5077946401987		8 ADGRG5	ENSG00000159618			0.10612	adhesion G protein-coupled receptor G5	
1.55056935674004	5		8 MAP3K7CL	ENSG00000156265	0.0104	coronary artery disease	3	0.673055	MAP3K7 C-terminal like
-2.4516805469563	8.37068740680722		9 EREG	ENSG00000124882	0.615417242050171	coronary artery calcification	23		epiregulin
-2.47029362828155	10.1910592145317		9 SPP1	ENSG00000118785	0.333324605482535	cardiovascular disease	603		secreted phosphoprotein 1
-2.70503262706167	9.84862294042934		9 BMP7	ENSG00000101144	0.327475610575549	cardiovascular disease	188		bone morphogenetic protein 7
-3.01932187497333	8.20945336562895		9 HOPX	ENSG00000171476	0.295285022222222	familial cardiomyopathy	29		HOP homeobox
-2.123437544963	9.72451385311995		9 KIF20A	ENSG00000112984	0.220411950722337	heart rate	9		kinesin family member 20A
-3.07905817081996	8.84235034341381		9 SERPINA3	ENSG00000196136	0.214303720238637	cardiovascular disease	29		serpin family A member 3
-2.63638953175741	8.60362634498619		9 KCNA7	ENSG00000104848	0.0334	Familial progressive cardiac conduction defect	6		potassium voltage-gated channel subfamily A member 7
-2.25768326776873	8.51569983828404		9 SERPINA5	ENSG00000188488	0.032291	hypertension	7		serpin family A member 5
-2.39849514745375	7.56985560833095		9 PCDH20	ENSG00000280165	0.0136	portal hypertension	2		protocadherin 20
-2.4257177445985	8.2807707701306		9 CNGA1	ENSG00000198515			5		cyclic nucleotide gated channel subunit alpha 1
-2.04612181352427	8.68299458368168		9 DNAH3	ENSG00000158486			1		dynein axonemal heavy chain 3
-2.31492872175805	8.01122725542325		9 ESRP2	ENSG00000103067			2		epithelial splicing regulatory protein 2
-2.1183927265089	8.32192809488736		9 GNMT	ENSG00000124713			8		glycine N-methyltransferase
-2.62286526104739	8.3264294871223		9 LRRN3	ENSG00000173114			6		leucine rich repeat neuronal 3
-2.138342830372	7.48381577726426		9 LINC00842	ENSG00000285294			0		long intergenic non-protein coding RNA 842
-2.9335371825788	7.6724253419715		9 PYGO1	ENSG00000171016			3		pygopus family PHD finger 1
-3.01128813875027	8.19967234483636		9 RBP4	ENSG00000138207			258		retinol binding protein 4
-2.27469537735965	7.467605550083		9 TMEM132B	ENSG00000139364			0		transmembrane protein 132B
-2.71964888573354	9.03617361255349		9 TMEM151B	ENSG00000178233			0		transmembrane protein 151B
-2.25399931985298	7.18982455888002		9 ZBED6	ENSG00000257315			5		zinc finger BED-type containing 6
-2.1370022250427	9.0389189892923		9 GATA5	ENSG00000130700		1 cardiovascular disease	120		GATA binding protein 5
-1.99193663664808	8.15987133677839		9 MYOT	ENSG00000120729		1 cardiomyopathy	32		myotilin
-2.34842334974533	9.51569983828404		9 FGF7	ENSG00000140285	0.0444722222222222	pulmonary arterial hypertension	54	0.048529	fibroblast growth factor 7

Log2 Fold Change Score	Scaled interactor number	Cluster labels	Gene symbol	Ensembl	Association overall score - OT	Disease association - OT	PubMed Report number for a gene in the context of any cardiovascular indication	GWAS standardised association score for cardiovascular indication	Description
-2.33952297988654	8.40514148313634		TRPC4	ENSG000001133107	0.0710127488599104	cardiovascular disease	155	0.053665	transient receptor potential cation channel subfamily C member 4
-3.08339785399008	8.91587937883577		WSCD2	ENSG00000075035			0	0.055893	WSC domain containing 2
-2.45308284528563	9.35535109642481		PCSK1	ENSG00000175426			27	0.056497	proprotein convertase subtilisin/kexin type 1
-1.95979104258837	8.24317398347295		CHDH	ENSG00000016391		1 cardiovascular disease	8	0.063677	choline dehydrogenase
-2.27042400382205	8.19475685442225		PHACTR3	ENSG000000087495			2	0.087486	phosphatase and actin regulator 3
-3.0425114971795	8.93369065495223		LRRIC7	ENSG000000033122			0	0.094918	leucine rich repeat containing 7
-2.77976147459426	8.09803208296053		SLC38A4	ENSG000001139209			9	0.109267	solute carrier family 38 member 4
-2.50826793602995	8.09803208296053		MARCO	ENSG00000019169			2982	0.111059	macrophage receptor with collagenous structure
-2.33722737642656	9.52552080909507		CHL1	ENSG000001134121	9.23637951231764e-05	congenital anomaly of cardiovascular system	13	0.191441	cell adhesion molecule L1 like
-2.43602481959331	8.68299458368168		TAOK1	ENSG000001160551			5	0.259799	TAO kinase 1
-2.83381178423145	8.18487534290828		GDA	ENSG00000119125	0.0637290136054422	myocardial infarction	562	0.501181	guanine deaminase
-2.46585983615093	8.9915218460757		ITGB6	ENSG00000115221	0.772906363010406	arterial stiffness measurement	11	0.503822	integrin subunit beta 6
-3.1607445031284	8.85486838326024		GRIP1	ENSG000001155974	0.2	Congenital vertebral-cardiac-renal anomalies syndrome	24	0.621951	glutamate receptor interacting protein 1
-1.97553121722681	8.65105169117893		WWC1	ENSG00000113645			13	0.663947	WW and C2 domain containing 1
-1.97369260369013	8.69348695749933		XRCC4	ENSG000001152422	0.0104	hypertension	16	0.696042	X-ray repair cross complementing 4
-2.54701327849869	9.31288295528435		ITGA2	ENSG000001164171			48	0.984909	integrin subunit alpha 2
-2.70656495930026	8.98868468677217		LRP1B	ENSG000001168702			18	1.28906	LDL receptor related protein 1B
-2.5462562518139	8.61057163474115		MLXIPL	ENSG00000009950	0.5190439389898994	cardiovascular disease	66	1.74508	MLX interacting protein like
-1.78566391473959	9.4858293087019		GPD1	ENSG000001167588	0.254230226448427	cardiovascular disease	19		glycerol-3-phosphate dehydrogenase 1
-1.8947520606397	9.09803208296053		KIAA0754	ENSG00000127603	0.216453298926353	peripheral arterial disease	1		KIAA0754
-1.55848652584082	7.98868468677217		HOOK1	ENSG000001134709	0.183	Aicardi-Goutières syndrome	4		hook microtubule tethering protein 1
-1.73832820045389	7.08746284125034		ART4	ENSG00000111339			4		ADP-ribosyltransferase 4 (Dombrock blood group)
-1.74518211303842	8.48784003382305		ADH1A	ENSG000001187758			2		alcohol dehydrogenase 1A (class I), alpha polypeptide
-1.80929413231189	8.04984854945056		ANKRD36	ENSG000001135976			3		ankyrin repeat domain 36
-1.77129083700519	7.33091687811462		B3GALT2	ENSG000001162630			0		beta-1,3-galactosyltransferase 2
-1.53470283149158	9.7125270043982		BLM	ENSG000001197299			168		BLM RecQ like helicase
-1.80089866167926	7.32192809488736		C1QL1	ENSG000001131094			1		complement C1q like 1
-1.62799955462344	6.62935662007961		CXXC4	ENSG000001168772			0		CXXC finger protein 4
-1.6053029879769	8.59245703728808		CDKL5	ENSG00000008086			11		cyclin dependent kinase like 5
-1.55064698428938	8.78463484555752		POLQ	ENSG000000051341			0		DNA polymerase theta
-1.67809759412146	8.29920801838728		ELK4	ENSG000001158711			9		ETS transcription factor ELK4
-1.92687304799156	8.17990909001493		FAM83D	ENSG00000101447			1		family with sequence similarity 83 member D
-1.90558165129112	7.8073549220576		GYG2	ENSG000000056998			5		glycogenin 2
-1.86428552933389	8.97441458980553		KNL1	ENSG000001137812			0		kinetochore scaffold 1
-1.62918229708674	8.37503943134693		MGST1	ENSG00000008394			8		microsomal glutathione S-transferase 1
-1.92494762061485	10.0014081943928		MYSM1	ENSG000001162601			0		Myb like, SWIRM and MPN domains 1
-1.6707743942261	9.50977500432694		MYO9A	ENSG00000066933			4		myosin IXA
-1.71594251759675	8.2336196767597		KCNK3	ENSG000001171303			95		potassium two pore domain channel subfamily K member 3

Log2 Fold Change Score	Scaled interactor number	Cluster labels	Gene symbol	Ensembl	Association overall score - OT	Disease association - OT	PubMed Report number for a gene in the context of any cardiovascular indication	GWAS standardised association score for cardiovascular indication	Description
-1.53542828238925	9.11113567023471	10	PDCD1	ENSG00000188389			221		programmed cell death 1
-1.79758699934353	7.82017896241519	10	SNORA40	ENSG00000210825			0		small nucleolar RNA, H/ACA box 40
-1.67615728561103	7.91288933622996	10	SMTNL2	ENSG00000188176			2		smoothelin like 2
-1.85051273249213	6.82017896241519	10	TMEM178B	ENSG00000261115			1		transmembrane protein 178B
-1.94961793585098	8.29462074889163	10	TROAP	ENSG00000135451			0		trophinin associated protein
-1.90333683157165	9.05528243550119	10	SSTR2	ENSG00000180616	1	cardiovascular disease	52		somatostatin receptor 2
-1.60926449202193	9.19967234483636	10	PROX1	ENSG00000117707			162	0.046901	prospero homeobox 1
-1.69812286368119	9.21916852046216	10	CENPF	ENSG00000117724	0.039485351305665	heart disease	7	0.047372	centromere protein F
-1.81782912656785	7.66533591718518	10	MFAP2	ENSG00000117122	0.1864	Aicardi-Goutières syndrome	6	0.048463	microfibril associated protein 2
-1.80266740398691	10.4051414631363	10	TFRC	ENSG00000072274			51	0.048463	transferrin receptor
-1.9343286442857	6	10	PRELID2	ENSG00000186314			2	0.048756	PRELI domain containing 2
-1.61656898626651	10.1774195379892	10	BUB1B	ENSG00000156970	0.18596	Aicardi syndrome	13	0.050256	BUB1 mitotic checkpoint serine/threonine kinase B
-1.64898401835846	6.10852445677817	10	VASH2	ENSG00000143494	0.0308	pulmonary arterial hypertension	12	0.052454	vasohibin 2
-1.90856215220625	9.27612440527424	10	ANLN	ENSG00000011426	0.0142	cardiovascular disease	6	0.056363	anillin actin binding protein
-1.76831359884661	7.59245703726808	10	MARVELD2	ENSG00000152939	0.0173777777777778	heart disease	6	0.057293	MARVEL domain containing 2
-1.56196042036997	7.76818432477693	10	ANKRD36B	ENSG00000196912			0	0.061468	ankyrin repeat domain 36B
-1.58709249103316	10.1811522568656	10	CDKN3	ENSG00000100526			6	0.064123	cyclin dependent kinase inhibitor 3
-1.70209659495742	8.44708322620965	10	OLFM4	ENSG00000102837			22	0.072193	olfactomedin 4
-1.63280881161153	7.85798099512757	10	RNF157	ENSG00000141576			1	0.106906	ring finger protein 157
-1.65681870879911	7.84549005094437	10	RNF152	ENSG00000176641			0	0.10847	ring finger protein 152
-1.6853046709455	9.54303182025524	10	CAMK1D	ENSG00000183049			6	0.516361	calcium/calmodulin dependent protein kinase ID
-1.93966307947016	8.13442632022093	10	ATP6V1C2	ENSG00000143882			0	0.520142	ATPase H+ transporting V1 subunit C2
-1.51306488364998	8.97727992349992	10	CACNB2	ENSG00000165995			76	0.543523	calcium voltage-gated channel auxiliary subunit beta 2
-2.01499539651828	7.06608919045777	10	FAM81A	ENSG00000157470			0	0.609717	family with sequence similarity 81 member A
-1.59908907235341	7.73470962022584	10	ADAMTS12	ENSG00000151388	0.109591687530794	cardiovascular disease	8	0.657565	ADAM metalloproteinase with thrombospondin type 1 motif 12
-1.74012845821396	7.49185309632967	10	CD5L	ENSG00000073754			14	0.760137	CD5 molecule like
-1.70219628873732	8.37068740680722	10	ART3	ENSG00000156219			8	0.923572	ADP-ribosyltransferase 3
-1.66482706058028	9.4655664048094	10	CACNA1D	ENSG00000157388			136	0.979092	calcium voltage-gated channel subunit alpha1 D
-1.50962918629019	10.4019461239765	10	APOB	ENSG00000084674			6214	2.72011	apolipoprotein B



Supplementary table 12. Dilated vs non-failing heart (GSE3585) contrast cluster cross-referencing with disease association databases and datasets

Log2 Fold Change Score	Scaled interactor number	Cluster labels	Gene symbol	Ensembl	Association overall score - OT	Disease association - OT	PubMed Report number for a gene in the context of any cardiovascular indication	GWAS standardised association score for cardiovascular indication	Description
-0.92246598816316	11.2461467746359	0	STAT3	ENSG00000168610		1 cardiovascular disease	2576	0.057388	signal transducer and activator of transcription 3
-0.57928944	10.1811522568656	0	CDKN3	ENSG000001100526	0.0294833333333333	hypertension	6	0.064123	cyclin dependent kinase inhibitor 3
-0.6129675	9.40087943628218	0	PPL	ENSG00000118898	0.210070863866955	cardiovascular disease	217	0.064408	periplakin
-0.568597398436409	10.4706588740606	0	ICAM1	ENSG00000090339	0.22875	Alcardi-Goulières syndrome	8986	2.73394	intercellular adhesion molecule 1
-0.787274825148	9.70217268536555	0	IDH2	ENSG00000182054		1 cardiovascular disease	50		isocitrate dehydrogenase (NADP(+)) 2
-0.60028744	10.6882503091332	0	H2AFZ	ENSG00000164032	0.004	cardiac hypertrophy	2		H2A.Z Variant Histone 1
-0.61079216	9.67771964164101	0	H1FO	ENSG00000189060			3		H1.0 Linker Histone
0.6517401	6.04439411935845	1	FAM216A	ENSG00000204856			0	1.96048	family with sequence similarity 216 member A
0.877676940951478	6.4594316186373	1	PPDPF	ENSG00000125534	0.0144197024176782	dilated cardiomyopathy	0		pancreatic progenitor cell differentiation and proliferation factor
0.4288969	7.15987133677839	1	RSN1	ENSG00000081019	0.2	coronary artery disease, autosomal dominant 2	1		round spermatid basic protein 1
0.336524	7.08746284125034	1	TMEM231	ENSG00000205084			4		transmembrane protein 231
0.5514841	7	1	HMG2	ENSG00000198830			8		high mobility group nucleosomal binding domain 2
0.4214306	7.17990909001493	1	ZSCAN18	ENSG00000121413			1		zinc finger and SCAN domain containing 18
1.2905462562525	8.93369065495223	2	KLHL3	ENSG00000146021		1 hypertension	77	0.067048	kelch like family member 3
1.6075773	8.39231742277876	2	ID4	ENSG00000172201	0.18308	Heart-hand syndrome type 3	15	0.107028	inhibitor of DNA binding 4, HLH protein
1.81482103301306	9.15228484230656	2	CFH	ENSG00000000971	0.341644941617632	cardiovascular disease	168	0.38501	complement factor H
1.766592503139	8.38801728534514	2	SPOCK1	ENSG00000152377	0.02352	gastric non-cardia carcinoma	5	1.24289	SPARC (osteonectin), cwcv and kazal like domains proteoglycan 1
1.24378905805	9.41996017784789	2	ODC1	ENSG00000115758	0.0557576975548353	cardiovascular disease	7		ornithine decarboxylase 1
1.2341019575186	10.0389189892923	2	MYH10	ENSG00000133026	0.32816138639532	dilated cardiomyopathy	28		myosin heavy chain 10
1.9229517	7.62205181945638	2	PHLDA1	ENSG00000139289	0.558664033835375	cardiovascular disease	9		pleckstrin homology like domain family A member 1
0.48857117	7.88874324889826	3	NCKIPSD	ENSG00000213672			3	0.051454	NCK interacting protein with SH3 domain
0.5544902656682	7.71424551766612	3	CHST3	ENSG00000122863	0.0304	pulmonary arterial hypertension	11	0.052944	carbohydrate sulfotransferase 3
0.39712238	8.37068740680722	3	TULP4	ENSG00000130338	0.0104	congenital heart disease	2	0.082924	TUB like protein 4
0.61286736	8.29001884693262	3	ROR1	ENSG00000185483	0.014	ischemic cardiomyopathy	19	0.082936	receptor tyrosine kinase like orphan receptor 1
0.81835365	8.61838550225861	3	LTBP1	ENSG00000049323	0.317106207893006	Genetic cardiac anomaly	35	0.179172	latent transforming growth factor beta binding protein 1
0.81306063504	8.44294349584873	3	EXT1	ENSG00000182197			26	0.636995	exostosin glycosyltransferase 1
0.74196243	9.07681559705083	3	LAMB1	ENSG00000091136	0.286596588043017	cardiovascular disease	11	0.739348	laminin subunit beta 1
0.87712765	7.99435343685886	3	NAV2	ENSG00000166833	0.641349049289366	cardiovascular disease	9	0.942098	neuron navigator 2
0.50399685	8.9915218460757	3	SEC31A	ENSG00000138674	0.10989486426115	hypertension	4	1.33	SEC31 homolog A, COPII coat complex component
0.56628895	8.3264294871223	3	SPRED2	ENSG00000198369	0.0443448186309749	cardiovascular disease	9	1.75078	sprouty related EVH1 domain containing 2
0.72861004	8.74819284958946	3	ATP13A3	ENSG00000133657		1 pulmonary arterial hypertension	6		ATPase 13A3
1.09076859307525	7.81378119121704	3	SSPN	ENSG00000123096		1 cardiovascular disease	17		sarcomer
0.42609978	7.85798099512757	3	CAMSAP2	ENSG00000118200	0.000363546899432006	Arteritis	1		calmodulin regulated spectrin associated protein family member 2
0.77229977	8.83605035505807	3	KIDINS220	ENSG00000134313	0.00267462982497156	heart disease	8		kinase D interacting substrate 220
0.7625923	8.54303182025524	3	ETV5	ENSG00000244405	0.0308	hypertension	29		ETS variant transcription factor 5

Log2 Fold Change Score	Scaled interactor number	Cluster labels	Gene symbol	Ensembl	Association overall score - OT	Disease association - OT	PubMed Report number for a gene in the context of any cardiovascular indication	GWAS standardised association score for cardiovascular indication	Description
0.75123596	7.94251450533924		SLC30A1	ENSG00000170385	0.0518	congenital heart disease	16		solute carrier family 30 member 1
0.52015495	8.62570884306447		CBFB	ENSG000000067955	0.1864	Aicardi-Goutières syndrome	10		core-binding factor subunit beta
0.600883209491	8.56985560833095		CLK1	ENSG000000013441	0.19572	Autosomal dominant progressive nephropathy with hypertension	4		CDC like kinase 1
0.50695592608	8.7279204545632		OGA	ENSG00000198408	0.221683075968955	cardiovascular disease	166		O-GlcNAcase
0.7160044	8.0389189892923		SLK	ENSG000000065613	0.613553443923593	mean arterial pressure	104		STE20 like kinase
0.6053648	8.66533591718518		ASMTL	ENSG00000169093			3		acetylserotonin O-methyltransferase like
0.8135338	8.2240016741981		PRSS23	ENSG00000150687			4		serine protease 23
0.5490122	7.92481250360578		RNF38	ENSG00000137075			4		ring finger protein 38
0.51859474	7.97154355395077		TRMT5	ENSG00000126814			2		tRNA methyltransferase 5
0.42307377	8.40087943628218		KMT5B	ENSG00000110066			1		lysine methyltransferase 5B
0.72925186	8.93957921431469		HNRNPH3	ENSG000000096746			0		heterogeneous nuclear ribonucleoprotein H3
0.5101156	8.07146236255662		AP3M2	ENSG000000070718			0		adaptor related protein complex 3 subunit mu 2
3.09602461642729	8.78135971352466		NPPA	ENSG00000175206		1 cardiovascular disease	437	0.170989	natriuretic peptide A
4.82623160625593	8.38801728534514		NPPB	ENSG00000120937	0.733629272738451	cardiovascular disease biomarker measurement	302		natriuretic peptide B
3.49044522241284	9.9128893622996		CCN2	ENSG00000118523	0.911169257585898	cardiovascular disease	513		cellular communication network factor 2
-0.33780430272	8.65105169117893		RNF5	ENSG00000204308	0.0178368888888889	cerebral artery occlusion	1	0.071333	ring finger protein 5
-0.52831745	8.16490692667569		MAPKAPK3	ENSG00000114738	0.0104	myocardial infarction	11	0.099883	MAPK activated protein kinase 3
-0.749969368	8.60362634498619		IMPA2	ENSG00000141401	0.198519743806109	heart disease	0	0.540609	inositol monophosphatase 2
-0.9175482	8.18487534290828		FCGBP	ENSG00000275395	0.0303797936422462	gastric non-cardia carcinoma	5		Fc fragment of IgG binding protein
-0.82914543	7.43462822763672		NSG1	ENSG00000168824	0.050427857786417	congenital heart disease	0		neuronal vesicle trafficking associated 1
-0.786302454432605	7.3037807481771		APOBEC2	ENSG00000124701	0.2	Early-onset myopathy with fatal cardiomyopathy	9		apolipoprotein B mRNA editing enzyme catalytic subunit 2
-0.5532818	8.16490692667569		CES2	ENSG00000172831	0.2008	hypertension	10		carboxylesterase 2
-0.686458019125	7.81378119121704		STEAP3	ENSG00000115107	0.240445	Dilated cardiomyopathy with ataxia	2		STEAP3 metalloproteinase
-0.5734087034931	8.32192809488736		GNMT	ENSG00000124713	0.28782815	cardiomyopathy	8		glycine N-methyltransferase
-0.7077122	8.73470962022584		HIST1H1C	ENSG00000187837			1		H1.2 Linker Histone, Cluster Member
-2.2874527	7.23840473932508		RARRES1	ENSG00000118849	0.052846398204565	hypertension	3	0.075029	retinoic acid receptor responder 1
-0.7333927	6.79441586635011		MTUS2	ENSG00000132938			1	0.136346	microtubule associated scaffold protein 2
-0.9536581	6.91886323727459		MID1IP1	ENSG00000165175			0		MID1 interacting protein 1
0.42668915	9.39446269461032		SYT11	ENSG00000132718			3	0.070328	synaptotagmin 11
0.5339651	9.65642486327778		DLG5	ENSG00000151208	0.0128	congenital heart disease	4	0.082071	discs large MAGUK scaffold protein 5
0.760953865343052	10.3026389237876		INSR	ENSG00000171105		1 cardiovascular disease	142	0.558493	insulin receptor
0.898537393858216	9.64385618977473		IGFBP3	ENSG00000146674		1 cardiovascular disease	562		insulin like growth factor binding protein 3
0.643863648256644	10.7739633684336		XPO1	ENSG00000082898		1 cardiovascular disease	14		exportin 1
0.476755493656	9.05799172275918		ARHGAP1	ENSG00000175220	0.2	Familial avascular necrosis of femoral head	5		Rho GTPase activating protein 1
0.3976755	9.27844945822048		YEATS2	ENSG00000163872	0.221854642033577	mean arterial pressure	1		YEATS domain containing 2
0.3478861	9.73978060977326		GTF2B	ENSG00000137947	0.532770344615976	cardiovascular disease	1		general transcription factor IIB
0.31162643	9.36413465500805		RPL17-C18orf32	ENSG00000215472			0		RPL17-C18orf32 Readthrough
-1.2389373373205	3.4594316186373		C1ORF105	ENSG00000180999	0.00491658424315499	dilated cardiomyopathy	1		Chromosome 1 Open Reading Frame 105

Supplementary table 13. Diabetic post-ischemic heart failure dataset (GSE26887) cluster cross-referencing with disease association databases and datasets

Log2 Fold Change Score	Scaled interactor number	Cluster labels	Gene symbol	Ensembl	Association overall score - OT	Disease association - OT	PubMed Report number for a gene in the context of any cardiovascular indication	GWAS standardised association score for cardiovascular indication	Description
0.9454193	7.82017896241519	0	AOC3	ENSG00000131471	0.0750835376215128	cardiovascular disease	90	0.046885	amine oxidase copper containing 3
0.944643	7.83920378809694	0	DEPTOR	ENSG000001155792	0.292551457981927	heart rate response to exercise	13	0.047463	DEP domain containing MTOR interacting protein
0.73079014	8.33539035469392	0	CC2D2A	ENSG00000048342	0.307139804635928	Genetic cardiac anomaly	7	0.05095	coiled-coil and C2 domain containing 2A
0.9590473	8.41362792902417	0	GZMK	ENSG00000113098	0.0324	Myocardial Ischemia	4	0.053505	granzyme K
0.75933266	7.70043971814109	0	HEG1	ENSG00000173706	0.23475	Genetic cardiac anomaly	23	0.056119	heart development protein with EGF like domains 1
1.0282478	7.8073549220576	0	CDH6	ENSG00000113361	0.483904927968979	resting heart rate	7	0.058737	cadherin 6
1.1346464	7.90689059560852	0	ECM2	ENSG00000106823	0.00895328337197348	dilated cardiomyopathy	3	0.059947	extracellular matrix protein 2
0.7281332	8.00562454919388	0	SLC16A9	ENSG00000165449	0.00209062141407026	Arterial stenosis	8	0.067287	solute carrier family 16 member 9
0.7169552	8.38370429247405	0	DCLK2	ENSG00000170390			1	0.079181	doublecortin like kinase 2
0.91115	8.07681559705083	0	RGS5	ENSG00000143248	0.104201728067114	cardiovascular disease	122	0.101839	regulator of G protein signaling 5
1.0446882	8.30833903013941	0	MEOX2	ENSG00000106511	0.296646094444445	cardiovascular disease	55	0.1062	mesenchyme homeobox 2
0.84842205	7.84549005094437	0	RNF152	ENSG00000176641			0	0.10847	ring finger protein 152
0.7298918	7.93663793900257	0	COLGALT2	ENSG00000198756	0.0560169778764248	arterial stiffness measurement	0	0.137666	collagen beta(1-O-galactosyltransferase 2
0.60214615	8.21431912080077	0	AFF3	ENSG00000144218			7	0.157888	AF4/FMR2 family member 3
0.7879982	7.84549005094437	0	ITGA11	ENSG00000137809	0.1927	Cardiodyrhythmic potassium-sensitive periodic paralysis	9	0.165139	integrin subunit alpha 11
0.83326626	7.97154355395077	0	CABLES1	ENSG00000134508	0.250273937479157	cerebrovascular disorder	2	0.182713	Cdk5 and Abl enzyme substrate 1
0.6383791	8.4757334309664	0	ANKRD6	ENSG00000135299	0.00311155714688393	hypertensive renal disease	2	0.229583	ankyrin repeat domain 6
0.6755018	8.2336196767597	0	OSBP10	ENSG00000144645	0.790188431739807	arterial stiffness measurement	3	0.501504	oxysterol binding protein like 10
0.86116314	7.4757334309664	0	PLA2R1	ENSG00000153246	0.0288055555555556	vasculitis	31	0.50389	phospholipase A2 receptor 1
0.86438084	8.00562454919388	0	FREM1	ENSG00000164946	0.58658787667272	cardiovascular disease	4	0.515734	FRAS1 related extracellular matrix 1
0.7468443	8.03342300153745	0	PLAGL1	ENSG00000118495	0.320887338712522	Genetic cardiac anomaly	25	0.594847	PLAG1 like zinc finger 1
0.7954359	8.39231742277876	0	PLXNA4	ENSG00000221866	0.23015	neurodevelopmental disorder with or without anomalies of the brain, eye, or heart	6	0.647016	plexin A4
0.6967001	7.90086680798075	0	APBB2	ENSG00000163697			1	0.713752	amyloid beta precursor protein binding family B member 2
0.773921	8.37068740680722	0	PLCH1	ENSG00000114805			1	0.810892	phospholipase C eta 1
0.91536427	8.4093909361377	0	FBLN5	ENSG00000140092		1 cardiovascular disease	73	1.12478	fibulin 5
0.7004652	7.97154355395077	0	SASH1	ENSG00000111961	0.00265413697168782	cardiac arrhythmia	9		SAM and SH3 domain containing 1
0.6522541	8.33091687811462	0	AMOT	ENSG00000126016	0.0104	chronic venous hypertension	26		angiomotin
0.95543957	7.8073549220576	0	SULT1C2	ENSG00000198203	0.05346	gastric non-cardia carcinoma	0		sulfotransferase family 1C member 2
0.91058826	8.22881869049588	0	PFKFB2	ENSG00000123836	0.183	Aicardi-Goutières syndrome	8		6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 2
1.007905	8.17492568250068	0	LPAR4	ENSG00000147145	0.270315277777778	familial cardiomyopathy	9		lysophosphatidic acid receptor 4
0.7229271	7.65105169117893	0	RIMKLB	ENSG00000166532	0.807798385620117	hypertension	0		ribosomal modification protein rimK like family member B
0.74335194	7.73470962022584	0	ABHD4	ENSG00000100439			0		abhydrolase domain containing 4
0.97845936	8.07146236255662	0	AP3M2	ENSG00000070718			0		adaptor related protein complex 3 subunit mu 2
0.68469715	7.72109918870718	0	DNAL1	ENSG00000163879			0		dynein axonemal light intermediate chain 1
0.8126869	8.00562454919388	0	EFHC1	ENSG00000096093			2		EF-hand domain containing 1
1.1073303	7.91886323727459	0	MNS1	ENSG00000138587			1		meiosis specific nuclear structural 1
1.0136509	7.78790255939143	0	MUC3A	ENSG00000169894			1		mucin 3A, cell surface associated

Log2 Fold Change Score	Scaled interactor number	Cluster labels	Gene symbol	Ensembl	Association overall score - OT	Disease association - OT	PubMed Report number for a gene in the context of any cardiovascular indication	GWAS standardised association score for cardiovascular indication	Description
0.61905193	8.09275714091985	0	RBM43	ENSG00000184898			3		RNA binding motif protein 43
0.8707514	8.05528243550119	0	SLC16A4	ENSG00000168679			16		solute carrier family 16 member 4
0.665926	7.6724253419715	0	SNX33	ENSG00000173548			0		sorting nexin 33
0.92883825	7.95419631038687	0	UGT2B10	ENSG00000109181			3		UDP glucuronosyltransferase family 2 member B10
0.9945507	8.02236781302845	0	XAF1	ENSG00000132530			6		XIAP associated factor 1
-1.6356926	9.01122725542325	1	GFPT2	ENSG00000131459	0.0104	myocardial infarction	4	0.046953	glutamine-fructose-6-phosphate transaminase 2
-1.0795231	9.51569983828404	1	FGF7	ENSG00000140285	0.0625452314481583	vascular disease	55	0.048529	fibroblast growth factor 7
-1.3777046	8.70735913208088	1	DHCR24	ENSG00000116133	0.02156	gastric non-cardia carcinoma	25	0.049039	24-dehydrocholesterol reductase
-1.70157549521529	9.27379559921426	1	CD163	ENSG00000177575	0.207319810015168	cardiovascular disease	493	0.051225	CD163 molecule
-1.15938	9.46964181723952	1	CTSC	ENSG00000109861	0.0487	hypertension	27	0.056076	cathepsin C
-0.7794523	9.04439411935845	1	PDE4D	ENSG00000113448		cardiovascular disease	164	0.175166	phosphodiesterase 4D
-1.0942574	8.61470984411521	1	STXBP6	ENSG00000168952	0.00259810056034269	Tachycardia	0	0.178052	syntaxin binding protein 6
-0.78551674	9.47167521439204	1	CD59	ENSG00000085063	0.324362417973247	cardiovascular disease	322	0.622682	CD59 molecule (CD59 blood group)
-1.3563299	8.49984588708321	1	SSR3	ENSG00000114850			0	0.917442	signal sequence receptor subunit 3
-1.523037	9.71938882094208	1	SELE	ENSG00000079087	0.213694829023607	cardiovascular disease	258	1.45274	selectin E
-0.8557415	8.77478705960117	1	FADS1	ENSG00000149485	0.610457862201664	heart rate	106	3.7884	fatty acid desaturase 1
-1.0573473	9.13185696060879	1	ATP1A1	ENSG00000163399		cardiovascular disease	174		ATPase Na <sup>+</sup> /K <sup>+</sup> transporting subunit alpha 1
-1.07635189301576	9.038918982923	1	TUBB6	ENSG00000176014		vascular disease	1		tubulin beta 6 class V
-1.3581958	8.85174904141606	1	TUBA3E	ENSG00000152086	0.004	cardiomyopathy	0		tubulin alpha 3e
-0.936636	8.74146698640115	1	RNASE2	ENSG00000169385	0.031424216014505	cardiovascular disease	2		ribonuclease A family member 2
-1.1835241	9.3151495622563	1	FPR1	ENSG00000171051	0.0435662720458554	cardiovascular disease	45		formyl peptide receptor 1
-1.2813988	9.13442632022093	1	DUSP5	ENSG00000138166	0.0619335443847817	cardiovascular disease	22		dual specificity phosphatase 5
-1.5813951	8.48784003382305	1	CNN1	ENSG00000130176	0.0708	dilated cardiomyopathy	34		calponin 1
-1.9903517	9.4178525148859	1	S100A8	ENSG00000143546	0.0888168957579326	cardiovascular disease	202		S100 calcium binding protein A8
-0.8386812	9.1548181090521	1	POLD2	ENSG00000106628	0.1	vasculitis	0		DNA polymerase delta 2, accessory subunit
-1.0181141	9.60547951806167	1	TUBA4A	ENSG00000127824	0.1884	Aicardi-Goutières syndrome	4		tubulin alpha 4a
-1.4041185	8.84549005094438	1	FGF18	ENSG00000156427	0.19036	Lethal faciocardiomeic dysplasia	16		fibroblast growth factor 18
-1.1663303	9.01959072835788	1	HAS2	ENSG00000170961	0.2	Familial progressive cardiac conduction defect	167		hyaluronan synthase 2
-1.3903275	9.37937836707126	1	KRT8	ENSG00000170421	0.28314666666666667	cardiomyopathy	10		keratin 8
-1.1936855	9.04165915163721	1	DLK1	ENSG00000185559	0.294136229434366	heart disease	119		delta like non-canonical Notch ligand 1
-1.2158594	8.62205181945638	1	PDPN	ENSG00000162493	0.295501223809524	Familial dilated cardiomyopathy	146		podoplanin
-1.9087296	9.2667865406949	1	KCNIP2	ENSG00000120049	0.500191615484548	heart disease	106		potassium voltage-gated channel interacting protein 2
-1.0804825	9.32192809488736	1	CSAR1	ENSG00000197405	0.75	anti-neutrophil antibody associated vasculitis	66		complement C5a receptor 1
-0.845314	8.72109918870719	1	FOSL2	ENSG00000075426	0.810660939917924	cardiovascular disease	29		FOS like 2, AP-1 transcription factor subunit
-2.168395	8.72109918870719	1	ANKRD2	ENSG00000165887			31		ankyrin repeat domain 2
-0.6821213	9.04439411935845	1	KLC2	ENSG00000174996			0		kinesin light chain 2
-0.72124004	9.10328780841202	1	SRM	ENSG00000116649			413		spermidine synthase
0.7420349	10.5774288280357	2	BMP4	ENSG00000125378	0.34218392413094	cardiovascular disease	626	0.050781	bone morphogenetic protein 4
1.3373194	10.5077946401987	2	ACTA2	ENSG00000107796		vascular disease	649	0.057922	actin alpha 2, smooth muscle
0.7945099	10.2691266791494	2	MAPK10	ENSG00000109339	0.0597055217129354	cardiovascular disease	13	0.063156	mitogen-activated protein kinase 10
0.8908138	10.6741922681457	2	NTRK2	ENSG00000148053	0.252278490923843	cardiovascular disease	60	0.114019	neurotrophic receptor tyrosine kinase 2

Log2 Fold Change Score	Scaled interactor number	Cluster labels	Gene symbol	Ensembl	Association overall score - OT	Disease association - OT	PubMed Report number for a gene in the context of any cardiovascular indication	GWAS standardised association score for cardiovascular indication	Description
0.990966969479851	10.1305705628054		ACE	ENSG00000159640		cardiovascular disease	21262	0.497336	angiotensin I converting enzyme
1.4016085	9.9901039638575		SLC6A1	ENSG00000157103	0.0368904939179823	arterial disorder	10	1.02476	solute carrier family 6 member 1
0.752505662322008	10.5877775163282		CD34	ENSG00000174059		heart rate	5335		CD34 molecule
1.7582741	9.85642552862553		HSPA2	ENSG00000126803	0.0088	hypertrophic cardiomyopathy	12		heat shock protein family A (Hsp70) member 2
0.9782324	10.0389189892923		MYH10	ENSG00000133026	0.32816138639532	dilated cardiomyopathy	28		myosin heavy chain 10
-1.4075699	8.21431912080077		PCDH7	ENSG00000169851	0.000704849016508335	retinal vascular disease	8	0.049753	protocadherin 7
-1.1952381	7.61470984411521		SLCO4A1	ENSG00000101187			3	0.058586	solute carrier organic anion transporter family member 4A1
-0.945035	7.49185309632967		PHTF2	ENSG00000006576			0	0.061346	putative homeodomain transcription factor 2
-0.66311646	8.18487534290828		TNIK	ENSG00000154310	0.883427858352661	arterial stiffness measurement	4	0.116986	TRAF2 and NCK interacting kinase
-0.62101555	7.93663793900257		COBL	ENSG00000106078			4	0.166725	cordon-bleu WH2 repeat protein
-1.0535727	7.876516946565		RDH10	ENSG00000121039	0.26294097222222	Conotruncal heart malformations	3	0.522865	retinol dehydrogenase 10
-0.6300564	7.36632221424582		METTL22	ENSG00000006785			0	0.683452	methyltransferase like 22
-1.5639668	8.17492568250068		ELL2	ENSG00000118985	0.006	Cognitive impairment-coarse facies-heart defects-obesity-pulmonary involvement-short stature-skeletal dysplasia syndrome	1	1.42467	elongation factor for RNA polymerase II 2
-1.47867868918135	7.82017896241519		ABRA	ENSG00000174429	0.00477660653187586	ischemic cardiomyopathy	40		actin binding Rho activating protein
-1.5306215	8.37503943134693		MGST1	ENSG000000008394	0.0136	congenital heart disease	8		microsomal glutathione S-transferase 1
-1.2137728	7.48381577726426		MT1A	ENSG00000205362	0.0346	cardiovascular disease	15		metallothionein 1A
-1.1985159	7.6724253419715		PXYLP1	ENSG00000155893	0.103423878550529	heart rate	0		2-phosphoxylose phosphatase 1
-0.96343840180172	8.29462074889163		LYVE1	ENSG00000133800	0.193557515740395	arterial stiffness measurement	105		lymphatic vessel endothelial hyaluronan receptor 1
-0.82270336	8.45532722030456		COTL1	ENSG00000103187	0.1948	Autosomal dominant progressive nephropathy with hypertension	2		coactosin like F-actin binding protein 1
-0.9445839	7.78790255939143		JPH1	ENSG00000104369	0.472529858350754	pericarditis	4		junctionophilin 1
-1.0970469	7.94251450533924		CENPV	ENSG00000166582			1		centromere protein V
-0.9231758	7.33091687811462		CSRNP1	ENSG00000144655			4		cysteine and serine rich nuclear protein 1
-1.3482113	8.09275714091985		RAB15	ENSG00000139998			4		RAB15, member RAS oncogene family
-0.8705368	8.43462822763673		SLA	ENSG00000155926			164		Src like adaptor
-1.1591511	7.97154355395077		SLC7A2	ENSG000000003989			20		solute carrier family 7 member 2
-0.7479849	7.89481776330794		SRPX	ENSG00000101955			5		sushi repeat containing protein X-linked
2.50118031365848	7.54689445988764		DSC1	ENSG00000134765	0.199424475243121	dilated cardiomyopathy	9	0.051188	desmocollin 1
3.59476301520953	8.78135971352466		NPPA	ENSG00000175206		cardiovascular disease	437	0.170989	natriuretic peptide A
2.472353	8.67948009950545		NEB	ENSG00000183091	0.302215434161525	cardiomyopathy	219	0.681983	nebulin
2.497488	8.98299357469431		FRZB	ENSG00000162998	0.056973109948679	dilated cardiomyopathy	15		frizzled related protein
2.759101	8.63299519714296		SFRP4	ENSG00000106483	0.0643642029123344	cardiovascular disease	40		secreted frizzled related protein 4
1.9342127	6.98868468677217		KLHL38	ENSG00000175946			0		kelch like family member 38
-1.3976002	10.4051414631363		TFRC	ENSG00000072274	0.314934748274738	cardiovascular disease	51	0.048463	transferrin receptor
-0.67367744	10.7960396088298		TLR2	ENSG00000137462	0.338338971159148	cardiovascular disease	961	0.065383	toll like receptor 2
-0.6312599	9.66711154207503		PHB	ENSG00000167085	0.0404	pulmonary arterial hypertension	122	0.067672	prohibitin
-1.4020166	10.6211361132746		CDKN1A	ENSG00000124762		cardiovascular disease	547	0.139652	cyclin dependent kinase inhibitor 1A
-1.4020166	10.6211361132746		CDKN1A	ENSG00000124762		cardiovascular disease	547	0.139652	cyclin dependent kinase inhibitor 1A
-0.9807377	9.77313920671969		TUBB4B	ENSG00000188229		heart disease	3	2.58984	tubulin beta 4B class IVb

Log2 Fold Change Score	Scaled interactor number	Cluster labels	Gene symbol	Ensembl	Association overall score - OT	Disease association - OT	PubMed Report number for a gene in the context of any cardiovascular indication	GWAS standardised association score for cardiovascular indication	Description
-1.2778900743889	11.5211096436513		IL6	ENSG00000136244		1 cardiovascular disease	2787		interleukin 6
-0.78205204	9.7176764230664		FASN	ENSG00000169710	0.0518848135528278	cardiovascular disease	430		fatty acid synthase
-0.718607957	10.1623913287569		GLUL	ENSG00000155821	0.0581662250285504	cardiovascular disease	20		glutamate-ammonia ligase
-0.7444143	10.0647427647503		AIF1	ENSG00000204472	0.322583395917535	cardiovascular disease	270		allograft inflammatory factor 1
-0.8504858	9.6599589242998		EIF3I	ENSG00000084623			1		eukaryotic translation initiation factor 3 subunit I
-0.6790819	9.65642486327778		TUBG1	ENSG00000131462			1		tubulin gamma 1
0.716502807383939	7.13955135239879		MTURN	ENSG00000180354	0.0168074907297785	dilated cardiomyopathy	0	0.048052	maturin, neural progenitor differentiation regulator homolog
0.71768475	7.09803208296053		CPXM2	ENSG00000121898	0.0124	cardiotoxicity	2	0.048572	carboxypeptidase X, M14 family member 2
0.66207695	7.4093909361377		SH3TC2	ENSG00000169247	0.888206541538239	hypertension	3	0.069967	SH3 domain and tetratricopeptide repeats 2
1.2451935	7.4178525148859		PAMR1	ENSG00000149090			8	0.186925	peptidase domain containing associated with muscle regeneration 1
0.8537178	6.88264304936184		CCDC113	ENSG00000103021			0	0.497276	coiled-coil domain containing 113
0.9098177	7.27612440527424		CCDC3	ENSG00000151468			2	0.528913	coiled-coil domain containing 3
0.81960773	7.10852445677817		THSD7A	ENSG00000005108	0.0358388888888889	hypertension	17	0.531226	thrombospondin type 1 domain containing 7A
1.223135	7.07681559705083		CFAP61	ENSG00000089101			1	0.637975	cilia and flagella associated protein 61
0.9419317	6.76818432477693		FAM13C	ENSG00000148541			0	0.669442	family with sequence similarity 13 member C
0.702137	7.4262647547021		TANGO2	ENSG00000183597		1 cardiovascular disease	10	0.670464	transport and golgi organization 2 homolog
0.73056316	6.90689059560852		JCAD	ENSG00000165757		1 cardiovascular disease	35		junctional cadherin 5 associated
1.34472127978344	6.2667865406949		TMEM140	ENSG00000146859	0.00534632110033744	ischemic cardiomyopathy	0		transmembrane protein 140
0.8866329	6.5077946401987		C1QTNF7	ENSG00000163145	0.0076	coronary artery disease	0		C1q and TNF related 7
1.3952188	6.85798099512757		SLC44A5	ENSG00000137968	0.00936438763471833	cardiovascular disease	1		solute carrier family 44 member 5
1.0713959	7.05528243550119		CRISPLD1	ENSG00000121005	0.0252	heart failure	2		cysteine rich secretory protein LCCL domain containing 1
0.736742	6.55458885167764		ZNF704	ENSG00000164684	0.120394639670849	arterial stiffness measurement	2		zinc finger protein 704
0.8174572	6.6724253419715		APBB3	ENSG00000113108			1		amyloid beta precursor protein binding family B member 3
1.1442862	7.08746284125034		ART4	ENSG00000111339			4		ADP-ribosyltransferase 4 (Dombrock blood group)
0.8407326	7.33091687811462		BCL6B	ENSG00000161940			4		BCL6B transcription repressor
0.6358671	6.61470984411521		BTN3A1	ENSG00000026950			0		butyrophilin subfamily 3 member A1
0.66088676	6.22881869049588		CCDC171	ENSG00000164989			0		coiled-coil domain containing 171
0.95116615	7.23840473932508		CDR1	ENSG00000184258			15		cerebellar degeneration related protein 1
0.7051039	6.52356195605701		CEP126	ENSG00000110318			0		centrosomal protein 126
1.0097842	7.09803208296053		KRTAP21-1	ENSG00000187005			0		keratin associated protein 21-1
1.0780964	6.18982455888002		NRK	ENSG00000123572			114		Nik related kinase
0.79331493	7.06608919045777		PCDH12	ENSG00000113555			5		protocadherin 12
1.0529556	6.79441586635011		PIK3IP1	ENSG00000100100			4		phosphoinositide-3-kinase interacting protein 1
0.7104025	7.04439411935845		RANBP3L	ENSG00000164188			2		RAN binding protein 3 like
1.1970367	7.33985000288462		SESN3	ENSG00000149212			10		sestrin 3
1.2122259	7.04439411935845		SULT1C4	ENSG00000198075			0		sulfotransferase family 1C member 4
0.87745094	7.4594316186373		YPEL1	ENSG00000100027			1		yippe like 1
1.6721287	8.99717948093762		FMOD	ENSG00000122176	0.048	heart failure	20	0.047068	fibromodulin

Log2 Fold Change Score	Scaled interactor number	Cluster labels	Gene symbol	Ensembl	Association overall score - OT	Disease association - OT	PubMed Report number for a gene in the context of any cardiovascular indication	GWAS standardised association score for cardiovascular indication	Description
1.15573458676483	8.24317398347295		FNDC1	ENSG00000164694	0.00365269911261096	ischemic cardiomyopathy	5	0.078697	fibronectin type III domain containing 1
1.3211765	8.29920801838728		DPT	ENSG00000143196	0.357486873865128	arterial stiffness measurement	345	0.0808	dermatopontin
1.7052364	8.14974711950468		COL14A1	ENSG00000187955	0.412141352891922	heart rate response to exercise	12	0.083512	collagen type XIV alpha 1 chain
1.8713417	8.10852445677817		NPR 3.00	ENSG00000113389		cardiovascular disease	58	0.101712	natriuretic peptide receptor 3
1.5688696	7.92481250360578		SVEP1	ENSG00000165124		cardiovascular disease	13	0.112483	sushi, von Willebrand factor type A, EGF and pentraxin domain containing 1
1.30003064704471	8.15987133677839		SMOC2	ENSG00000112562	0.0182297974996283	vascular disease	13	0.147347	SPARC related modular calcium binding 2
1.33498466496198	8.34429590791582		LTBP2	ENSG00000119681	0.199063555555556	heart disease	26	0.184024	latent transforming growth factor beta binding protein 2
1.5091677	8.45532722030456		KCNJ3	ENSG00000162989	0.2723875	cardiac arrhythmia	36	0.57039	potassium inwardly rectifying channel subfamily J member 3
1.4191274692	8.62935662007961		PDE5A	ENSG00000138735		cardiovascular disease	380	0.917442	phosphodiesterase 5A
1.3353043	7.99435343685886		HTR4	ENSG00000164270		acute myocardial infarction	11	1.09022	5-hydroxytryptamine receptor 4
1.98462521436403	8.3037807481771		PRELP	ENSG00000188783	0.0256	heart failure	7		proline and arginine rich end leucine rich repeat protein
1.5719681	7.76818432477693		FAXDC2	ENSG00000170271	0.112923523411155	electrocardiography	0		fatty acid hydroxylase domain containing 2
1.22277252337422	8.4757334309664		SLC40A1	ENSG00000138449	0.2	coronary artery disease, autosomal dominant 2	18		solute carrier family 40 member 1
1.6482477	8.56985560833095		P2RY14	ENSG00000174944	0.292136933333333	heart disease	4		purinergic receptor P2Y14
2.0397625	8.59991284218713		ENPP2	ENSG00000136960	0.32104693638164	cardiovascular disease	18		ectonucleotide pyrophosphatase/phosphodiesterase 2
1.7696838	8.10852445677817		IGSF10	ENSG00000152580			1		immunoglobulin superfamily member 10
-0.8760967	5.58496250072116		LBH	ENSG00000213626	0.0183457542628309	cerebrovascular disorder	72		LBH regulator of WNT signaling pathway
-1.51342762019995	6.89481776330794		FCN3	ENSG00000142748	0.0623	heart failure	11		ficolin 3
-0.6520672	6.49185309632967		C19orf47	ENSG00000160392	0.107490286231041	resting heart rate	0		chromosome 19 open reading frame 47
-1.0985508	4.8073549220576		CDRT15	ENSG00000223510			0		CMT1A duplicated region transcript 15
-0.79782104	6.89481776330794		JAGN1	ENSG00000171135			2		jagunal homolog 1
-1.3613062	7.12928301694497		MEDAG	ENSG00000102802			0		mesenteric estrogen dependent adipogenesis
-1.238616	6.84549005094437		RTL9	ENSG00000243978			0		retrotransposon Gag like 9
0.8970251	8.62570884306447		TIE1	ENSG00000066056	0.261246111111111	hypertrophic cardiomyopathy	120	0.049117	tyrosine kinase with immunoglobulin like and EGF like domains 1
1.0541258	8.48784003382305		ANO1	ENSG00000131620	0.443062752485275	aortic root size	91	0.053002	anoctamin 1
1.178196	8.76818432477693		CPE	ENSG00000109472	0.2	coronary artery disease, autosomal dominant 2	280	0.056875	carboxypeptidase E
0.8587904	8.74483383749955		SEMA6A	ENSG00000092421	0.119487526988983	Anti-neutrophil cytoplasmic antibody-associated vasculitis	14	0.059624	semaphorin 6A
1.0189104	9.39660478118186		PRDM1	ENSG00000057657	0.2	Genetic cardiac anomaly	32	0.061484	PR/SET domain 1
0.96929158248	9.47167521439204		MME	ENSG00000196549		cardiovascular disease	84	0.100575	membrane metalloendopeptidase
0.8499718	8.88874324889826		SMAD9	ENSG00000120693		vascular disease	34	0.108775	SMAD family member 9
0.77372265	8.74483383749955		CENPC	ENSG00000145241	0.023279536715657	cardiomyopathy	0	0.112658	centromere protein C
1.0676622	8.59991284218713		HMCN1	ENSG00000143341	0.201	dilated cardiomyopathy	4	0.153732	hemicentin 1
0.81258965	8.61838550225861		LTBP1	ENSG00000049323	0.317106207893006	Genetic cardiac anomaly	35	0.179172	latent transforming growth factor beta binding protein 1
0.79587173	8.90989308377004		ARHGEF28	ENSG00000214944	0.0359871399596876	vascular disease	0	0.230931	Rho guanine nucleotide exchange factor 28
0.74583626	9.05528243550119		LDB2	ENSG00000169744			14	0.235095	LIM domain binding 2

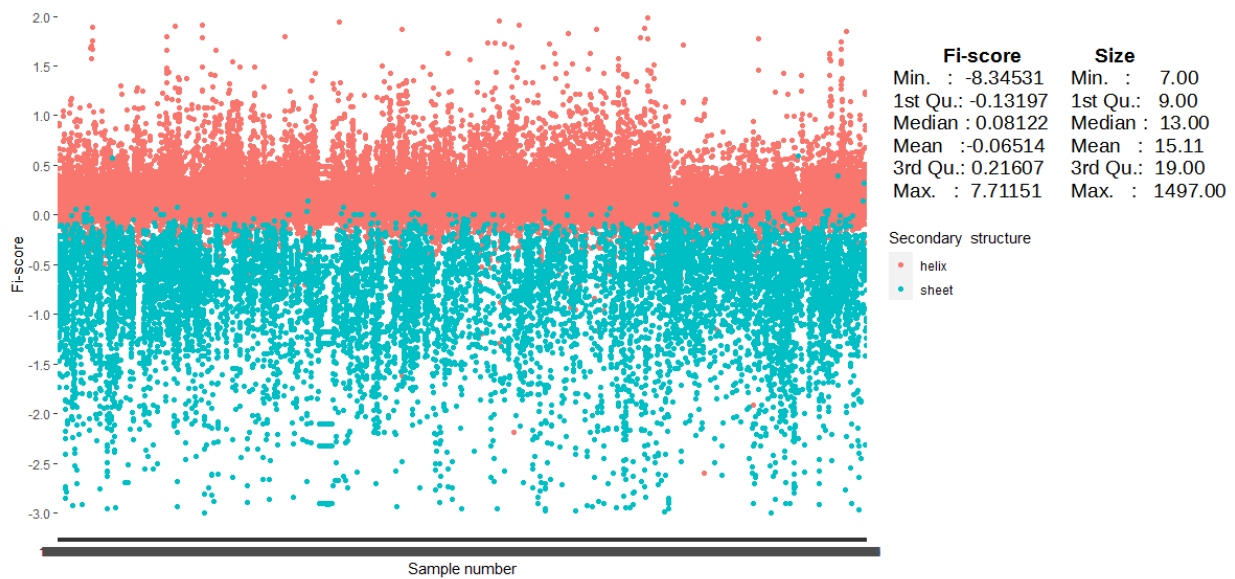
Log2 Fold Change Score	Scaled interactor number	Cluster labels	Gene symbol	Ensembl	Association overall score - OT	Disease association - OT	PubMed Report number for a gene in the context of any cardiovascular indication	GWAS standardised association score for cardiovascular indication	Description
1.1536655	8.75822321472672		PLCE1	ENSG00000138193		1 cardiovascular disease	17	0.242876	phospholipase C epsilon 1
0.71289444	9.04712391211403		PCSK5	ENSG000000099139	0.334084807112215	1 cardiovascular disease	15	0.279217	proprotein convertase subtilisin/kexin type 5
0.9334259	9.43879185257826		NRG1	ENSG00000157168	0.83868237015398	1 cardiovascular disease	202	0.302508	neuregulin 1
0.9914799	8.59991284218713		KCNN3	ENSG00000143603		1 cardiovascular disease	95	0.324822	potassium calcium-activated channel subfamily N member 3
0.61952114	9.38801728534514		HLA-B	ENSG00000234745	0.936991677736148	1 cardiovascular disease	567	0.336932	major histocompatibility complex, class I, B
0.85855675	8.94251450533924		BMP6	ENSG00000153162	0.018	1 congenital heart disease	72	0.510348	bone morphogenetic protein 6
0.8962126	8.58871463558226		PREX2	ENSG00000046889		1 Cerebral arteriovenous malformation	2	0.580283	phosphatidylinositol-3,4,5-trisphosphate dependent Rac exchange factor 2
1.1325579	8.58871463558226		ITGA8	ENSG00000077943	0.2	1 Congenital vertebral-cardiac-renal anomalies syndrome	11	0.660239	integrin subunit alpha 8
0.69878006	9.24792751344359		MYO10	ENSG00000145555	0.1948	1 Autosomal dominant progressive nephropathy with hypertension	7	0.670943	myosin X
0.67784977	8.91288933622996		MECOM	ENSG00000085276		1 cardiovascular disease	23	0.77375	MDS1 and EVI1 complex locus
1.0650034	8.8569637333939		SGIP1	ENSG00000118473	0.991233631968498	1 heart rate	2	1.04846	SH3GL interacting endocytic adaptor 1
0.90437603	9.55074678538324		SLIT3	ENSG00000184347	0.298000860997732	1 Familial dilated cardiomyopathy	28	1.18762	slit guidance ligand 3
1.0028601	8.38801728534514		SPOCK1	ENSG00000152377	0.02352	1 gastric non-cardia carcinoma	5	1.24289	SPARC (osteonectin), cwcv and kazal like domains proteoglycan 1
0.7893839	9.14974711950468		ENTPD1	ENSG00000138185	0.339209168172997	1 cardiovascular disease	32	1.61717	ectonucleoside triphosphate diphosphohydrolase 1
0.83325577	9.14974711950468		TIMP2	ENSG000000035862	0.102861488480588	1 cardiovascular disease	294	1.87707	TIMP metalloproteinase inhibitor 2
0.6729946	9.00281501560705		FZD4	ENSG00000174804		1 retinal vascular disease	57		frizzled class receptor 4
0.8965254	9.20945338652895		KCNAA5	ENSG00000130037		1 cardiac arrhythmia	235		potassium voltage-gated channel subfamily A member 5
0.7015095	9.40087943628218		VEGFC	ENSG00000150630		1 cardiovascular disease	490		vascular endothelial growth factor C
0.7166195	9.52552080909507		C/CG2	ENSG00000138764	0.00753244040668435	1 dilated cardiomyopathy	3		cyclin G2
0.852013228649132	9.06608919045777		NAP1L3	ENSG00000186310	0.00912202178593987	1 dilated cardiomyopathy	1		nucleosome assembly protein 1 like 3
1.14453691610264	8.53915881110803		LMO3	ENSG00000048540	0.0104	1 hypertension	4		LIM domain only 3
1.00368299364409	8.70390357344466		FZD7	ENSG00000155780	0.0128077432134922	1 ischemic cardiomyopathy	14		frizzled class receptor 7
1.03604149258193	8.60362634498619		TM7SF2	ENSG00000149809	0.0444267380952381	1 cardiovascular disease	3		transmembrane 7 superfamily member 2
1.0330381	9.09539702279256		IGFBP5	ENSG00000115461	0.0484757037273693	1 vascular disease	29		insulin like growth factor binding protein 5
1.1550779	8.90989308377004		P2RY13	ENSG00000181631	0.0614463333333333	1 cardiovascular disease	12		purinergic receptor P2Y13
0.6970482	9.05799172275918		ARHGAP1	ENSG00000175220	0.2	1 Familial avascular necrosis of femoral head	5		Rho GTPase activating protein 1
0.76843166	9.41362792902417		EFNB2	ENSG00000125266	0.340913482063927	1 cardiovascular disease	63		ephrin B2
0.8135996	9.55074678538324		TCF4	ENSG00000196628	0.498447239398956	1 heart rate response to exercise	131		transcription factor 4
0.7375641	8.58498250072116		TP53INP1	ENSG00000164938	0.559791449461632	1 cardiovascular disease	10		tumor protein p53 inducible nuclear protein 1
0.6842003	8.9915218460757		HEY1	ENSG00000164683			115		hes related family bHLH transcription factor with YRPW motif 1



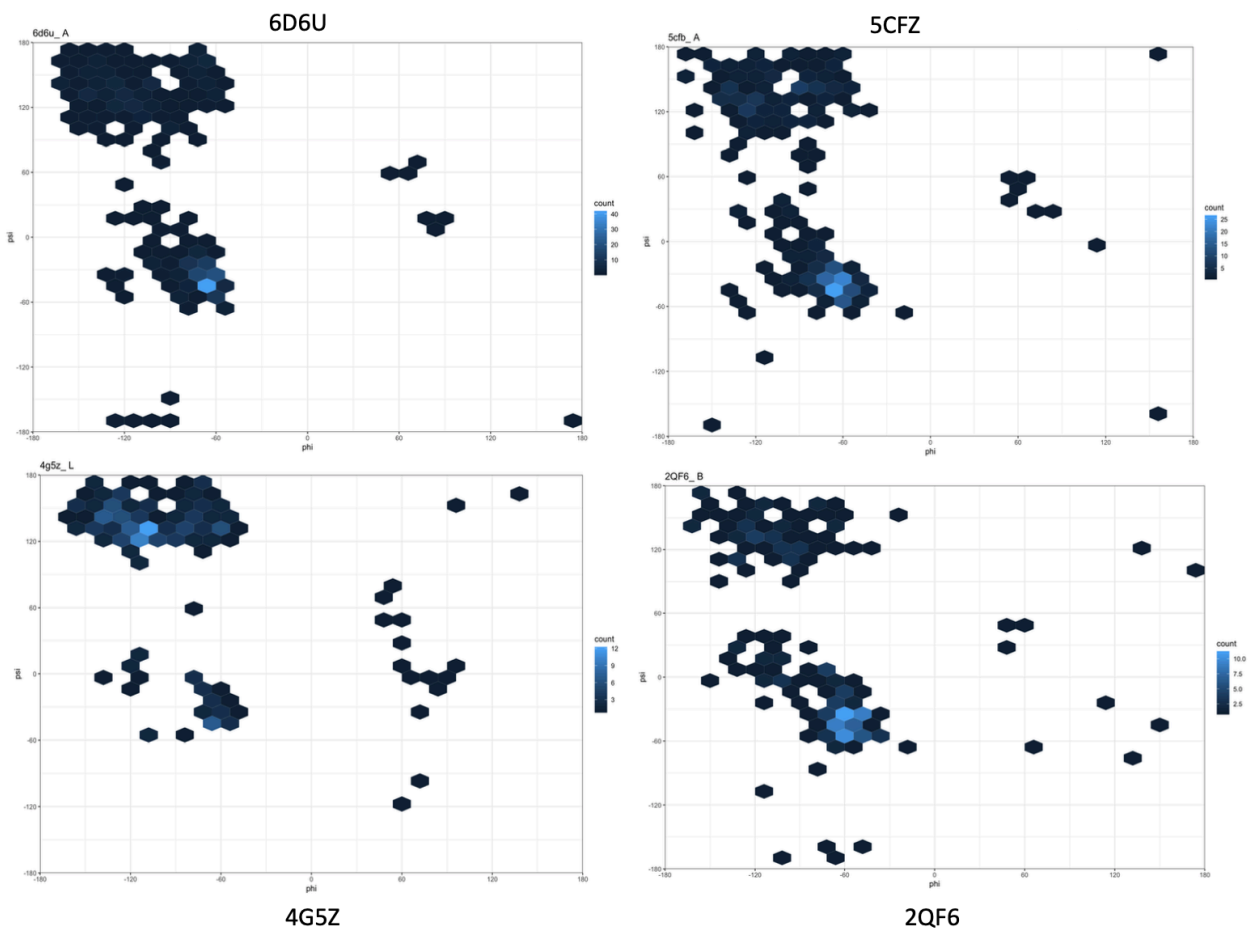
## **Integrative *omics* approaches for new target identification and therapeutics development**

### **10.2. Fi-score: a novel approach to characterise protein topology and aid in drug discovery studies**

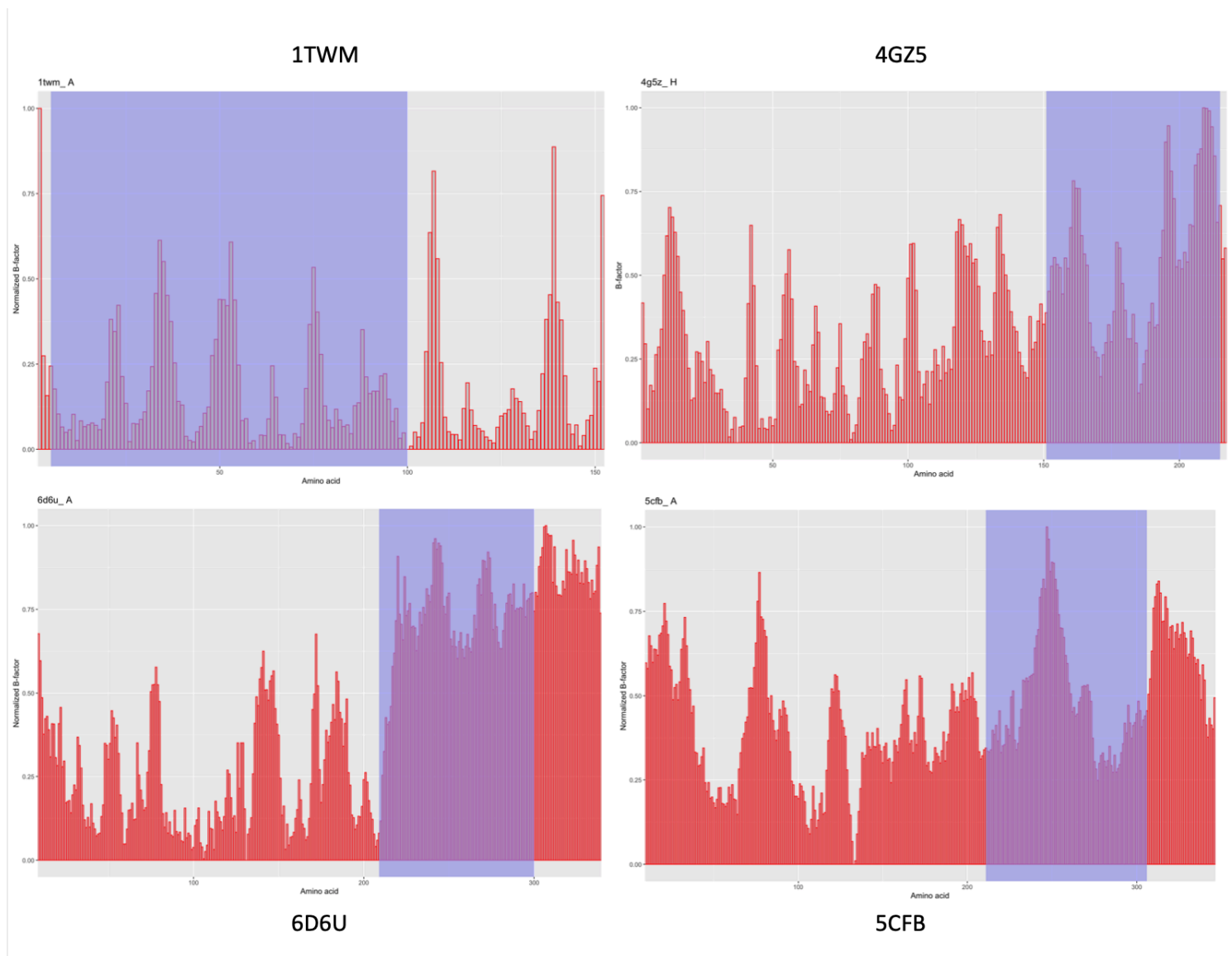
## Supplementary materials



**Supplementary Figure 1.** 3352 protein test set (total 50,043 secondary structure elements, Suppl. Table 1) was scored based on the Fi-score capturing  $\alpha$ -helices and  $\beta$ -sheets as well as rarer structural elements. Summary table for the dot plot shows the main distribution parameters for the investigated Fi-scores and structure sizes. Graph and summary table were created with R/RStudio.

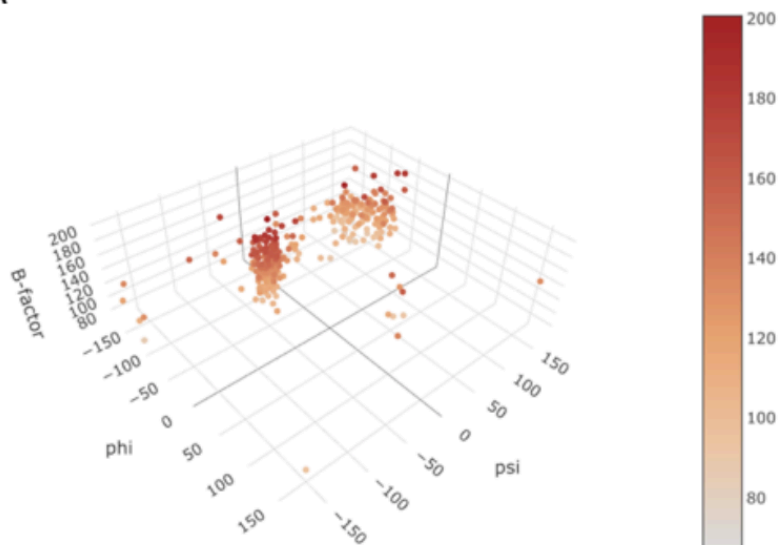


**Supplementary Figure 2.** Representative examples of Ramachandran plots for representative proteins PDB ID: 6D6U(A chain), 5CFZ (A chain), 4G5Z (L chain), 2QF6 (chain B). Graphs created with with R/RStudio.

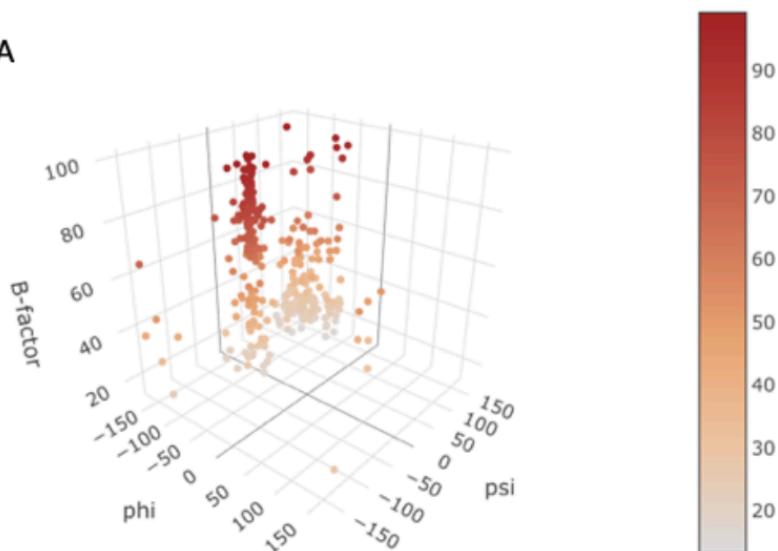


**Supplementary Figure 3.** Representative examples of normalised B-factor value distribution (from 0 to 1) for proteins PDB ID: 1TWM (chain A), 4G5Z (H chain), 6D6U (A chain), 5CFZ (A chain), where shaded blue region represents analysed regions (Table 1). Graphs created with with R/RStudio.

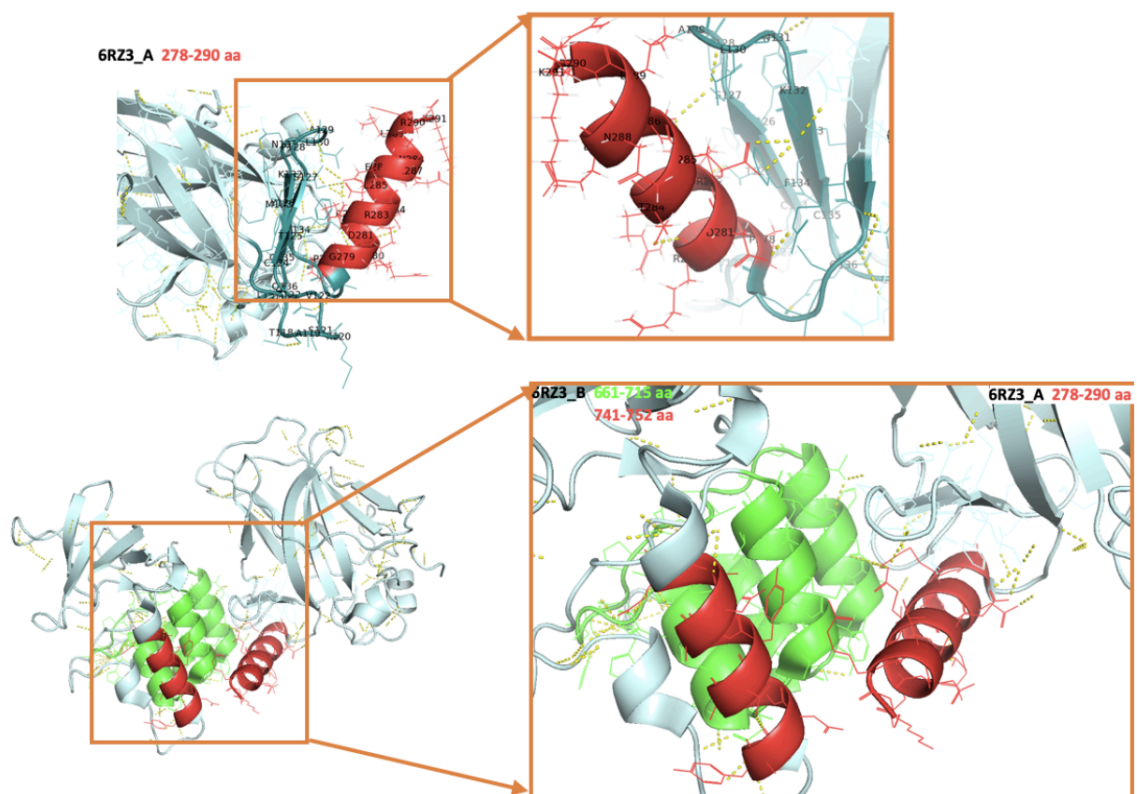
5CFB\_A



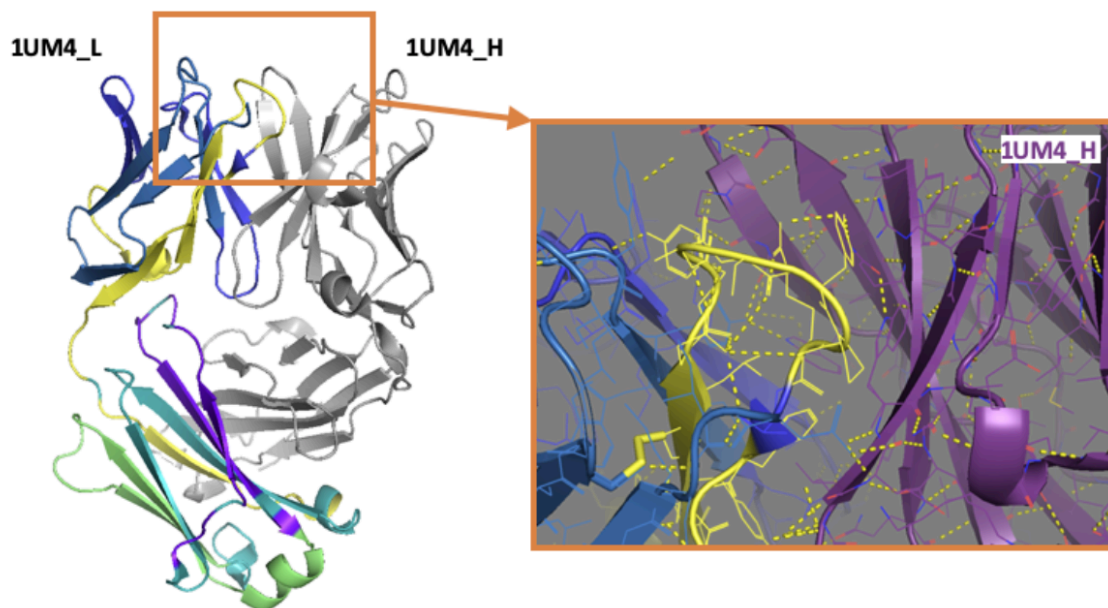
6D6U\_A



**Supplementary Figure 4.** Representative examples of torsion angle and B-factor value distribution for human GABA-A receptor, subunit beta-2 (PDB ID: 6D6U, chain A) and human glycine receptor alpha-3 (PDB ID: 5CFB, chain A) where a colour scale represents B-factor value without normalisation. Graphs created with with R/RStudio.



**Supplementary Figure 5.** Region of a single outer  $\alpha$ -helix of cellular tumour antigen p53 (PDB ID: 6RZ3, chain A) (top panel) and a contact site between the outer  $\alpha$ -helix of cellular tumour antigen p53 (PDB ID: 6RZ3, chain A) and the carboxyl-terminal conserved region of inhibitor of apoptosis-stimulating protein of p53 (iASPP) (PDB ID: 6RZ3, chain B) where yellow dotted lines represent interchain polar contacts (bottom panel). 3D molecule images rendered with PyMol.



**Supplementary Figure 6.** Catalytic antibody 21H3 with haptens (PDB ID: 1UM4, chain H and L) where N-terminal heavy and light chain contact site are shown in a close-up with polar contacts depicted in a dashed yellow line. 3D molecule images rendered with PyMol.

## **Integrative *omics* approaches for new target identification and therapeutics development**

### **10.3. *Fiscore* package: effective protein structural data visualisation and exploration**

**Supplementary Table 1.** PSI-BLAST alignment results.

<b>PDB ID</b>	<b>Description</b>	<b>Scientific Name</b>	<b>Max Score</b>	<b>Total Score</b>	<b>Query Cover</b>	<b>E value</b>	<b>Per. Ident</b>	<b>Acc. Len</b>	<b>Accession</b>
<b>3KMR, 1FBY</b>	retinoic acid receptor alpha isoform 1 [Homo sapiens]	Homo sapiens	96.7	96.7	80%	4E-23	32.64%	462	NP_000955.1
<b>2GPU, 6KNR</b>	estrogen-related receptor gamma isoform 1 [Homo sapiens]	Homo sapiens	83.2	83.2	85%	2E-18	30.73%	458	NP_001429.2



**Supplementary Table 2.** Student T-test (two-sided, unpaired) results.

<b>Fi-score distribution 1</b>	<b>Fi-score distribution 2</b>	<b>T-value</b>	<b>p-value</b>
Nur77	Retinoic acid receptor alpha	0.62868	0.5367
Nur77	Estrogen-related receptor gamma	-0.49413	0.6279
Estrogen-related receptor gamma	Retinoic acid receptor alpha	-0.55116	0.5891

**Supplementary Figure 1.** Nur77 ligand binding domain PSI-BLAST alignment with the retinoic acid receptor alpha.

retinoic acid receptor alpha isoform 1 [Homo sapiens]

Sequence ID: [NP\\_000955.1](#) Length: 462 Number of Matches: 1

**Range 1: 227 to 417** [GenPept](#) [Graphics](#)

[▼ Next Match](#) [▲ Previous Match](#)

Score	Expect	Method	Identities	Positives	Gaps
96.7 bits(239)	4e-23	Compositional matrix adjust.	63/193(33%)	105/193(54%)	4/193(2%)
Query 46	QFYDLLSGSLEVIRKWA EKIPGFAELSPADQDLLLESFALELFILRLAYRSKPGEGKLIF				105
	+F +L + + ++A+++PGF L+ ADQ LL++A L++ ILR+ R P + + F				
Sbjct 227	KFSELSTKCI IKTVEFAKQLPGFTTLTIADQITLLKAAACLDILILRICTRYTPEQDTMTF				286
Query 106	CSGLVLHRLQCAR-GFGDWIDSILAFSRSLSLHLLVDVPAFACLSALVLIT-DRHGLQEPR				163
	GL L+R Q GFG D + AF+ L L +D LSA+ LI DR L++P				
Sbjct 287	SDGLTLNRTQMHNAGFGPLTDLVFAFANQLLPLEMDDAETGLLSAICLICGDRQDLEQPD				346
Query 164	RVEELQNRIASCLKEHVA AVAGEPQPASCLSRLGKLP ERLCTQGLQRI FYLKLEDLV				223
	RV+ LQ + LK +V P ++L K+ +LR++ +G +R+ LK+E				
Sbjct 347	RVDMLQEPLLEALKVYVR - -KRRPSRPHMFPKMLMKITDLRSISAKGAERVITLKMEIPG				404
Query 224	PPPIIDKIFMDT 236				
	PP+I ++ ++				
Sbjct 405	SMPPLIQEMLENS 417				

**Supplementary Figure 2.** Nur77 ligand binding domain PSI-BLAST alignment with the estrogen-related receptor gamma.

estrogen-related receptor gamma isoform 1 [Homo sapiens]

Sequence ID: [NP\\_001429.2](#) Length: 458 Number of Matches: 1

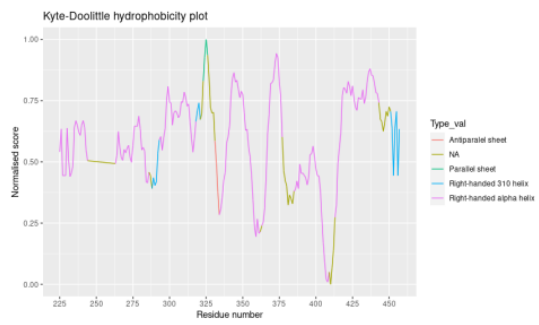
Range 1: 255 to 452 [GenPept](#) [Graphics](#)

[▼ Next Match](#) [▲ Previous Match](#)

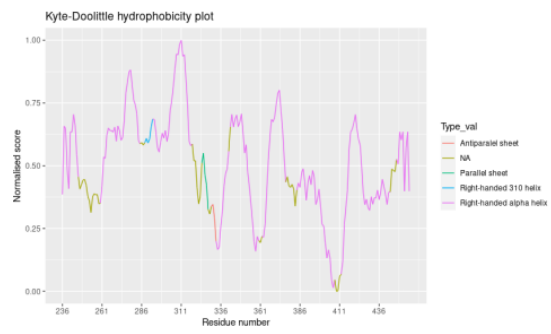
Score	Expect	Method	Identities	Positives	Gaps
83.2 bits(204)	2e-18	Compositional matrix adjust.	63/205(31%)	105/205(51%)	10/205(4%)
Query 34		PHFGKEDAGDVQQFYDLLSGSLEVIRKWAEEKIPGFAELSPADQDLLLESFALELFILRLA			93
		P D + DL L VI WA+ IPGF+ LS ADQ LL+SA++E+ IL +			
Sbjct 255		PTVPDSDIKALTTLCDLADRELVVIIGWAKHIPGFSTLSLADQMSSLQSAWMEILILGVV			314
Query 94		YRSKPGEGKLIFCSGLVLRHQCA-RGFGDWIDSILAFSRLHSLLDVPAFACLSALVL			152
		YRS E +L++ ++ Q G D ++IL + S+ ++ F L A+ L			
Sbjct 315		YRSLSFEDLVYADDYIMDEDQSKLAGLLDLNNAIQLVKKYKSMKLEKEEFVTLKAIAL			374
Query 153		I-TDRHGLQEPRRVEELQNRIASCLKEHVAAVAGE-PQPASCLSRLLGKLPRLTCTQG			210
		+D +++ V++LQ+ + L+++ A E P+ A ++L LP LR T+			
Sbjct 375		ANSDSMHIEDVEAVQKLQDVLHEALQDYEAGQHMEDPRRA--GKMLMTPLLRQTSTKA			431
Query 211		LQRIFYLKLEDLVPPPIIDKIFMD	235		
		+Q + +KLE VP + K+F++			
Sbjct 432		VQHFYNIKLEGKVP---MHKLFLE	452		

Supplementary Figure 3. Hydrophobicity plots.

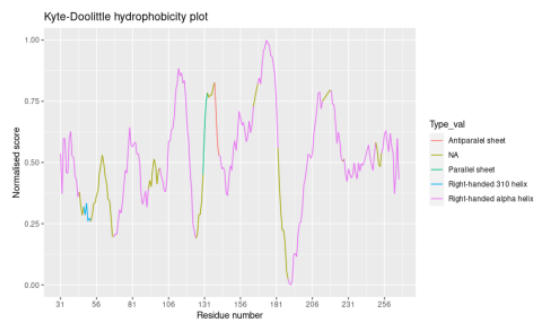
**1FBY - retinoic acid receptor alpha**



**6KNR - estrogen-related receptor gamma**



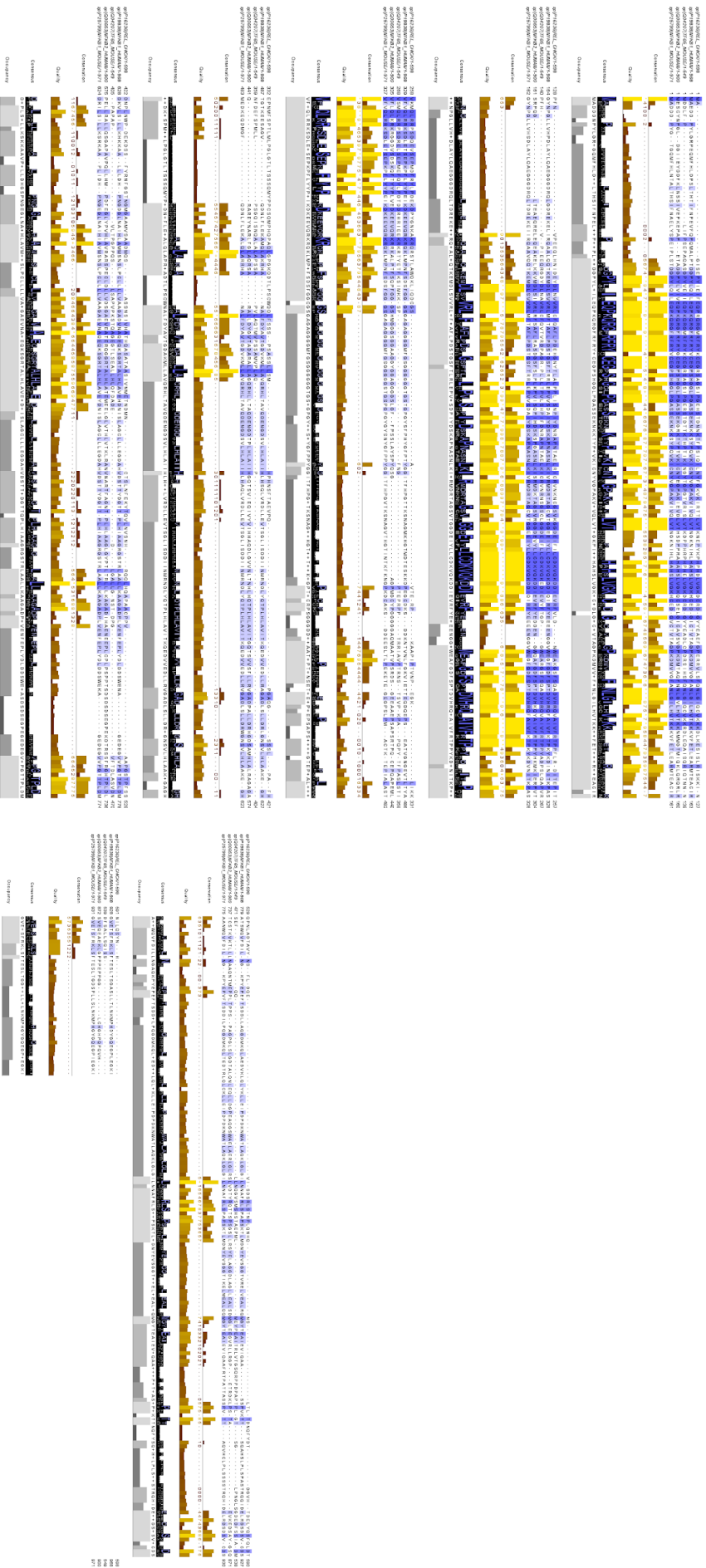
**6KZ5 - NUR77 Ligand binding Domain**



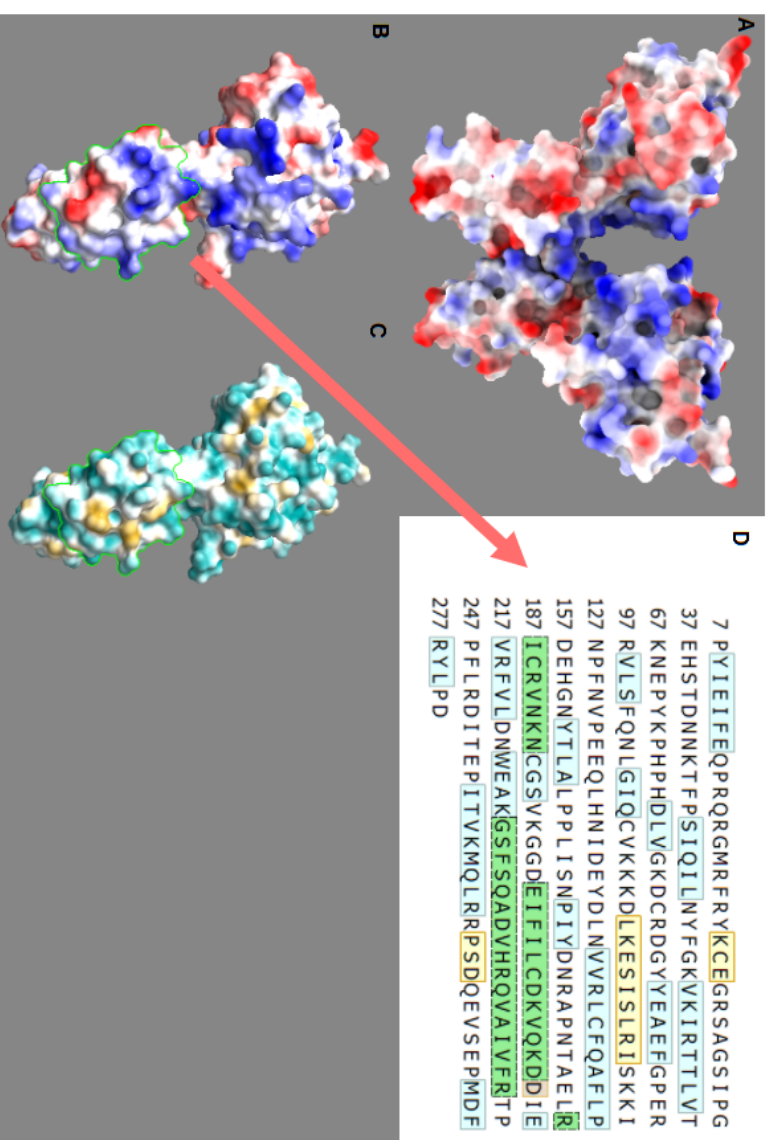
## **Integrative *omics* approaches for new target identification and therapeutics development**

### **10.4. *In silico* drug discovery for a complex immunotherapeutic target - human c-Rel protein**

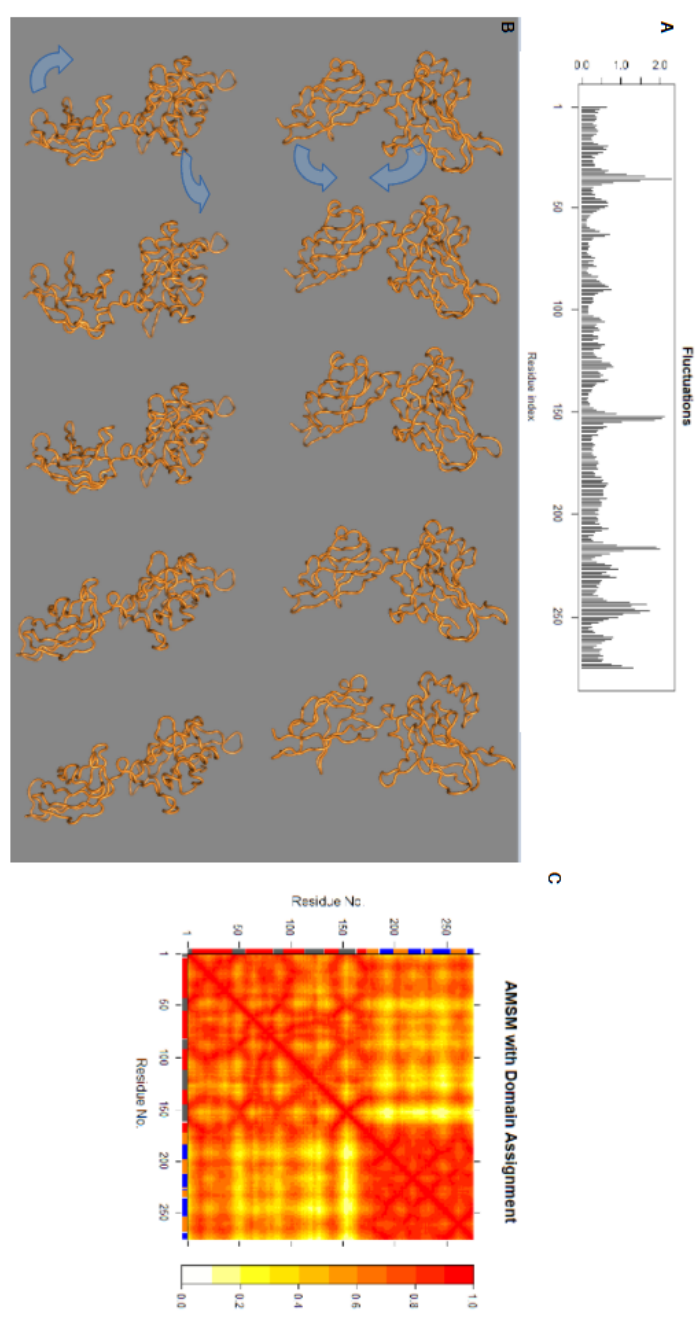
**Supplementary Figure 1.** T-Coffee sequence alignment for REL proteins using default settings where the higher identity percentage is represented with a more intense blue colour; additional parameters, such as the alignment quality score, conservation score, occupancy and consensus sequence are also provided with the alignment.



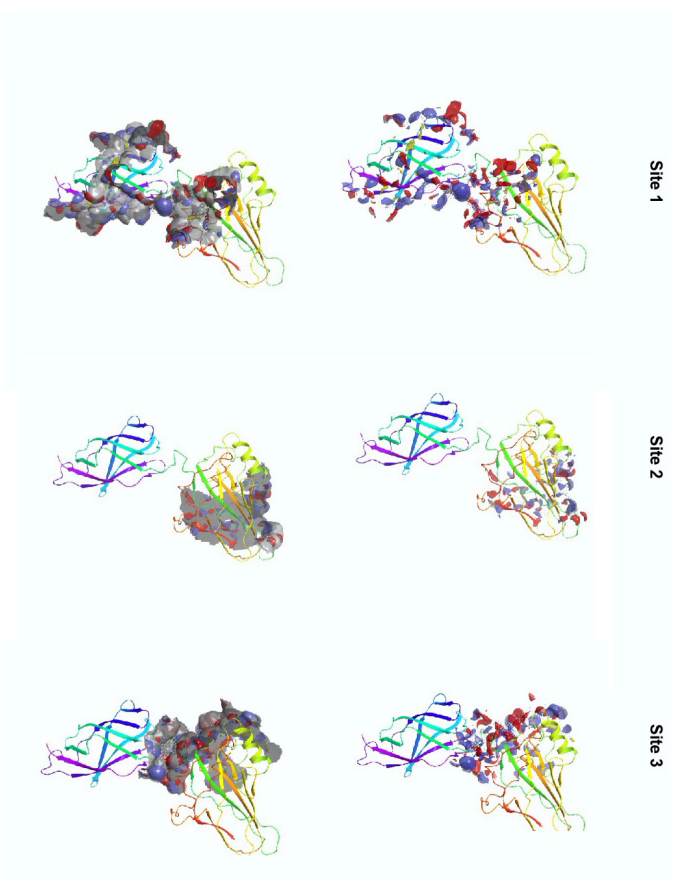
**Supplementary Figure 2.** c-Rel protein (PDB ID: 1GJ1) dimer and monomer visualization (A&B) where red-blue spectrum represents Coulombic electrostatic potential ranging from negative to positive, respectively. The monomer coloring ranges from dark cyan for the most hydrophilic region through white to dark golden for the most hydrophobic site (C). Protein sequence panel (D) shows helix regions in yellow, beta-strands in blue and selected dimer lock region in green which is also contoured around contact sites (B&C).



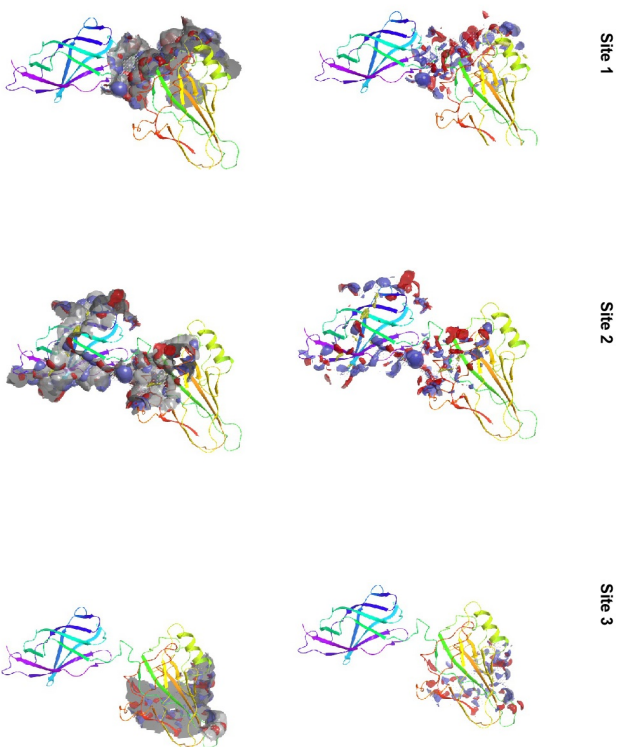
**Supplementary Figure 3.** c-Rel protein (PDB ID: 1G1j) atomic movement fluctuations per residue (A) and snapshots of the highest frequency (0.004) movements (B) based on the normal mode analysis (NMA); atomic movement similarity matrix (AMSM) provides a specific domain assignment (C).



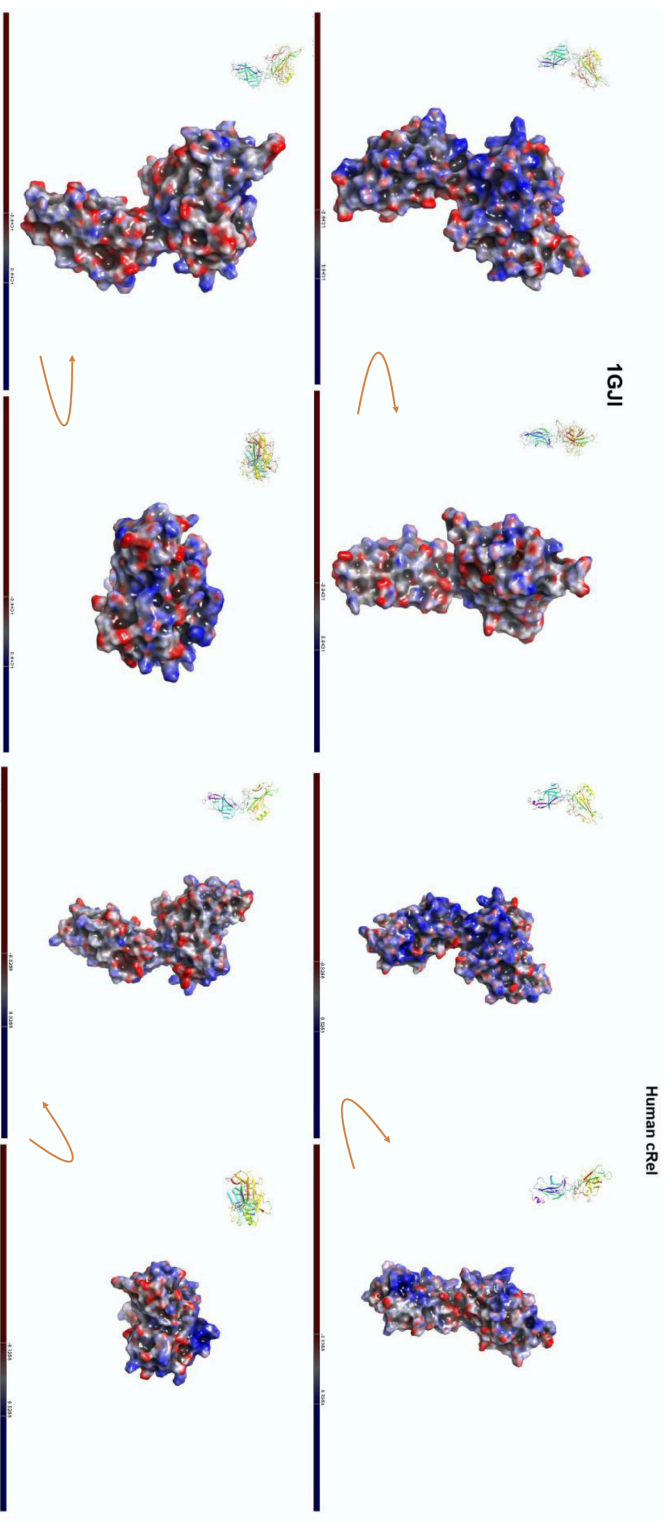




**Supplementary Figure 4.** Surface distribution for chicken c-Rel (PDB ID: 1G11) of selected three sites. Top panels represent contact surface of hydrogen-bond donor (blue), hydrogen-bond acceptor (red), hydrophobic sites are coloured in yellow. Bottom panels represent filled surface.

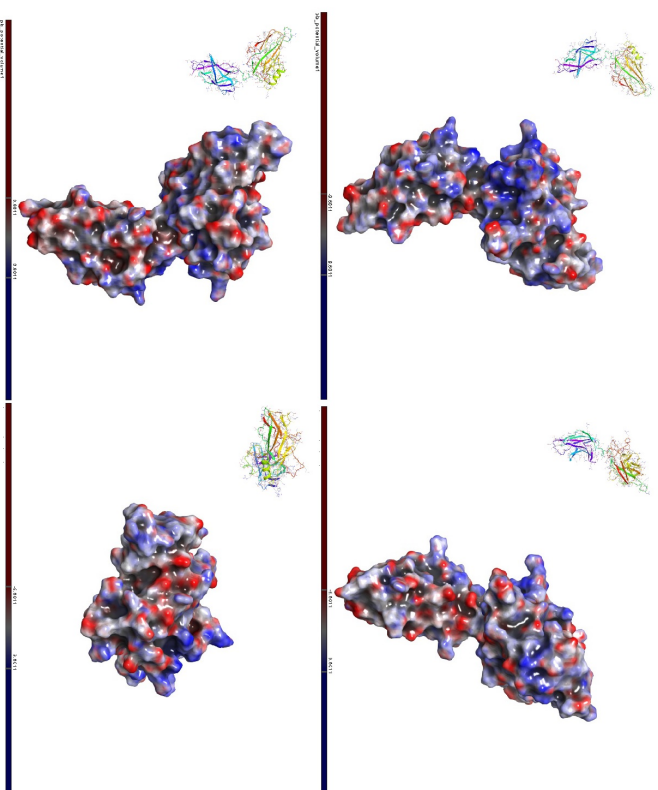


**Supplementary Figure 5. Three binding sites for mouse p65 (PDB ID: 5U01).** Top panels represent the contact surface for hydrogen-bond donor (blue), hydrogen-bond acceptor (red) and hydrophobic sites are coloured in yellow. Bottom panels represent filled surface.

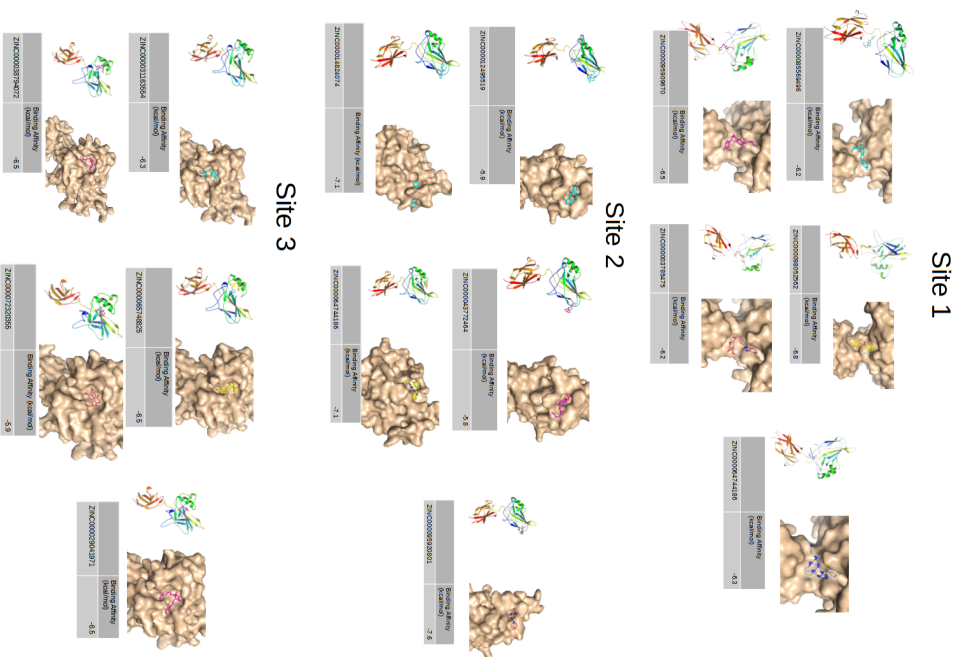


**Supplementary Figure 6.** Poisson-Boltzmann (APBS) electrostatic surface distribution for the chicken c-Rel (PDB ID: 1GJI) and human modelled c-Rel protein. Scale for each distribution provided individually, arrows indicate rotation direction; blue colour represents more electronegative, while red more electropositive regions.

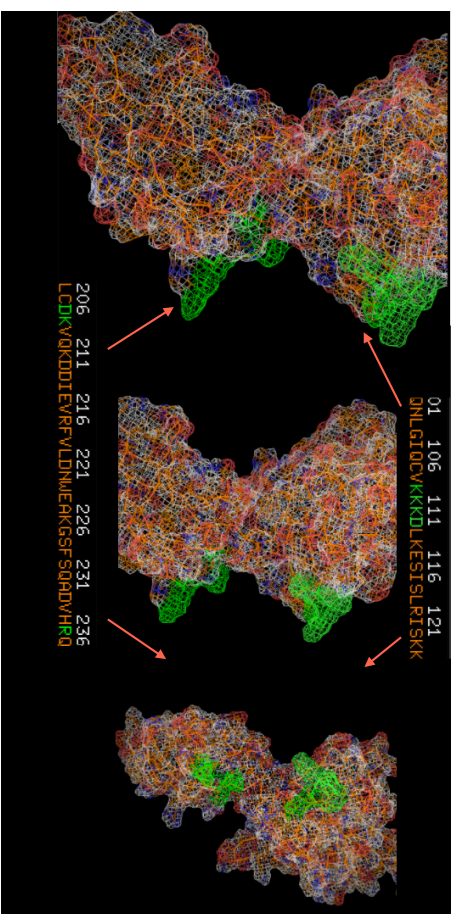
## 5uo1



**Supplementary Figure 7. Poisson-Boltzmann (APBS) electrostatic surface distribution for mouse p65 (PDB ID: 5U01).** Scale for each distribution provided individually, with blue colour representing more electropositive, while red more electronegative regions.



**Supplementary Figure 8.** AutoDock Vina docked and scored compounds for each binding site matching earlier screening sites and compounds per each site.



**Supplementary Figure 9.** GROMACS 1 ns length simulation snapshots revealing the local movement of residues around the binding site (highlighted). Sequence elements are shown for the specific regions.

**Supplementary Table 1.** REL family structures and sequence information

Protein name	PDB ID	Resolution, Å	Resolution, R	Species	Sequence Ref, UniProt
p65/RelA	2I9T chain A	2.80	0.288	Mus musculus	Q04207-1 (canonical)
p50	2I9T chain B	2.80	0.288	Mus musculus	P25799-1 (canonical)
p52	1A3Q chain A and B	2.10	0.320	Homo sapiens	Q00653-1 (canonical)
p50	1SVC chain P	2.60	0.286	Homo sapiens	P19838-1 (canonical)
c-Rel	1GJI chain B	2.85	0.279	Gallus gallus	P16236-1 (canonical)
p65/RelA	5U01, chain B	2.50	0.274	Mus musculus	Q04207-1 (canonical)

**Supplementary Table 2.** SiteMap analysis for the mouse p65 (PDB:5U01) protein dividing the protein into 5 regions.

Name	SiteScore	size, Å <sup>2</sup>	Dscore	volume, Å <sup>3</sup>	exposure	enclosure	phobic	philic
Site 1	0.886	674	1	360.493	0.783	0.373	0.119	0.504
Site 2	0.877	570	0.999	277.487	0.838	0.347	0.119	0.458
Site 3	0.895	304	1.023	178.017	0.832	0.359	0.164	0.406
Site 4	0.879	293	0.991	184.534	0.774	0.367	0.062	0.519
Site 5	0.882	289	0.999	161.896	0.797	0.362	0.107	0.484

**Supplementary Table 3.** Chicken c-Rel (PDB ID:1GJI) site screening results showing the number of compounds entering each round of the screening.

<b>HTVS screening mode <math>\Delta G &lt; -2</math> kJ/mol</b>	<b>SP screening mode <math>\Delta G &lt; -2</math> kJ/mol</b>	<b>XP screening mode <math>\Delta G &lt; -3</math> kJ/mol</b>
<b>Site 1</b>		
<b>34 M</b>	<b>338</b>	<b>11</b>
<b>Site 2</b>		
<b>34 M</b>	<b>163</b>	<b>33</b>
<b>Site 3</b>		
<b>34 M</b>	<b>1007</b>	<b>206</b>



## Integrative *omics* approaches for new target identification and therapeutics development

### 10.5. *Chemexpy* documentation

**The supplementary chapter is based on the published software package**

Kanapeckaitė A. *Chemexpy*: Cheminformatics package for compound feature evaluation. PyPi. 2021 Oct. 07. Version 1.0.10; <https://pypi.org/project/chemexpy/>

#### **Conclusion of this chapter**

My developed *Chemexpy* package provides a user-friendly and organised approach to explore chemical libraries and identify key features. The information generated by the package functions can be easily integrated into other pipelines or downstream processing. The package provides exploratory plots as well as compound similarity assessment allowing to search for similar compounds. Moreover, there are several additional functions helping to easily extract chemical descriptors and evaluate chemical libraries.

#### **Contribution to this chapter (100%)**

- Developed new programmatic features to accompany the related publication and make cheminformatics analyses more streamlined.
- Performed software package development and testing.
- Conceptualised and wrote the documentation files and vignettes, including the figure preparation.
- Corresponding author and maintainer.

## Documentation for the *Chemexpy* package

Package version: v1.0.10

Date: 10/06/2021

Author: Auste Kanapeckaite

The package contains the following functions:

1. data\_prep
2. molecule\_check
3. scatter\_plot
4. correlation\_plot
5. feature\_plot
6. normality\_check
7. feature\_check
8. feature\_violinplots
9. similarity\_search
10. similarity\_dendogram
11. similarity\_heatmap

**Dependencies:**

rdkit, pandas, numpy, scipy, seaborn, matplotlib, collections

### Introduction.

*Chemexpy* package provides a user-friendly and organised approach to explore chemical libraries and identify key features. The information generated by the package functions can be easily integrated into other pipelines or downstream processing. The package provides exploratory plots as well as compound similarity assessment allowing to search for similar compounds. Moreover, there are several additional functions helping to easily extract chemical descriptors and evaluate chemical libraries.

### Function Description.

#### 1. Function data\_prep

Function call example: data\_prep(data,\*args)

#Function provides a snapshot of the input data as well as returns a processed data file to include information on chemical descriptors, atomic composition, chemical structure features.

#Input values: path to a csv file that contains compound ID 'CID' and smiles 'SMILES' columns. These columns have to be named as described above. Additional columns can be passed as arguments if, for example, the data file contains other columns of interest.

#Output values: data frame with added chemical descriptors. The output could be integrated into downstream analyses and databases or used to visualise the structures.

## 2. Function `molecule_check`

Function call example: `molecule_check(data,*args)`

#Function allows to visualise molecules of interest as well as returns a data frame that contains information on the selected list of molecules. It is recommended not to select more than 20 molecules at a time to draw the structures.

#Input values: data frame with "CID" (compound ID) and "SMILES" (smiles column). Please note, the IDs for columns need to match the examples.

#Additional input: arguments for "CID", e.g., "2821293". If none is selected first 10 structures will be drawn. Names for compounds have to be in a string format.

#Output values: structure visualisation and a data frame that can be used for further visualisations.

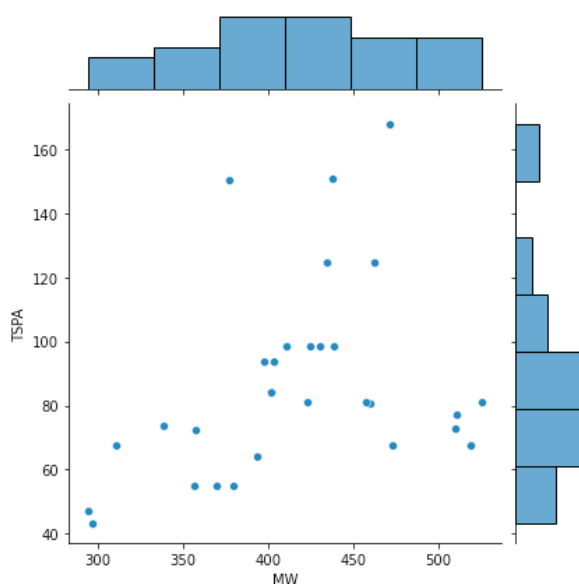
## 3. Function `scatter_plot`

Function call example: `scatter_plot(data,var1=None,var2=None)`

#Function takes the data file provided by the `data_prep` function and plots analytical scatter plots for selected variables.

#Input values: data frame generated by the `data_prep` function, as well as variables to plot, e.g. "MW" and "TSPA".

#Output values: scatter plot.



#### 4. Function correlation\_plot

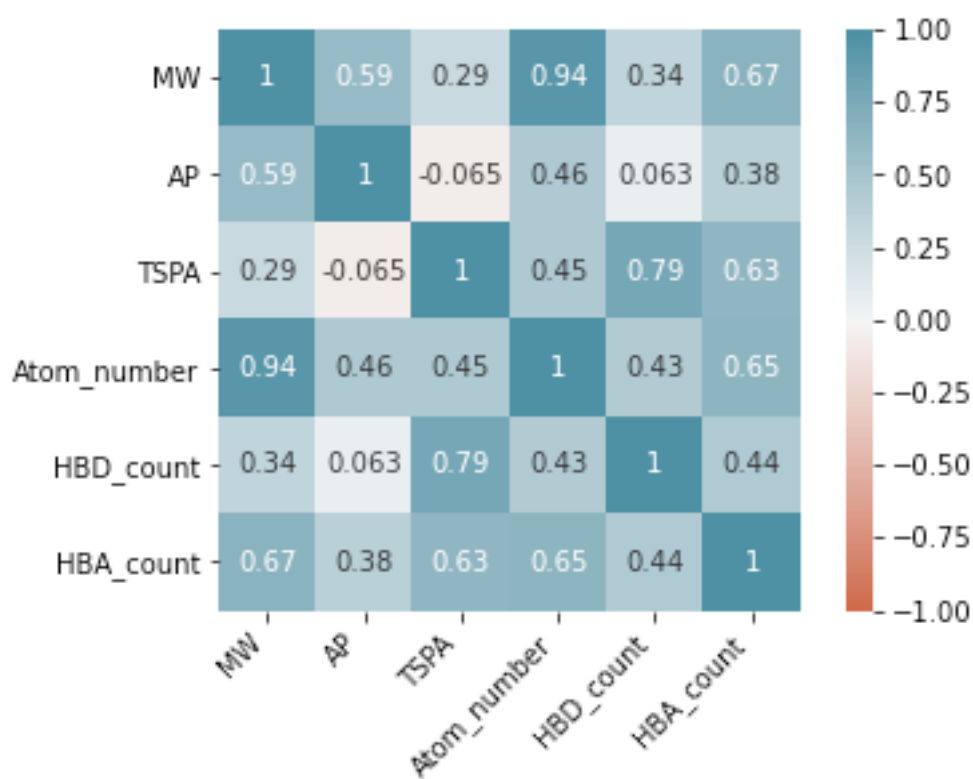
Function call example: `correlation_plot(data,*args)`

#Function takes the data file provided by the `data_prep` function and plots a correlation heatmap.

#Input values: data frame generated by the `data_prep` function, as well as variables to calculate correlation and plot selected values, e.g., "MW" and "TSPA".

If the user does not select args, the default values will be used:  
"Atom\_number", "MW", "TSPA", "HBD\_count", "HBA\_count", "Rotatable\_bond\_count", "MolLogP", "Ring\_number", "AP".

#Output values: plot for correlation visualisation and a data frame with correlation values.



#### 5. Function feature\_plot

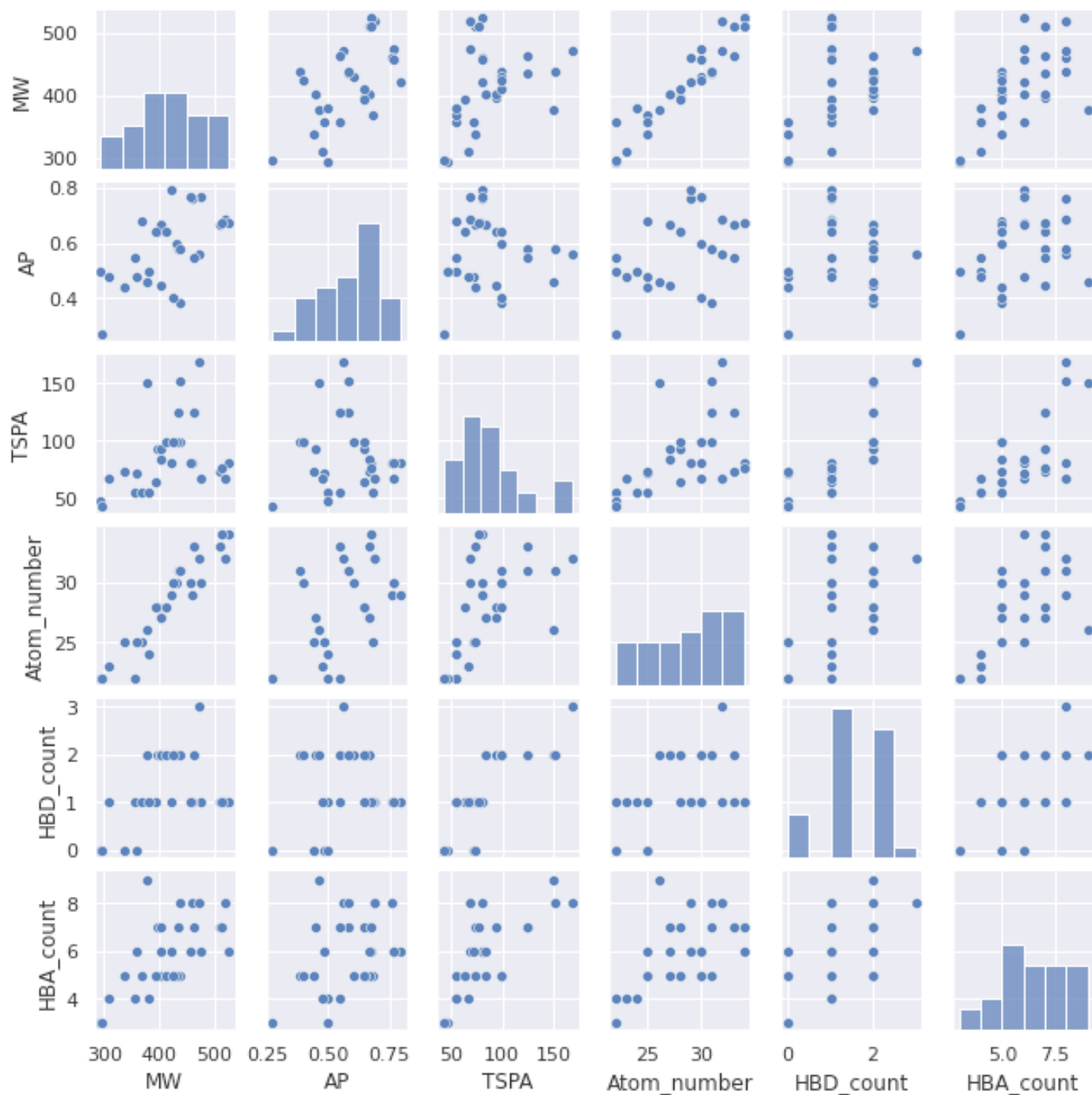
Function call example: `feature_plot(data,*args)`

#Function takes the data file provided by the `data_prep` function and plots analytical scatter plots for multiple features.

#Input: data frame generated by the `data_prep` function, as well as variables for feature plotting, e.g. "MW" and "TSPA".

#If the user does not select args, the default values will be used:  
"Atom\_number", "MW", "TSPA", "HBD\_count", "HBA\_count",  
"Rotatable\_bond\_count", "MolLogP", "Ring\_number", "AP".

#Output: scatter plot of multiple feature visualisation.



## 6. Function normality\_check

Function call example: `normality_check(data,var=None)`

#Function takes the data file provided by the `data_prep` function and plots a histogram with an estimated probability density function.

#Input: data frame generated by the `data_prep` function, as well as a single variable to check the normality for, e.g., "MW" and "TSPA".

#Output: bar plot with an estimated normal distribution line plot based on distribution probability.

## 7. Function feature\_check

Function call example: `feature_check(data,var1=None, var2=None, type=None)`

#Function takes the data file provided by the `data_prep` function and plots analytical contour plots to assess chemical feature distribution when considering a specific chemical entity category. That is, a categorical type data needs to be provided, such as active or inactive, etc.

#Input: data frame generated by the `data_prep` function, two variables for distribution check, e.g., "MW" and "TSPA", and a column name to select categorical data from.

#Output: contour plot with feature distribution.

## 8. Function feature\_violinplots

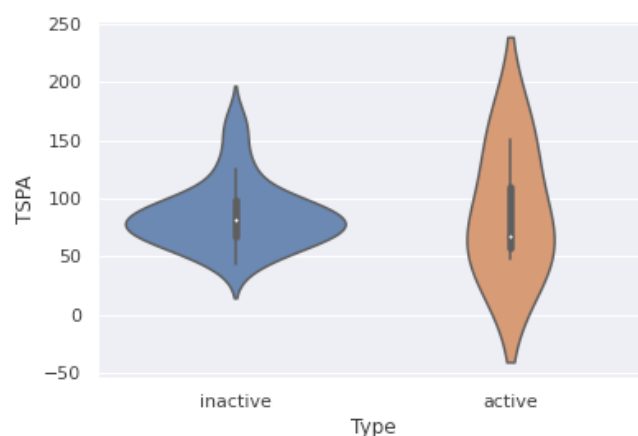
Function call example: `feature_violinplots(data,var1=None,type=None)`

#Function takes the data file provided by `data_prep` function and plots analytical violin plots to assess the type distribution for selected fetatures.

#Note categorical type data needs to be provided, such as active or inactive, etc.

#Input: data frame generated by the `data_prep` function as well as a variable name for the distribution check, e.g., "MW" and "TSPA"; also a column name is required to select categorical data specified through the "type" designation.

#Output: contour plot with feature distribution.



## 9. Function similarity\_search

Function call example: `similarity_search(data, target=None)`

#Function takes the data file provided by the `data_prep` function and searches for similar structures based on the target molecule.

#Fingerprinting is based on Morgan fingerprints and the similarity search is based on Tanimoto similarity.

#Input: data frame generated by the `data_prep` function as well as a SMILE string (e.g., the "target" variable) for a molecule to search in the database.

#Output: data frame of similar structures.

## 10. Function similarity\_dendogram

Function call example: `similarity_dendogram(data)`

#Function takes the data file provided by the `data_prep` function and plots a dendogram based on compound similarity.

#Fingerprinting is based on Morgan fingerprints and the similarity search is based on Tanimoto similarity.

#Input: data frame generated by the `data_prep` function.

#Output: dendogram and a data frame with similarity values.

## 11. Function similarity\_heatmap

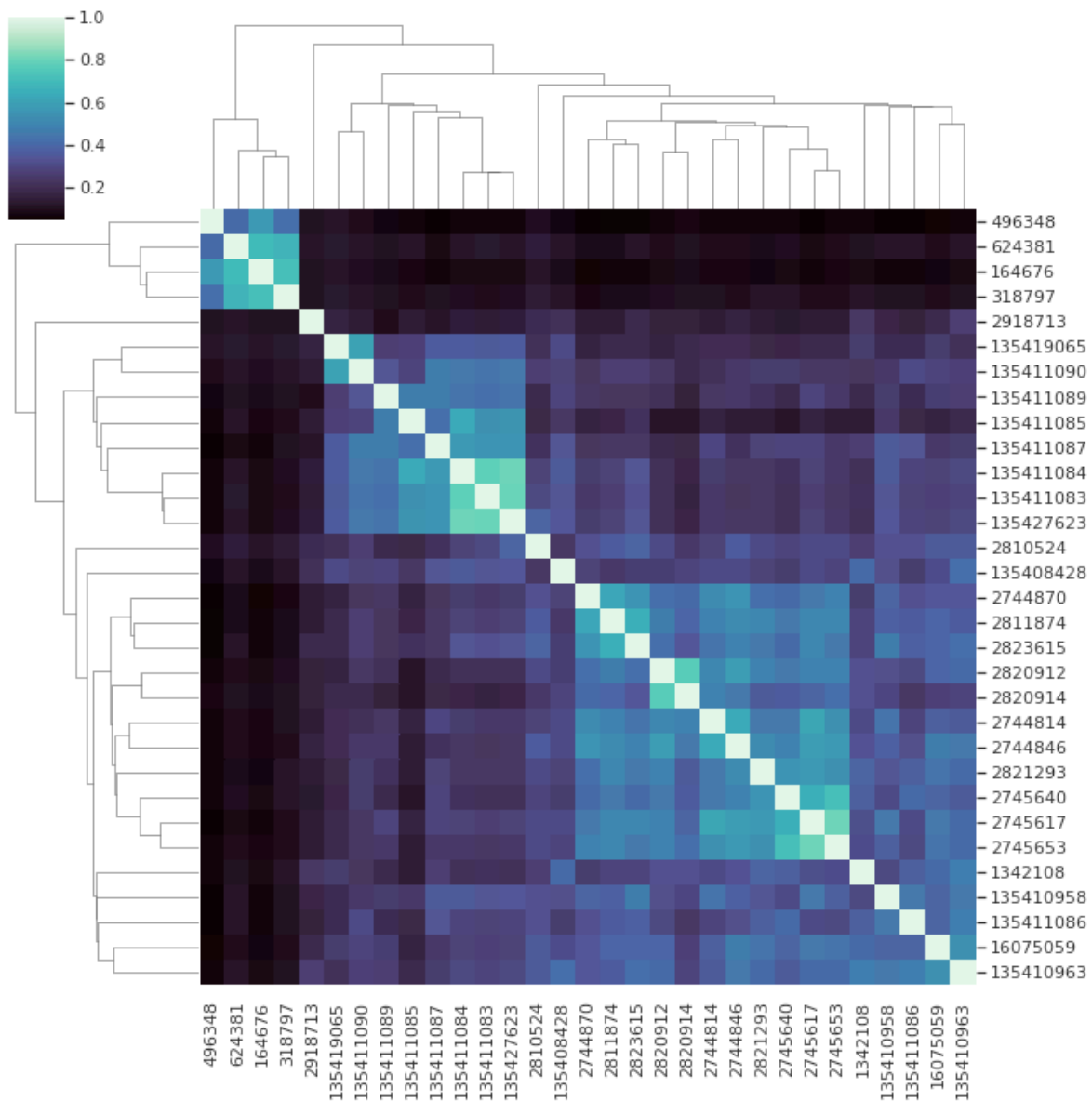
Function call example: `similarity_heatmap(data)`

#Function takes the data file provided by the `data_prep` function and plots a heatmap based on compound similarity.

#Fingerprinting is based on Morgan fingerprints and the similarity search is based on Tanimoto similarity.

#Input: data frame generated by the `data_prep` function.

#Output: heatmap and a data frame with similarity values.



## Running tests and example use cases.

The working directory should contain example data sets (provided with packages `PATH="./tests"`). There are several datasets to choose from, namely `data_1.csv` and `data_2.csv`.

```
#initiative variables to data
data="./test/data_1.csv"
```

```
#prepare the data for subsequent use
#we are selecting an additional column to assess the activity based on a categorical value
data=data_prep(data, "Type")
```



```
#assess a selected set of molecules and retrieve a data frame that contains information about these molecules
```

```
data_eval=molecule_check(data,"2821293")
```

```
#evaluate exploratory plots
```

```
scatter_plot(data,"MW","TSPA")
```

```
corr=correlation_plot(data,"MW","TSPA","AP","HBD_count","HBA_count")
```

```
#perform multiple feature assessment
```

```
feature_plot(data,"MW","TSPA","AP","HBD_count","HBA_count")
```

```
#check the normality of the data distribution
```

```
normality_check(data,"MW")
```

```
#since we have one categorical value we can perform a feature check
```

```
feature_check(data,var1="MW", var2="AP", type="Type")
```

```
#similarity assessment
```

```
target='COC(=O)c1c[nH]c2cc(OC(C)C)c(OC(C)C)cc2c1=O'
```

```
target_matches=similarity_search(data, target)
```

```
#similarity value generation for all pairwise comparisons
```

```
#dendogram plotting and a data frame preparation with similarity values
```

```
#both functions produce the same data frame
```

```
similarity_data=similarity_dendogram(data)
```

```
similarity_data=similarity_heatmap(data)
```

**Integrative *omics* approaches for new target identification and  
therapeutics development**



**School of Chemistry, Food and Pharmacy**

**Department of Pharmacology**