UNIVERSITÀ DEGLI STUDI DI MILANO

UNIVERSITAT DE BARCELONA

**Double Degree Doctorate between**

**UNIVERSITÀ DEGLI STUDI DI MILANO**

FACOLTÀ DI SCIENZE DEL FARMACO

Department of Pharmaceutical Sciences

PhD Course in Pharmaceutical Sciences (XXXI Cycle)

and

**UNIVERSITAT DE BARCELONA**

INSTITUT D'INVESTIGACIONS BIOMÈDIQUES AUGUST PI I SUNYER

Hospital Clinic Barcelona

PhD in Biotechnology

# Exploring the Space of Computer-Aided Drug Design

Modulator Designs Targeting PFKFB3 and
Computational Workflow Developments


PhD thesis of

**Xiao Hu**

Matricola: R11495


**Supervisors:**

Prof. Alessandro Contini

Dr. Joan Albert Barberà

**Doctorate Coordinator:**

Prof. Aldini Giancarlo

**Academic years 2016/2019**



This PhD programme was founded in the frame of

# HORIZON 2020 - Marie Skłodowska-Curie

# ITN-European Joint Doctorate

# MOGLYNET-programme

"MOGLYNet: "Modulation of glycolytic flux as a new approach for treatment of atherosclerosis and plaque stabilization: a multidisciplinary study"

*The project leading to this application has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No.67552*

# Abstract

This thesis involves works relating to the application and development steps within the computer-aided drug design (CADD) process. Tackling atherosclerosis via the path of metabolism modulation by targeting PFKFB3 is a novel idea regarding atherosclerosis treatment. The disease has been connected to abnormal up-regulation in the cellular metabolism process, predominantly through an elevation in glycolysis. Part of the work elucidated here applied a multi-strategy approach targeting an important bifunctional enzyme, PFKFB3, which regulates the glycolysis process. With the object of modulating PFKFB3 activity, multiple strategies have been attempted, resulting in several potential hit compounds for further testing and developments. Moreover, a virtual screening workflow was developed and tested. This workflow incorporated an additional rescoring procedure to re-evaluate the screening outcome. A series of test sets were generated and used for benchmark study. The rescoring procedure was further dissected, and statistical analyses were performed. Furthermore, it was suggested that water molecules play important roles in bridging ligand-receptor interactions. A C++ program was developed to select important binding site water positions in both *holo* and *apo* protein structures. The program can be easily incorporated into a virtual screening process and initial testing has shown promis-

ing performance. Additionally, the applicability of network analysis in biomarker detection was also preliminarily investigated. The early stage of a network analysis workflow development was carried out and was demonstrated with promising results. This thesis can provide further insights into, and also showed bright future prospects of, CADD studies.

# Declaration

I hereby declare that this Ph.D. thesis entitled *"Exploring the Space of Computer-Aided Drug Design - Modulator Designs Targeting PFKFB3 and Computational Workflow Developments"* was carried out by me for the degree of ITN-European Joint Doctorate in Pharmaceutical Science and Biotechnology under the guidance and supervision of Prof. Alessandro Contini, Department of Pharmaceutical Sciences, Università degli Studi di Milano, Italy, and Dr. Joan Albert Barberà, iDiBAPS, Hospital Clinic Barcelona, Universitat de Barcelona, Spain.

The works and discussions put forth are based on my research activities of the last three years carried out both in Università degli Studi di Milano, Italy, and iDiBAPS, Hospital Clinic Barcelona, Spain. The books, articles and websites, which I have made use of are acknowledged at the respective place in the text. For the present thesis, which I am submitting to the University, no degree or diploma or distinction has been conferred on me before, either in this or in any other University.

# Acknowledgements

During the period of the three-year research, I have received support from multiple individuals of significant who have contributed one way or another to this study. My deepest gratitude to the principal investigator of this work, Prof. Alessandro Contini, for his unconditional patience, kindness, and care. Without his insightful suggestions, his heart-warming encouragements, and all the fruitful discussions between us, the work would not have achieved its present form. I have seen in him an unpretentious and devoted scholar. It had been a memorable experience working with him and would influence me in the long future.

I also wish to express my sincere gratitude to one of my mentors, Dr Olga Tura, iDiBAPS, Hospital Clinic Barcelona, and her group for the supports in the training and research in a field that I had not set foot in before. The assistance extended further after my departure and have continued up to this date.

I'd like to also thank my colleague, Helena Macut, as a collaborator, and more importantly, a close friend, during my PhD studies. I worked closely with her and the scientific discussions had become a daily routine between us throughout. The experimental supports from her was evaluable. Without them, much of my work would

render meaningless. The friendship from her has also saved me from multiple crises along the way.

I also thank Dr Sara Pellegrino, Prof. Maria Luisa Gelmi, Dr Sabrina Pavan for their important contributions to the Moglynet programme, and numerable situations where I required urgent (or sometimes not so urgent) assistance. Their kindness has smoothed much of my research work during the period. I thank Simone Assanelli for his assistance in part of the work that it would be only possible for me to manage alone if I miraculously had grown another brain.

I appreciated my dear friends, Giada, Antoniata, Francesco, and Massimiliano for saved me with the occasional chilling hours and long chats we had together during some most stressed periods of the last three years. I want to thank my best friend, David, for accompanying and encouraging me throughout the struggling writing hours leading to the formation of this thesis. Lastly but not least, my gratitude to my parents for their life long support of me pursuing my future career.

# Contents

# Part I:

# Introduction

# CHAPTER 1

# Computer-aided drug design

## 1.1 In drug development process

The process of drug discovery and development is costly; it can take an average of 10-15 years and hundreds of millions to several billion US dollars for a new drug to reach the markets.[1–3] The developments in combinatorial chemistry and high-throughput screening (HTS) have allowed generating and screening of large compound databases in experimental settings.[4,5] HTS, benefited from its high success rate, is one of the most applied steps within the process of drug discovery in the pharmaceutical industries. However, it holds true under most of the circumstances that the detail information regarding the binding mechanism and dynamics of the hit compounds on the molecular scale are still lacking. This shortage in knowledge can easily affect the later lead or drug development.[6,7] Moreover, both combinatorial chemistry and HTS require a large sum of financial input and resources and might not always result in a satisfactory outcome. Hence, in comparison, techniques applied within

computer-aided drug design (CADD) not only reduces the time and cost of the drug development process but also provides rational information regarding molecular or biological mechanisms from the early steps.

CADD can be involved in multiple stages that consist in the drug discovery and development process. Cheminformatics and bioinformatics tools have shown good potentials in target identifications and validations. Some promising outcomes have been demonstrated with the applications of chemical structure similarity search,[8] data mining,[9] bioactivity spectra based algorithms,[10] interaction fingerprints,[11,12] and network-based methods.[13–16] Moreover, CADD methods play important roles in the hit or lead identifications. Two categories of computational approaches, ligand-based and structure-based methods, are widely applied for early drug discovery.

Ligand-based drug design (LBDD) is mostly applied when the structural information of the identified target is lacking. Ligand information such as binding affinities, chemical structure, and physicochemical properties are frequently exploited for configuring ligand-target interactions. Similarity search using 2D or 3D fingerprint methods is one class of viable solutions for discovering new compounds and have been well reviewed elsewhere.[17] Little information required for a query renders similarity search a very useful class of methods at the very beginning of a drug discovery project, and it is in general time/cost-efficient for huge chemical databases. The methods are also in general diverse in implementations. Similar in concepts, pharmacophore modelling is also frequently used for identifying molecules with similar 3D arrangements of important functional groups to known active compounds.[18] Furthermore, another appealing method, namely the quantitative structure-activity relationship (QSAR), has also gained its reputation in predicting the biological property of novel compounds. QSAR is based on the presumption that the bioactivity of a compound against a target is associated with its structural and molecular properties.[19] This correlation among a set of compounds can be modelled statistically and mathematics activity predictions of novel compounds can also be derived.[19]

On the other hand, structure-based drug design (SBDD) benefited from the upsurging

available structures of macromolecules with the advances in X-ray crystallography and nuclear magnetic resonance (NMR) techniques. The number of available crystal structures in the Protein Data Bank has almost tripled for the last decade (54479 crystal structures in 2008 and 147609 in 2018).[20] Moreover, homology modelling and molecular dynamic (MD) simulations are also valuable tools for generating the desired macromolecule structures. Two strategies can be taken in SBDD, the *de novo* design and the virtual screening (VS) approaches, both require the 3D information from the targeted receptors.[7,21] In a *de novo* drug design process, small fragments are fitted into the binding site of the receptor and linked according to rules that allow rational synthesis. This fragment-based approach has good potentials in producing novel and potent lead compounds.[22] Alternatively, libraries of small molecules can also be virtually screened for identifying hits targeting the receptor of interests. The prediction of binding poses and ranking of the fitness are generally accomplished by using docking methods. Due to the application of small molecule libraries, VS can also be combined with LBDD methods, such as similarity search and pharmacophore modelling, to improve the efficiency of the overall CADD process.

## 1.2 Structure-based virtual screening

As a computational alternative to HTS, the concept of VS was established in the 1990s. Using computer programs, prediction of binding to a target macromolecule (i.e. a receptor) can be performed for a list of compounds at a very low cost. Moreover, VS also have an advantage over HTS for that the screening compounds do not necessarily exist. Application of VS also is not affected by factors that in general exert circumstantial limitations on experimental techniques such as solubility and aggregation. All these suggest VS a useful tool for discovering novel active compounds.

One key prerequisite of VS is the knowledge regarding the spatial and energetic profiles crucial for binding. This, in general, can be achieved by using the 3D structures

of the receptor of interest, either from X-ray crystallography, NMR, or homology modelling. An important consideration at the beginning of a VS run is the choice of the receptor 3D geometry that meet the circumstance of the screening. Specific cares should be taken depending on the source of receptor structures. Resolution and the diffraction data are crucial in determining the quality of a crystal structure, while the strong dependence of NMR data on the local distribution of the Nuclear Overhauser Effects along the protein chain suggest that careful examinations are also required. Furthermore, homology models generated from crystal structures of related proteins have also been successfully applied in multiple VS studies.[23–25]

Another crucial factor to be considered for VS is the quality of the screening library. There are multiple public and commercial chemical databases available, such as ChEMBL, ZINC, and Asinex to name a few. The sizes of the libraries ranging from tens of thousands to over a billion. Though it is possible to screen all the compounds within a database with the gradual advancing in computer power, the active compounds are more likely to share some common properties that span only part of the chemical space. Excessive screening can lead to unnecessary wasting of computational resources. A too large library is also more likely to introduce noises that complicate later compound selection. Hence, it is always recommendable that some pre-screening filtering to be applied. Common strategies include filters according to Lipinski's rules[26,27] or similarity search based on known active ligands.[28] More specifically, pharmacophore filtering can also be applied if desired ligand-receptor interactions are available. Multiple studies have demonstrated successful applications of pharmacophoric constraints shrinking large compound libraries from several folds to hundreds of folds smaller.[29–31]

Docking methods lie at the very core of virtual screening. The aim of a docking program in VS is to provide accurate predictions in both the structural model and biological activity. However, modelling of the molecular interactions is an intricate and difficult task to allow sensitivity in biological recognition. To overcome some of the difficulties, the simulated binding (i.e. docking) generally involves a multi-step

process. Docking starts with the posing step, which uses algorithms to position small molecules within the binding site. Conformational sampling is crucial to ensure accurate pose predictions in this step and an ensemble of poses are first generated. The fitness of poses is evaluated with scoring functions both during the posing simulations and a later ranking stage.

Scoring functions are designed to estimate binding affinity between the ligands and receptors with various assumptions and simplifications in physical phenomena and interaction terms. During the early stages of docking runs, relatively simple scoring schemes are applied to roughly screen the conformers at higher speeds. The selected poses are then evaluated with more sophisticated scoring functions and ranked with more confidence in finding meaningful conformers. The comparison and summary of search algorithms and scoring functions in docking programs are out of the scope of this thesis and are better reviewed elsewhere.[32–34] Current scoring functions vary in performance due to the differences their design and the accuracy is still generally far from satisfactory.[35] Hence, the development of a highly accurate scoring function of good efficiency still remains a major challenge. Efforts have been taken for improved scoring performance by including entropic effects,[36] incorporating quantum mechanic calculations,[37] consensus scoring,[38,39] and applying machine learning methods.[40]

Other environmental factors are also frequently discussed for VS applications. Most docking process only considers one static conformation of the receptor while in the physiological environment molecular targets are always at a dynamical equilibrium. It has been demonstrated that consideration of receptor flexibility can potentially improve VS outcomes.[41] Docking using an ensemble of receptor conformers can partially resolve the issue related to target flexibility. MD simulations[42] or crystal structures[43,44] can both be the source of the receptor ensembles. However, this comes at a higher computational cost since VS about each conformer should be performed separately.

Explicit water molecules at the binding site is another hot topic for developing better

VS programs. Water is well-acknowledged to bridging ligand-receptor interactions through hydrogen bonds at the binding interface.[45,46] Analysis of thousands of crystal structures have shown that in general one or more water molecules that are facilitating ligand binding.[47,48] Another study on 392 high-resolution complexes also suggested that more than 50% of the active site water molecules are mediating ligand-target interactions.[49] This recognition is reflected by the development and adaptation of various software to consider explicit water during docking or VS process.[50–52]

The thesis is mainly framed as a CADD expedition. Two major parts of the work have involved steps within a CADD process. The first part involves the drug designing process targeting a glycolytic enzyme, PFKFB3, for atherosclerosis treatment. The SBDD methods were applied, taking advantages of mainly the VS and the MD simulations. (Chapters 6 to 8) Different designing strategies were devised and the molecular mechanisms of the hit compounds were investigated with the application of CADD techniques.

On the other hand, the second part concerns the development and assessment of computational methods and workflows that lie at the basis of the CADD techniques. A workflow incorporating MM-PBSA rescoring to VS outcomes was also developed and assessed. (Chapter 10 and 11) A water selection program was developed to identify important water molecules prior to docking experiments and have shown some potentials due to its high computational efficiency. (Chapter 12) The testing of the workflow was carried out in-depth with several different test sets and the outcome was analysed in detail. Moreover, the application of bioinformatic tools in target identification was also explored.(Chapter 13)

## Chapter references

[1] J.A. DiMasi, R.W. Hansen, and H.G. Grabowski. The price of innovation: new estimates of drug development costs. *Journal of health economics*, 22(2):151–185, 2003.

[2] C.M. Song, S.J. Lim, and J.C. Tong. Recent advances in computer-aided drug design. *Briefings in bioinformatics*, 10(5):579–591, 2009.

[3] S.M. Paul, D.S. Mytelka, C.T. Dunwiddie, C.C. Persinger, B.H. Munos, S.R. Lindborg, and

A.L. Schacht. How to improve R&D productivity: the pharmaceutical industry's grand challenge., 2010.

[4] A. Lavecchia and C. Di Giovanni. Virtual screening strategies in drug discovery: a critical review. *Current medicinal chemistry*, 20(23):2839–2860, 2013.

[5] H. Jhoti, S. Rees, and R. Solari. High-throughput screening and structure-based approaches to hit discovery: is there a clear winner?, 2013.

[6] R. Macarron, M.N. Banks, D. Bojanic, D.J. Burns, D.A. Cirovic, T. Garyantes, D.V.S. Green, R.P. Hertzberg, W.P. Janzen, J.W. Paslay, U. Schopfer, and G.S. Sittampalam. Impact of high-throughput screening in biomedical research. *Nature Reviews Drug Discovery*, 10:188, 2011.

[7] E. Lionta, G. Spyrou, D.K. Vassilatis, and Z. Cournia. Structure-based virtual screening for drug discovery: principles, applications and recent advances. *Current topics in medicinal chemistry*, 14(16):1923–1938, 2014.

[8] M.J. Keiser, V. Setola, J.J. Irwin, C. Laggner, A.I. Abbas, S.J. Hufeisen, N.H. Jensen, M.B. Kuijer, R.C. Matos, T.B. Tran, R. Whaley, R.A. Glennon, J. Hert, K.L.H. Thomas, D.D. Edwards, B.K. Shoichet, and B.L. Roth. Predicting new molecular targets for known drugs. *Nature*, 462:175, 2009.

[9] Nidhi, M. Glick, J.W. Davies, and J.L. Jenkins. Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. *Journal of chemical information and modeling*, 46(3):1124–1133, 2006.

[10] T. Cheng, Q. Li, Y. Wang, and S.H. Bryant. Identifying compound-target associations by combining bioactivity profile similarity search and public databases mining. *Journal of chemical information and modeling*, 51(9):2440–2448, 2011.

[11] I. Eberini, S. Daniele, C. Parravicini, C. Sensi, M.L. Trincavelli, C. Martini, and M.P. Abbracchio. In silico identification of new ligands for GPR17: a promising therapeutic target for neurodegenerative diseases. *Journal of computer-aided molecular design*, 25(8):743–752, 2011.

[12] R. Cao and Y. Wang. Predicting Molecular Targets for Small-Molecule Drugs with a Ligand-Based Interaction Fingerprint Approach. *ChemMedChem*, 11(12):1352–1361, 2016.

[13] S. Zhao and S. Li. Network-Based Relating Pharmacological and Genomic Spaces for Drug Target Identification. *PLOS ONE*, 5(7):e11764, 2010.

[14] A.-L. Barabasi, N. Gulbahce, and J. Loscalzo. Network medicine: a network-based approach to human disease. *Nature reviews. Genetics*, 12(1):56–68, 2011.

[15] X. Wang, X. Wei, B. Thijssen, J. Das, S.M. Lipkin, and H. Yu. Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nature biotechnology*, 30 (2):159–164, 2012.

[16] H.B. Engin, A. Gursoy, R. Nussinov, and O. Keskin. Network-based strategies can help mono- and poly-pharmacology drug discovery: a systems biology view. *Current pharmaceutical design*, 20(8):1201–1207, 2014.

[17] R.P. Sheridan and S.K. Kearsley. Why do we need so many chemical similarity search methods? *Drug Discovery Today*, 7(17):903–911, 2002.

[18] A. Vuorinen and D. Schuster. Methods for generating and applying pharmacophore models as virtual screening filters and for bioactivity profiling. *Methods*, 71:113–134, 2015.

[19] C.C. Melo-Filho, R.C. Braga, and C.H. Andrade. 3D-QSAR approaches in drug design: perspectives to generate reliable CoMFA models. *Current computer-aided drug design*, 10(2): 148–159, 2014.

[20] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

[21] S. Kalyaanamoorthy and Y.-P.P. Chen. Structure-based drug design to augment hit discovery. *Drug Discovery Today*, 16(17):831–839, 2011.

[22] A. Kumar, A. Voet, and K.Y.J. Zhang. Fragment based drug design: from experimental to computational approaches. *Current medicinal chemistry*, 19(30):5128–5147, 2012.

[23] Caterina Bissantz, Philippe Bernard, Marcel Hibert, and Didier Rognan. Protein-based virtual screening of chemical databases. II. Are homology models of G-Protein Coupled Receptors suitable targets? *Proteins*, 50(1):5–25, 2003.

[24] A. Evers and G. Klebe. Successful virtual screening for a submicromolar antagonist of the neurokinin-1 receptor based on a ligand-supported homology model. *Journal of medicinal chemistry*, 47(22):5381–5392, 2004.

[25] A. Evers and T. Klabunde. Structure-based drug discovery using GPCR homology modeling: successful virtual screening for antagonists of the alpha1A adrenergic receptor. *Journal of medicinal chemistry*, 48(4):1088–1097, 2005.

[26] C.A. Lipinski, F. Lombardo, B.W. Dominy, and P.J. Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, 46(1):3–26, 2001.

[27] C.A. Lipinski. Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discovery Today: Technologies*, 1(4):337–341, 2004.

[28] R. Perez-Pineiro, A. Burgos, D.C. Jones, L.C. Andrew, H. Rodriguez, M. Suarez, A.H. Fairlamb, and D.S. Wishart. Development of a Novel Virtual Screening Cascade Protocol to Identify Potential Trypanothione Reductase Inhibitors. *Journal of Medicinal Chemistry*, 52(6): 1670–1680, 2009.

[29] R. Heinke, A. Spannhoff, R. Meier, P. Trojer, I. Bauer, M. Jung, and W. Sippl. Virtual screening and biological characterization of novel histone arginine methyltransferase PRMT1 inhibitors. *ChemMedChem*, 4(1):69–77, 2009.

[30] D. Kireev, T.J. Wigle, J. Norris-Drouin, J.M. Herold, W.P. Janzen, and S.V. Frye. Identification of non-peptide malignant brain tumor (MBT) repeat antagonists by virtual screening of commercially available compounds. *Journal of medicinal chemistry*, 53(21):7625–7631, 2010.

[31] K. Lee, K.-W. Jeong, Y. Lee, J.Y. Song, M.S. Kim, G.S. Lee, and Y. Kim. Pharmacophore modeling and virtual screening studies for new VEGFR-2 kinase inhibitors. *European Journal of Medicinal Chemistry*, 45(11):5420–5427, 2010.

[32] Douglas B Kitchen, Helene Decornez, John R Furr, and Jurgen Bajorath. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature reviews. Drug discovery*, 3(11):935–949, 2004.

[33] N. Moitessier, P. Englebienne, D. Lee, J. Lawandi, and C.R. Corbeil. Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. *British journal of pharmacology*, 153 Suppl 1:S7–26, 2008.

[34] S.-Y. Huang, S.Z. Grinter, and X. Zou. Scoring functions and their evaluation methods for protein-ligand docking: recent advances and future directions. *Physical chemistry chemical physics : PCCP*, 12(40):12899–12908, 2010.

[35] T. Cheng, X. Li, Y. Li, Z.i Liu, and R. Wang. Comparative Assessment of Scoring Functions on a Diverse Test Set. *Journal of Chemical Information and Modeling*, 49(4):1079–1093, 2009.

[36] S.-Y. Huang and X. Zou. Inclusion of Solvation and Entropy in the Knowledge-Based Scoring Function for Protein−Ligand Interactions. *Journal of Chemical Information and Modeling*, 50 (2):262–273, 2010.

[37] K. Raha and K.M. Merz. Large-Scale Validation of a Quantum Mechanics Based Scoring Function: Predicting the Binding Affinity and the Binding Mode of a Diverse Set of Protein−Ligand Complexes. *Journal of Medicinal Chemistry*, 48(14):4558–4575, 2005.

[38] P.S. Charifson, J.J. Corkery, M.A. Murcko, and W.P. Walters. Consensus Scoring: A Method for Obtaining Improved Hit Rates from Docking Databases of Three-Dimensional Structures into Proteins. *Journal of Medicinal Chemistry*, 42(25):5100–5109, 1999.

[39] R. Wang and S. Wang. How Does Consensus Scoring Work for Virtual Library Screening? An Idealized Computer Experiment. *Journal of Chemical Information and Computer Sciences*, 41 (5):1422–1426, 2001.

[40] D.H. Fogel and Gary B. Computational Intelligence Methods for Docking Scores. *Current Computer-Aided Drug Design*, 5(1):56–68, 2009.

[41] M. Rueda, G. Bottegoni, and R. Abagyan. Consistent improvement of cross-docking results using binding site ensembles generated with elastic network normal modes. *Journal of chemical information and modeling*, 49(3):716–725, 2009.

[42] M. Okamoto, K. Takayama, T. Shimizu, K. Ishida, O. Takahashi, and T. Furuya. Identification of death-associated protein kinases inhibitors using structure-based virtual screening. *Journal of medicinal chemistry*, 52(22):7323–7327, 2009.

[43] M. Rueda, G. Bottegoni, and R. Abagyan. Recipes for the Selection of Experimental Protein Conformations for Virtual Screening. *Journal of Chemical Information and Modeling*, 50(1): 186–193, 2010.

[44] R.B. Murphy, M.P. Repasky, J.R. Greenwood, I. Tubert-Brohman, S. Jerome, R. Annabhimoju, N.A. Boyles, C.D. Schmitz, R. Abel, R. Farid, and R.A. Friesner. WScore: A Flexible and Accurate Treatment of Explicit Water Molecules in Ligand-Receptor Docking. *Journal of Medicinal Chemistry*, 59(9), 2016.

[45] J.E Ladbury. Just add water! The effect of water on the specificity of protein-ligand binding sites and its potential application to drug design. *Chemistry & Biology*, 3(12):973–980, 1996.

[46] S. de Beer, N. Vermeulen, and C. Oostenbrink. The Role of Water Molecules in Computational Drug Design. *Current Topics in Medicinal Chemistry*, 10(1):55–66, 2010.

[47] J. Gunther, A. Bergner, M. Hendlich, and G. Klebe. Utilising structural knowledge in drug design strategies: applications using Relibase. *Journal of molecular biology*, 326(2):621–636, 2003.

[48] Manfred Hendlich, Andreas Bergner, Judith Günther, and Gerhard Klebe. Relibase: Design and development of a database for comprehensive analysis of protein-ligand interactions. *Journal of Molecular Biology*, 326(2):607–620, 2003.

[49] Y. Lu, R. Wang, C.Y. Yang, and S. Wang. Analysis of ligand-bound water molecules in high-resolution crystal structures of protein-ligand complexes. *Journal of Chemical Information and Modeling*, 47(2):668–675, 2007.

[50] S.E. Wong and F.C. Lightstone. Accounting for water molecules in drug design. *Expert opinion on drug discovery*, 6(1):65–74, 2011.

[51] F. Spyrakis and C.N. Cavasotto. Open challenges in structure-based virtual screening: Receptor modeling, target flexibility consideration and active site water molecules description. *Archives of biochemistry and biophysics*, 583:105–119, 2015.

[52] X. Hu, I. Maffucci, and A. Contini. Advances in the Treatment of Explicit Water Molecules in Docking and Binding Free Energy Calculations, 2018.

# CHAPTER 2

# Atherosclerosis

## 2.1 In cardiovascular diseases

Cardiovascular diseases (CVDs) are the number 1 cause of death globally.[1] In 2016, it was estimated that 31 % of all global death (about 17.9 million people) were due to CVDs, 85 % of which are caused by heart attack and stroke.[1] The low- and middle-income countries are the most affected by CVDs (over three-quarters of the global CVD mortalities), and CVDs are exhibiting an increase in prevalence each year. A heavy economic burden from CVDs treatment impacts heavily on the low- and middle-income countries. Moreover, pressing demands of better CVD management is also called for in the developed countries.

CVDs can occur without symptoms and remain undiagnosed for a long period. The first warning, however, are in general severe - such as a heart attack or stroke - is can lead to irrevocable consequences. The main causations of heart attacks and strokes are blockages obstructing blood flow to the heart or brain, commonly started

from the accumulation of lipids on the inner walls of the blood vessels. The term plaque, or atheromatous plaque, refers to the abnormal material accumulation on the blood intima, most commonly arteries. The formation of plaque is closely connected to the development of atherosclerosis, a disease depicts the luminal narrowing of arteries.

Atherosclerosis can happen at arbitrary arteries in the body, and disrupt circulation to the heart, brain, or peripheral parts. It is generally developed as a chronic inflammatory process that leads to lipid-rich plaque formation (also containing cell debris and calcium) within the layer of the arterial wall. The plaque typically has a lipid core and a fibrous cap, which are elements defining the vulnerability status of the plaque. A thick fibrous cap and a small lipid core represent a low risk or stable plaque. A vulnerable plaque, on the other hand, is prone to rupture due to the thinner fibrous cap and the larger lipid-rich necrotic core. Both types of plaques can lead to artery stenosis (the narrowing of the artery) or occlusion through different mechanisms.

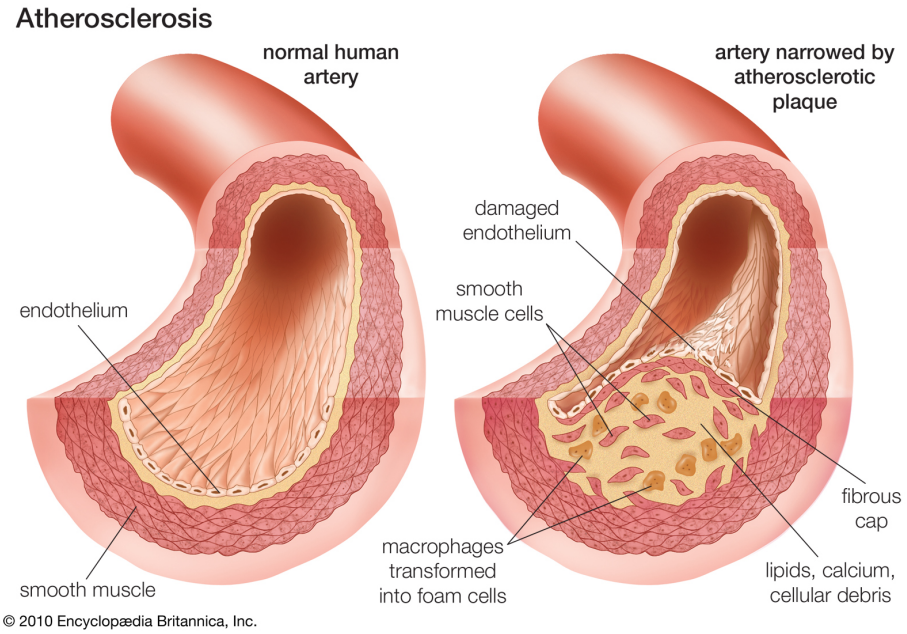



**Figure 2.1:** Atherosclerosis. From: Atherosclerosis. Britannica Online Encyclopaedia. By courtesy of Encyclopaedia Britannica, Inc., copyright © 2010 Encyclopaedia Britannica, Inc; used with permission.[2]

## 2.2 Pathogenesis

One key element in the development of atherosclerosis is the lipid contents in the bloodstream. One major cause of atherosclerosis is related to the increase in the concentration of apolipoprotein B (apoB) containing lipoproteins in the plasma and their accumulation the arterial intima.[3] However, other risk factors can also come into play, such as smoking, hypertension, and diabetes mellitus.

The atherosclerotic lesions, in general, are triggered by inflammatory responses initialised by the retention of atherogenic lipoproteins within the walls of blood vessels.[4,5] These lipoproteins attach to the proteoglycans in the extracellular matrix mostly through electrostatic interactions.[6] ApoB 100 has multiple sites responsible for proteoglycan binding with site B playing the major role.[7] Lipoproteins other than apoB, such as apoE and serum amyloid A, are also capable of binding to proteoglycans.[8–10] Lipoprotein retention in the intima can lead to the oxidisation and aggregation of low-density lipoprotein (LDL). Physical alterations of LDL (change in both configuration and lipid organisation) can occur after binding to proteoglycans, which evokes macrophage engulfment.[11–13] The structural change also further accelerates LDL oxidation and aggregation.[14,15] Moreover, it has been suggested that lipoprotein retention is a self-accelerating process; retained LDL triggers cellular responses that facilitate additional entrapment of LDL.[16–18] Overall, lipoprotein retention is the initiating step of developing atherosclerotic lesions.

Inflammation responses is another important process during atherosclerotic progression. The aggregated LDL particles are pro-inflammatory and can activate the endothelium on the vessel inner walls. The activated endothelial cells can result in the recruitment of blood-borne monocytes into sub-endothelial space, which is an important early inflammatory response.[19,20] Once entered the intima, the monocytes further promote the accumulation of modified LDL by secreting lipoprotein binding proteoglycans, sustaining the inflammation process.[4,21] Monocytes can also transform into macrophages or dendritic cells under the influence of monocyte colony stimulating

factor.[22] During the ingestion of lipoproteins by macrophages, the balance between uptake and efflux can be easily disrupted, leading to lipid accumulation forming the foam cells. In established lesions, the growth of plaque is promoted by foam cells, which is also accelerated by increased retention of lipoproteins.[10]

In atherosclerosis, adventitial and intimal angiogenesis can be considered as responses to the hypoxic environment due to the thickening of the intima and progressive growth of plaque.[23–25] The intimal hyperplasia progression and necrotic core development are also facilitated by angiogenesis, which is also likely one of the risk factors of plaque rapture.[26] Evidence, including the observation of intra-plaque angiogenesis, support the connection between hypoxia-inducible factor (HIF) and atherosclerosis.[27–29] Correlation between hypoxia, the presence of macrophages, and the expression of HIF and VEGF in advanced human atherosclerosis has also been demonstrated.[30] The new micro-vessels allow the migration of inflammatory cells to the deeper areas of advanced plaques, perpetuating the chronic inflammation process.[32] These inflammatory cells can also induce intra-plaque haemorrhage by means of neovessel rupture.[32]

The rupturing risks of a vulnerable plaque are controlled by micro-vessel structure and endothelial integrity. Thin-walled micro-vessels are prone to collapse and leakage and are frequently observed in the atherosclerotic plaque, indicating that the angiogenesis here is likely pathological.[33–35] Additionally, studies have also shown that endothelial integrity of micro-vessels within atherosclerosis is compromised, likely related to the constant expression of VEGF during atherogenesis.[30,36,37] Due to the above distinctions to the physiological angiogenesis, the plaque is also related to bio-mechanical destabilisation. High wall stresses and plaque strains are two key elements leading to the rupture of plaque; the former can be induced by lumen narrowing, resulting in increased local blood pressure and pulsate flow.[38,39] Plaque characteristics and arterial wall remodelling, on the other hand, determine the status of plaque strains.[40,41]

## 2.3 Treatments and prospectives

Current treatment for atherosclerosis include lifestyle changes, medicines to lower risks of heart attacks and strokes, and medical procedures and surgeries.[42] Different medicines can be prescribed to slow down the progress of atherosclerosis. Controlling plasma cholesterol level and platelet clumping are the viable resolutions. Other treatments include administration of beta blocker, angiotensin-converting enzyme (ACE) inhibitors, calcium channel blockers, and diuretics.[43] For severe symptoms, surgical procedures are also recommendable, such as angioplasty and stent placement or bypass surgery.

Recent studies related to pathogenesis and diagnosis of cardiovascular diseases has also provided further implications into atherosclerosis treatments. Promising outcomes have been demonstrated in targeting vascular dysfunction, inflammatory processes, angiogenesis, and metabolism.[44–47] In this thesis, the modulator design targeting PFKFB3, an important regulator in glycolytic metabolism, has been attempted in the context of modulating atherosclerosis metabolism. The connection of the enzyme to atherosclerosis will be elaborated in Chapter 3. Using novel design strategies of modulations will likely provide additional insights into the field of atherosclerosis therapy.

## Chapter references

[1] World Health Organization (WHO). Fact sheets - cardiovascular diseases. `https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)`, 2017. Accessed: Dec 2018.

[2] Encyclopaedia Britannica Inc. Encyclopaedia britannica - atherosclerosis. `https://www.britannica.com/science/atherosclerosis/media/40908/95216`, 2010. Accessed: January 4, 2019.

[3] Y. Nakashima, T.N. Wight, and K. Sueishi. Early atherosclerosis in humans: role of diffuse intimal thickening and extracellular matrix proteoglycans. *Cardiovasc. Res.*, 79(1):14–23, 2008.

[4] K.J. Williams and I. Tabas. The response-to-retention hypothesis of early atherogenesis. *Arter. Thromb. Vasc. Biol.*, 15(5):551–561, 1995.

[5] Y. Nakashima, H. Fujii, S. Sumiyoshi, T.N. Wight, and K. Sueishi. Early Human Atherosclerosis. *Arter. Thromb. Vasc. Biol.*, 27(5):1159–1165, 2007.

[6] G. Camejo. The Interaction of Lipids and Lipoproteins with the Intercellular Matrix of Arterial Tissue: Its Possible Role in Atherogenesis. *Adv. Lipid Res.*, 19:1–53, 1982.

[7] P. Fogelstrand and J. Borén. Retention of atherogenic lipoproteins in the artery wall and its role in atherogenesis. *Nutr. Metab. Cardiovasc. Dis.*, 22(1):1–7, 2012.

[8] B.H. Chung, G. Tallis, V. Yalamoori, G.M. Anantharamaiah, and J.P. Segrest. Liposome-like particles isolated from human atherosclerotic plaques are structurally and compositionally similar to surface remnants of triglyceride-rich lipoproteins. *Arterioscler. Thromb. Vasc. Biol.*, 14(4):622–635, 1994.

[9] T. Twickler, G.M. Dallinga-Thie, M.J. Chapman, and J.S. Cohn. Remnant lipoproteins and atherosclerosis. *Curr. Atheroscler. Rep.*, 7(2):140–147, 2005.

[10] A. Usman, D. Ribatti, U. Sadat, and J.H. Gillard. From Lipid Retention to Immune-Mediate Inflammation and Associated Angiogenesis in the Pathogenesis of Atherosclerosis. *J. Atheroscler. Thromb.*, 22(8):739–749, 2015.

[11] L. Mateu, E.M. Ávila, G. Camejo, V. León, and N. Liscano. The structural stability of low-density lipoprotein: A kinetic X-ray scattering study of its interaction with arterial proteoglycans. *Biochim. Biophys. Acta*, 795(3):525–534, 1984.

[12] E. Hurt, G. Bondjers, and G. Camejo. Interaction of LDL with human arterial proteoglycans stimulates its uptake by human monocyte-derived macrophages. *J. Lipid Res.*, 31(3):443–454, 1990.

[13] G. Camejo, E. Hurt, O. Wiklund, B. Rosengren, F. López, and G. Bondjers. Modifications of low-density lipoprotein induced by arterial proteoglycans and chondroitin-6-sulfate. *Biochim. Biophys. Acta*, 1096(3):253–261, 1991.

[14] E. Hurt-Camejo, G. Camejo, B. Rosengren, F. Lopez, C. Ahlstrom, G. Fager, and G. Bondjers. Effect of arterial proteoglycans and glycosaminoglycans on low density lipoprotein oxidation and its uptake by human macrophages and arterial smooth muscle cells. *Arterioscler. Thromb. Vasc. Biol.*, 12(5):569–583, 1992.

[15] J.E. Upritchard and W.H.F. Sutherland. Oxidation of heparin-treated low density lipoprotein by peroxidases. *Atherosclerosis*, 146(2):211–219, 1999.

[16] G. Camejo, G. Fager, B. Rosengren, E. Hurt-Camejo, and G. Bondjers. Binding of low density lipoproteins by proteoglycans synthesized by proliferating and quiescent human arterial smooth muscle cells. *J. Biol. Chem.*, 268(19):14131–14137, 1993.

[17] M.Y. Chang, S. Potter-Perigo, C. Tsoi, A. Chait, and T.N. Wight. Oxidized Low Density Lipoproteins Regulate Synthesis of Monkey Aortic Smooth Muscle Cell Proteoglycans That Have Enhanced Native Low Density Lipoprotein Binding Properties. *J. Biol. Chem.*, 275(7): 4766–4773, 2000.

[18] M. Gustafsson, M. Levin, K. Skalen, J. Perman, V. Friden, P. Jirholt, S.-O. Olofsson, S. Fazio, M.F. Linton, C.F. Semenkovich, G. Olivecrona, and J. Boren. Retention of low-density lipoprotein in atherosclerotic lesions of the mouse: evidence for a role of lipoprotein lipase. *Circ. Res.*, 101(8):777–783, 2007.

[19] C.K. Glass and J.L. Witztum. Atherosclerosis: The Road Ahead. *Cell*, 104(4):503–516, 2001.

[20] J. Mestas and K. Ley. Monocyte-Endothelial Cell Interactions in the Development of Atherosclerosis. *Trends Cardiovasc. Med.*, 18(6):228–232, 2008.

[21] I. Tabas. Macrophage death and defective inflammation resolution in atherosclerosis. *Nat. Rev. Immunol.*, 10(1):36–46, 2010.

[22] J.L. Johnson and A.C. Newby. Macrophage heterogeneity in atherosclerotic plaques. *Curr. Opin. Lipidol.*, 20(5):370–378, 2009.

[23] J.K. Williams, M.L. Armstrong, and D.D. Heistad. Vasa vasorum in atherosclerotic coronary arteries: responses to vasoactive stimuli and regression of atherosclerosis. *Circ. Res.*, 62(3): 515–523, 1988.

[24] H.M. Kwon, G. Sangiorgi, E.L. Ritman, C. McKenna, D.R.Jr Holmes, R.S. Schwartz, and A. Lerman. Enhanced coronary vasa vasorum neovascularization in experimental hypercholesterolemia. *J. Clin. Invest.*, 101(8):1551–1556, 1998.

[25] O. Galili, K.J. Sattler, J. Herrmann, J. Woodrum, M. Olson, L.O. Lerman, and A. Lerman. Experimental hypercholesterolemia differentially affects adventitial vasa vasorum and vessel structure of the left internal thoracic and coronary arteries. *J. Thorac. Cardiovasc. Surg.*, 129 (4):767–772, 2005.

[26] T. Nakano, T. Ninomiya, S. Sumiyoshi, M. Onimaru, H. Fujii, H. Itabe, Y. Nakashima, K. Sueishi, K. Tsuruya, Y. Oda, T. Kitazono, and Y. Kiyohara. Chronic kidney disease is associated with neovascularization and intraplaque hemorrhage in coronary atherosclerosis in elders: results from the Hisayama Study. *Kidney Int.*, 84(2):373–380, 2013.

[27] G. Pasterkamp, A.H. Schoneveld, D.J. Hijnen, D.P.V. de Kleijn, H. Teepen, A.C. van der Wal, and C. Borst. Atherosclerotic arterial remodeling and the localization of macrophages and matrix metalloproteases 1, 2 and 9 in the human coronary artery. *Atherosclerosis*, 150(2): 245–253, 2000.

[28] F.L. Celletti, J.M. Waugh, P.G. Amabile, A. Brendolan, P.R. Hilfiker, and M.D. Dake. Vascular endothelial growth factor enhances atherosclerotic plaque progression. *Nat. Med.*, 7:425, 2001.

[29] C. Ihling, T. Szombathy, B. Bohrmann, M. Brockhaus, H.E. Schaefer, and B.M. Loeffler. Coexpression of endothelin-converting enzyme-1 and endothelin-1 in different stages of human atherosclerosis. *Circulation*, 104(8):864–869, 2001.

[30] J.C. Sluimer, J.-M. Gasc, J.L. van Wanroij, N. Kisters, M. Groeneweg, M.D. Sollewijn Gelpke, J.P. Cleutjens, L.H. van den Akker, P. Corvol, B.G. Wouters, M.J. Daemen, and A.-P.J. Bijnens. Hypoxia, Hypoxia-Inducible Transcription Factor, and Macrophages in Human Atherosclerotic Plaques Are Correlated With Intraplaque Angiogenesis. *J. Am. Coll. Cardiol.*, 51(13):1258–1265, 2008.

[31] T. Nakano, Y. Nakashima, Y. Yonemitsu, S. Sumiyoshi, Y.-X. Chen, Y. Akishima, T. Ishii, M. Iida, and K. Sueishi. Angiogenesis and lymphangiogenesis and expression of lymphangiogenic factors in the atherosclerotic intima of human coronary arteries. *Hum. Pathol.*, 36(4): 330–340, 2005.

[32] R. Virmani, F.D. Kolodgie, A.P. Burke, A.V. Finn, H.K. Gold, T.N. Tulenko, S.P. Wrenn, and J. Narula. Atherosclerotic plaque progression and vulnerability to rupture: angiogenesis as a source of intraplaque hemorrhage. *Arter. Thromb. Vasc. Biol.*, 25(10):2054–2061, 2005.

[33] J.A. Fryer, P.C. Myers, and M. Appleberg. Carotid intraplaque hemorrhage: the significance of neovascularity. *J. Vasc. Surg.*, 6(4):341–349, 1987.

[34] M. Hoshiga, C.E. Alpers, L.L. Smith, C.M. Giachelli, and S.M. Schwartz. Alpha-v beta-3 integrin expression in normal and atherosclerotic artery. *Circ. Res.*, 77(6):1129–1135, 1995.

[35] J.J. Boyle, B. Wilson, R. Bicknell, S. Harrower, P.L. Weissberg, and T.P. Fan. Expression of angiogenic factor thymidine phosphorylase and angiogenesis in human atherosclerosis. *J. Pathol.*, 192(2):234–242, 2000.

[36] S. Suarez and K. Ballmer-Hofer. VEGF transiently disrupts gap junctional communication in endothelial cells. *J. Cell Sci.*, 114(6):1229 LP – 1235, 2001.

[37] J.C. Sluimer, F.D. Kolodgie, A.P.J.J. Bijnens, K. Maxfield, E. Pacheco, B. Kutys, H. Duimel, P.M. Frederik, V.W.M. van Hinsbergh, R. Virmani, and M.J.A.P. Daemen. Thin-Walled Microvessels in Human Coronary Atherosclerotic Plaques Show Incomplete Endothelial Junctions: Relevance of Compromised Structural Integrity for Intraplaque Microvascular Leakage. *J. Am. Coll. Cardiol.*, 53(17):1517–1527, 2009.

[38] P.D. Richardson, M.J. Davies, and G.V.R. Born. Influence of plaque configuration and stress distribution on fissuring of coronary atherosclerotic plaques. *Lancet*, 334(8669):941–944, 1989.

[39] A.P.G. Hoeks, K.D. Reesink, E. Hermeling, and R.S. Reneman. Local Blood Pressure Rather Than Shear Stress Should Be Blamed for Plaque Rupture. *J. Am. Coll. Cardiol.*, 52(13): 1107–1108, 2008.

[40] F.J.H. Gijsen, J.J. Wentzel, A. Thury, F. Mastik, J.A. Schaar, J.C.H. Schuurbiers, C.J. Slager, W.J. van der Giessen, P.J. de Feyter, A.F.W. van der Steen, and P.W. Serruys. Strain distribution over plaques in human coronary arteries relates to shear stress. *Am. J. Physiol. Heart Circ. Physiol.*, 295(4):H1608–14, 2008.

[41] J. Ohayon, G. Finet, A.M. Gharib, D.A. Herzka, P. Tracqui, J. Heroux, G. Rioufol, M.S. Kotys, A. Elagha, and R.I. Pettigrew. Necrotic core thickness and positive arterial remodeling index: emergent biomechanical factors for evaluating the risk of plaque rupture. *Am. J. Physiol. Heart Circ. Physiol.*, 295(2):H717–27, 2008.

[42] The National Heart Lung and Blood Institute (NHLBI). Atherosclerosis. `https://www.nhlbi.nih.gov/health-topics/atherosclerosis`, 2018. Accessed: Dec 2018.

[43] Mayo Clinic. Arteriosclerosis / atherosclerosis - diagnosis & treatment. `https://www.mayoclinic.org/diseases-conditions/arteriosclerosis-atherosclerosis/diagnosis-treatment/drc-20350575`, 2018. Accessed: Dec 2018.

[44] J.C. Sluimer and M.J. Daemen. Novel concepts in atherogenesis: angiogenesis and hypoxia in atherosclerosis. *J. Pathol.*, 218(1):7–29, 2009.

[45] Andreas Daiber, Sebastian Steven, Alina Weber, Vladimir V Shuvaev, Vladimir R Muzykantov, Ismail Laher, Huige Li, Santiago Lamas, and Thomas Munzel. Targeting vascular (endothelial) dysfunction. *Br. J. Pharmacol.*, 174(12):1591–1619, 2017.

[46] Nihed Draoui, Pauline de Zeeuw, and Peter Carmeliet. Angiogenesis revisited from a metabolic perspective: role and therapeutic implications of endothelial cell metabolism. *Open Biol.*, 7 (12), 2017.

[47] Neil Ruparelia, Joshua T Chai, Edward A Fisher, and Robin P Choudhury. Inflammatory processes in cardiovascular disease: a route to targeted therapies. *Nat. Rev. Cardiol.*, 14(3): 133–144, 2017.

# CHAPTER 3

# PFKFB3 and its relation to atherosclerosis

## 3.1 PFKFB3: an overview

6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 3 (PFKFB3 hereafter) enzyme is one of the four known PFKFB isozymes found in humans.[1,2] *PFKFB3* is the gene that encodes the human enzyme iPFK2 however it is also frequently been referred to the enzyme itself. The first crystal structure of the enzyme was resolved in 2006[3] but relevant studies had already commenced as early as 1982.[4,5] From 2012, it underwent an increase as reflected by the increasing numbers of publications and the number of the times the articles have been cited during more recent years (Figure 3.1). This is likely due to the increased input into researches within the field of cancer and cardiovascular diseases, and the advances in biotechnology. This allowed a better understanding of PFKFB3's biological and structural properties, hence highlighted the important connection of the enzyme to health and disease-related issues.

The PFKFB family is responsible for modulating the intracellular level of fructose-2,6-

bisphosphate (F2,6BP);[1,7,8] the latter is an important activator of the rate-limiting glycolytic enzyme 6-phosphofructo-1-kinase (PFK1).[9,10] PFBFKs are bifunctional and catalyse the phosphorylation of fructose-6-phosphate (F6P) to F2,6BP and also the reversed reaction.[1,7,8] Four isozymes of the family have been so far determined (PFKFB1-4), and are expressed in different tissues.[1,2] More than 85% similarities were observed within catalytic domains between different isozymes. However, the high divergent in amino acid sequences at the termini likely provided much different isozyme properties. Originally discovered to be expressed in the cells of liver/muscle, heart, and testes, respectively, the PFKFB1, PFKFB2, and PFKFB4 all display almost equivalent activity between kinase and phosphatase.[7] PFKFB3, on the other hand, was demonstrated with an exceptionally high kinase activity with the kinase-to-phosphatase ratio around 700:1.[11] Initially isolated from cells of the brain and human placenta,[12] the gene expression of the 3 isozyme was later discovered ubiquitous and inducible.[13,14] Some transformed cells, whole organs, and primary human epithelial cells were suggested to co-express all four PFKFBs.[15,16] However, the sig-

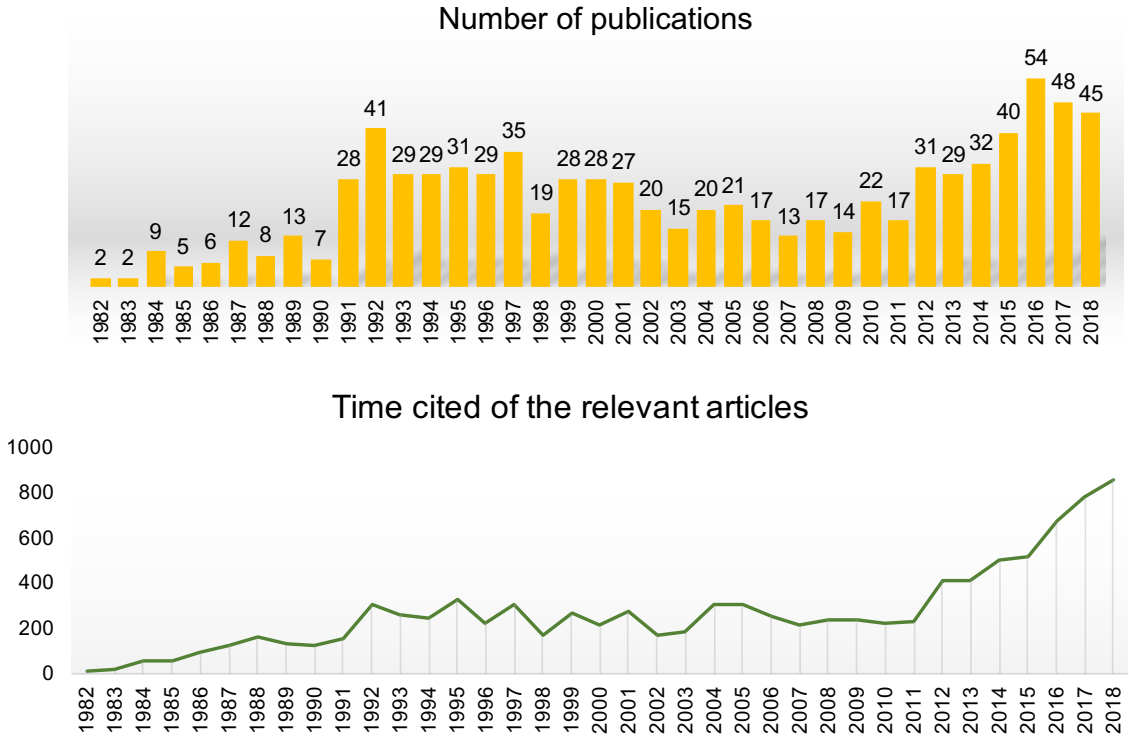



**Figure 3.1:** Statistics of PFKFB3 related studies. Data retrieved from Web of Science.[6]

nificantly high kinase activity of PFKFB3 is likely to play a major role in determining the intracellular level of F2,6P.[16]

## 3.2 Biological pathways and relevant diseases

PFK-1 catalyses an irreversible phosphorylation reaction converting F6P to fructose-1,6-bisphosphate (F1,6BP), an early step of glycolysis.[17,18] F2,6BP can inhibit fructose-1,6-bisphosphatase (F1,6BPase) and activating PFK-1.[9,10] More remarkably, F2,6P exhibits much higher positive effect on glycolysis than F1,6BP, and is still by far the most potent PFK-1 activator and an important regulator of glycolysis or gluconeogenesis.[19,20] Upon binding, it also alters the conformation of PFK-1 to acquire an increased affinity for F6P,[10] and weakened negative feedback from ATP.[21] Hence, PFKFBs, especially the 3 isozyme, is closely connected to the up-regulation of cellular glycolytic flux through regulating the intracellular F2,6BP concentration.

Cancer cells are known for abnormally up-regulated aerobic glycolysis, as described

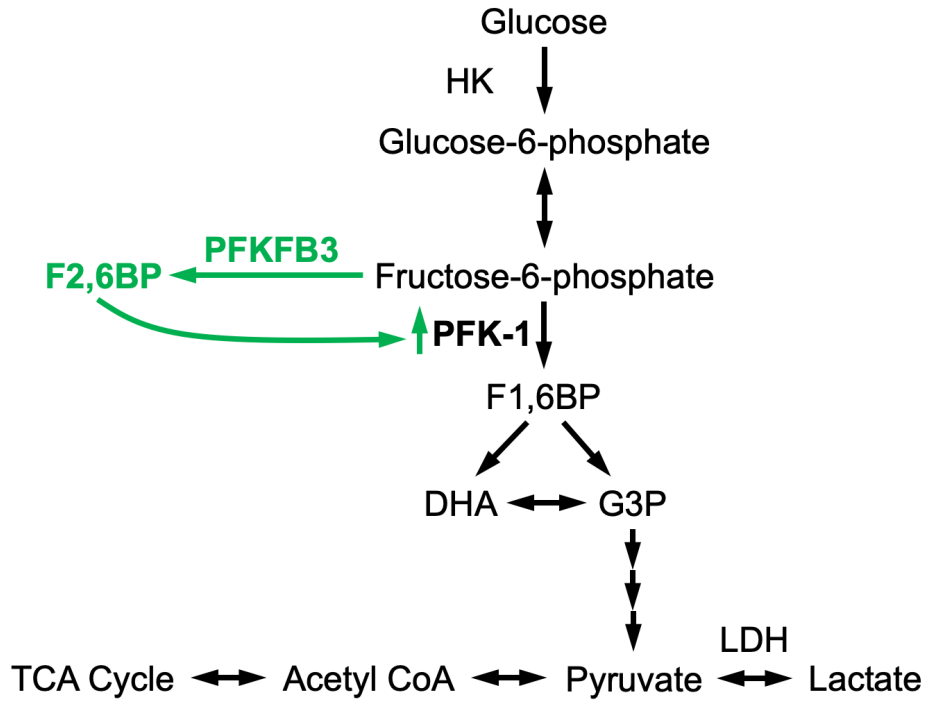



**Figure 3.2:** PFKFB3 regulates glycolysis pathway through F2,6P and PFK-1.

by the Warburg effect.[22] The increased rate of glycolysis also provides precursors for molecules necessary for cell growth and proliferations.[23] A significantly elevated amount of F2,6BP during steady state was discovered for multiple transformed cell lines.[2,24,25] The expression of PFKFB3 mRNA has been determined in HeLa, breast cancer, leukaemia, colon and colorectal adenocarcinoma, lung carcinoma, gastric cancer, and pancreatic cancer cell lines *in vitro*.[14,25–30] Moreover, *in situ* studies also confirmed that PFKFB3 is over-expressed in multiple tissues isolated from tumours comparing to normal controls.[31] These findings constitutionally confirmed the connection of PFKFB3 to abnormal cancer metabolism.

One extensively studied factor leading to PFKFB3 over-expression in cancer cells is its connection to hypoxia.[15,32–36] It has been well established that the HIF-1 complex is a key mediator of the hypoxia response and highly expressed in cancers.[37–40] It can recognise two transcription factor binding sites that are close to each other, consisting in the namely hypoxia-response elements (HREs).[37,38] Studies have revealed that PFKFB3 is induced by hypoxia to a greater extent than the other isozymes.[15,32,34] Like other glycolytic enzymes promoted by hypoxia, expression of PFKFB3 is stimulated through HREs in its promoters.[32,34,41] To enhance the activity of PFKFB3 under hypoxia, the AMP-activated protein kinase (AMPK) is also involved. AMPK is activated by an elevated cellular AMP/ATP ratio and phosphorylates PFKFB3, increasing the enzymatic $V_{max}$.[42,43] Emerging evidence suggested the connection between phosphorylated PFKFB3 to tumour cells.[44] The hypoxic over-expression and activation of PFKFB3, hence, lead to the accumulation of glycolytic stimulating F2,6BP within the cytosolic environment.

More recent studies have also revealed the relation between PFKFB3 to obesity and diabetes.[45–48] Insulin can increase intracellular F2,6BP concentration by activating PFKFB3 through phosphorylation in adipocytes.[45,49] Moreover, *in vivo* study using PFKFB3$^{+/-}$ mice developed more severe insulin resistance, decreased insulin signalling, and increased adipose tissue inflammatory response when put on a high-fat diet (HFD).[46] On the other hand, over-expression of PFKFB3 in the transgenic mice

under HFD displayed decreased adipose inflammatory response, and improved insulin sensitivity.[47] A kinome screen study further determined that PFKFB3 is associated with the insulin/insulin-like growth factor (IGF)-1 signalling pathway.[48] The multi-directional studies show that PFKFB3 is a crucial regulatory node in glycolysis and cellular metabolism. The involvement in multiple biological processes renders it an important therapeutic target.

## 3.3 Connecting PFKFB3 and atherosclerosis

A decade after the new millennium, a collection of studies have focused on connecting angiogenesis to the metabolism profile, especially to PFKFB3.[50–55] PFKFB3 has been reported to relate to vessel sprouting and angiogenesis in both endothelial cells and tumours.[50,51] Tumour angiogenesis is proven relating to cancer proliferation and metastasis.[56–58] Tumour endothelial cells (TECs) supplies the fundamental structural and functional abnormality of tumour vessels.[59] Moreover, normal endothelial cells (ECs), though more regular in shape and size, have also higher dependence on glycolytic metabolism.[50,60–62] It was suggested that the EC glycolytic level is comparable to that of cancer cells, generating >85% of the total cellular ATP.[50,62] Silencing and knockdown of PFKFB3 impair both vessel sprouting and angiogenesis.[50–52]

As early as 1936, microvessels were observed in atherosclerotic plaques.[63] Mainly originated from the adventitia (the connective tissue outside the blood vessels), these micro-vessels appears with more disorganised and excessive networks accompanied by enlarged and irregular vessel diameters and prone to inflammation.[64,65] The plaque vessels also gradually increase with lesion progression and hypoxia, presented in advanced atherosclerotic lesions, also stimulate angiogenesis.[66] HIF, with the potential promoting PFKFB3 transcription and translation, is also associated with vascular endothelial growth factor(VEGF), endothelin-1, and matrix metalloproteinase-2 in atherosclerosis.[67–69] Direct evidence has confirmed in advanced human atherosclerosis that the expression of HIF and VEGF is correlated to hypoxia and the presence

of macrophages.[70] VEGF/VEGF receptors (VEGFRs) signalling pathway is closely related to angiogenesis; in atherosclerotic mice models, administration of VEGF or anti-VEGF antibody can promote or decreases angiogenesis and macrophage infiltration.[68,71,72] Hence, the connection between angiogenesis and atherosclerosis has been well established.

Only recently PFKFB3 starts to be relating to atherosclerosis.[73–76] Doddaballapur et al. discovered that fluid shear stress quiesce EC metabolism through inducing the transcription factor Krüppel-like factor 2 (KLF2).[73] KLF2, as a known inhibitor of HIF-1 and VEGF,[77,78] can thus indirectly suppress the expression of PFKFB3; this is reflected by the significant decrease in PFKFB3 expression induced by laminar flow.[73] Baek et al. suggested a flow-mediated VEGFR-PKC(protein kinase C) -PFKFB3 signalling; both pulsatile shear stress and oscillatory shear stress regulate glycolytic metabolism through PKC -dependent endothelial PFKFB3 expression.[76]

On the other hand, a key study by Tawakol et al. demonstrated that both HIF-1 and PFKFB3 relate closely to pro-inflammatory and metabolic activity of macrophage within atherosclerosis.[74] Both *in vitro* and *in vivo*, silencing HIF-1 or PFKFB3 down-regulate glycolysis and pro-inflammatory activation.[74] Moreover, in a clinical setting, PFKFB3 showed higher expression level in symptomatic atherosclerotic patients when compared to control.[75] Though more studies should be expanded in the context of atherosclerosis, PFKFB3 is emerging as a potential therapeutic target for atherosclerosis treatment.

## Chapter references

[1] D. A. Okar, A. Manzano, A. Navarro-Sabate, L. Riera, R. Bartrons, and A. J. Lange. PFK-2/FBPase-2: maker and breaker of the essential biofactor fructose-2,6-bisphosphate. *Trends Biochem. Sci.*, 26(1):30–35, 2001.

[2] M. H. Rider, L. Bertrand, D. Vertommen, P. A. Michels, G. G. Rousseau, and L. Hue. 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase: head-to-head with a bifunctional enzyme that controls glycolysis. *Biochem. J*, 381(3):561–579, 2004.

[3] S.-G. Kim, N.P. Manes, M.R. El-Maghrabi, and Y.-H. Lee. Crystal structure of the hypoxia-inducible form of 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase (pfkfb3): A possible new target for cancer therapy. *J. Biol. Chem.*, 281(5):2939–2944, 2006.

[4] M.R. el Maghrabi, T.H. Claus, J. Pilkis, and S.J Pilkis. Regulation of 6-phosphofructo-2-kinase activity by cyclic amp-dependent phosphorylation. *Proc. Natl. Acad. Sci. U.S.A.*, 79 (2):315–319, 1982.

[5] M.R. El-Maghrabi, E. Fox, J Pilkis, and S.J. Pilkis. Cyclic AMP-dependent phosphorylation of rat liver 6-phosphofructo 2-kinase/fructose 2,6-bisphosphatase. *Biochem. Biophys. Res. Commun.*, 106(3):794–802, 1982.

[6] Clarivate Analytics (United States). Web of science, 2018. URL `https://www.webofknowledge.com`. Accessed: 2018-12-04.

[7] David A. Okar and Alex J. Lange. Fructose-2,6-bisphosphate and control of carbohydrate metabolism in eukaryotes. *BioFactors*, 10(1):1–14, 1999.

[8] D. A. Okar, C. Wu, and A. J. Lange. Regulation of the regulatory enzyme, 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase. *Adv. Enzyme Regul.*, 44(1):123 – 154, 2004.

[9] E. Van Schaftingen, L. Hue, and H.G. Hers. Control of the fructose 6-phosphate/fructose 1,6-bisphosphate cycle in isolated hepatocytes by glucose and glucagon. role of a low-molecular-weight stimulator of phosphofructokinase. *Biochem. J*, 192(3):887–895, 1980.

[10] E. Van Schaftingen, L. Hue, and H.G. Hers. Fructose 2,6-bisphosphate, the probably structure of the glucose- and glucagon-sensitive stimulator of phosphofructokinase. *Biochem. J*, 192(3): 897–901, 1980.

[11] R. Sakakibara, M. Kato, N. Okamura, T. Nakagawa, Y. Komada, N. Tominaga, M. Shimojo, and M. Fukasawa. Characterization of a human placental fructose-6-phosphate, 2-kinase/fructose- 2,6 - bisphosphatase1. *J. Biochem.*, 122(1):122–128, 1997.

[12] A. Sakai, M. Kato, M. Fukasawa, M. Ishiguro, E. Furuya, and R. Sakakibara. Cloning of cdna encoding for a novel isozyme of fructose 6-phosphate,2-kinase/fructose 2,6-bisphosphatase from human placenta1. *J. Biochem.*, 119(3):506–511, 1996.

[13] A. Manzano, J.L. Rosa, F. Ventura, J.X. Perez, M. Nadal, X. Estivill, S. Ambrosio, J. Gil, and R. Bartrons. Molecular cloning, expression, and chromosomal localization of a ubiquitously expressed human 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase gene (PFKFB3). *Cytogenet. Cell Genet.*, 83(3-4):214–217, 1998.

[14] J. Chesney, R. Mitchell, F. Benigni, M. Bacher, L. Spiegel, Y. Al-Abed, J.H. Han, C. Metz, and R. Bucala. An inducible gene product for 6-phosphofructo-2-kinase with an au-rich instability element: Role in tumor cell glycolysis and the warburg effect. *Proc. Natl. Acad. Sci. U.S.A.*, 96(6):3047–3052, 1999.

[15] O. Minchenko, I. Opentanova, and J. Caro. Hypoxic regulation of the 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase gene family (pfkfb-1–4) expression in vivo. *FEBS Lett.*, 554 (3):264–270, 2003.

[16] S. Telang, A. Yalcin, A.L. Clem, R. Bucala, A.N. Lane, J.W. Eaton, and J. Chesney. Ras transformation requires metabolic control by 6-phosphofructo-2-kinase. *Oncogene*, 25:7225, 2006.

[17] G. Weber. Enzymology of cancer cells. *N. Engl. J. Med.*, 296(10):541–551, 1977.

[18] M.J. Marshall, D.M. Goldberg, F.E. Neal, and D.R. Millar. Enzymes of glucose metabolism in carcinoma of the cervix and endometrium of the human uterus. *Br. J. Cancer*, 37(6):990–1001, 1978.

[19] H.G. Hers. Fructose 2,6-bisphosphate. *Biochem Soc Trans.*, 11(3):250–251, 1983.

[20] H.G. Hers. The discovery and the biological role of fructose 2,6-bisphosphate. *Biochem Soc Trans.*, 12(5):729–735, 1984.

[21] C. Wu, S. A Khan, L.-J. Peng, and A.J. Lange. Roles for fructose-2,6-bisphosphate in the control of fuel metabolism: Beyond its allosteric effects on glycolytic and gluconeogenic enzymes. *Adv. Enzyme Regul.*, 46(1):72–88, 2006.

[22] O. Warburg. On the origin of cancer cells. *Science*, 123(3191):309–314, 1956.

[23] T. Bui and C.B Thompson. Cancer's sweet tooth. *Cancer Cell*, 9(6):419–420, 2006.

[24] D. Colomer, J.L. Vives-Corrons, A. Pujades, and R. Bartrons. Control of phosphofructokinase by fructose 2,6-bisphosphate in b-lymphocytes and b-chronic lymphocytic leukemia cells. *Cancer Res.*, 47(7):1859–1862, 1987.

[25] L. Riera, A. Manzano, A. Navarro-Sabaté, J.C. Perales, and R. Bartrons. Insulin induces PFKFB3 gene expression in HT29 human colon adenocarcinoma cells. *Biochim. Biophys. Acta, Mol. Cell. Res.*, 1589(2):89–92, 2002.

[26] J.A. Hamilton, M.J. Callaghan, R.L. Sutherland, and C.K.W. Watts. Identification of prg1, a novel progestin-responsive gene with sequence homology to 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase. *Mol. Endocrinol.*, 11(4):490–502, 1997.

[27] A.Y. Bobarykina, D.O. Minchenko, I.L. Opentanova, M. Moenner, J. Caro, H. Esumi, and O.H. Minchenko. Hypoxic regulation of PFKFB-3 and PFKFB-4 gene expression in gastric and pancreatic cancer cell lines and expression of PFKFB genes in gastric cancers. *Acta Biochim. Pol.*, 53(4):789–799, 2006.

[28] M.N. Calvo, R. Bartrons, E. Castaño, J.C. Perales, A. Navarro-Sabaté, and A. Manzano. Pfkfb3 gene silencing decreases glycolysis, induces cell-cycle delay and inhibits anchorage-independent growth in hela cells. *FEBS Lett.*, 580(13):3308–3314, 2006.

[29] L.L.C. Marotta, V. Almendro, A. Marusyk, M.l Shipitsin, J. Schemme, S.R. Walker, N. Bloushtain-Qimron, J.J. Kim, S.A. Choudhury, R. Maruyama, Z. Wu, M. Gönen, L.A. Mulvey, M.O. Bessarabova, S.J. Huh, S.J. Silver, S.Y. Kim, S.Y. Park, H.E. Lee, Karen S Anderson, Andrea L Richardson, Tatiana Nikolskaya, Yuri Nikolsky, X Shirley Liu, David E Root, William C. H., D.A. Frank, and K. Polyak. The JAK2/STAT3 signaling pathway is required for growth of CD44+CD24– stem cell–like breast cancer cells in human tumors. *J. Clin. Investig*, 121(7):2723–2735, 2011.

[30] L. Novellasdemunt, M. Obach, L. Millán-Ariño, A. Manzano, F. Ventura, J.L. Rosa, A. Jordan, À. Navarro-Sabate, and R. Bartrons. Progestins activate 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase 3 (PFKFB3) in breast cancer cells. *Biochem. J*, 442(2):345–356, 2012.

[31] T. Atsumi, J. Chesney, C. Metz, L. Leng, S. Donnelly, Z. Makita, R. Mitchell, and R. Bucala. High Expression of Inducible 6-Phosphofructo-2-Kinase/Fructose-2,6-Bisphosphatase (iPFK-2; PFKFB3) in Human Cancers. *Cancer Res.*, 62(20):5881–5887, 2002.

[32] A. Minchenko, I. Leshchinsky, I. Opentanova, N. Sang, V. Srinivas, V. Armstead, and J. Caro. Hypoxia-inducible Factor-1-mediated Expression of the 6-Phosphofructo-2-kinase/fructose-2,6-bisphosphatase-3 (PFKFB3) Gene: ITS POSSIBLE ROLE IN THE WARBURG EFFECT. *J. Biol. Chem.*, 277(8):6183–6187, 2002.

[33] M. Fukasawa, T. Tsuchiya, E. Takayama, N. Shinomiya, K. Uyeda, R. Sakakibara, and S. Seki. Identification and Characterization of the Hypoxia-Responsive Element of the Human Placental 6-Phosphofructo-2-Kinase/Fructose-2,6-Bisphosphatase Gene. *J. Biochem.*, 136(3):273–277, 2004.

[34] M. Obach, À. Navarro-Sabaté, J. Caro, X. Kong, J. Duran, M. Gómez, J.C. Perales, F. Ventura, J.L. Rosa, and R. Bartrons. 6-Phosphofructo-2-kinase (pfkfb3) Gene Promoter Contains Hypoxia-inducible Factor-1 Binding Sites Necessary for Transactivation in Response to Hypoxia. *J. Biol. Chem.*, 279(51):53562–53570, 2004.

[35] R. Bartrons and J. Caro. Hypoxia, glucose metabolism and the Warburg's effect. *J. Bioenerg. Biomembr.*, 39(3):223–229, 2007.

[36] J. Chesney, J. Clark, A.C. Klarer, Y. Imbert-Fernandez, A.N. Lane, and S. Telang. Fructose-2,6-bisphosphate synthesis by 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase 4 (PFKFB4) is required for the glycolytic response to hypoxia and tumor growth. *Oncotarget*, 5(16):6670–6686, 2014.

[37] J. Caro. Hypoxia Regulation of Gene Transcription. *High Altitude Med. Biol.*, 2(2):145–154, 2001.

[38] P.H. Maxwell, C.W. Pugh, and P.J. Ratcliffe. Activation of the HIF pathway in cancer. *Curr. Opin. Genet. Dev.*, 11(3):293–299, 2001.

[39] G.L. Semenza. Development of novel therapeutic strategies that target HIF-1. *Expert Opin. Ther. Targets*, 10(2):267–280, 2006.

[40] M.C. Brahimi-Horn, J. Chiche, and J. Pouysségur. Hypoxia and cancer. *J. Mol. Med.*, 85(12): 1301–1307, 2007.

[41] C.V. Dang and G.L. Semenza. Oncogenic alterations of metabolism. *Trends Biochem. Sci.*, 24 (2):68–72, 1999.

[42] S.-P. Hong, F.C. Leiper, A. Woods, D. Carling, and M. Carlson. Activation of yeast Snf1 and mammalian AMP-activated protein kinase by upstream kinases. *Proc. Natl. Acad. Sci. U.S.A.*, 100(15):8839–8843, 2003.

[43] A.-S. Marsin, C. Bouzin, L. Bertrand, and L. Hue. The Stimulation of Glycolysis by Hypoxia in Activated Monocytes Is Mediated by AMP-activated Protein Kinase and Inducible 6-Phosphofructo-2-kinase. *J. Biol. Chem.*, 277(34):30778–30783, 2002.

[44] H. Bando, T. Atsumi, T. Nishio, H. Niwa, S. Mishima, C. Shimizu, N. Yoshioka, R. Bucala, and T. Koike. Phosphorylation of the 6-Phosphofructo-2-Kinase/Fructose 2,6-Bisphosphatase/PFKFB3 Family of Glycolytic Regulators in Human Cancer. *Clin. Cancer Res.*, 11(16):5784–5792, 2005.

[45] T. Atsumi, T. Nishio, H. Niwa, J. Takeuchi, H. Bando, C. Shimizu, N. Yoshioka, R. Bucala, and T. Koike. Expression of Inducible 6-Phosphofructo-2-Kinase/Fructose-2,6-Bisphosphatase/PFKFB3 Isoforms in Adipocytes and Their Potential Role in Glycolytic Regulation. *Diabetes*, 54(12):3349–3357, 2005.

[46] Y. Huo, X. Guo, H. Li, H. Wang, W. Zhang, Y. Wang, H. Zhou, Z. Gao, S. Telang, J. Chesney, Y.E. Chen, J. Ye, R.S. Chapkin, and C. Wu. Disruption of Inducible 6-Phosphofructo-2-kinase Ameliorates Diet-induced Adiposity but Exacerbates Systemic Insulin Resistance and Adipose Tissue Inflammatory Response. *J. Biol. Chem.*, 285(6):3713–3721, 2010.

[47] Y. Huo, X. Guo, H. Li, H. Xu, V. Halim, W. Zhang, H. Wang, Y.-Y. Fan, K.T. Ong, S.-L. Woo, R.S. Chapkin, D.G. Mashek, Y. Chen, H. Dong, F. Lu, L. Wei, and C. Wu. Targeted Overexpression of Inducible 6-Phosphofructo-2-kinase in Adipose Tissue Increases Fat Deposition but Protects against Diet-induced Insulin Resistance and Inflammatory Responses. *J. Biol. Chem.*, 287(25):21492–21500, 2012.

[48] S. Trefely, P.-S. Khoo, J.R. Krycer, R. Chaudhuri, D.J. Fazakerley, B.L. Parker, G. Sultani, J. Lee, J.-P. Stephan, E. Torres, K. Jung, C. Kuijl, D.E. James, J.R. Junutula, and J. Stöckli.

Kinome Screen Identifies PFKFB3 and Glucose Metabolism as Important Regulators of the Insulin/Insulin-like Growth Factor (IGF)-1 Signaling Pathway. *J. Biol. Chem.*, 290(43):25834–25846, 2015.

[49] F Sobrino and A Gualberto. Hormonal regulation of fructose 2,6-bisphosphate levels in epididymal adipose tissue of rat. *FEBS Lett.*, 182(2):327–330, 1985.

[50] K. De Bock, M. Georgiadou, S. Schoors, A. Kuchnio, B.W. Wong, A.R. Cantelmo, A. Quaegebeur, B. Ghesquière, S. Cauwenberghs, G. Eelen, L.-K. Phng, I. Betz, B. Tembuyser, K. Brepoels, J. Welti, I. Geudens, I. Segura, B. Cruys, F. Bifari, I. Decimo, R. Blanco, S. Wyns, J. Vangindertael, S. Rocha, R.T. Collins, S. Munck, D. Daelemans, H. Imamura, R. Devlieger, M. Rider, P.P. Van Veldhoven, F. Schuit, R. Bartrons, J. Hofkens, P. Fraisl, S. Telang, R.J. DeBerardinis, L. Schoonjans, S. Vinckier, J. Chesney, H. Gerhardt, M. Dewerchin, and P. Carmeliet. Role of PFKFB3-Driven Glycolysis in Vessel Sprouting. *Cell*, 154(3):651–663, 2013.

[51] Y. Xu, X. An, X. Guo, T.G. Habtetsion, Y. Wang, X. Xu, S. Kandala, Q. Li, H. Li, C. Zhang, R.B. Caldwell, D.J. Fulton, Y. Su, M.N. Hoda, G. Zhou, C. Wu, and Y Huo. Endothelial PFKFB3 Plays a Critical Role in Angiogenesis. *Arter. Thromb. Vasc. Biol.*, 34(6):1231–1239, 2014.

[52] A.R. Cantelmo, L.-C. Conradi, A. Brajic, J. Goveia, J. Kalucka, A. Pircher, P. Chaturvedi, J. Hol, B. Thienpont, L.-A. Teuwen, S. Schoors, B. Boeckx, J. Vriens, A. Kuchnio, K. Veys, B. Cruys, L. Finotto, L. Treps, T.E. Stav-Noraas, F. Bifari, P. Stapor, I. Decimo, K. Kampen, K. De Bock, G. Haraldsen, L. Schoonjans, T. Rabelink, G. Eelen, B. Ghesquière, J. Rehman, D. Lambrechts, A.B. Malik, M. Dewerchin, and P. Carmeliet. Inhibition of the Glycolytic Activator PFKFB3 in Endothelium Induces Tumor Vessel Normalization, Impairs Metastasis, and Improves Chemotherapy. *Cancer Cell*, 30(6):968–985, 2016.

[53] M. Gu, L. Li, Z. Zhang, J. Chen, W. Zhang, J. Zhang, L. Han, M. Tang, B. You, Q. Zhang, and Y. You. PFKFB3 promotes proliferation, migration and angiogenesis in nasopharyngeal carcinoma. *J. Cancer*, 8(18):3887–3896, 2017.

[54] A. Trenti, S. Tedesco, C. Boscaro, N. Ferri, A. Cignarella, L. Trevisi, and C. Bolego. The Glycolytic Enzyme PFKFB3 Is Involved in Estrogen-Mediated Angiogenesis via GPER1. *Journal of Pharmacology and Experimental Therapeutics*, 361(3):398–407, 2017.

[55] F. Peng, Q. Li, J.-Y. Sun, Y. Luo, M. Chen, and Y. Bao. PFKFB3 is involved in breast cancer proliferation, migration, invasion and angiogenesis. *Int. J. Oncol.*, 52(3):945–954, 2018.

[56] R. Kinkade, P. Dasgupta, A. Carie, D. Pernazza, M. Carless, S. Pillai, N. Lawrence, S.M. Sebti, and S. Chellappan. A Small Molecule Disruptor of Rb/Raf-1 Interaction Inhibits Cell Proliferation, Angiogenesis, and Growth of Human Tumor Xenografts in Nude Mice. *Cancer Res.*, 68(10):3810–3818, 2008.

[57] K.H. Jung, H.-M. Zheng, Y. Jeong, M.-J. Choi, H. Lee, S.-W. Hong, H.-S. Lee, M.K. Son, S. Lee, S. Hong, and S.-S. Hong. Suppression of tumor proliferation and angiogenesis of hepatocellular carcinoma by HS-104, a novel phosphoinositide 3-kinase inhibitor. *Cancer Lett.*, 328(1):176–187, 2013.

[58] H. Mukai, A. Muramatsu, R. Mashud, K. Kubouchi, S. Tsujimoto, T. Hongu, Y. Kanaho, M. Tsubaki, S. Nishida, G. Shioi, S. Danno, M. Mehruba, R. Satoh, and R. Sugiura. PKN3 is the major regulator of angiogenesis and tumor metastasis in mice. *Sci. Rep.*, 6:18979, 2016.

[59] R.K. Jain. Antiangiogenesis Strategies Revisited: From Starving Tumors to Alleviating Hypoxia. *Cancer Cell*, 26(5):605–622, 2014.

[60] M. Quintero, S.L. Colombo, A. Godfrey, and S. Moncada. Mitochondria as signaling organelles in the vascular endothelium. *Proc. Natl. Acad. Sci. U.S.A.*, 103(14):5379–5384, 2006.

[61] J.R. Merchan, K. Kovács, J.W. Railsback, M. Kurtoglu, Y. Jing, Y. Piña, N. Gao, T.G. Murray, M.A. Lehrman, and T.J. Lampidis. Antiangiogenic Activity of 2-Deoxy-D-Glucose. *PLoS One*, 5(10):e13699, 2010.

[62] K. De Bock, M. Georgiadou, and P. Carmeliet. Role of Endothelial Cell Metabolism in Vessel Sprouting. *Cell Metab.*, 18(5):634–647, 2013.

[63] J.C. Paterson. Vascularization and hemorrhage of the intima of arteriosclerotic arteries. *Arch. Pathol.*, 22:312–324, 1936.

[64] M.P.W. Moos, Nicole John, R. Gräbner, S. Noßmann, B. Günther, R. Vollandt, C.D. Funk, B. Kaiser, and A.J.R. Habenicht. The Lamina Adventitia Is the Major Site of Immune Cell Accumulation in Standard Chow-Fed Apolipoprotein E–Deficient Mice. *Arter. Thromb. Vasc. Biol.*, 25(11):2386–2391, 2005.

[65] J. Herrmann, L.O. Lerman, D. Mukhopadhyay, C. Napoli, and A. Lerman. Angiogenesis in atherogenesis. *Arter. Thromb. Vasc. Biol.*, 26(9):1948–1957, 2006.

[66] J.C. Sluimer and M.J. Daemen. Novel concepts in atherogenesis: angiogenesis and hypoxia in atherosclerosis. *J. Pathol.*, 218(1):7–29, 2009.

[67] G. Pasterkamp, A.H. Schoneveld, D.J. Hijnen, D.P.V. de Kleijn, H. Teepen, A.C. van der Wal, and C. Borst. Atherosclerotic arterial remodeling and the localization of macrophages and matrix metalloproteases 1, 2 and 9 in the human coronary artery. *Atherosclerosis*, 150(2): 245–253, 2000.

[68] F.L. Celletti, J.M. Waugh, P.G. Amabile, A. Brendolan, P.R. Hilfiker, and M.D. Dake. Vascular endothelial growth factor enhances atherosclerotic plaque progression. *Nat. Med.*, 7:425, 2001.

[69] C. Ihling, T. Szombathy, B. Bohrmann, M. Brockhaus, H.E. Schaefer, and B.M. Loeffler. Coexpression of endothelin-converting enzyme-1 and endothelin-1 in different stages of human atherosclerosis. *Circulation*, 104(8):864–869, 2001.

[70] J.C. Sluimer, J.-M. Gasc, J.L. van Wanroij, N. Kisters, M. Groeneweg, M.D. Sollewijn Gelpke, J.P. Cleutjens, L.H. van den Akker, P. Corvol, B.G. Wouters, M.J. Daemen, and A.-P.J. Bijnens. Hypoxia, Hypoxia-Inducible Transcription Factor, and Macrophages in Human Atherosclerotic Plaques Are Correlated With Intraplaque Angiogenesis. *J. Am. Coll. Cardiol.*, 51(13):1258–1265, 2008.

[71] A. Luttun, M. Tjwa, L. Moons, Y. Wu, A. Angelillo-Scherrer, F. Liao, J.A. Nagy, A. Hooper, J. Priller, B. De Klerck, V. Compernolle, E. Daci, P. Bohlen, M. Dewerchin, J.-M. Herbert, R. Fava, P. Matthys, G. Carmeliet, D. Collen, H.F. Dvorak, D.J. Hicklin, and P. Carmeliet. Revascularization of ischemic tissues by PlGF treatment, and inhibition of tumor angiogenesis, arthritis and atherosclerosis by anti-Flt1. *Nat. Med.*, 8:831, 2002.

[72] T. Nakano, Y. Nakashima, Y. Yonemitsu, S. Sumiyoshi, Y.-X. Chen, Y. Akishima, T. Ishii, M. Iida, and K. Sueishi. Angiogenesis and lymphangiogenesis and expression of lymphangiogenic factors in the atherosclerotic intima of human coronary arteries. *Hum. Pathol.*, 36(4): 330–340, 2005.

[73] A. Doddaballapur, K.M. Michalik, Y. Manavski, T. Lucas, R.H. Houtkooper, X. You, W. Chen, A.M. Zeiher, M. Potente, S. Dimmeler, and R.A. Boon. Laminar shear stress inhibits endothelial cell metabolism via klf2-mediated repression of pfkfb3. *Arter. Thromb. Vasc. Biol.*, 35(1):137–145, 2015.

[74] A. Tawakol, P. Singh, M. Mojena, M. Pimentel-Santillana, H. Emami, M. MacNabb, J.H.F. Rudd, J. Narula, J.A. Enriquez, P.G. Través, M. Fernández-Velasco, R. Bartrons, P. Martín-Sanz, Z.A. Fayad, A. Tejedor, and L. Boscá. Hif-1α and pfkfb3 mediate a tight relationship between proinflammatory activation and anerobic metabolism in atherosclerotic macrophages. *Arter. Thromb. Vasc. Biol.*, 35(6):1463–1471, 2015.

[75] S. Bekkering, I. van den Munckhof, T. Nielen, E. Lamfers, C. Dinarello, J. Rutten, J. de Graaf, L.A.B. Joosten, M.G. Netea, M.E.R. Gomes, and N.P. Riksen. Innate immune cell activation and epigenetic remodeling in symptomatic and asymptomatic atherosclerosis in humans in vivo. *Atherosclerosis*, 254:228–236, 2016.

[76] K.I. Baek, R. Li, N. Jen, H. Choi, A. Kaboodrangi, P. Ping, D. Liem, T. Beebe, and T.K. Hsiai. Flow-Responsive Vascular Endothelial Growth Factor Receptor-Protein Kinase C Isoform Epsilon Signaling Mediates Glycolytic Metabolites for Vascular Repair. *Antioxid. Redox Signal.*, 28(1):31–43, 2017.

[77] R. Bhattacharya, S. SenBanerjee, Z. Lin, S. Mir, A. Hamik, P. Wang, P. Mukherjee, D. Mukhopadhyay, and M.K. Jain. Inhibition of Vascular Permeability Factor/Vascular Endothelial Growth Factor-mediated Angiogenesis by the Kruppel-like Factor KLF2. *Journal of Biological Chemistry*, 280(32):28848–28851, 2005.

[78] D. Kawanami, G.H. Mahabeleshwar, Z. Lin, G.B. Atkins, A. Hamik, S.M. Haldar, K. Maemura, J.C. LaManna, and M.K. Jain. Kruppel-like Factor 2 Inhibits Hypoxia-inducible Factor 1$\alpha$ Expression and Function in the Endothelium. *Journal of Biological Chemistry*, 284(31):20522–20530, 2009.

# Part II:

# Designing Therapeutic Modulators Targeting PFKFB3 for Atherosclerosis Treatment

# CHAPTER 4

# PFKFB3: Previous studies on structures and drug designs

## 4.1 Knowledge from crystal structures

PFKFBs are homodimeric and bifunctional; each monomer contains a kinase (PFK-2) and a bisphosphatase (FBPase) active sites(Figure 4.1). Functionally, each polypeptide chain of PFKFB can be divided into two halves. The N terminal half is responsible for the kinase reaction, catalysing F6P to F2,6BP. The sequence close to the C-terminus, on the other hand, is forming the enzymatic active half responsible for dephosphorylation of F2,6BP. Structural properties of PFK-2 is more connected with the mononucleotide-binding protein family, in which *ras* and adenylate kinase also belong, whereas the FBPase domain is similar to the phosphoglycerate mutase family.[1]

In 2006, Kim and co-workers resolved the first crystal of PFKFB3 with enzymatic products incorporated within the active sites.[2] A $\beta$-hairpin structure at the N-terminus in proximity to the FBPase site, differing from other forms of PFKFBs,

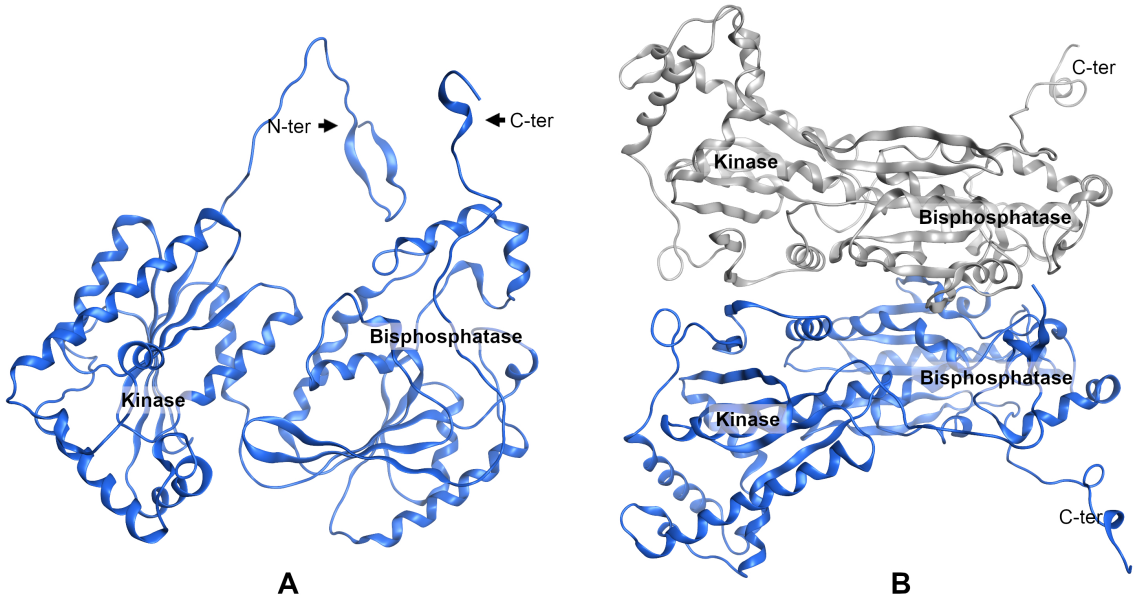



**Figure 4.1:** The structure of PFKFB3 as a monomer (A) and a dimer (B). The kinase and bisphosphatase halves are as indicated. The termini are also marked.

was observed the first time.[2] The interactions from the N-terminus could cause an increase in affinities for both substrate (F2,6BP) and product (F6P). In addition, an alteration of an arginine in the FBPase active site of the liver isozyme to serine in PFKFB3 further promote the binding preferring the F6P, hence, impairing product releasing.[2] Moreover, N-terminal splicing forms of liver PFKFB can achieve near 10-fold increase of bisphosphatase/kinase activity ratio.[3] This suggests that perturbing the N terminal $\beta$-hairpin can be a viable way for phosphatase activation, indirectly lowering PFKFB3 kinase activity.

Three more crystal models were generated later to further investigate the enzymatic mechanism of bisphosphatase.[4] Though lacking the arginine at position 302 comparing to the liver PFKFB, the PFKFB3 is still able to partially transfer the phosphate group from F2,6BP.[4] Nonetheless, the conserved arginine among other isozymes provides higher bisphosphatase activity than in PFKFB3,[5–7] supporting the observation of the exceptionally high kinase-to-bisphosphatase activity ratio in the PFKFB3. However, potential interference from C-terminus has also been suggested as it has a unique conformation in PFKFB3.[4]

On the other hand, differences in amino acid sequence in the kinase site contribute to a conformational alteration compared to the other isozymes, providing improved substrate affinities.[2] In a later study, Kim et al. suggested an associative enzymatic mechanism for the PFKFB3 kinase, assisted by Lys168.[8] This was based on the observation that the two kinase substrates, F6P and ATP, have direct interactions through multiple hydrogen bonds.[8]

## 4.2 Drug design targeting PFKFB3

Due to the close connection of PFKFB3 to cancer cell metabolism, multiple studies have been devoted to kinase inhibitors targeting PFKFB3. One of the most historical but controversial study is the discovery of 3-(3-pyridinyl)-1-(4-pyridinyl)-2-propen-1-one (3PO, compound **1** in Figure 4.2).[9] The virtual screening method was applied using a PFKFB3 homology model based on the rat testes PFKFB4. 3PO was tested both in vitro and in vivo. later, a novel 3PO derivative was also developed, namely PFK15 (compound **2**), with improved pharmacokinetic properties.[10] However, a test of 3PO on cellular lactate production showed no difference to control, i.e. 3PO failed to inhibit glycolysis in the cell line tested within the non-cytotoxic concentration.[11] PFK15 also showed $IC_{50} > 1000$ $\mu$M when tested in vitro on PFKFB3, inconsistent to the previous studies. Boyd et al. also challenged the validity of 3PO to effective PFKFB3 inhibition.[12] Using a kinase kit different from the original testing method by Clem et al., they demonstrated that 3PO was inactive with a $IC_{50} > 100$ $\mu$M.[12] A more recent study also failed to determine the $IC_{50}$ for 3PO.[13] Moreover, no crystal structures are yet resolved for 3PO or PFK15, further mystifying the mechanism of 3PO activities discovered along its development and validation.

Within the current MOGLYNET programme in which the author is a member, we also determined that 3PO is somehow a non-effective compound in PFKFB3 kinase inhibition. Mixed effects on different cell lines have also been observed. However, these assessments are still under investigation and are also out of the scope of this

thesis. Hence, the detailed information is not elucidated herein.

Other classes of compounds other than 3PO have also been determined using PFKFB3 as the therapeutic target. Targeting the F6P kinase pocket, Seo et al. devised a workflow combining virtual screening, experimental validation, and structural similarity search for developing novel PFKFB3 inhibitors. N4A, YN1 and YZ9 (compound **3-5** in Figure 4.2) from ZINC database were determined inhibiting PFKFB3 with low $\mu$M $IC_{50}$ with complex crystal structures resolved for N4A and YN1.[14] In another study, a series of triazolophenylpyridazinone compounds were synthesised and tested on PFKFB3.[11] Compound **6** (in Figure 4.2) was first identified with 12 $\mu$M $IC_{50}$ using a library of 87,500 compounds in HTS.[11] Multiple derivatives were also synthesised and the best compound (compound **6**) achieved a $IC_{50}$ of 6.1 $\mu$M.[11] The high po-

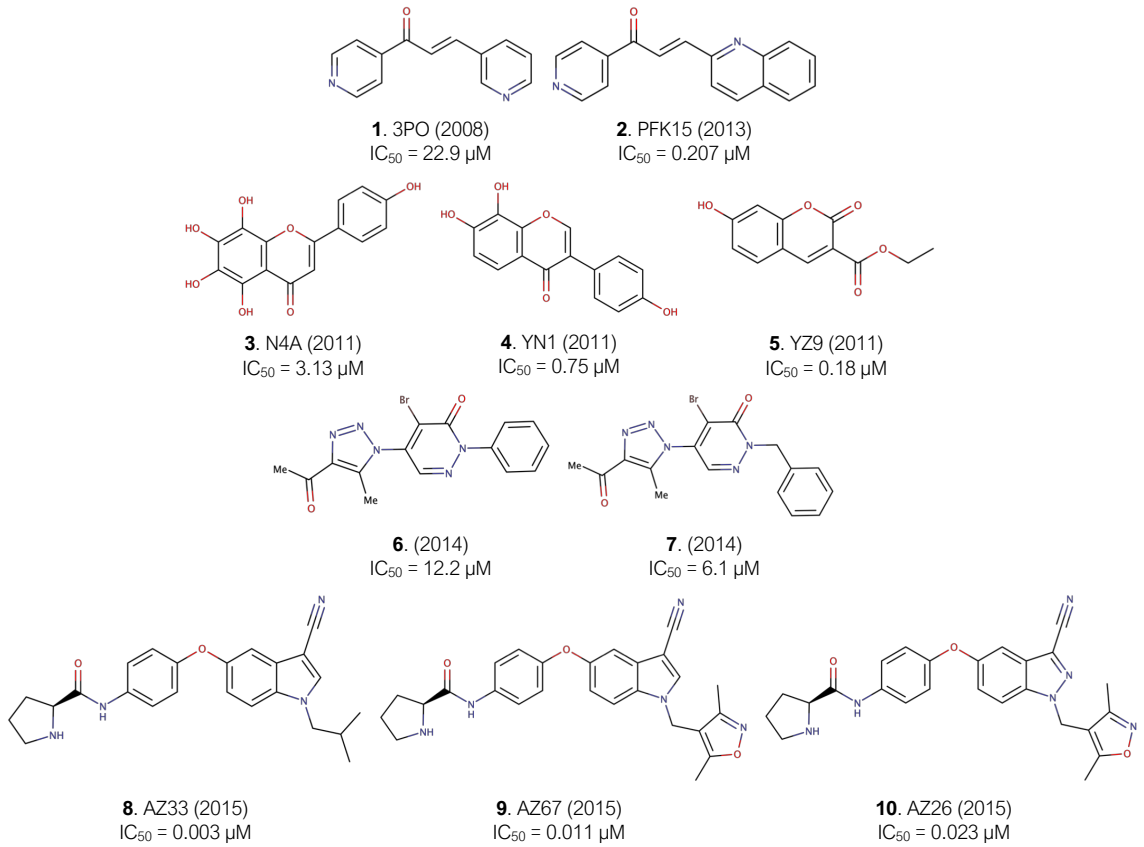
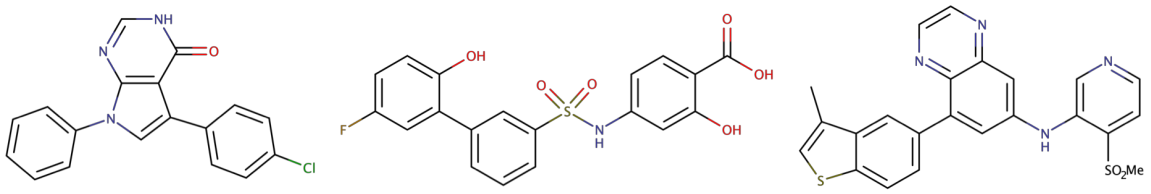



**Figure 4.2:** The PFKFB3 inhibitors published before the commence of MOGLYNET programme. $IC_{50}$ for kinase activity are as indicated.




**11**. (2017)
$IC_{50}$ = 0.371 μM

**12**. KAN0438241 (2018)
$IC_{50}$ = 0.019 μM

**13**. (2018)
$IC_{50}$ = 0.014μM

**Figure 4.3:** The PFKFB3 inhibitors published after the commence of MOGLYNET programme (2016-2018). $IC_{50}$ for kinase activity are as indicated.

tential of targeting PFKFB3 as a therapeutic strategy for cancer has also drawn the attention of pharmaceutical industries. In 2015, a group of scientist from Oncology Innovative Medicines Unit, AstraZeneca in collaboration with other institutes published multiple indole and indazole derivatives with up to low nM PFKFB3 kinase $IC_{50}$.[12] The compounds (examples include **8-10** in Figure 4.2) show some selectivities against PFKFB1 and PFKFB2, and down-regulation of F1,6BP and inhibition of cellular lactate production were also demonstrated.[12]

During the recent 3 years during which this work has been carrying out, more PFKFB3 kinase inhibitors have been discovered and developed. As a separate series from those discovered by Boyd et al. in 2015, 6 more PFKFB3 inhibitors (compound **11** as an example, Figure 4.3) were published with the HTS workflow applied.[15] These compounds, less potent than the previous series, also shown worse selectivity among the PBKFBs tested.[15] Using also HTS, Gustafsson et al. discovered KAN0438241 (compound **12**) and its ester derivative KAN0438757 as effective PFKFB3 with low $\mu$M $IC_{50}$ and cellular responses.[13] Compound **13** and others were developed by Boutard et al. with nM PFKFB3 inhibition both on the recombinant enzyme and in cells.[16] All the studies above provided evidence of binding using the crystallography techniques and structures have been deposited in the Protein Data Bank (PDB bank).[17]

In this thesis, two strategies have been taken to lower up-regulated PFKFB3 activity under pathological conditions. In one attempt, we attempted activating bisphosphatase by perturbing the interaction from the auto-regulatory domain. By targeting

at the auto-regulatory domain binding site directly and an alternate allosteric site discovered in this study, libraries of compounds were virtually screened and then tested in experimental settings. In the second approach, the kinase inhibition was also carried out. We targeted at the bigger F6P and ATP pocket and selected compounds with the preference towards F6P binding site. The compounds were screened for interference in kinase activity and preliminarily tested in cell viabilities. The studies can provide some further insights into modulator designs for PFKFB3.

## Chapter references

[1] M. H. Rider, L. Bertrand, D. Vertommen, P. A. Michels, G. G. Rousseau, and L. Hue. 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase: head-to-head with a bifunctional enzyme that controls glycolysis. *Biochem. J*, 381(3):561–579, 2004.

[2] Song-Gun Kim, Nathan P Manes, M Raafat El-Maghrabi, and Yong-Hwan Lee. Crystal Structure of the Hypoxia-inducible Form of 6-Phosphofructo-2-kinase/fructose-2,6-bisphosphatase (PFKFB3): A POSSIBLE NEW TARGET FOR CANCER THERAPY. *J. Biol. Chem.*, 281 (5):2939–2944, 2006.

[3] I.J. Kurland, B. Chapman, and M.R. El-maghrabi. N- and C-termini modulate the effects of pH and phosphorylation on hepatic 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase. *Biochem. J.*, 347(2):459–467, 2000.

[4] M.C. Cavalier, S.-G. Kim, D. Neau, and Y.-H. Lee. Molecular basis of the fructose-2,6-bisphosphatase reaction of PFKFB3: Transition state and the C-terminal function. *Proteins*, 80(4):1143–1153, 2011.

[5] C.A. Hasemann, E.S. Istvan, K. Uyeda, and J. Deisenhofer. The crystal structure of the bifunctional enzyme 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase reveals distinct domain homologies. *Structure*, 4(9):1017–1029, 1996.

[6] Y.-H. Lee, Y. Li, K. Uyeda, and C.A. Hasemann. Tissue-specific Structure/Function Differentiation of the Liver Isoform of 6-Phosphofructo-2-kinase/Fructose-2,6-bisphosphatase. *J. Biol. Chem.*, 278(1):523–530, 2003.

[7] N.P. Manes and M.R. El-Maghrabi. The kinase activity of human brain 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase is regulated via inhibition by phosphoenolpyruvate. *Arch. Biochem. Biophys.*, 438(2):125–136, 2005.

[8] S.-G. Kim, M. Cavalier, M.R. El-Maghrabi, and Y.-H. Lee. A Direct Substrate–Substrate Interaction Found in the Kinase Domain of the Bifunctional Enzyme, 6-Phosphofructo-2-kinase/Fructose-2,6-bisphosphatase. *J. Mol. Biol.*, 370(1):14–26, 2007.

[9] B. Clem, S. Telang, A. Clem, A. Yalcin, J. Meier, A. Simmons, M.A. Rasku, S. Arumugam, W.L. Dean, J. Eaton, A. Lane, J.O. Trent, and J. Chesney. Small-molecule inhibition of 6-phosphofructo-2-kinase activity suppresses glycolytic flux and tumor growth. *Mol. Canc. Therapeut.*, 7(1):110–120, 2008.

[10] Brian F Clem, Julie O'Neal, Gilles Tapolsky, Amy L Clem, Yoannis Imbert-Fernandez, Daniel A Kerr, Alden C Klarer, Rebecca Redman, Donald M Miller, John O Trent, Sucheta

Telang, and Jason Chesney. Targeting 6-Phosphofructo-2-Kinase (PFKFB3) as a Therapeutic Strategy against Cancer. *Mol. Cancer Ther.*, 12(8):1461–1470, aug 2013.

[11] D.G. Brooke, E.M. van Dam, C.K.W. Watts, A. Khoury, M.A. Dziadek, H. Brooks, L.-J.K. Graham, J.U. Flanagan, and W.A. Denny. Targeting the Warburg Effect in cancer; relationships for 2-arylpyridazinones as inhibitors of the key glycolytic enzyme 6-phosphofructo-2-kinase/2,6-bisphosphatase 3 (PFKFB3). *Bioorganic Med. Chem*, 22(3):1029–1039, 2014.

[12] S. Boyd, J.L. Brookfield, S.E. Critchlow, I.A. Cumming, N.J. Curtis, J. Debreczeni, S.L. Degorce, C. Donald, N.J. Evans, S. Groombridge, P. Hopcroft, N.P. Jones, J.G. Kettle, S. Lamont, H.J. Lewis, P. MacFaull, S.B. McLoughlin, L.J.M. Rigoreau, J.M. Smith, S. St-Gallay, J.K. Stock, A.P. Turnbull, E.R. Wheatley, J. Winter, and J. Wingfield. Structure-Based Design of Potent and Selective Inhibitors of the Metabolic Kinase PFKFB3. *J. Med. Chem.*, 58(8): 3611–3625, 2015.

[13] N.M.S. Gustafsson, K. Färnegårdh, N. Bonagas, A.H. Ninou, P. Groth, E. Wiita, M. Jönsson, K. Hallberg, J. Lehto, R. Pennisi, J. Martinsson, C. Norström, J. Hollers, J. Schultz, M. Andersson, N. Markova, P. Marttila, B. Kim, M. Norin, T. Olin, and T. Helleday. Targeting PFKFB3 radiosensitizes cancer cells and suppresses homologous recombination. *Nat. Commun.*, 9(1): 3872, 2018.

[14] M. Seo, J.-D. Kim, D. Neau, I. Sehgal, and Y.-H. Lee. Structure-Based Development of Small Molecule PFKFB3 Inhibitors: A Framework for Potential Cancer Therapeutic Agents Targeting the Warburg Effect. *PLoS One*, 6(9):e24179, 2011.

[15] S.A. St-Gallay, N. Bennett, S.E. Critchlow, N. Curtis, G. Davies, J. Debreczeni, N. Evans, I. Hardern, G. Holdgate, N.P. Jones, L. Leach, S. Maman, S. McLoughlin, M. Preston, L. Rigoreau, A. Thomas, A.P. Turnbull, G. Walker, J. Walsh, R.A. Ward, E. Wheatley, and J. Winter-Holt. A High-Throughput Screening Triage Workflow to Authenticate a Novel Series of PFKFB3 Inhibitors. *SLAS Discov.*, 23(1):11–22, 2017.

[16] N. Boutard, A. Białas, A. Sabiniarz, P. Guzik, K. Banaszak, A. Biela, M. Bień, A. Buda, B. Bugaj, E. Cieluch, A. Cierpich, Ł. Dudek, H.-M. Eggenweiler, J. Fogt, M. Gaik, A. Gondela, K. Jakubiec, M. Jurzak, A. Kitlińska, P. Kowalczyk, M. Kujawa, K. Kwiecińska, M. Leś, R. Lindemann, M. Maciuszek, M. Mikulski, P. Niedziejko, A. Obara, H. Pawlik, T. Rzymski, M. Sieprawska-Lupa, M. Sowińska, J. Szeremeta-Spisak, A. Stachowicz, M.M. Tomczyk, K. Wiklik, Ł. Włoszczak, S. Ziemiańska, A. Zarebski, K. Brzózka, M. Nowak, and C.-H. Fabritius. Discovery and Structure–Activity Relationships of N-Aryl 6-Aminoquinoxalines as Potent PFKFB3 Kinase Inhibitors. *ChemMedChem*, (Early View), 2018.

[17] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The Protein Data Bank. *Nucleic Acids Res.*, 28(1):235–242, 2000.

# CHAPTER 5

# PFKFB3 starting structures

## 5.1 Choice of crystal structure

A good starting structure of the targeted protein is crucial for *in silico* molecular studies. Using a monomeric structure of PFKFB3 can be problematic. On single monomer, the auto-regulatory domain is only interacting with the C-terminal half of PFKFB3 through the loop of the hairpin structure. It seems an odd scenario that the non-interacting part of the domain is actually forming a stable $\beta$-structure (Figure 5.1A). The dimeric structure revealed that the stable secondary structure is probably contributed by the other monomer providing an interacting surface for stabilising the $\beta$-hairpin structure (Figure 5.1B). Hence, to target the binding site of the auto-regulatory domain requires a relatively intact structure of the homo-dimeric PFKFB3. This intact structure is also important for a correct performance in MD simulations. A relatively large interacting surface between the two monomers suggests a very stable dimeric structure. A large continuous $\beta$-sheet connecting the

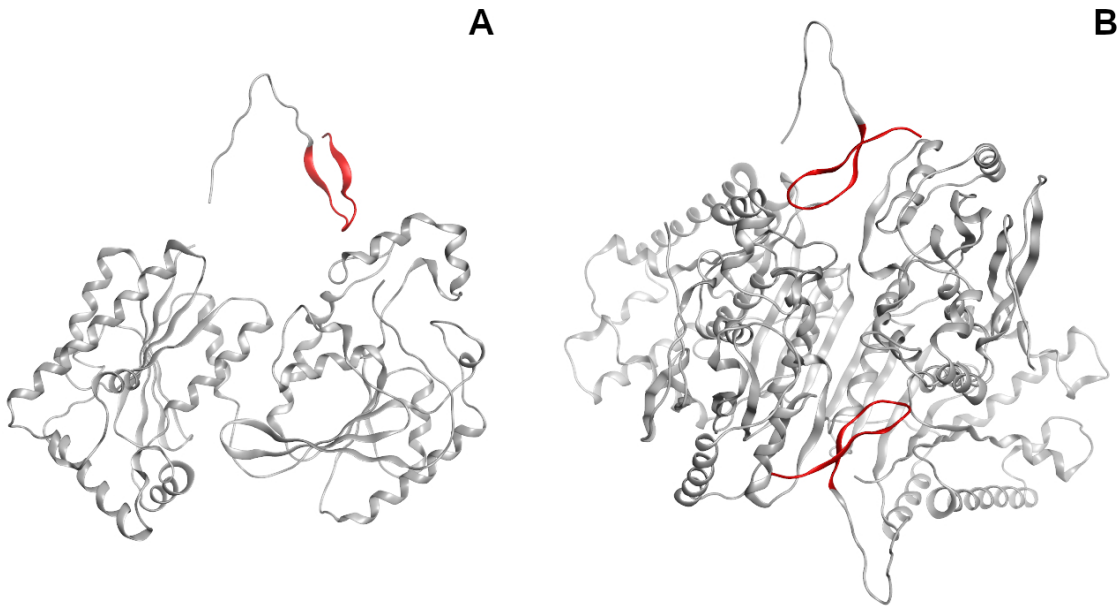



**Figure 5.1:** The position of the $\beta$-hairpin on PFKFB3 in the monomer and the dimer. The hairpin is highlighted as red. In the dimeric form, the $\beta$-hairpin is located at the interface of the two monomers, likely stabilising the dimeric structure of PFKFB3. Structure with the PDB ID 2i1v[1] is displayed.

**Table 5.1:** The 18 crystal structures of PFKFB3 available in the PDB bank. The selected crystal structure for PFKFB3 for this work is highlighted in green. The ligand ID for natural substrates / products: ATP (ACP is with one of the connecting oxygens between phosphates replaced by a carbon atom), ADP, FDP (fructose-2,6-bisphosphate), and F6P (fructose-6-phosphate).

| PDB ID | Resolution (Å) | R-value Free | R-value Work | Ligand Info.(ID) | | Length (monomer) | Biological Assembly (Source) |
|---|---|---|---|---|---|---|---|
| | | | | Kinase | Phsphotase | | |
| 2axn | 2.10 | 0.233 | 0.209 | ADP, EDT | F6P | 451/520 | Monomer (Author) |
| 2dwo | 2.25 | 0.243 | 0.214 | ADP, PEP | F6P | 456/520 | Dimer (Author, PISA) |
| 2dwp | 2.70 | 0.264 | 0.224 | ACP, F6P | F6P | 431/520 | Dimer (Author) |
| 2i1v | 2.50 | 0.262 | 0.214 | ADP, FDP | F6P, PHS | 449/520 | Dimer (Author, PISA) |
| 3qpu | 2.30 | 0.236 | 0.184 | EDO, 2×POP | POP, SRT | 439/520 | Dimer (Author) |
| 3qpv | 2.50 | 0.248 | 0.194 | ADP, FDP | F6P, PHS | 440/520 | Dimer (Author) |
| 3qpw | 2.50 | 0.248 | 0.194 | ADP, PEP | ALF | 431/520 | Dimer (Author, PISA) |
| 4d4j | 3.00 | 0.207 | 0.190 | BKI, PO4 | F6P, PHS | 442/449 | Dimer (Author, PISA) |
| 4d4k | 3.24 | 0.210 | 0.199 | BKE, PO4 | F6P, PHS | 439/449 | Dimer (Author, PISA) |
| 4d4l | 3.16 | 0.219 | 0.207 | BKS, PO4 | F6P, PHS | 442/449 | Dimer (Author, PISA) |
| 4d4m | 2.32 | 0.209 | 0.191 | BKV, 2×PO4 | F6P, PHS | 439/449 | Dimer (Author, PISA) |
| 4ma4 | 2.23 | 0.242 | 0.199 | ADP, MLA | F6P | 444/520 | Dimer (Author, PISA) |
| 5ajv | 3.01 | 0.225 | 0.182 | 8R2, PO4 | FDP | 435/520 | Dimer (Author, PISA) |
| 5ajw | 2.50 | 0.223 | 0.199 | S6L, PO4 | F6P, PHS | 439/520 | Dimer (Author, PISA) |
| 5ajx | 2.58 | 0.206 | 0.198 | FD9, PO4 | F6P, PHS | 441/520 | Dimer (Author, PISA) |
| 5ajy | 2.37 | 0.212 | 0.198 | 87T, PO4 | F6P, PHS | 439/520 | Dimer (Author, PISA) |
| 5ajz | 2.35 | 0.220 | 0.183 | MJP. PO4 | F6P, PHS | 440/520 | Dimer (Author, PISA) |
| 5ak0 | 2.03 | 0.204 | 0.179 | 8V1, PO4 | F6P, PHS | 442/520 | Dimer (Author, PISA) |

two kinase half of PFKFB3 further confirm that the dimeric form is more closely to the natural structure under the actual physiological conditions. Therefore, in this

thesis, a dimeric PFKFB3 structure is essential for accurate *in silico* modelling.

Multiple crystal structures in the dimeric form are available in PDB bank[2] (Table 5.1). The structures were all obtained through X-ray diffraction method. However, all of the structures have certain parts of the protein missing, mainly around both termini. Some further processes are required for generating the intact structure. To select the most suitable crystal structure for later studies, multiple criteria were taken into account. Firstly, the resolution of the crystal structure should be acceptable, as a poor resolution suggests an increased likelihood of wrong directions of amino acid side chains. Therefore, the structures with a resolution higher than 3.0 Å were excluded. Secondly, we consider the reproduction of the PFKFB3 structure under the natural condition as the priority for later computational investigations. Therefore, this implies that the crystal structure should be either in the *apo* state or, with the natural substrates or products bound at the active sites. Hence, this further narrowed the choices to two crystal structures with the PDB ID 2i1v[1] and 3qpv[3] respectively.

Furthermore, ensemble docking is excluded as a viable solution for the VS studies in this thesis. Perturbing the auto-regulatory domain in terms of modulating PFKFB3 kinase activity was never been attempted before. Hence, very little structure information is available other than the existing crystal structures with the same $\beta$-hairpin domain bound. For the kinase inhibition study discussed in a later chapter (Chapter 7), multiple crystal structures with inhibitors binding at the ATP binding site are available[? ?] . However, all available non-natural crystal ligands bound in F6P binding site are much smaller in size than F2,6P. Thus, it is also unreasonable to use starting structures with much narrower binding space than the natural product, which will likely hamper the performance of VS investigations. However, the possibility should not be excluded in a later study if more structural information becomes available.

Moreover, the flexible parts of the protein should be at a minimum, i.e. the resolved length of the protein should be as long as possible. Therefore, the structure with

the PDB ID 2i1v[1] stands out among the list of candidates. It contains the natural products of the enzyme in both of the binding sites and an especially extended loop region adjacent to the auto-regulatory domain. The numbers of missing residues are minimised to 5 on the N-terminus. Additionally, it also includes part of the C-terminus up to 8 resolved residues, which is missing in most of the other structures in the selection list. Therefore, by using this structure, it will greatly simplify our later modelling of the missing parts, which only requires a direct and standardised loop building process.

## 5.2 Loop building and structural preparation

The MOE software package[4] was applied for the structure preparations. All crystal water and the phosphoric acid are deleted from the structure. The missing loops (P28-N32 and S445-N453) of the crystal structure were constructed through a standard loop building procedure included in the "Structure Preparation" functionalities of MOE.[4] The absence of the longer sequence on the C-terminus was not included in the modelling. Both N- and C-termini were capped to prevent terminal artefacts. All

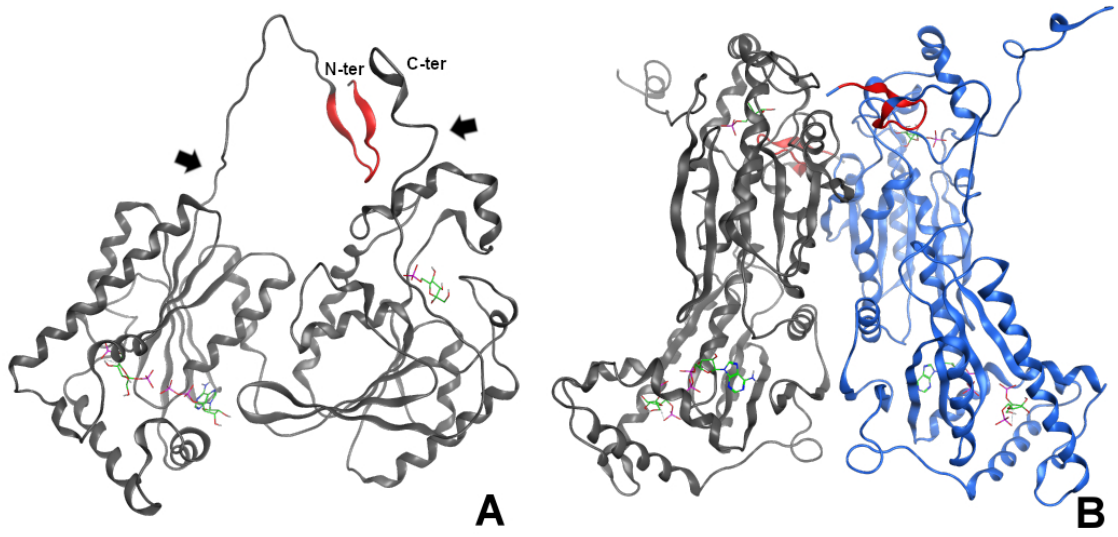



**Figure 5.2:** The processed structure for PFKFB3. The auto-regulatory domains are shown in red. A: monomer, B: homo-dimer. The arrows indicated the loops built using MOE.[4]

missing hydrogens are added using Protonate 3D method provided in MOE. Charges are assigned using Amber10:EHT force field. The resulted structure is illustrated in Figure 5.2.

For VS studies on phosphatase modulators, the generated structure with the natural products within the active sites was used. The auto-regulatory domain was modified according to the strategies specified in the later section. The model for kinase inhibitor design had the ligands within the kinase active sites removed, leaving the binding sites vacant for docking experiments. The detailed preparation of starting structures for MD simulation are elucidated in the later chapter as more conditions of PFKFB3 were considered and more specific steps were required.

## Chapter references

[1] S.-G. Kim, M. Cavalier, M.R. El-Maghrabi, and Y.-H. Lee. A Direct Substrate–Substrate Interaction Found in the Kinase Domain of the Bifunctional Enzyme, 6-Phosphofructo-2-kinase/Fructose-2,6-bisphosphatase. *J. Mol. Biol.*, 370(1):14–26, 2007.

[2] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

[3] M.C. Cavalier, S.-G. Kim, D. Neau, and Y.-H. Lee. Molecular basis of the fructose-2,6-bisphosphatase reaction of PFKFB3: Transition state and the C-terminal function. *Proteins*, 80(4):1143–1153, 2011.

[4] Chemical Computing Group ULC. Moe, molecular operating environment, 2018. 1010 Sherbooke St. West, Suite 910, Montreal, QC, Canada, H3A 2R7.

# CHAPTER 6

# Design of PFKFB3 phosphatase modulator

## 6.1 Screening strategies and libraries

Limited information is available regarding strategies for activating bisphosphatase activity. To increase the success rate of the VS, two strategies for PFKFB3 bisphosphatase activity modulation were devised. One strategy is by targeting directly at the binding site of the $\beta$-hairpin. Majority of the crystal structures available all have the $\beta$-hairpin in-place, occupying the binding site to be targeted. Hence, to fully expose the binding site to allow protein-restrained docking or VS, the $\beta$-hairpin and the additional connecting loop were stripped (Leu2-Pro28), and a representation of the binding site surface is shown in Figure 6.1A. This structure was applied for virtual screening (the *apo* PFKFB3 hereafter).

Some initial concerns regarding the above VS strategy were noted during structural inspections. Firstly, this binding site has a relatively large and flat interaction surface. The lacking in concavity is likely due to the structure of $\beta$-hairpin that is relatively

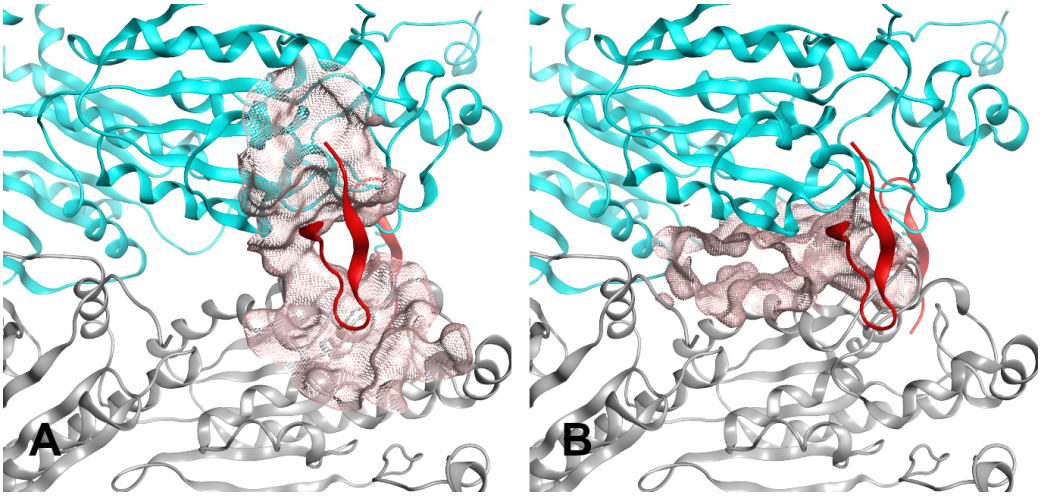



**Figure 6.1:** The molecular surfaces of the targeted binding sites for PFKFB3 bisphosphatase modulator screening in this thesis. The surfaces are shown in pink. Each monomer is shown in different colour (grey and blue). The auto-regulatory domains are shown in red. A: targeting the auto-regulatory domain binding surface. B: targeting the secondary binding site close to the auto-regulatory domain.

sheet-like. This site bears some characteristics of the interface for protein-protein interaction and can be a site difficult to tackle. Secondly, the targeted site, in this case, is pre-occupied with the auto-regulatory domain under the physiological conditions. The actual process of ligand binding would require the displacement of the $\beta$-hairpin. However, no definite information is yet available that the auto-regulatory domain is in an "on-off" dynamic, i.e. it is not yet determined the $\beta$-hairpin is in a relatively loose bound-unbound equilibrium in solution. Hence, it is also dangerous to target this site solely as it is already implying a high off-target rate. Moreover, even if the $\beta$-hairpin dynamic equilibrium exists, the site would suffer from a high solvent exposure that would likely disrupt the hydrophobic interactions for ligand binding. Therefore, it is rational to use a secondary allosteric site with a different modulating mechanism targeting PFKFB3 FBPase activity.

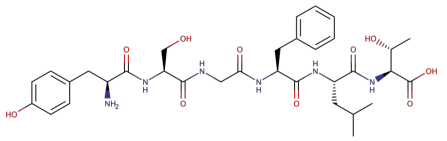
During some initial VS tests using a larger docking sphere, we noticed that most of the high scored ligands moved into a secondary pocket close to the intended $\beta$-hairpin binding site (Figure 6.1B). This site is right at the interface of the two PFKFB3 monomers. Further analysis and visual inspection suggested this secondary binding site have the potential to stabilise the binding of ligands in close vicinity to the

$\beta$-hairpin region. It provides a channel-like space to accommodate ligands and is decorated by multiple negatively charged and polar amino acid residues (such as Lys318, Glu360, Asp364, Gln367, and Arg368). Furthermore, the insertion of a ligand at this side can potentially lead to binding destabilisation of the -hairpin domain and even further delocalisation. It will likely further merit the FBPtase activation if the ligand is also capable of binding at the exposed $\beta$-hairpin binding site due to the delocalisation. Additionally, an intact PFKFB3 structure will be used for VS for targeting at this secondary side. No parts of the protein would be virtually spliced. Hence, less biased factors will be introduced and the complete form of PFKFB3 was adapted for this strategy (the *holo* PFKFB3 hereafter).

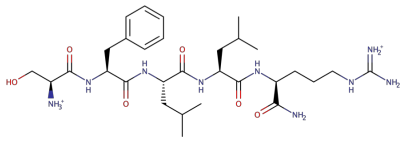
Two screening libraries were derived from ZINC database[1] with the sizes of 644 and 4819 compounds, respectively. The smaller library was manually collected from the active compounds with ADMET information on certain targets. The second libraries are selected from the subsets shown functional activities less than 10 $\mu$M. Both libraries have excluded kinase targets as the ligands can potentially interact with PFKFB3 kinase. This may mask the experimental testing targeting the bisphosphatase half. Repetitions were removed by screen out the compounds with the same ZINC ID and the 3D structures were generated and minimised using MOE.[2]

## 6.2 Virtual screening

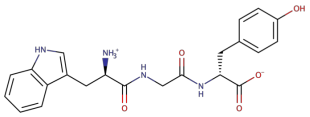
Each of the derived ZINC libraries was screened using both the *apo* and *holo* structures of PFKFB3 focusing on different docking centres. The radius for docking was set as 15 Å for both, chosen based on the longest inter-atomic distance found in the screening libraries. Different ring conformations were generated and evaluated by SPORES included in the PLANTS docking package.[3,4] SDWASH[2] was applied to determine reasonable tautomerism states of the ligand and stereoisomers were evaluated using SDSTEREO.[2] The structures were minimised briefly using the db_Minimize function within the MOE package.[2] The docking of the generated conformers was performed



**14.** $K_d$ = 18 ± 6 μM



**15.** $K_d$ = 44 ± 14 μM



**16.** $K_d$ = 3 ± 1 μM

Enzymatic kinetic measurements fitted to Hill equation:

$$V = \frac{V_{max} \times [S]^h}{K_{half}{}^h + [S]^h}$$

The approximate $K_d$ of substrate (F6P) is calculated as:

$$Approx.K_d = Mean(K_{half})^{Mean(h)}.$$

| | $V_{max}$ ± S.E. | $K_{half}$ ± S.E. | $h$ ± S.E. | Approx. substrate $K_d$ |
|---|---|---|---|---|
| **blank** | 0.16 ± 0.03 | 21.51 ± 5.57 | 2.03 ± 0.66 | 507.29 |
| **14** | 0.18 ± 0.04 | 26.54 ± 5.71 | 2.39 ± 0.68 | 2530.01 |
| **15** | 0.81 ± 0.29 | 58.42 ± 25.88 | 1.50 ± 0.21 | 446.52 |
| **16** | 0.91 ± 0.43 | 71.72 ± 44.97 | 1.30 ± 0.20 | 258.42 |

**Figure 6.2:** Experimental measurements and fitting results for the discovered hits. Data provide by Macut[5] and Regazzoni.[6]

using PLANTS with default settings and search speed set to 1.[3,4] The ligands were ranked according to the score using $PLANTS_{chemplp}$ scoring method.[4]

The ligand-receptor interaction was then used to select the list of compounds to be tested experimentally. Score cut-offs were applied to select the better-ranked ligands for further interaction analysis. Due to differences in properties of the two targeted sites, the score cut-offs were set differently for the two strategies adapted; the total $PLANTS_{chemplp}$ cut-off of -110 for the screening using the *holo* receptor and -100 for the *apo* receptor. The selected compounds from each library underwent a further selection step according to their interactions to the receptor. This interaction analyses were performed and visualised using *ligand interactions* analysis included within the MOE package.[2]

Separate lists of testing compounds were generated for each library per receptor structure, and the selected compounds were firstly screened with Nanotemper Microscale Thermophoresis (MST) and the dissociation constants ($K_d$) were then measured with a gradient of concentrations. 3 compounds belong to the peptide category were determined with sub-$\mu$M $K_d$ (compound **14-16** in Figure 6.2).[5] The enzymatic kinetics of PFKFB3 were measured using liquid chromatography tandem mass spectrometry(LC-MS/MS) with a Surveyor HPLC system connected to a TSQ Quan-

tum Ultra triple quadrupole mass spectrometer by a Finnigan IonMax electrospray ionisation (ESI) source assembled with a high flow stainless steel emitter (Thermo Scientific, Rodano, Milan, Italy).[6] Two out of the three peptide hits provided improved PFKFB3 bisphosphatase activity and, likely, substrate affinity (Figure 6.2).

## 6.3 Binding mode prediction using MM-PBSA rescoring

The workflow discussed in Chapter 10 was developed at a later stage of this project. It is hence only applied to determine the *in silico* binding modes of the hit peptides to PFKFB3. Moreover, in the computational studies, all three peptides were predicted capable of binding at both the $\beta$-hairpin binding site and the secondary allosteric site. However, this VS was performed considering all the stereoisomers. In the experimental setting, all stereoisomers of compound **16** were tested, and only the isomeric form shown in Figure 6.2 were determined with micro-molar binding affinity.[5] More interestingly, visual inspection of binding poses showed that in the secondary binding site the stereoisomer in the form of the compound **16** was determined binder *in silico*. Other stereoisomers of peptide Trp-Gly-Tyr were not binding to PFKFB3 under experimental conditions. Hence, it is safe to suggest that the hit peptides are probably interacting with PFKFB3 at the secondary allosteric site and only these binding poses were further analysed using MM-PBSA rescoring method.

The docking and rescoring experiments were performed in 12 replicates. Compound **16** gave the best reproducibilities in predicted binding modes, probably due to its smaller size and lower number of rotatable bonds. Eleven out of twelve repeats resulted in the binding pose represented in Figure 6.3A from the initial VS. The interactions stabilising the binding include two hydrogen bonds to backbones of Arg368 and Ala319 on the other monomer (Ala780 indicated on the graph). Moreover, three CH-$\pi$ interactions are also contributing to the binding (Figure 6.3A). After MM-PBSA rescoring, which allows slight ligand and receptor positional re-adjustment, one additional ligand-to-backbone hydrogen bond established to the lysine next to

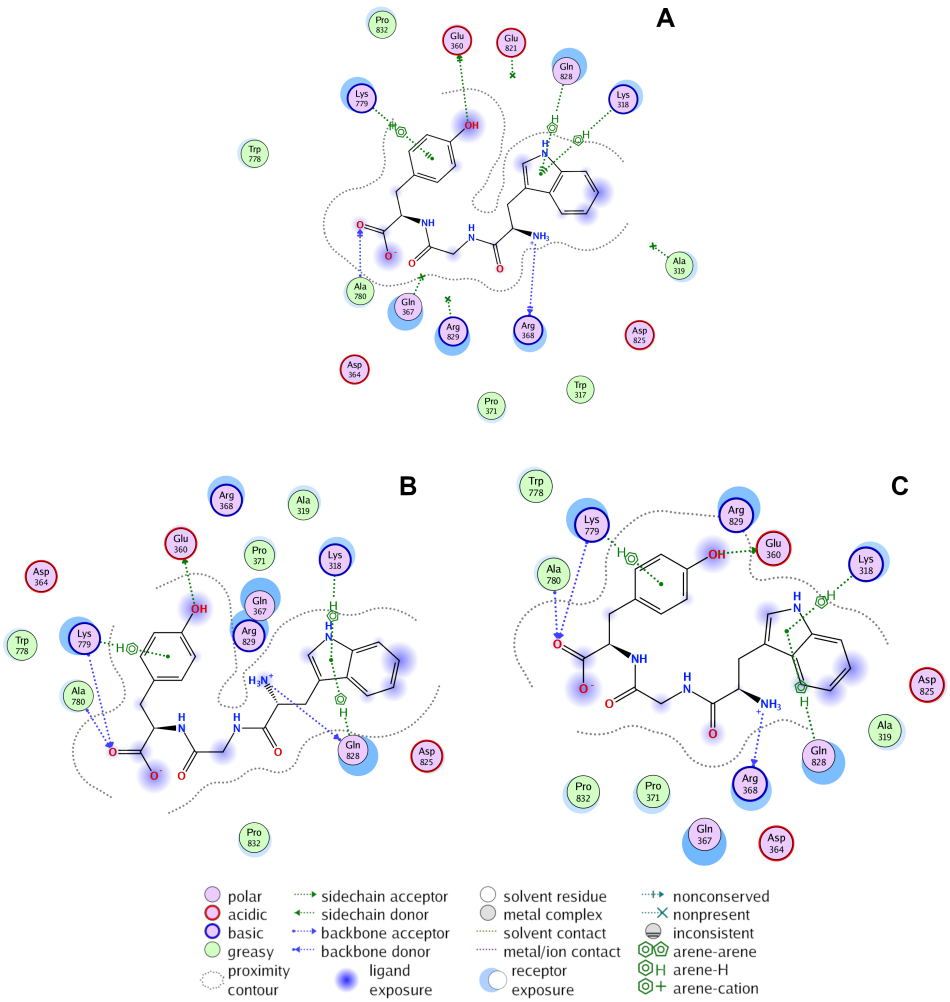



**Figure 6.3:** The predicted poses of compound **16** before (A) and after (B and C) MM-PBSA rescoring. The rescoring process adjusted the original poses to have one additional hydrogen bond between C-terminus of peptide **16** to Lys779 backbone. Each monomer of the applied PFKFB3 structure contains 461 amino acid residues. Hence, the residues numbered from 462 and up are amino acids from the second chain in the structure.

the Ala780. Some dynamic in binding was also implied from the rescoring outcomes. The charged N-terminus of compound **16** alternates its interaction between Arg368 and Gln828, and 11 out of 12 replicates were populated with the two conformations illustrated in Figure 6.3B and C.

Compound **15**, on the other hand, adopted poses with much higher variations. This is likely due to the larger amount of rotatable bonds within the longer peptide. However, interaction analysis revealed that the binding-contributing amino acid residues are backbone hydrogen bonds from Gln367 (10/12), Lys318 (9/12), Ala319 (8/12),

and Gln828 (8/12). One additional electrostatic interaction from the Arginine in peptide **15** to Glu315 was also frequented at 9 out of the 12 repeats. Additionally, from the enthalpic perspective, compound **15** is more likely to have a better binding affinity against PFKFB3 due to the size and the charge of the arginine residue. Indeed, compound **15** scored higher than **16** both in VS and MM-PBSA rescoring. However, the lower $K_d$ exhibit by the shorter peptide **16** suggested that the entropic contribution is one of the determinant factors when considering ligand binding at the allosteric site. The binding site lies in-between the two monomers of PFKFB3, making it a highly restrained place for ligand positioning. Larger compounds may suffer much larger entropic penalties than smaller binders. Hence, peptide **16** could be a better scaffold for later hit-to-lead development.

Some features of hit peptide binding can be summarised about the allosteric site. The peptide poses, under most of the circumstances, are contributed by hydrogen bonding to the receptor backbone. This indicates a good potential of targeting this allosteric site since the more stable backbone structure is contributing to the interacting framework. The binding of compound **15** also suggesting that the site can also accommodate larger molecules, though at the cost of higher entropic penalties. To further investigate the dynamic properties of the allosteric site, MD simulations were also performed and are discussed in Chapter 8.

## Chapter references

[1] J.J. Irwin, T. Sterling, M.M. Mysinger, E.S. Bolstad, and R.G. Coleman. ZINC: A Free Tool to Discover Chemistry for Biology. *Journal of Chemical Information and Modeling*, 52(7):1757–1768, 2012.

[2] Chemical Computing Group ULC. Moe, molecular operating environment, 2013.08, 2018. 1010 Sherbooke St. West, Suite 910, Montreal, QC, Canada, H3A 2R7.

[3] O. Korb, T. Stützle, and T.E. Exner. An ant colony optimization approach to flexible protein–ligand docking. *Swarm Intell.*, 1(2):115–134, 2007.

[4] O. Korb, T. Stüttzle, and T.E. Exner. Empirical scoring functions for advanced protein-ligand docking with plants. *J. Chem. Inf. Model.*, 49(1):84–96, 2009. doi: 10.1021/ci800298z.

[5] H. Macut. personal communication, 2016-2018.

[6] L.G. Regazzoni. personal communication, 2018.

# CHAPTER 7

# Design of novel PFKFB3 kinase inhibitors

## 7.1 Design strategy

Previous kinase inhibitor designs have been mainly focused on the ATP binding pocket of PFKFB3.[1–4] The reason for this trend is likely to be historical; ATP as a universal substrate for kinase has been the major blocking point. Moreover, a larger category of kinases undergo auto-phosphorylation or phosphorylating a downstream protein for the signalling cascade. The difficulties related to protein-protein interactions make the drug design targeting a small molecule binding pocket (i.e. ATP binding site) a more appealing approach, even though at the cost of a higher off-target rate. Larger "kinase-ready" compound libraries are also available from various databases, designed for ATP binding sites.

PFKFB3 is, on the other hand, a non-classical case of kinases. The enzyme is bifunctional, with the additional phosphatase activity counteracting the kinase reaction. More importantly, the kinase catalyses phosphorylation of fructose-6-phosphate

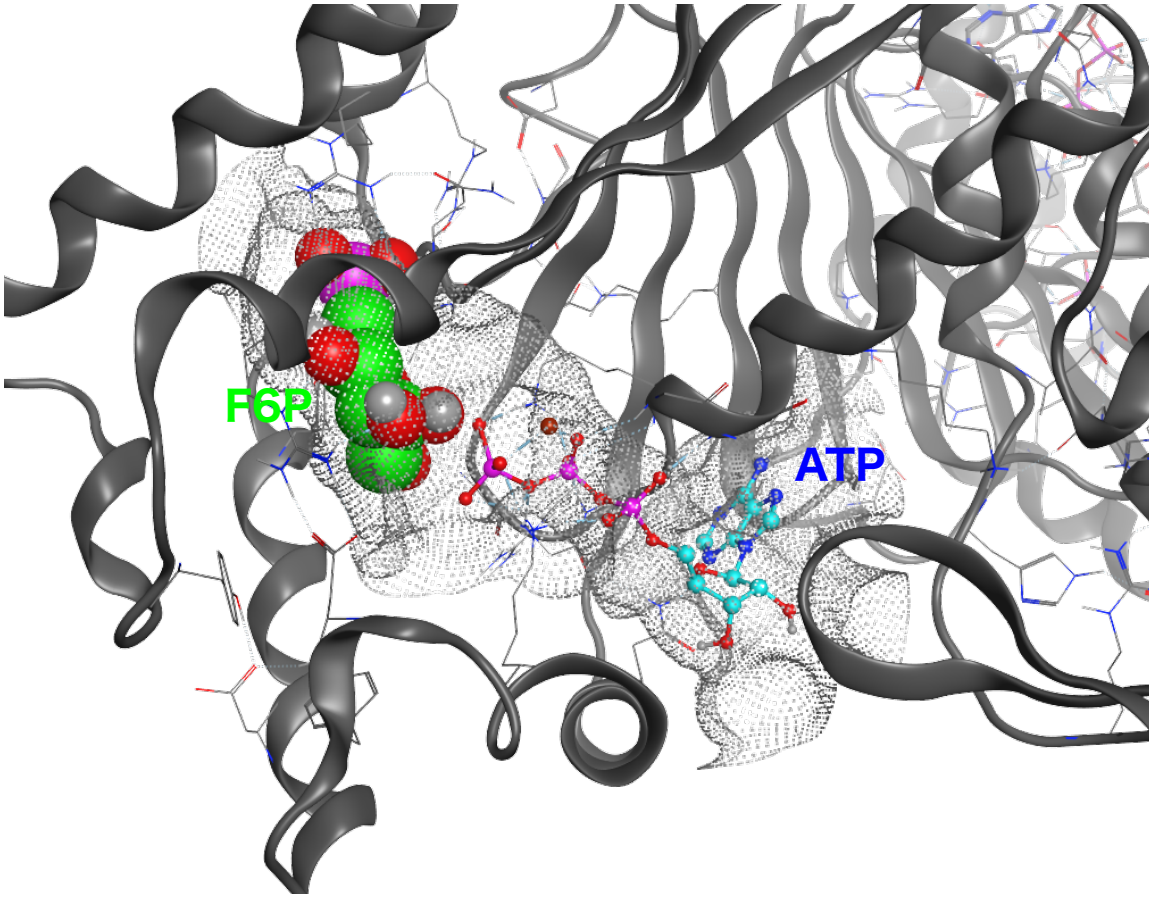



**Figure 7.1:** The active site of kinase on PFKFB3. The F6P molecule is coloured in green and ATP in cyan. The meshed surface shows the interface between the bound ligands to PFKFB3 kinase.

(F6P), an additional small molecule with a well-defined binding site (figure 7.1). It is likely this additional site of interest could provide improved inhibition specificity for PFKFBs. Moreover, the proximity of ATP and F6P active sites forms a larger channel that can comfortably accommodate larger molecules. The depth of this channel, though potentially increasing the entropic penalty upon ligand binding, can also provide stronger enthalpic contributions with a larger protein-ligand interface. Targeting F6P pocket has been attempted before, Seo et al. has discovered three molecules with low $\mu$M IC50. In this thesis, both F6P and ATP binding sites are taken into consideration. By considering both binding sites, the ligands can have a free choice of their enthalpic preference between the two sites. Moreover, this can also introduce better ligand adaptability within a larger binding space. Combining the properties of both

pockets would likely provide improved affinity.

## 7.2 Methods and materials

### 7.2.1 Screening library preparations

Two ligand libraries from Asinex were applied for virtual screening.[6] The kinase pre-plate set and the "lead-like" set (Elite and Synergy libraries) were downloaded from Asinex website. The libraries were processed using MOE and 3D-structures of ligands were generated and minimised briefly. The kinase library was then applied directly to the later virtual screening (a total of 6502 ligands). However, the "lead-like" set contains a total of 103325 compounds, which is too many to be handled using the computational resources available. Hence, further criteria were applied to shrink the library further. All compounds with a molecular weight bigger than 350 and smaller than 250 were excluded from the library. Since the purpose was to find hit compounds for further structural modifications later, the relatively low molecular range was considered reasonable. Moreover, compounds with hydrogen bond donors less than 5 and rotatable bonds less than 10 were retained and a logP cut-off at 3 was also applied. These selection steps shrank the size of the library to 6354 compounds. Multiple tautomerisms and protonation states were also considered for each ligand. Different states were generated and evaluated using the UNICON program[7] and only the top-scored were preserved.

### 7.2.2 Virtual screening and MM-PBSA rescoring

The screenings of the kinase set and lead-like set were carried out separately due to the different nature of the ligands. With the concerns that the ligands potentially targeting kinase were designed primarily for ATP/ADP binding site, the docking region was defined as including both F6P and ATP binding site on PFKFB3 to allow the open choices for ligands during the docking process. Later, after ranking with

MM-PBSA score, the poses were visually inspected and any ligands with interactions concentrated within the ATP binding site were excluded. This strategy is considered reasonable since it will provide *in silico* filtering of the screening ligands to exclude the potential hits at the ATP binding site. On the other hand, the docking of lead-like compounds was restricted to F6P binding sites.

The PLANTS docking software[8,9] was applied for the docking process. To improve the pose sampling, parameters were re-adjusted in PLANTS.[8,9] The evaporation factor was set to 0.15 and iteration scaling factor was changed to 7.0. Moreover, the RMSD similarity threshold for clustering was adjusted to 1.0 Å. Other parameters were applied as default settings. The final outcomes were evaluated using the total $PLANTS_{chemplp}$ scoring function with the un-normalised scores (TOTAL_SCORE).[8,9] The top 1000 ranked ligands are passed to the MM-PBSA rescoring as detailed in Chapter 10. The solute (internal) dielectric value 3 was applied for the calculation.

The virtual screenings were performed in three replicates for each library. Consortium selections were performed for the top 50 ranked ligands for each replica per library and the poses were also visually inspected. A total of 12 compounds were purchased from Asinex, among which 6 were from the kinase set and 6 were from the lead-like set. All 12 compounds were screened using Promega ADP-Glo™ Kinase Assay and showed more than 80% of PFKFB3 kinase inhibition at 10 $\mu$M, comparable to the inhibition activity by AZ33.[1](Figure 7.2A)

### 7.2.3 Experimental testing: iDiBAPS, Hospital Clinic Barcelona

Three of the hit ligands (shown in Figure 7.2B-D) were tested preliminarily in Hospital Clinic Barcelona, iDiBAPS, on human pathological endothelial cells (EC). Two diseases related to the cell lines applied in this work are pulmonary arterial hypertension (PAH) and chronic thromboembolic pulmonary hypertension (CTEPH). Both diseases belong to the category of pulmonary hypertension and defined by a mean pulmonary artery pressure above 25 mmHg.[11] The disease-specific lesions within the

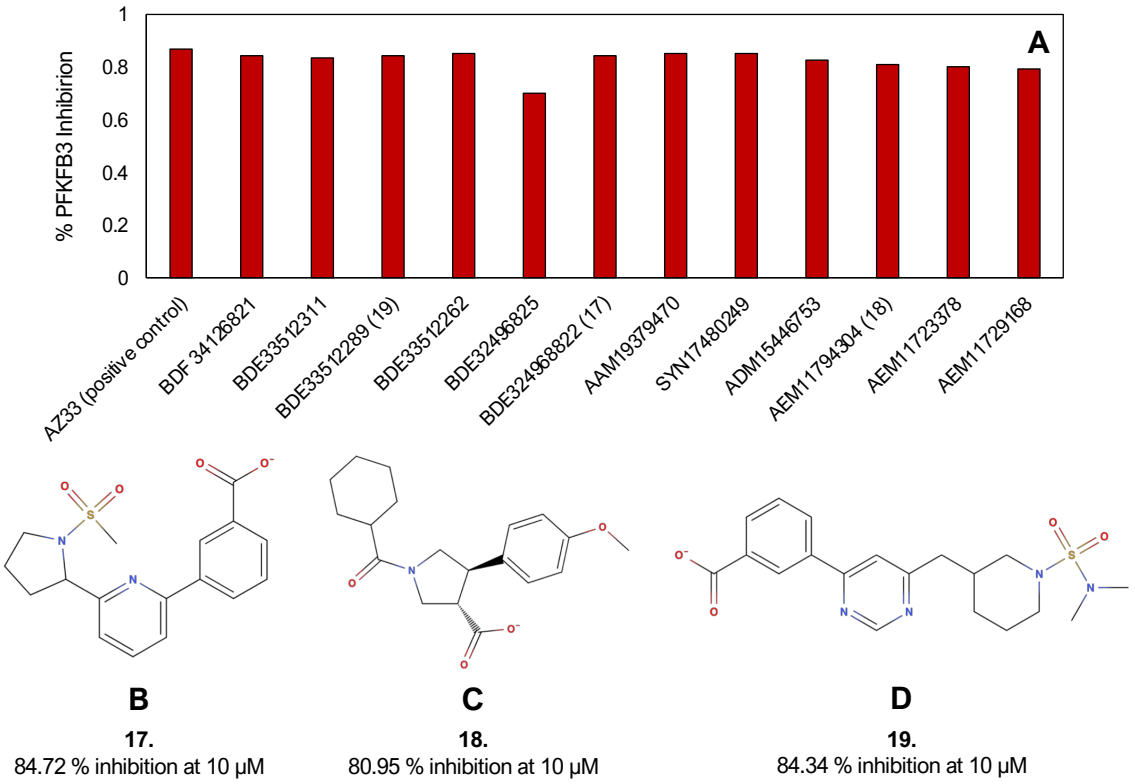



**Figure 7.2:** The PFKFB3 kinase inhibition level of the selected Asinex compounds (A) and compounds been tested on human or patient EC cell lines (B-D). Result for AZ33 inhibition at 10 $\mu$M is also shown. (Data provided by Macut[10]) B: Asinex ID BDE324968822. C: Asinex ID AEM11794304. D: Asinex ID BDE33512289.

smaller pulmonary arteries (<500 $\mu$m in diameter) is an important determinant of PAH.[11] On the other hand, CTEPH is characterised by prominent unresolved thrombi obstructions in the main pulmonary arteries.[12,13]

The prevalence of aerobic glycolysis has been observed among PAH patients.[14] Some signs of increased aerobic glycolysis were also observed in human pulmonary microvascular endothelial cells transfected with bone morphogenic protein receptor type 2 (BMPR2) vector.[15] Mutations in BMPR2 has been well-acknowledged relating to PAH.[16–20] However, glycolysis up-regulation is not the only deranged metabolic pathways in PAH.[21] Studies in mice and PAH patients demonstrate that the disease is also associated with normoxic activation of HIF-1$\alpha$.[15,22,23] Thus, as a downstream factor of HIF-1$\alpha$ activation, it can be inferred that PFKFB3 is also likely up-regulated in PAH. On the other hand, the connection between PFKFB3 and CTEPH disease is scarcely

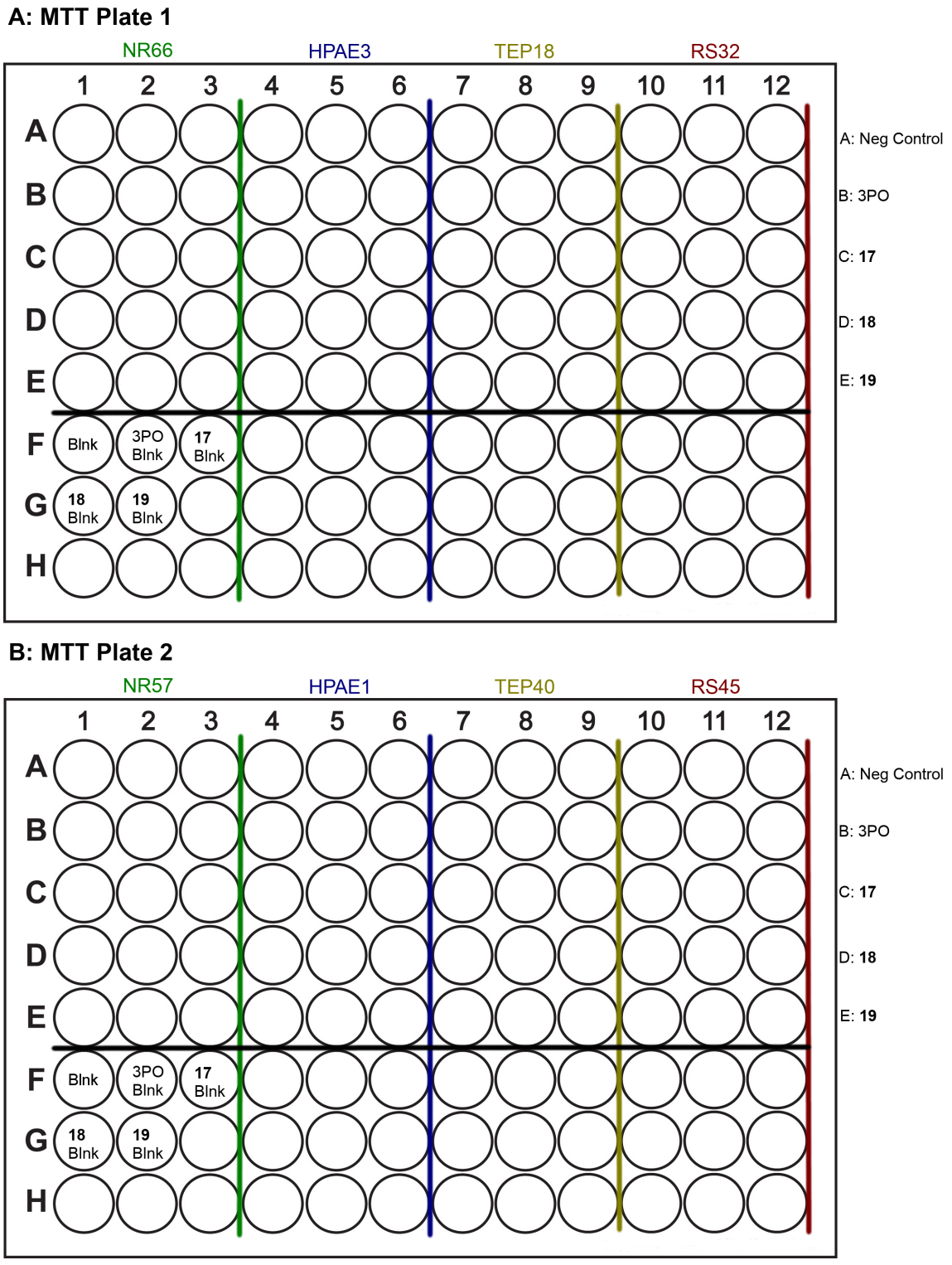



**Figure 7.3:** The plate layouts for MTT assays performed in this work. Negative control indicate that the normal media was applied. Blank controls (Blnks) were wells with MTT reagents but no cells were seeded.

explored. The implications on metabolism are mainly from studies on PAH.[24–26] PFKFB3 is a potential new target for CTEPH treatment as currently available therapies for CTEPH are mainly focused on symptom management. Inhibiting PFKFB3 activity can potentially modulate the activation of glycolysis cascade and decrease cell proliferation or viability. Hence, MTT test was chosen for determine the effects of compounds **17**, **18**, and **19** on the selected cell lines. A positive control using 3PO, the known glycolysis inhibitor, was also included.

Two types of human pathological endothelial cell lines were adopted for the experiment and compared with healthy human pulmonary artery ECs (HPAE, healthy control cell lines). EC cells from pulmonary artery of subjects undergoing lung transplant were also used as another healthy control (labelled as NR). Cell lines labelled with RS and TEP are patient cell lines; RS is a cell line isolated from lung transplant tissue of PAH patients by Dr D. Szulcek[27] (VU University Medical Center Amsterdam, Department of Pulmonary Diseases, Amsterdam Cardio-vascular Science, the Netherlands). TEP cells were isolated from endarterectomy tissue extracted from pulmonary arteries of CTEPH patients provided by Dr O. Tura[28] (Hospital Clinic Barcelona, iDiBAPS, Spain).

The cells were seeded onto 0.2% gelatine-coated T25 flasks in EGM$^{TM}$-2 phenol-red-free endothelial cell growth medium supplemented with 10% of fetal bovine serum. The cells were cultivated until confluence at 37 °C. The cells were then trypsinised and transferred onto 0.2% gelatine-coated 96-well plate. EGM2 + 10% FBS media with or without testing compounds were added to form the control and experiments set and are cultivated at 37 °C for 48 hrs in a 5%-$CO_2$ environment. The concentrations of compounds in media were 20 $\mu$M for 3PO, and 50 $\mu$M for **17**, **18**, and **19**. The plate layouts are as shown in Figure 7.3. The Vybrant®MTT Cell Proliferation Assay Kit (Thermo Fisher Scientific Inc.) was applied and the standard protocol was followed using the SDS-HCl solution. The media were firstly exchanged for fresh EGM2 + 10% FBS phenol-red-free media with or without testing compounds before the experiments. The absorbance signals were recorded at 570 nm (Synergy).

## 7.3 Results and discussions

### 7.3.1 Virtual screening

The predicted binding poses from the two Asinex library sets were illustrated in Figure 7.4A and C. Multiple amino acid residues interacting with the hit ligands are preserved for both library sets. These residues are Arg98, Arg132, and Tyr193, all three are presented within the F6P binding sites. In comparison to the binding pose of F6P, Arg132 also interact closely with the phosphate group while Arg98 could be the alternative binding candidate in case of dynamic pose adjustment.

More interestingly, the two types of hit ligands provide different binding modes within

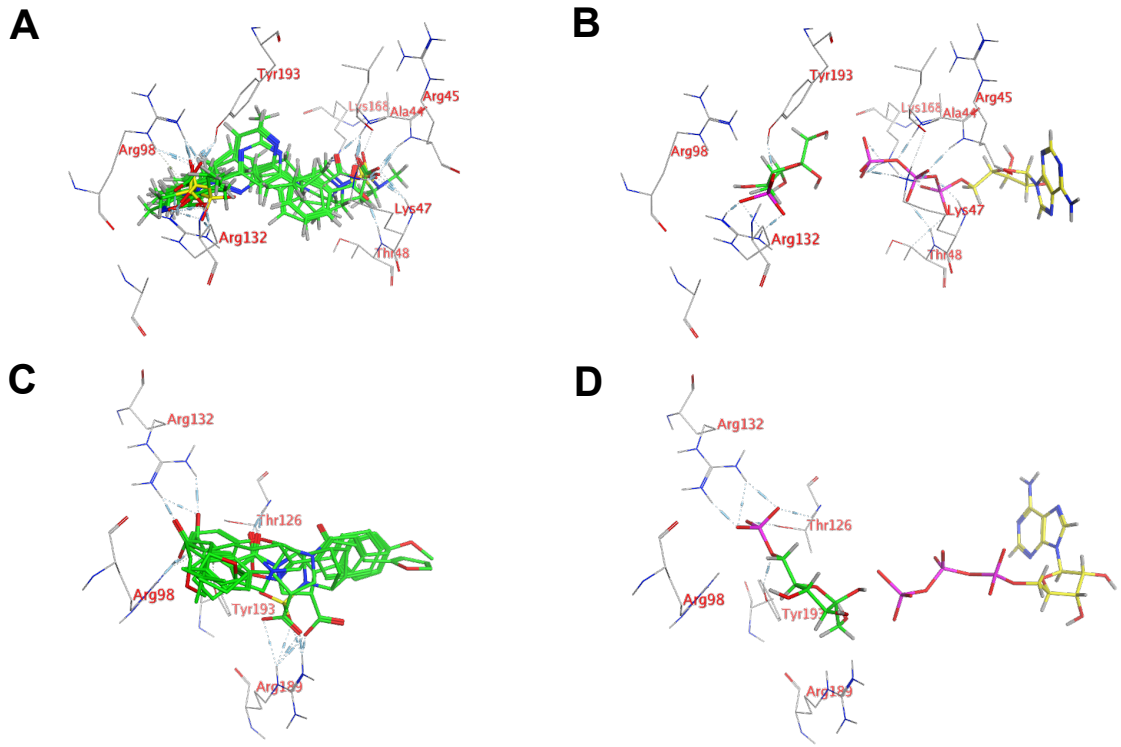



**Figure 7.4:** The binding poses and the involved amino acid interactions. Figure A represents the favourable binding poses of the 6 tested ligands selected from Asinex kinase set. Figure B shows how the interacting amino acids from A interacting with F6P and ATP. Figure C represents the favourable binding poses of the 6 tested ligands selected from Asinex lead-like set. D gives the relevant amino acid interactions to F6P and ATP in connection to C.

the joint F6P-ATP active site. Ligands from the lead-like set concentrate their interactions to PFKFB3 mostly within the F6P binding site; lacking interactions other than the potential non-polar interactions were observed on the ATP side of the pocket. However, the ligands from the kinase set have more extensive interaction modes. Both F6P and ATP binding pockets were perturbed upon ligand binding and larger numbers of amino acid residues are involved. Other than the three pre-mentioned residues within F6P site, 5 more from the ATP sites were also observed (Ala44, Arg45, Lys47, Thr48, and Lys168). Moreover, the backbone secondary amine groups from Ala44, Arg45, Lys47, and Thr48 forming a hub of hydrogen bond donors, which can comfortably accommodate electro-rich functional groups. These imply that the hit ligands from kinase set could have higher ligand-efficiency targeting the F6P-ATP binding site, likely resulting in higher binding affinity. However, this presumption is yet to be determined using experimental binding assays.

### 7.3.2 MTT assay

To directly compare the cell viability responses after different treatment to control, the relative response was calculated by normalising over the absorbance signal from the control wells with testing-compound-free media. The results are shown in Figure 7.5. All signals from treatments with PFKFB3 hit compounds were similar to 3PO treatment for all pathological cell lines. This indicates the hit compounds impair cell viability on a scale similar to 3PO, which was demonstrated with biological activity *in vitro* in previous studies. More interestingly, 3PO in one set of experiments seems to affect the viability of the healthy cell lines (Figure 7.5A for NR66). However, future studies should be performed to confirm the above observations.

Additionally, both RS and TEP cells were observed with higher responses to hit compound treatments to control. It was suggested that RS and TEP cells possess an elevated level of cell proliferations *in vitro* (unpublished data[29]). This up-regulation requires that more active metabolism processes should incur. This is likely accompanied by an upturn in the glycolysis process through increased expression or activity

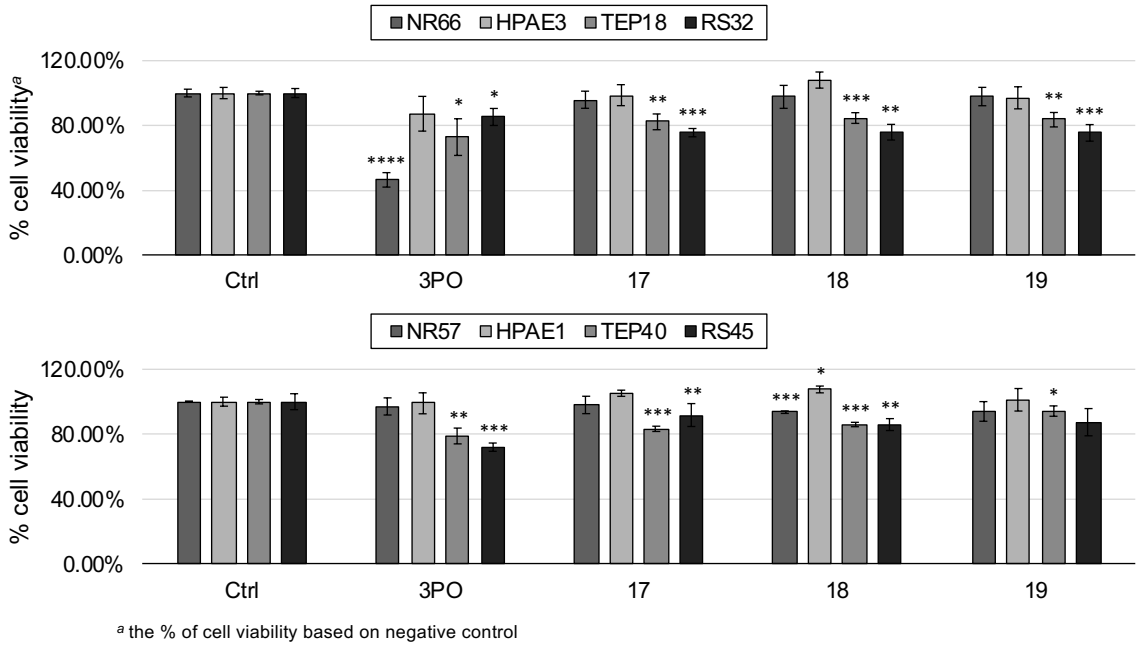



**Figure 7.5:** The cell viability level using different treatment in different human endothelial cell lines. HPAE and NR are healthy pulmonary artery endothelial cells. TEP samples are pulmonary artery ECs from CTEPH patients. RS are ECs isolated from pulmonary arteries of PAH patients. Error bars are represented as the standard deviations between the three replicates of each experiment. The null hypothesis testing $p$-values were calculated when comparing to controls.

of PFKFB3. Hence, inhibiting glycolysis using 3PO or inhibiting PFKFB3 kinase activity would likely provide a higher response in RS and TEP cells, as shown in Figure 7.5.

However, some further studies should be carried out in the future to confirm the biological activities of the hit compounds. This can be accomplished by using more sets of healthy and patient cell lines for testing in MTT assays. Moreover, the relatively weak responses of the cell lines to the hit compounds implied that a higher compound concentration should also be assessed. Other experimental methods for assessing cell growth and proliferation can also be tested for the hit compounds. Limitations still remain that small sample size was applied in this thesis. Nonetheless, the preliminary tests included in this thesis have shown some promises of the novel hits and have shed lights on targeting PFKFB3 for metabolism modulation in treatment for cardiovascular diseases.

# Chapter references


[1] S. Boyd, J.L. Brookfield, S.E. Critchlow, I.A. Cumming, N.J. Curtis, J. Debreczeni, S.L. Degorce, C. Donald, N.J. Evans, S. Groombridge, P. Hopcroft, N.P. Jones, J.G. Kettle, S. Lamont, H.J. Lewis, P. MacFaull, S.B. McLoughlin, L.J.M. Rigoreau, J.M. Smith, S. St-Gallay, J.K. Stock, A.P. Turnbull, E.R. Wheatley, J. Winter, and J. Wingfield. Structure-Based Design of Potent and Selective Inhibitors of the Metabolic Kinase PFKFB3. *J. Med. Chem.*, 58(8): 3611–3625, 2015.

[2] S.A. St-Gallay, N. Bennett, S.E. Critchlow, N. Curtis, G. Davies, J. Debreczeni, N. Evans, I. Hardern, G. Holdgate, N.P. Jones, L. Leach, S. Maman, S. McLoughlin, M. Preston, L. Rigoreau, A. Thomas, A.P. Turnbull, G. Walker, J. Walsh, R.A. Ward, E. Wheatley, and J. Winter-Holt. A High-Throughput Screening Triage Workflow to Authenticate a Novel Series of PFKFB3 Inhibitors. *SLAS Discov.*, 23(1):11–22, 2017.

[3] N.M.S. Gustafsson, K. Färnegårdh, N. Bonagas, A.H. Ninou, P. Groth, E. Wiita, M. Jönsson, K. Hallberg, J. Lehto, R. Pennisi, J. Martinsson, C. Norström, J. Hollers, J. Schultz, M. Andersson, N. Markova, P. Marttila, B. Kim, M. Norin, T. Olin, and T. Helleday. Targeting PFKFB3 radiosensitizes cancer cells and suppresses homologous recombination. *Nat. Commun.*, 9(1): 3872, 2018.

[4] N. Boutard, A. Białas, A. Sabiniarz, P. Guzik, K. Banaszak, A. Biela, M. Bień, A. Buda, B. Bugaj, E. Cieluch, A. Cierpich, Ł. Dudek, H.-M. Eggenweiler, J. Fogt, M. Gaik, A. Gondela, K. Jakubiec, M. Jurzak, A. Kitlińska, P. Kowalczyk, M. Kujawa, K. Kwiecińska, M. Leś, R. Lindemann, M. Maciuszek, M. Mikulski, P. Niedziejko, A. Obara, H. Pawlik, T. Rzymski, M. Sieprawska-Lupa, M. Sowińska, J. Szeremeta-Spisak, A. Stachowicz, M.M. Tomczyk, K. Wiklik, Ł. Włoszczak, S. Ziemiańska, A. Zarebski, K. Brzózka, M. Nowak, and C.-H. Fabritius. Discovery and Structure–Activity Relationships of N-Aryl 6-Aminoquinoxalines as Potent PFKFB3 Kinase Inhibitors. *ChemMedChem*, (Early View), 2018.

[5] M. Seo, J.-D. Kim, D. Neau, I. Sehgal, and Y.-H. Lee. Structure-Based Development of Small Molecule PFKFB3 Inhibitors: A Framework for Potential Cancer Therapeutic Agents Targeting the Warburg Effect. *PLoS One*, 6(9):e24179, 2011.

[6] Asinex Corporation. Asinex - screening libraries. `https://http://www.asinex.com/`, 2018. Accessed: Apr 2018.

[7] K. Sommer, N.-O. Friedrich, S. Bietz, M. Hilbig, T. Inhester, and M. Rarey. Unicon: A powerful and easy-to-use compound library converter. *J. Chem. Inf. Model.*, 56(6):1105–1111, 2016.

[8] O. Korb, T. Stützle, and T.E. Exner. An ant colony optimization approach to flexible protein–ligand docking. *Swarm Intell.*, 1(2):115–134, 2007.

[9] O. Korb, T. Stüttzle, and T.E. Exner. Empirical scoring functions for advanced protein-ligand docking with plants. *J. Chem. Inf. Model.*, 49(1):84–96, 2009. doi: 10.1021/ci800298z.

[10] H. Macut. personal communication, 2016-2018.

[11] N. Galie, M. Humbert, J.-L. Vachiery, S. Gibbs, I. Lang, A. Torbicki, G. Simonneau, A. Peacock, A. Vonk Noordegraaf, M. Beghetti, A. Ghofrani, M.A. Gomez Sanchez, G. Hansmann, W. Klepetko, P. Lancellotti, M. Matucci, T. McDonagh, L.A. Pierard, P.T. Trindade, M. Zompatori, and M. Hoeper. 2015 ESC/ERS Guidelines for the diagnosis and treatment of pulmonary hypertension: The Joint Task Force for the Diagnosis and Treatment of Pulmonary Hypertension of the European Society of Cardiology (ESC) and the European Respiratory Society (ERS): Endorsed by: Association for European Paediatric and Congenital Cardiology (AEPC), International Society for Heart and Lung Transplantation (ISHLT). *European heart journal*, 37(1): 67–119, 2016.

[12] M. Humbert. Pulmonary arterial hypertension and chronic thromboembolic pulmonary hypertension: pathophysiology. *European respiratory review : an official journal of the European Respiratory Society*, 19(115):59–63, 2010.

[13] I. Lang. Chronic thromboembolic pulmonary hypertension: a distinct disease entity. *European respiratory review : an official journal of the European Respiratory Society*, 24(136):246–252, 2015.

[14] J. Rehman and S.L. Archer. A proposed mitochondrial–metabolic mechanism for initiation and maintenance of pulmonary arterial hypertension in fawn-hooded rats: The warburg model of pulmonary arterial hypertension. In J.X.-J. Yuan and J.P.T. Ward, editors, *Membrane Receptors, Channels and Transporters in Pulmonary Circulation*, pages 171–185, Totowa, NJ, 2010. Humana Press.

[15] G. Marsboom, C. Wietholt, C.R. Haney, P.T. Toth, J.J. Ryan, E. Morrow, T. Thenappan, P. Bache-Wiig, L. Piao, J. Paul, C.-T. Chen, and S.L. Archer. Lung $^{18}$F-Fluorodeoxyglucose Positron Emission Tomography for Diagnosis and Monitoring of Pulmonary Arterial Hypertension. *Am. J. Respir. Crit. Care Med.*, 185(6):670–679, 2012.

[16] K.B. Lane, R.D. Machado, M.W. Pauciulo, J.R. Thomson, J.A. Phillips, J.E. Loyd, W.C. Nichols, and R.C. Trembath. Heterozygous germline mutations in BMPR2, encoding a TGF-$\beta$ receptor, cause familial primary pulmonary hypertension. *Nat. Genet.*, 26:81, 2000.

[17] Z. Deng, J.H. Morse, S.L. Slager, N. Cuervo, K.J. Moore, G. Venetos, S. Kalachikov, E. Cayanis, S.G. Fischer, R.J. Barst, S.E. Hodge, and J.A. Knowles. Familial Primary Pulmonary Hypertension (Gene PPH1) Is Caused by Mutations in the Bone Morphogenetic Protein Receptor–II Gene. *Am. J. Hum. Genet.*, 67(3):737–744, 2000.

[18] K. Teichert-Kuliszewska, M.J.B. Kutryk, M.A. Kuliszewski, G. Karoubi, D.W. Courtman, L. Zucco, J. Granton, and D.J. Stewart. Bone Morphogenetic Protein Receptor-2 Signaling Promotes Pulmonary Arterial Endothelial Cell Survival. *Circ. Res.*, 98(2):209–217, 2006.

[19] X. Yang, L. Long, P.N. Reynolds, and N.W. Morrell. Expression of Mutant BMPR-II in Pulmonary Endothelial Cells Promotes Apoptosis and a Release of Factors that Stimulate Proliferation of Pulmonary Arterial Smooth Muscle Cells. *Pulm. Circ.*, 1(1):103–110, 2011.

[20] E.D. Austin and J.E. Loyd. The genetics of pulmonary arterial hypertension. *Circ. Res.*, 115 (1):189–202, 2014.

[21] T.R. Assad and A.R. Hemnes. Metabolic Dysfunction in Pulmonary Arterial Hypertension. *Curr. Hypertens. Rep.*, 17(3):20, 2015.

[22] R. Kluge, H. Barthel, H. Pankau, A. Seese, J. Schauer, H. Wirtz, H.-J. Seyfarth, J. Steinbach, O. Sabri, and J. Winkler. Different mechanisms for changes in glucose uptake of the right and left ventricular myocardium in pulmonary hypertension. *J. Nucl. Med.*, 46(1):25–31, 2005.

[23] E.L. Lundgrin, M.M. Park, J. Sharp, W.H.W. Tang, J.D Thomas, K. Asosingh, S.A. Comhair, F.P. DiFilippo, D.R. Neumann, L. Davis, B.B. Graham, R.M. Tuder, I. Dostanic, and S.C. Erzurum. Fasting 2-deoxy-2-[18F]fluoro-D-glucose positron emission tomography to detect metabolic changes in pulmonary arterial hypertension hearts over 1 year. *Ann. Am. Thorac. Soc.*, 10(1):1–9, 2013.

[24] M.C. van de Veerdonk, H.J. Bogaard, and N.F. Voelkel. The right ventricle and pulmonary hypertension. *Heart Fail. Rev.*, 21(3):259–271, 2016.

[25] D. Gopalan, M. Delcroix, and M. Held. Diagnosis of chronic thromboembolic pulmonary hypertension. *Eur. Respir. Rev.*, 26(143):160108, 2017.

[26] G. Simonneau, A. Torbicki, P.r Dorfmüller, and N. Kim. The pathophysiology of chronic thromboembolic pulmonary hypertension. *Eur. Respir. Rev.*, 26(143):160112, 2017.

[27] D. Szulcek. personal communication, 2018.

[28] O. Tura. personal communication, 2018.

[29] V. Smolders. personal communication, 2018.

# CHAPTER 8

# MD simulations of PFKFB3

## 8.1 Methodology

To investigate the dynamic behaviours of PFKFB3, MD simulations were carried out for a total of 3 forms of the enzyme. In Chapter 5, the choice of PFKFB3 from the hub of crystal structures available in PDB bank has been elucidated in detail. The chosen structure, PDB ID 2i1v,[1] contains the natural enzymatic products present within the kinase and phosphatase active sites. However, both *apo* (i.e. structure with empty binding sites) and substrate-bound crystal structures of PFKFB3 were unavailable. Hence, some pre-processing based on the product-bound PFKFB3 was required.

The *apo* structure was generated by deleting the bound ligands from the modified structure of 2i1v[1] from Chapter 5. The substrate-bound PFKFB3 was generated by modifying the structure of the bound products in the starting structure. In the kinase active site, the phosphate group presented at 2-position of F2,6BP was deleted

and an additional phosphate was added on the phosphate end of ADP structure. The F6P ligand within the FBPase active site was modified to F2,6BP. The bound ligands in the kinase/bisphosphatase active sites were adjusted and local brief minimisations were carried out using MOE (Amber10:EHT, Born implicit solvation, and gradient cut-off at 0.1 RMS kcal·mol$^{-1}$·Å$^{-2}$).[2]

Moreover, the foreign ligands (ATP, ADP, F2,6BP, and F6P) were not included within the standard AMBER parameter sets. Thus, re-parameterisation was also required. ATP and ADP parameters compatible with AMBER simulation package were published elsewhere[3] and were downloaded from Ref. 4. The re-parameterisations of F2,6BP and F6P were performed with the RESP ESP charge Derive program (R.E.D.)[5] and antechamber.[6,7] The R.E.D. program was applied to assign partial charges to atoms presented within F2,6BP and F6P. The structures of F2,6BP and F6P were first submitted to conformational search in MOE[2] using the default settings. The two conformations with the lowest calculated conformational energies were saved for partial charge assignment. For each conformer, two orientations were also adopted during the charging process by rotating the molecule by 180°. The terminal oxygens within the same phosphate groups in both F2,6BP and F6P were forced to be equivalently charged within each functional group. The atom types of *gaff* were assigned using antechamber.[6,7]

The *ff14SB*[8] and the *gaff*[6] force field were applied for PFKFB3 protein and the ligands, respectively. The topology and coordinates files required for MD simulation were generated using *tleap*.[9] The protein / complexes were firstly neutralised with sodium or chloride ions; 6 $Cl^-$ were added for the *apo* PFKFB3 and 10 $Na^+$ were added for the product- and substrate-bound structures. The protein or complexes were then solvated within TIP3P explicit water molecules in a cubic box 12 Å from protein surface to the barrier. The neutralised and solvated structures were used as the starting structures for MD simulations.

MD simulations were run for *apo*, product-bound, and substrate-bound PFKFB3 structures, 3 replicates for each state. The *pmemd.MPI* module was applied for all

the simulations.[9] The energy minimisation, heating, and equilibration procedures prior to MD simulations were all performed in multi-step manners, with the same procedures for every run. A non-bonded cut-off of 8 Å was applied throughout.

Energy minimisation started with relaxing hydrogen positions with other atoms restrained (100 kcal·mol$^{-1}$·Å$^{-2}$) with 1000 steepest-decent (SD) followed by the conjugate gradient (CG) method to a maximum cycle of 5000. This step is followed by a 5000-max-cycle minimisation (2000 SD+ CG) of only water and ions (other atoms restrained at 50 kcal·mol$^{-1}$·Å$^{-2}$). The minimisation was then extended onto side chains of amino acid for another 5000 maximum cycles (2500 SD + CG, 25 kcal·mol$^{-1}$·Å$^{-2}$ restraints).

The system then underwent 6 steps of heating steps to a final temperature of 300 K under the NVT condition (constant volume and temperature with the total number of atoms unchanged). From this step on, the SHAKE algorithm is applied to constrain bonds involving hydrogen atoms and Langevin dynamics was applied for temperature scaling. Each heating step was performed with controlled heating of 50 K gap over 5 ps with a 0.0005 ps time step, while protein backbone was weakly restrained (10 kcal·mol$^{-1}$·Å$^{-2}$). After each 50 K heating process, the system was equilibrated for an additional 5 ps at the targeted temperature.

After the temperature of the simulated system reached 300 K, a 200-ps equilibration was firstly performed under NVT condition with weak backbone restraints (5 kcal·mol$^{-1}$·Å$^{-2}$). The simulation condition is then switched to NPT (constant pressure and temperature with the total number of atoms unchanged) and equilibrated for another 200 ps with the 5 kcal·mol$^{-1}$·Å$^{-2}$ constraints still applied for protein backbones. Following this, 5 steps of 500-ps equilibration runs were performed to reduce the constraint weight on protein backbones gradually. Each step reduces the constraint by 1 kcal·mol$^{-1}$·Å$^{-2}$ until the constraint was completely removed.

Lastly, MD simulations were performed under NPT condition at 300 K for 200 ns for each PFKFB3 state. The hydrogen-involved bond SHAKE constraint and Langevin temperature scaling were applied. The non-bonded cutoff was continued as 8 Å and

the simulation time step was set to 0.002 ps. Frames were recorded at a 1-ps frequency and the trajectories were written in the binary NetCDF format. The MD simulations of each PFKFB3 were performed for 3 replicates.

## 8.2 Results and discussions

The trajectories from the 9 MD simulations were firstly assessed with root-mean-squared displacement (RMSD) analysis. The RMSD was calculated about the backbone atoms (C and N within the peptide bond, and C$\alpha$) of the protein to evaluate the time period when the dynamic of the structure reached equilibrium. The plot of RMSD over simulation time showed that all trajectories reached equilibrium roughly after 100 ns (Figure 8.1). Hence, 100 - 200 ns was the time range chosen for further atomic correlation analysis.

Atomic correlation analysis was performed in a per-residue manner with both the backbone and side chain atoms included. The analysis was performed with *cpptraj*[9] and heat map plots were generated using *gnuplot*.[10] This analysis evaluates the average correlations between the motions of residues about the input trajectories. The plots of the 9 MD trajectories were shown in Figure 8.2 and 8.3. Since the dimeric form of PFKFB3 applied in this thesis contains a total of 922 amino acid, each PFKFB3 monomer contains 461 residues. Thus each correlation graph can be divided into two parts: inter- (between residue 1-461 and 462-922) and intra-monomeric (among residue 1-461 or 462-922) correlations. Moreover, the intra-monomeric correlation can be further separated to the correlation between kinase and bisphosphatase halves and correlation among different functional half of the PFKFB3.

The inter-residual correlation between kinase and bisphosphatase halves of the PFKFB3 was of particular interest for the studies related to this thesis. The auto-regulatory domain interfering with FBPase activity was based on the interaction between the N-terminus to the bisphosphatase half of the enzyme. Moreover, activity test of FBPase using the strong kinase inhibitor, AZ33, also shown around a 6-fold increase

in bisphosphatase activity (unpublished data provided by Regazzoni),[11] similar to the FBPase activators discovered (Figure 6.2, Chapter 6). These experimental observations imply that perturbing the ATP binding site through kinase inhibitors and binding of the modulator at the allosteric site for bisphosphatase result in the same

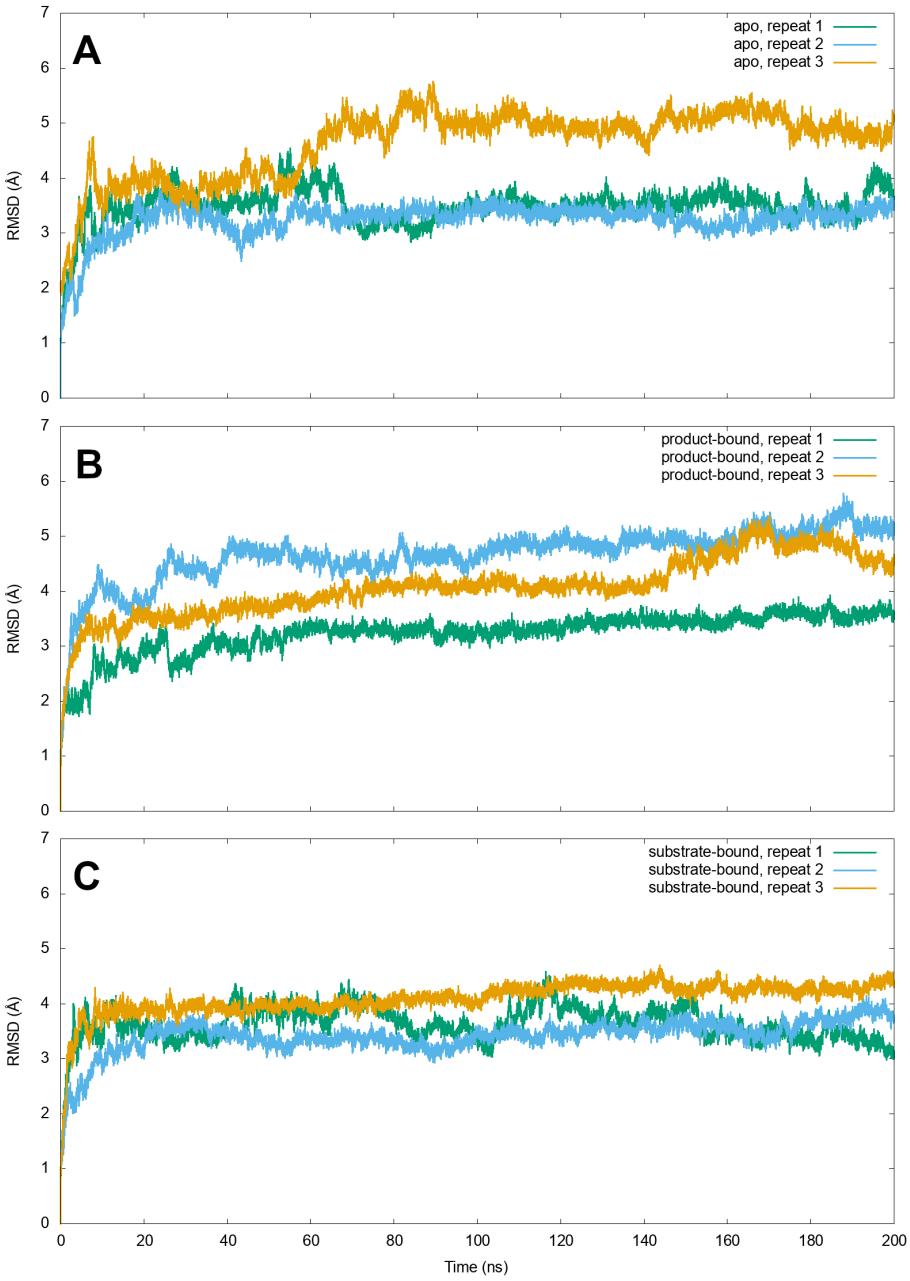



**Figure 8.1:** RMSD plots of the trajectories from the MD simulations. A: The three repeats of MD simulations using the *apo* PFKFB3 structure. B: The three repeats of MD simulations using the product-bound PFKFB3 structure. C: The three repeats of MD simulations using the substrate-bound PFKFB3 structure.

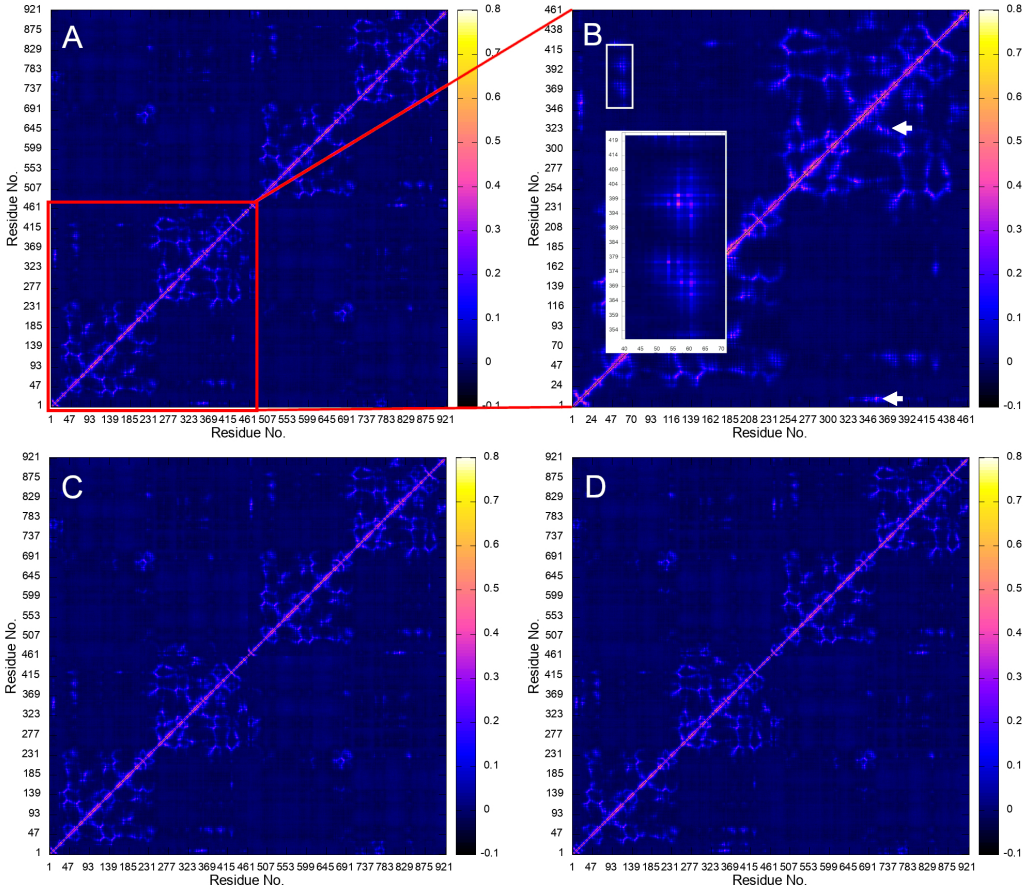



**Figure 8.2:** Atomic correlation plots of the last 100 ns MD simulations of the three repeats using the *apo* PFKFB3 structure. A, C, and D are the plots considering the complete homodimer. B is the zoomed-in plot of one of the monomers shown in the red box. In B, the box and the zoomed-in sub-plot show the three weakly correlated -helices – $\alpha$1, $\alpha$17, and $\alpha$18. The two arrows in B show the stronger correlations of the auto-regulatory domain ( -hairpin) to the E322-A325 loop (shown in yellow) through part of the 17-helix (Y362-E370).

outcome of FBPase activation.

The trajectories from MD simulation provide a possible explanation regarding the underlying molecular mechanisms. A weak correlating zone that connects the ATP binding site to the bisphosphatase active site was observed for all the simulated trajectories analysed. This is through the chained interactions between three $\alpha$-helices – $\alpha$1, $\alpha$17, and $\alpha$18. Moreover, $\alpha$17 is the essential element to perturb activity of the bisphosphatase (Figure 8.4A); part of the helix lies parallel to the Glu322-Ala325 loop flooring the bisphosphatase site. More interestingly, this loop region is also

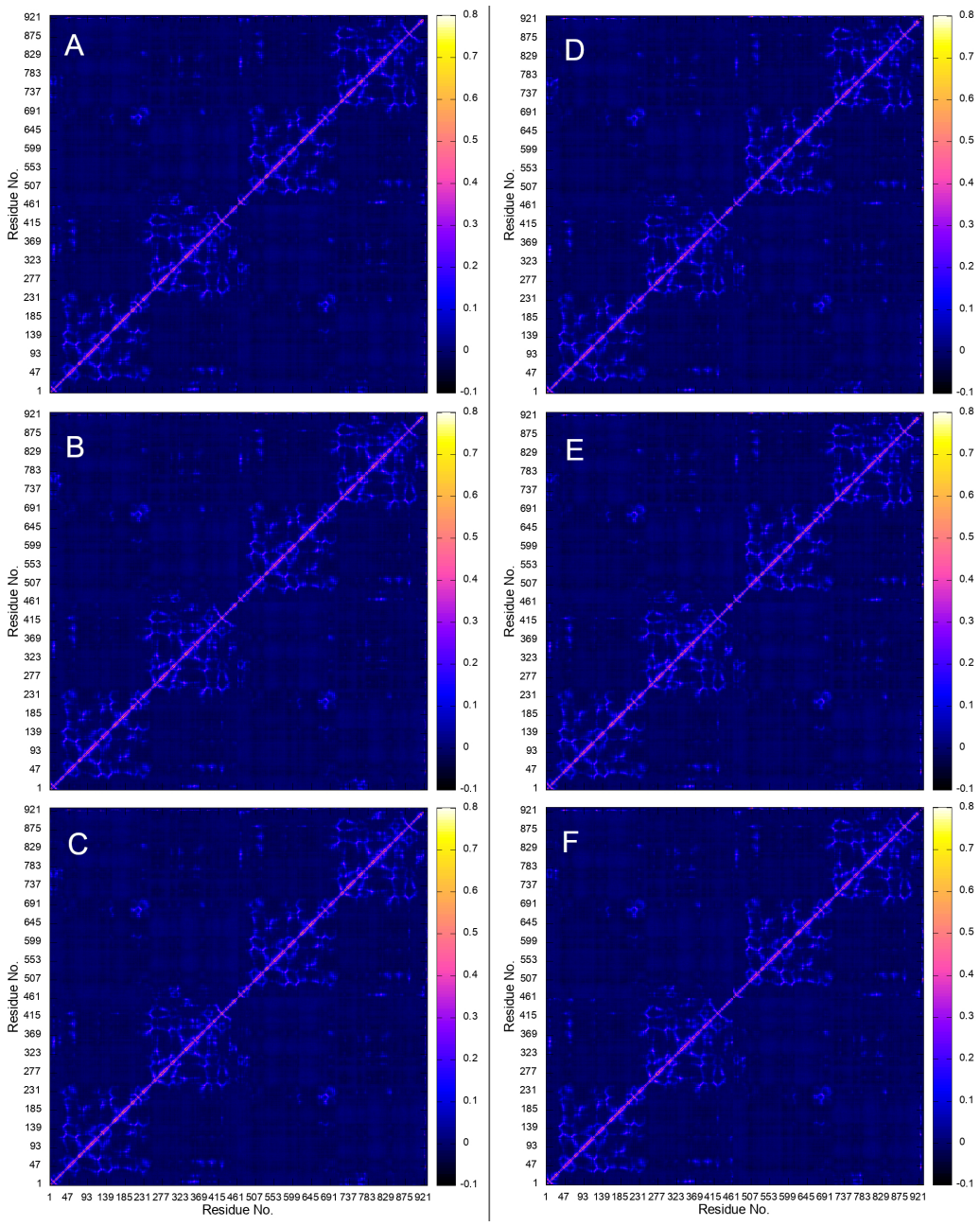



**Figure 8.3:** Atomic correlation plots of the last 100 ns MD simulations of the three repeats using the the product-bound and the substrate-bound PFKFB3 structures. A-C: MD simulations using the product-bound PFKFB3. D-F: MD simulations using the substrate-bound PFKFB3.

indirectly correlated to the auto-regulatory domain that modulates bisphosphatase activity(Figure 8.4B). The similar fold-changes in the activity due to perturbation from AZ33 and by removing the auto-regulatory domain also support that the crucial connection point is through $\alpha$17.

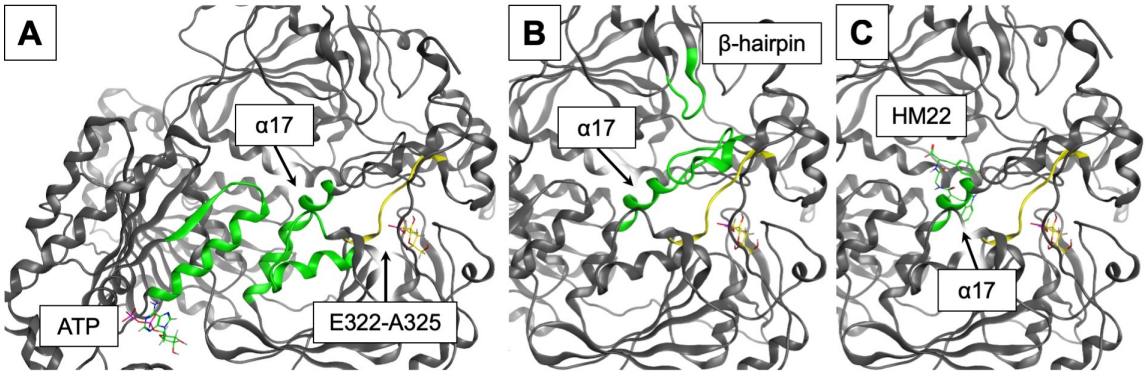



**Figure 8.4:** The highlights of different mechanisms discussed in this section. The loop affected within the FBPase pocket is highlighted as yellow in A-C. In A, the three weakly correlated -helices connecting ATP binding site to FBPase active site are highlighted as green. The -hairpin-to-phosphatase correlated regions are illustrated in B. The green region shows the stronger correlations of the auto-regulatory domain ( -hairpin) to the E322-A325 loop (show in yellow) through part of the 17-helix (Y362-E370). In C, the docked pose of HM22 indicates that the peptide is interfering with E322-A325 loop directly through 17-helix.

On the other hand, compound **15** and **16** (Figure 6.2, Chapter 6) took a more direct approach by affecting the conformation of the $\alpha$17 helix directly. However, this interference required a more well-adapted binding pose. Both **15** and **16** adapted more kinked binding poses after docking, suggesting that much more volume of the binding space is occupied. This leads to the similar activation outcome as was accomplished by ATP and the $\beta$-hairpin region. Compound **14**, however, adopted a more extended and relaxed binding pose, likely resulting in the un-perturbed FBPase activity while still capable of binding to PFKFB3.

## Chapter references

[1] S.-G. Kim, M. Cavalier, M.R. El-Maghrabi, and Y.-H. Lee. A Direct Substrate–Substrate Interaction Found in the Kinase Domain of the Bifunctional Enzyme, 6-Phosphofructo-2-kinase/Fructose-2,6-bisphosphatase. *J. Mol. Biol.*, 370(1):14–26, 2007.

[2] Chemical Computing Group ULC. Moe, molecular operating environment, 2018. 1010 Sherbooke St. West, Suite 910, Montreal, QC, Canada, H3A 2R7.

[3] K.L. Meagher, L.T. Redman, and H.A. Carlson. Development of polyphosphate parameters for use with the AMBER force field. *Journal of computational chemistry*, 24(9):1016–1025, 2003.

[4] Bryce R. Amber parameter database. `http://research.bmh.manchester.ac.uk/bryce/amber/`, 2018. Accessed: April 2016.

[5] F.-Y. Dupradeau, C. Cezard, R. Lelong, E. Stanislawiak, J. Pecher, J. C. Delepine, and P. Cieplak. R.E.DD.B.: a database for RESP and ESP atomic charges, and force field libraries. *Nucleic acids research*, 36(Database issue):D360–7, 2008.

[6] J. Wang, R.M. Wolf, J.W. Caldwell, P.A. Kollman, and D.A. Case. Development and testing of a general amber force field. *Journal of computational chemistry*, 25(9):1157–1174, 2004.

[7] J. Wang, W. Wang, P.A. Kollman, and D.A. Case. Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of Molecular Graphics and Modelling*, 25(2):247–260, 2006.

[8] J.A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K.E. Hauser, and C. Simmerling. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.*, 11(8):3696–3713, 2015.

[9] D. A Case, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, D. Ghoreishi, M.K. Gilson, H. Gohlke, A.W. Goetz, D. Greene, R Harris, N. Homeyer, S. Izadi, A. Kovalenko, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, D.J. Mermelstein, K.M. Merz, Y. Miao, G. Monard, C. Nguyen, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R. Qi, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C.L. Simmerling, J. Smith, R. Salomon-Ferrer, J. Swails, R.C. Walker, J. Wang, H. Wei, R.M. Wolf, X. Wu, L. Xiao, D.M. York, and P.A. Kollman. Amber 2018, 2018.

[10] T. Williams, Kelley C., and many others. Gnuplot 5.2: an interactive plotting program. `http://gnuplot.sourceforge.net/`, 2018.

[11] L.G. Regazzoni. personal communication, 2018.

# CHAPTER 9

# Conclusions and prospects

In this thesis, the two strategies adopted have shown good initial results in terms of PFKFB3 activity modulation. The design of FBPase modulators resulted in 2 hit compounds with low $\mu$M binding affinity. The compounds also achieved a similar outcome in activity fold change as N-terminus spliced PFKFB.[1] Moreover, a novel allosteric site for activating bisphosphatase was also discovered.

However, it is still a long way to achieve a satisfactory influence on the biological behaviours of the PFKFB3. The hit compounds still require structure optimisations to further improve affinity, as some later attempts by modifying the amino acids within the hit peptides were unfruitful (results not shown). This is likely due to the fact that the binding of the hit peptides was mainly contributed by backbone-to-backbone hydrogen bonds; changes in the peptide side chains would have likely resulted in disruptions of binding poses. More crucially, the influences of FBPase activation on the kinase activity are yet to be confirmed. As PFKFB3 bears an exceptionally high kinase-to-bisphosphatase activity ratio, it is possible that the effects

of FBPase activation remain inconspicuous. However, the finding can still provide insights in PFKFBs with a lower kinase-to-bisphosphatase ratio (such as the liver isozyme, PFKFB2).

The hit molecules discovered for PFKFB3 kinase inhibition also require further optimisations. The $IC_{50}$ in kinase inhibition is yet to be determined and the binding affinity testing is also recommendable. Moreover, the choice of the biological assay can also be expanded to other more metabolism related experiments. Intracellular F1,6BP or F2,6BP level can be good indicating parameters to be tested for PFKFB3 inhibition in cells. Nonetheless, the good success rate of the VS method applied to render it a decent protocol for future hit discovery targeting PFKFB3 kinase.

## Chapter references

[1] I.J. Kurland, B. Chapman, and M.R. El-maghrabi. N- and C-termini modulate the effects of pH and phosphorylation on hepatic 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase. *Biochem. J.*, 347(2):459–467, 2000.

# Part III:

# Method and Protocol Developments

# CHAPTER 10

# Workflow for MM-PB/GBSA rescoring

## 10.1 Workflow overview

The workflow was designed as a single *bash* script incorporating multiple software including Open Babel,[1] UNICON,[2] AutoDock,[3] PLANTS[4,5], and AMBER packages.[6] The overall workflow is illustrated in Figure 10.2. In brief, the script starts with generating reasonable tautomerism and protonation states of the ligands within the screening library. Each state is then passed to VS procedures targeting receptor of interest by PLANTS using an adjusted procedure. The top 3 ranked conformers per state are saved and ligand parameters are generated. Combined with receptor structures, the final complexes firstly undergo a brief energy minimisation of the binding site while other parts of the structure restrained. The adjusted poses are then submitted for the MM-PB/GBSA re-evaluations and outcomes are analysed with statistical methods. The detailed procedures are elucidated in the following sections of this chapter.

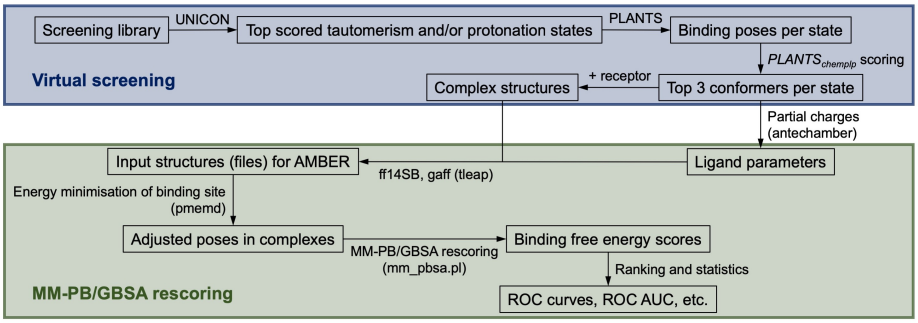



**Figure 10.1:** A graphic overview of the MM-PB/GBSA rescoring workflow.

## 10.2 Virtual screening

This part includes the initial screening library processing and docking experiments for each ligand. The screening libraries were firstly split into separate docking jobs for individual ligands. The jobs were then run with gnu parallel[7] on multiple-core CPU processors. Each job contains steps as specified as follow. The UNICON program[2] was applied to generate reasonable tautomeric and protonation states of ligands (the top scored ones defined within the program) for VS. The docking procedures of each determined state were then carried out using PLANTS.[4,5] The ligand atoms were kept flexible while the ones for receptors were fixed. The cluster RMSD was adjusted to 1.0 and the sampling factor $\sigma$ to 7.0 in the PLANTS configure files to improve numbers of poses searched. The poses were initially scored with the $PLANTS_{chemplp}$ scoring function.[4,5] The three top-ranked docking poses were saved for each tautomeric and/or protonation state to be input into later MM-PBSA rescoring. In Chapter 11, all virtual screening and rescoring were initially performed in triplicate. However, negligible differences between replicates were observed. Hence, the results of MM-PBSA rescoring assessment are shown for only one replicate.

## 10.3 Energy minimisation

The energy minimisation was introduced for occasional minor structural clashes were observed in some of the docked poses. These clashes were likely due to the empirical

nature of the $PLANTS_{chemplp}$ scoring function[4,5] and would have resulted in errors in MM-PBSA rescoring. Moreover, the fixed receptor structures would possibly introduce errors to the final pose accuracy. Thus, the brief energy minimisation of the binding poses was included for structure relaxation prior to the MM-PBSA single-frame calculation. The minimisation was performed using the AMBER simulation package.[6] The ligands were assigned with the AM1-BCC partial charges[8] using antechamber[9] or the Gasteiger charges from Autodock.[3] The *ff14SB* force field[10] was applied to the proteins and *gaff*[9] to the ligands.

The minimisation was performed in a modified *GBn* implicit solvent model (*igb* = 8).[11] The recommended *mbondi3* radii[11] were applied during the preparation with *tleap*. The maximum number of minimisation cycles was set to 5000 and the method switched from steepest descent to conjugate gradient after 500 cycles. The convergence criterion was relaxed to 0.1 kcal·mol$^{-1}$·Å$^{-1}$. A positional restraint with a weight of 100 kcal·mol$^{-1}$·Å$^{-2}$ was exerted onto atoms with a distance of more than 4.5 Å away from the ligand. These procedures were carried out for all the saved poses from the docking procedures.

## 10.4 MM-PBSA rescore

The binding free energy score was calculated using the MM-PB/GBSA (molecular mechanics (MM) with Poisson-Boltzmann (PB) or generalized Born (GB) and surface area solvation) methods. The method was developed by Kollman et al. and the binding free energy,[14] $\Delta G_{binding}$, is estimated as below:

$$\Delta G_{binding} = G_{complex} - G_{receptor} - G_{ligand}$$

The free energies of complex ($G_{complex}$), receptor ($G_{receptor}$), and ligand ($G_{ligand}$) are calculated by a summation of multiple energy terms:

$$G_x = E_{MM} + G_{PB\text{-}polar} + G_{PB\text{-}nonpolar} - TS$$

The $E_{MM}$ represents the gas phase molecular mechanic contributions calculated using force field definition. $G_{polar}$ is the polar solvation term obtained by numerically solving either the Poisson-Boltzmann or the generalised Born equations. The non-polar solvation term ($G_{non\text{-}polar}$), on the other hand, is calculated using a linear relation involving solvent accessible surface area (SASA) in the form as

$$G_{PB\text{-}nonpolar} = \gamma * SASA + b$$

with $\gamma$ and $b$ taking different values according to radii applied to atoms within the system.

The entropic contribution ($-TS$) has been suggested important for estimating the absolute affinities. However, it has also been a debatable element to be included within the MM-PB/GBSA calculation. It was suggested that similar entropic contributions would likely result from similar ligand structures targeting the same receptor.[14–16] Oehme et al., in a later study, further confirmed this entropic indiscrimination also extended to ligand with big structural variations.[17] Moreover, the high computational expenses will very likely impair the efficiency of virtual screening. Hence, the entropic term is not included in this work.

The Perl script *mm_pbsa.pl* provided by AMBER suite was applied for MM-PB/GBSA single-framed calculation.[6] The *sander pbsa* program was used to calculate the electrostatic contribution by solving the Poisson-Boltzmann equation. The external dielectric value was set to 80.0. The rescoring processes were carried out with multiple internal/solute dielectric values as 1, 2, 4, and 6, respectively. Dielectric value for molecular mechanic (MM) part of energy evaluation was also adjusted accordingly. The ionic strength of 150 mM was used. The Bondi radii were applied for MM-PBSA calculation and missing or wrong radii were added or corrected.[18,19] The mbondi2 set of radii were applied for the MM-GBSA calculations using the GB$^{OBC}$ model ($igb = 5$).[20] The $\gamma$ and $b$ values were selected as 0.00542 and 0.92 for PB and 0.005 and 0.0 for GB.

## 10.5 Data analysis

For VS experiments performed for PFKFB3 study in Chapter 7, data analysis only included the final ranking of ligands according to the lowest calculated binding free energy scores from all the rescored poses. However, for VS assessment performed in Chapter 11, multiple analysis were included after VS and MM-PBSA rescoring steps.

Receiver operating characteristic area under curve(ROC AUC) value is one common type of parameter to assess the enrichment performance of a VS method. The ROC curve is widely applied for estimate diagnostic ability of a binary classifier system. It is generated by plotting the true positive rate (TPR) against the false positive rate (FPR) at continuous threshold settings, where

$$TPR = \frac{\sum True\ Positive}{\sum Condition\ Positive}$$

and

$$FPR = \frac{\sum False\ Positive}{\sum Condition\ Negative}$$

In ROC plots, a random guess with no discrimination between true positive and true negative would give a diagonal line from left bottom to the top right corner. (Figure 1) The best (or perfect) prediction of the classifier system is the (0,1) point, representing no false positives and no false negatives were detected. Points above the diagonal lines suggesting a better classification than random guess while below indicate a worsened performance. Accordingly, the area under curve (AUC) of the ROC can be classified as perfect (when ROC AUC = 1), better (ROC AUC $> 0.5$), or worse (ROC AUC $< 0.5$) than a random guess.

The ROC AUC calculation was programmed both in C++11 and python3 standards depending on application scenario. In both versions, the TPR and FPR were calculated by moving down one at each threshold in the final ranking list of the screened ligands. The ROC AUC values were then derived by calculating the total area of

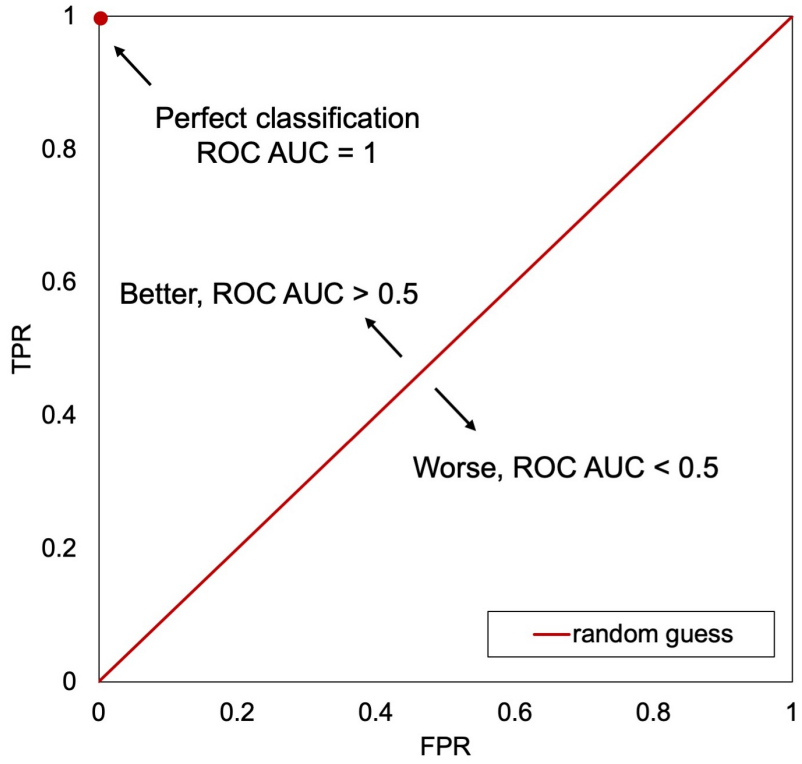



**Figure 10.2:** A graphic representation of the ROC space.

trapezoids, each formed by the two TPR values enclosing FPR step. This AUC calculation was considered accurate and efficient since only relatively small numbers of ligands in the screened libraries were involved. No curve fitting was performed in the AUC calculation to introduce additional errors. Moreover, no differences in values were observed between the two programmed versions. Hence, the ROC AUC analysis was applied consistently in all the testing systems.

Furthermore, the enrichment performance of total non-polar and polar calculated energies consisting in the total MM-PB/GBSA calculated energies were also considered separately. The total MM-PB/GBSA calculated energies can be represented as:

$$\Delta G_{total} = \Delta G_{polar} + \Delta G_{non\text{-}polar}$$

where

$$\Delta G_{polar} = \Delta G_{ele(gas\ phase)} + \Delta G_{polar(PB/GB)}$$

$$\Delta G_{non\text{-}polar} = \Delta G_{non\text{-}polar(gas\ phase)} + \Delta G_{non\text{-}polar(PB/GB)}$$

The energies were calculated by combining the gas phase and the corresponding PB/GB terms. This decomposition process gives the non-polar, electrostatic-polar, and total calculated energies from MM-PBSA estimation in the later enrichment analyses.

To estimate the distributions of calculated energies among active and inactive/decoy ligands, the kernel density estimation (KDE) was adopted. KDE provides a non-parametric estimation of the probability density function based on a finite data sample. This renders the method suitable to represent a rough binding free energy score distribution of the finite numbers of actives and inactives/decoys in this thesis. One parameter within KDE, called the bandwidth $h$, is important in the estimation process and in general, should be optimised. Too small or too big the bandwidth would lead to an undersmoothed or oversmoothed outcome in density estimation (Figure 10.3). However, in this thesis, the accuracy of the KDE method was not desired as the analysis was applied to provide a more graphical illustration of bind-

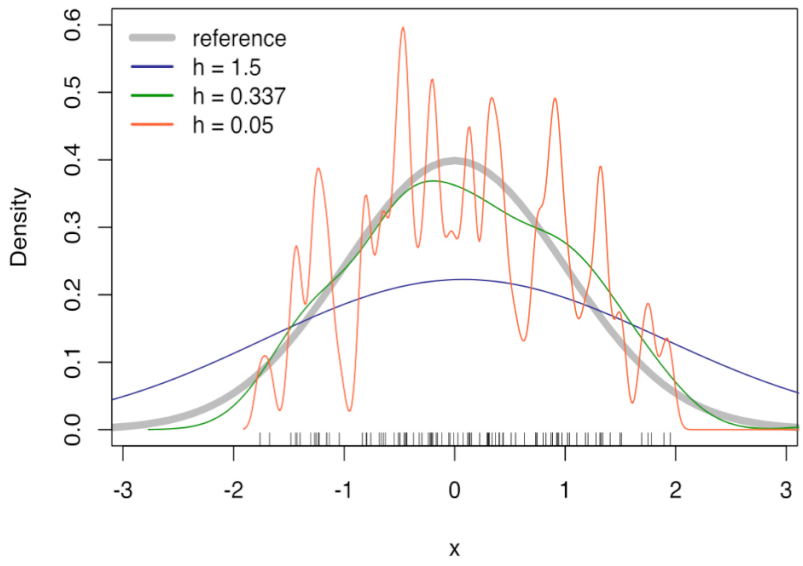



**Figure 10.3:** Kernel density estimation of 100 normally distributed random numbers using different smoothing bandwidths. When bandwidth h=0.05, KDE provides an undersmoothed estimation (orange). While h = 1.5, an oversmoothed density estimation is shown in blue. The optimal bandwidth is 0.337 as the estimation is close to the true density (green).

ing free energy score distribution differences between actives and inactives/decoys. Hence, an oversmoothed bandwidth (0.5) was applied throughout. The analysis was performed using *seaborn.kdeplot* function included in seaborn python data visualization library.[21]

## Chapter references


[1] N.M. O'Boyle, M. Banck, C.A. James, C. Morley, T. Vandermeersch, and G.R. Hutchison. Open Babel: An open chemical toolbox. *J. Cheminformatics*, 3(1):33, 2011.

[2] K. Sommer, N.-O. Friedrich, S. Bietz, M. Hilbig, T. Inhester, and M. Rarey. Unicon: A powerful and easy-to-use compound library converter. *J. Chem. Inf. Model.*, 56(6):1105–1111, 2016.

[3] G.M. Morris, R. Huey, W. Lindstrom, M.F. Sanner, R.K. Belew, D.S. Goodsell, and A.J. Olson. Autodock4 and autodocktools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.*, 30(16):2785–2791, 2009.

[4] O. Korb, T. Stützle, and T.E. Exner. An ant colony optimization approach to flexible protein–ligand docking. *Swarm Intell.*, 1(2):115–134, 2007.

[5] O. Korb, T. Stüttzle, and T.E. Exner. Empirical scoring functions for advanced protein-ligand docking with plants. *J. Chem. Inf. Model.*, 49(1):84–96, 2009. doi: 10.1021/ci800298z.

[6] D. A Case, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, D. Ghoreishi, M.K. Gilson, H. Gohlke, A.W. Goetz, D. Greene, R Harris, N. Homeyer, S. Izadi, A. Kovalenko, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, D.J. Mermelstein, K.M. Merz, Y. Miao, G. Monard, C. Nguyen, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R. Qi, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C.L. Simmerling, J. Smith, R. Salomon-Ferrer, J. Swails, R.C. Walker, J. Wang, H. Wei, R.M. Wolf, X. Wu, L. Xiao, D.M. York, and P.A. Kollman. Amber 2018, 2018.

[7] Ole Tange. GNU Parallel 2018. 2018. URL `https://zenodo.org/record/1146014`.

[8] A. Jakalian, D.B. Jack, and C.I. Bayly. Fast, efficient generation of high-quality atomic charges. am1-bcc model: Ii. parameterization and validation. *J. Comput. Chem.*, 23(16):1623–1641, 2002.

[9] J. Wang, R.M. Wolf, J.W. Caldwell, P.A. Kollman, and D.A. Case. Development and testing of a general amber force field. *J. Comput. Chem.*, 25(9):1157–1174, 2004.

[10] J.A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. Hauser, and C. Simmerling. ff14sb: Improving the accuracy of protein side chain and backbone parameters from ff99sb. *J. Chem. Theory Comput.*, 11(8):3696–3713, 2015.

[11] H. Nguyen, D.R. Roe, and C. Simmerling. Improved generalized born solvent model parameters for protein simulations. *J. Chem. Theory Comput.*, 9(4):2020–2034, 2013.

[12] I. Massova and P.A. Kollman. Combined molecular mechanical and continuum solvent approach (mm-pbsa/gbsa) to predict ligand binding. *Perspect. Drug Discovery Des.*, 18(1):113–135, 2000.

[13] Samuel G. and Ulf R. The mm/pbsa and mm/gbsa methods to estimate ligand-binding affinities. *Expert Opin. Drug Discovery*, 10(5):449–461, 2015.

[14] P.A. Kollman, I. Massova, C. Reyes, B. Kuhn, S. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, O. Donini, P. Cieplak, J. Srinivasan, D.A. Case, and T.E. Cheatham. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc. Chem. Res.*, 33(12):889–897, 2000.

[15] J. Wang, P. Morin, W. Wang, and P.A. Kollman. Use of MM-PBSA in reproducing the binding free energies to HIV-1 RT of TIBO derivatives and predicting the binding mode to HIV-1 RT of efavirenz by docking and MM-PBSA. *J. Am. Chem. Soc.*, 123(22):5221–5230, 2001.

[16] S. Wong, R.E. Amaro, and J.A. McCammon. MM-PBSA Captures Key Role of Intercalating Water Molecules at a Protein-Protein Interface. *J. Chem. Theory Comput.*, 5(2):422–429, 2009.

[17] D.P. Oehme, R.T.C. Brownlee, and D.J.D. Wilson. Effect of atomic charge, solvation, entropy, and ligand protonation state on mm-pb(gb)sa binding energies of hiv protease. *J. Comput. Chem.*, 33(32):2566–2580, 2012.

[18] A Bondi. van der waals volumes and radii. *J. Phys. Chem.*, 68(3):441–451, 1964.

[19] S.S. Batsanov. Van der waals radii of elements. *Inorg. Mater.*, 37(9):871–885, 2001.

[20] A. Onufriev, D. Bashford, and D.A. Case. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins Struct. Funct. Bioinf.*, 55(2):383–394, 2004.

[21] M. Waskom, O. Botvinnik, P. Hobson, J.B. Cole, Y. Halchenko, S. Hoyer, A. Miles, T. Augspurger, T. Yarkoni, T. Megies, L.P. Coelho, D. Wehner, Cynddl, E. Ziegler, Diego0020, Y.V. Zaytsev, T. Hoppe, S. Seabold, P. Cloud, M. Koskinen, K. Meyer, A. Qalieh, and D. Allan. seaborn: v0.5.0 (November 2014). 2014. URL `https://doi.org/10.5281/zenodo.12710{#}.XGrwdASZ6IV.mendeley`.

# CHAPTER 11

# MM-PBSA rescoring assessment and potential issues

## 11.1 Rescoring with MM-PBSA method

The MM-PB/GBSA free energy calculation, in essence, was not developed for evaluating docking performance but relative binding free energies of ligands. However, the concept is in principle similar. The method was designed with more complicated parameterisations for molecules, plus the conventional application on trajectories from MD simulations. Hence, it is much more computationally expensive than traditional scoring functions but compensated by improved ranking accuracy.

Since its early debut,[1,2] the method has undergone validations using a wide variety of biomolecular systems.[3–16] The studies were mainly involved binding assessment correlations to experimental data such as binding affinity or $IC_{50}$. Moderate to good outcomes were recorded for both MM-PBSA and MM-GBSA methods. However, the overall performance still varies among different type of receptors. Hou et al. compared the MM-GBSA model to 11 scoring functions, and the model was demonstrated with

improved Spearman correlation coefficients to experimentally determined $IC_{50}$.[7]

Despite the original designing purpose of the MM-PB/GBSA model, its performance in virtual screening has also been investigated in multiple studies.[3,4,11,13,15] The significant improvements in enrichment was determined with dihydrofolate reductase,[4] tyrosine kinases,[11,15] aldose reductase, and factor X[11] comparing to virtual screening using Autodock program.[17,18] In multiple DUD-E[19] protein receptors, the MM-GBSA method also outperforms Autodock Vina (vina hereafter).[13] More interestingly, in a separate study using mineralocorticoid receptor from DUD data set[20], Thompson and co-workers showed that rescoring using molecular mechanic (MM) part of the energy achieved very similar enrichment outcome to using MM-PBSA.[3]

Conventionally, MM-PB/GBSA calculations are carried out by averaging the estimated binding free energy of multiple conformations, typically from an MD simulation trajectory. However, several studies have also suggested that the improvements from calculations on short MD simulation trajectories, comparing to single-framed calculation, was marginal. These works have shown that similar outcomes can be provided by energy minimisation to including short MD simulation in both docking and virtual screening.[4,14,15] It was also suggested that short time-scale MD simulation is required for some systems where the docking process failed to determine the correct binding poses.[7,12]

In addition, as the parameterisation of molecules in MM-PB/GBSA method take into consideration of charging, different partial charge models have also been investigated and some dependency was demonstrated. Multiple atomic partial charge methods (including empirical (Gasteiger), semi-empirical (AM1-BCC), and various restrained electrostatic potential (RESP) charges) were assessed by Oehme et al.. The MM-PB/GBSA binding free energies were calculated for 6 HIV protease inhibitors and the Gasteiger, HF/STO-3G, and B3LYP/cc-pVTZ charging methods provided the best correlations to experimental data.[10] In addition, similar predicted relative affinity using AM1-BCC, RESP, and xAvESP methods was shown with MM-GBSA rescoring in cucuerbit[8]uril and $\alpha$-cyclodextrin systems.[9] The RESP charges with HF-SCF/6-

31G* level of calculation was also resulted as the optimal method across 5 different biomolecular systems, while AM1-BCC and ESP also performed competently.[12]

One parameter within the MM-PB/GBSA model, the solute dielectric constant, has also drawn some attention. Using ligands targeting 7 classes of proteins, Yang and co-workers showed that increasing internal dielectric value can lead to improved correlations to experimental data.[8] Agreeable results were also demonstrated by Sun et al. in a work later in both docking and virtual screening among three tyrosine kinases.[15] However, different systems were suggested to have a varied preference of optimal internal dielectric values.[15]

Considering explicit solvation has also become a more recent attempt to improve MM-PB/GBSA calculations. Maffucci and co-workers developed a workflow, namely Nwat-MM/GBSA, by including water molecules close to the ligands in short MD simulation trajectories of ligand-receptor complexes.[16,21] The assessment included several systems, such as penicillopepsin, HIV1-protease, BCL-XL, Rac1, and $\beta$-lactamase. The studies demonstrated improvements in both experimental correlations and virtual screening using Nwat-MM/GBSA to standard MM-GBSA rescoring.[16,21]

In this chapter, 3 test sets were firstly generated to assess the performance of MM-PB/GBSA method included in the AMBER package in VS rescoring. To ensure a relatively thorough assessment, the test sets consist of protein receptors from different categories and ligand libraries of different sources. Some receptors were also repeatedly appeared in the test sets to assess variation caused by using different libraries. The test set underwent VS and the generated poses were rescored using MM-PB/GBSA with the applications of different internal dielectric values. The quality of VS performance was evaluated according to the capability of the rescoring method to distinguish actives from inactives or decoys. Comparisons in enrichment outcomes were performed for the polar-electrostatic contributions, the non-polar energetics, and the total binding free energy. Two popular and computationally efficient charging methods (Gasteiger and AM1-BCC) were also investigated for their influences on the electrostatics. The outcomes provide further insights in applications using the

MM-PB/GBSA model for VS rescoring.

## 11.2 Methodology

### 11.2.1 Test set preparations

Test sets used for different assessments in this chapter are listed in Table 11.1. Since previous studies regarding MM-PBSA rescoring were mainly compared to Autodock[11,15] or vina,[26] the comparison in performances of the docking program used in this thesis, PLANTS[24,25], and vina was performed. The purpose was to guarantee that PLANTS should at least perform as good as vina in reproducing crystal poses of a set of benchmarked complexes. The test set adopted, the test set 0, was derived from the CCDC-Astex "clean" set.[22] The MOE software[23] was used to pre-process all the complexes. Missing loops with more than 4 amino acids were capped with acetyl or N-methyl amide groups. Otherwise, the loops were rebuilt using MOE software[23] during the structure preparation process. The missing loops were visually inspected to ensure the capping would not affect the binding site. All spliced termini were also capped to avoid effects from artificial charges. The part of the complexes within 4.5 Å range from the ligand was minimised briefly with MOE (AMBER 10EHT force field and the Born implicit solvent).[23] The ligands were then stripped from the binding sites. Using PLANTS[24,25] or vina[26], the original ligands were docked back into the pockets for 12 repeated experiments and the binding poses were analysed according to its RMSD to the original crystal poses.

The rest 3 test sets were generated for virtual screening assessments. In the test set 1, the receptors were selected from DUD-E database[19] according to the number of true inactive ligands available; if there are more than 100 inactive ligands available (except HIVPV which have an inactive count of 96) the receptor is included. Moreover, membrane proteins or receptors have metal presented within the binding pocket were excluded. The ligand libraries consist of the active and inactive compounds from DUD-E database[19]. The SMILES format were downloaded. 3D structures were

generated and minimised using MOE.[23]

The ChEMBL database[29] is the source of active and inactive ligands of the receptor proteins in the test set 2. Actives and inactives were both extracted according to the experimental data. The inhibition constant ($K_i$) values are more connected to the binding affinities of the compounds to the receptors thus were considered as the priority factor. Some compounds are flagged as inactive or the results were shown as too large that exceeds the typical detectable range of the experimental method of choice. These are evidence showing they are likely not binders to the receptor of interest. Hence, they are categorised as inactives in the ligand libraries. Furthermore, the total number of inactive compounds for each receptor was one of the criteria for receptor selection in the test set 2 and was set at a minimum of 200 inactive ligands after clustering processing of the library with MOE.[23] This is to allow a statistically meaningful active-to-inactive ratio no larger than 1:10 and at least 10 and no more than 20 active compounds can be considered in the final library. However, for some receptors, such as CTSD and PPARG, the inactives derived from the $K_i$ lists has less than 200 compounds. Thus for these two receptors, the ligands with

**Table 11.1:** Test sets applied for the MM-PBSA rescoring assessment.

| Index | Size | Source | Protein contained (PDB ID/gene name) |
|---|---|---|---|
| 0 | 73 | The CCDC-Astex "clean" set for pose prediction validation | 1a28, 1a4q, 1abe, 1abf, 1aoe, 1apu, 1aqw, 1atl, 1bma, 1byb, 1c5c, 1c5x, 1c83, 1cbs, 1cil, 1cle, 1d0l, 1d3h, 1ejn, 1eta, 1f3d, 1flr, 1glp, 1glq, 1hfc, 1hpv, 1hsb, 1hsl, 1hvr, 1hyt, 1ida, 1jap, 1kel, 1lcp, 1lic, 1lst, 1mld, 1mmq, 1mrg, 1mrk, 1mts, 1nco, 1ppc, 1pph, 1qbr, 1rnt, 1rob, 1slt, 1snc, 1srj, 1tmn, 1tng, 1tnh, 1tni, 1tnl, 1tyl, 1ukz, 1xid, 1xie, 2ak3, 2cmd, 2ctc, 2fox, 2gbp, 2h4n, 2qwk, 2tmn, 3cla, 3ert, 4dfr, 5abp, 6rnt, 7tim |
| 1 | 18 | Active and inactive ligands from DUD-E | ACES, CDK2, DPP4, EGFR, ESR1, FA10, FGFR1, FNTA, GRIA2, GRIK1, HIVPR, KPCB, LCK, PTN1, SRC, THRB, TRY1, VGFR2 |
| 2 | 19 | Clustered active and inactive ligands from ChEMBL | ACES, ADORA1, AURKA, CA1, CASP1, CDK2, CTSD, EGFR, ELANE, ESR1, FGFR1, GSK3B, LCK, MAPK1, MMP1, NTRK1, PPARG, SRC, VGFR2 |
| 3 | 18 | Same proteins as in test set 1 but using the 50 best active compounds and a drug-like 1000 decoys set[a] | Same as test set 1 |

[a] A decoys set from previous studies[27,28].

determined $IC_{50}$ were also considered. All ligand libraries were filtered with molecular weights between 250 and 550. Clustering of the compound libraries was performed to remove repetitions and compounds that are too similar to each other. The clustering procedures were carried out using the Tanimoto superset/subset similarity metrics on the bit-packed MACCS structural fingerprints included in the MOE package.[23]

Same receptors from the test set 1 were included in the test set 3 as well. Libraries in this test set included a 1000-decoy set with an average molecular weight of 400 from PDB bank applied in a previous study[27,28], plus the 50 most active compounds of each receptor from DUD-E database[19]. The decoys from the DUD-E database[19] was not considered due to the size of the libraries require more computational time. Moreover, the universal decoy application across different receptors would emphasise more on the scoring performance under a more generalised circumstance.

The active ligands, on the other hand, were all selected as the $n$-most active compounds of the receptor of interest from the processed active ligand list. The $n$ here should satisfy the active-to-inactive ratio accordingly. The 3D structures were generated from SMILES and minimised using MOE.[23]

All receptor crystal structures were downloaded from the Protein Data bank.[30] Any water molecule, salt, and co-crystal reagent were removed for all test sets. Additionally, test set 1-3 contains receptors without metals participating in ligand binding. Hence, all metals were also removed for receptors included in these three test sets. Through visual inspection, all post-translational modifications were determined far from the binding site to be participating in direct ligand-receptor interaction. Hence, all PTMs were also removed. With the concerns of compatibility with AMBER package[31], some amino acid residue names were also corrected. The *Protonate 3D* function within the MOE package[23] was applied to add hydrogen atoms to the receptor structures. The binding sites were also visually inspected to confirm the corrected placement of the hydrogens. The PLANTS program[24,25] was used to determine the centres of the binding sites. An additional 2 Å was added to the half of the longest intramolecular atom-to-atom distance determined from the processed screening library

and applied as the radii of the docking search sphere.

### 11.2.2 Molecular weight adjustment procedures

In a later analysis of the rescoring results, the molecular weight (MW) composition of the test set 2 and 3 were adjusted. This process is to lower the biases caused by the differences in molecular size distribution between the active and inactive ligands (or decoys). The bias was not initially excluded since assessing MW-biased systems is also important. Combining both biased and unbiased studies can provide a more complete picture of different scenarios in virtual screening. Additionally, this processing procedure was not applied to test set 1 since the sizes of the libraries are relatively small. The further shrinking of the library would make the statistical analysis meaningless. Some receptors form test set 2 were also not included for a similar reason.

The procedure focused on lowering the ROC AUC values of the high-to-low MW ligand rankings to about 0.5 or lower. Firstly, the active ligand MW range was determined. The 30 and 100 percentiles of this range were used as the lower and upper bound of the MW cut-offs for inactive ligands. The inactive ligands with MW outside of the cut-offs were then excluded in the new adjusted library. The size of the new inactive list was then determined. Receptors with the inactive ligand size less than 70 were not included in later analysis. The sizes of the active ligands in the new MW-unbiased lists were then determined according to the size of the new inactive lists to give active-to-inactive ratios around 1:10 or 1:20. From the original active list, random active ligands with the determined sizes were selected. Final results in a later section have shown that this procedure can effectively reduce the MW ROC AUC to about 0.5 or less.

## 11.3 Results and discussions

### 11.3.1 Validation of PLANTS performance

It is a crucial requirement for the docking software to be able to provide satisfactory binding poses for MM-PBSA rescoring. Hence, PLANTS program[24,25] was firstly assessed for pose sampling in terms of reproducing experimental binding poses. 73 crystal complexes from the CCDC-Astex "clean" set[22] were used as the test set and the performance was tested against vina.[26] Some crystal structure also contains multiple monomers. The monomers were also considered separately. This resulted in a total of 97 binding sites for the test set 0. Moreover, exhaustiveness values at both 10 and 12 were also tested.

Results have shown that overall PLANTS performs significantly better than vina in reproducing crystal poses. When considering the top-ranked conformation, PLANTS can reproduce 87.2% of the crystal ligand binding conformations (RMSD $<$ 2.0 Å). Vina,[26] one the other hand, provided the same 62.9% success rate for both exhaustiveness 10 and 12. The detailed per-receptor comparisons are shown in Figure 11.1.

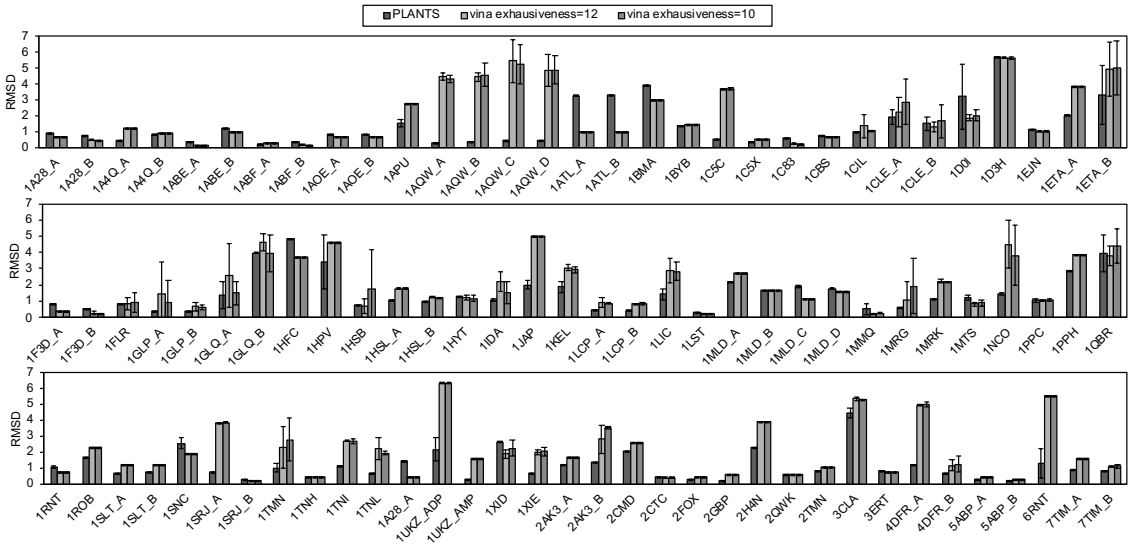



**Figure 11.1:** Comparison of pose predictions between PLANTS and vina (with exhaustiveness = 12 and 10). Only the top 1st ranked poses were considered. The error bars were shown as the standard deviation between the 12 docking experiments.

PLANTS also gave an impressive 95.1% reproducibility when including the top-three conformations. Hence, in the following rescoring assessments the top three poses generated by PLANTS for each protonation or tautomerisation states were applied.

### 11.3.2 Rescoring using different energy compositions from MM-PBSA

Previous studies have suggested that solute dielectrics could have an impact on estimated binding free energies using MM-PB/GBSA methods; increasing the value could improve the outcome likely due to the change in electrostatic calculations.[8,15] To investigate the influence of solute / internal dielectrics on virtual screening performance in this study, the calculated MM-PBSA score was evaluated according to total electrostatic-polar energies, total non-polar energies, and total MM-PBSA energies and the ROC AUC values were considered separately. By changing the internal dielectric values, the effects on ROC AUC values were demonstrated in the form of squared Pearson correlation coefficient ($r_p{}^2$) and linear fitting of higher internal dielectric values to internal dielectric equals 1 (Figure 11.2A-F). Both electrostatic-polar and non-poler energies provide near perfect linear correlations across different solute dielectric values. However, the correlations drastically worsen when considering the total MM-PBSA score between different internal dielectrics(Figure 11.2G-I).

Further information can be interpreted from direct comparisons of the ROC AUC values.(Tables 11.2 to 11.4) Non-polar interactions give over-0.7 ROC AUC at 61.1 - 66.7% rate across all the internal dielectric tested (Table 11.2), much better comparing to the 44.4% when using $PLANTS_{chemplp}$ scoring methods.[24,25] A better performance of non-polar energies was observed for test set 2, 73.7 - 78.9% of the libraries resulted in ROC AUC over 0.7. This rate is 47.4% using $PLANTS_{chemplp}$ in test set 2(Table 11.3). However, the over-0.7 ROC AUC rates are much worse when only consider electrostatic and polar interactions; the rates are merely 5.6% and 15.79% for test set 1 and 2, respectively. This trend does not vary when change the internal dielectric values. Furthermore, the ROC AUC results remain similar among non-polar or electrostatic-polar energies even different solute dielectric values were

applied.

As for the total MM-PBSA scores, the variations increase when different internal dielectrics were applied. For the test set 1 and 2, the results are much worse when solute dielectric equals to 1. When the internal dielectric increases, the ROC AUC

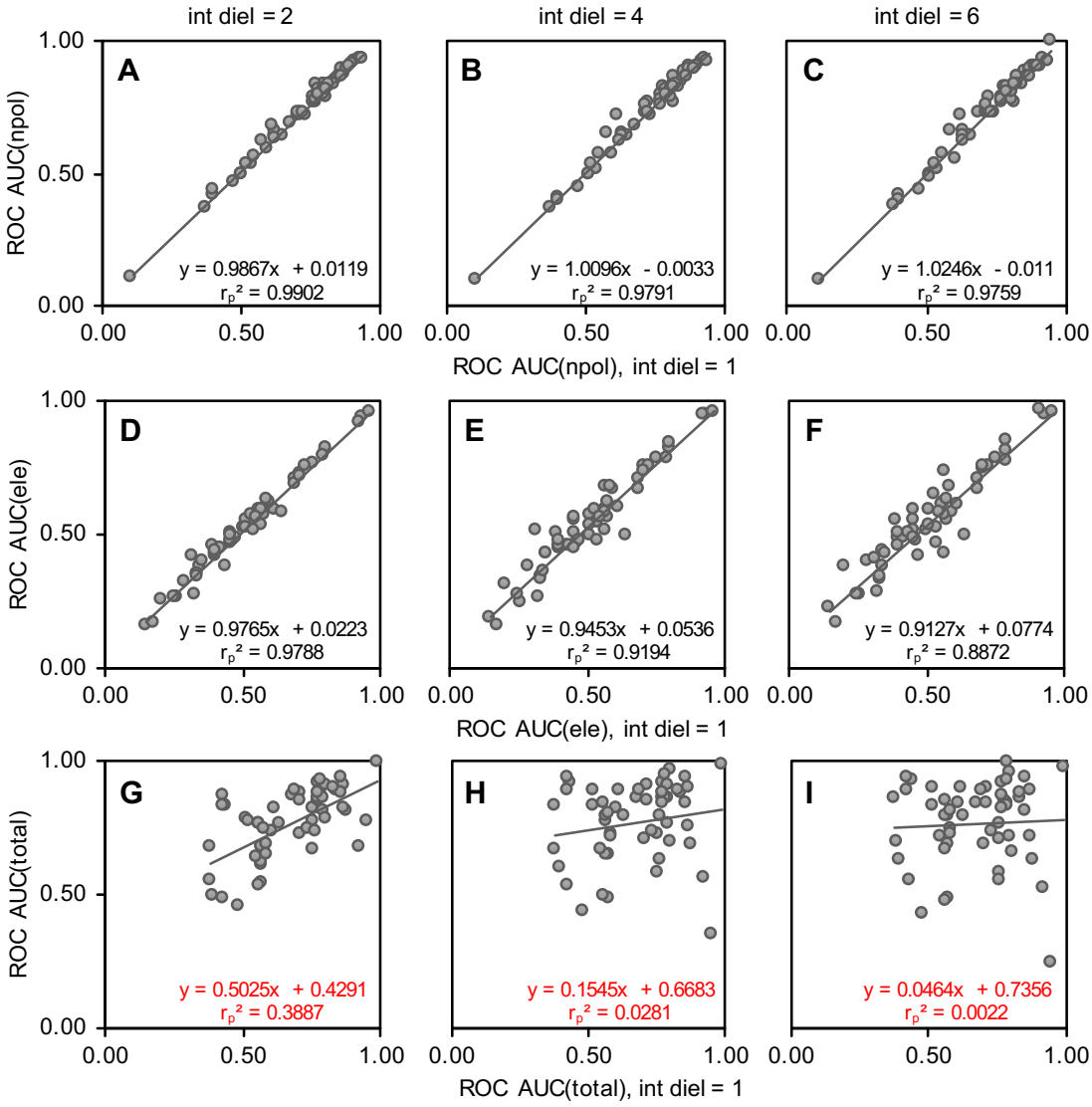



**Figure 11.2:** The correlations of ROC AUC values using internal dielectrics 2, 4, and 6 to the ones calculated with the internal dielectric value of 1 with outcomes from all test sets combined. Results from different energy compositions of MM-PBSA calculations are shown in separate rows. AM1-BCC was the charging method. A-C: the ROC AUCs calculated from non-polar energies. D-F: the ROC AUCs calculated from electrostatic and polar energies. G-I: the ROC AUCs calculated from total MM-PBSA binding free energies.

**Table 11.2:** The ROC AUC results using different internal dielectric values with MM-PBSA virtual screening on test set 1 (DUD-E test set). Squared Pearson correlation coefficients to results from internal dielectric = 1 are also shown. AM1-BCC was the charging method.

| Gene ID | ROC AUC | ROC AUC, internal dielectric = 1 | | | ROC AUC, internal dielectric = 2 | | | ROC AUC, internal dielectric = 4 | | | ROC AUC, internal dielectric = 6 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | (PLANTS) | non-polar | Ele + polar | total | non-polar | Ele + polar | total | non-polar | Ele + polar | total | non-polar | Ele + polar | total |
| ACES | 0.8656 | 0.9209 | 0.2008 | 0.4384 | 0.9175 | 0.2464 | 0.8284 | 0.9253 | 0.3071 | 0.9172 | 0.9283 | 0.3730 | 0.9300 |
| CDK2 | 0.3931 | 0.5136 | 0.5068 | 0.5707 | 0.5000 | 0.5170 | 0.5419 | 0.4949 | 0.4881 | 0.4836 | 0.4921 | 0.5130 | 0.4802 |
| DPP4 | 0.6807 | 0.7756 | 0.4684 | 0.5704 | 0.7873 | 0.4676 | 0.6766 | 0.7929 | 0.4657 | 0.7688 | 0.7877 | 0.4736 | 0.7877 |
| EGFR | 0.6883 | 0.6361 | 0.5403 | 0.7052 | 0.6408 | 0.5506 | 0.7262 | 0.6500 | 0.5375 | 0.7049 | 0.6562 | 0.5177 | 0.6833 |
| ESR1 | 0.5498 | 0.4050 | 0.9282 | 0.9180 | 0.4191 | 0.9338 | 0.6793 | 0.4123 | 0.9480 | 0.5600 | 0.4202 | 0.9446 | 0.5181 |
| FA10 | 0.6149 | 0.8185 | 0.5128 | 0.7867 | 0.7904 | 0.5467 | 0.8820 | 0.7705 | 0.5683 | 0.8611 | 0.7657 | 0.5920 | 0.8367 |
| FGFR1 | 0.8009 | 0.9364 | 0.1487 | 0.4281 | 0.9320 | 0.1531 | 0.8704 | 0.9271 | 0.1791 | 0.9344 | 0.9247 | 0.2167 | 0.9340 |
| FNTA | 0.6049 | 0.6603 | 0.3380 | 0.3945 | 0.6404 | 0.3508 | 0.4930 | 0.6428 | 0.3351 | 0.5962 | 0.6387 | 0.3304 | 0.6235 |
| GRIA2 | 0.7128 | 0.8498 | 0.3940 | 0.6073 | 0.8417 | 0.4485 | 0.7330 | 0.8455 | 0.4972 | 0.8192 | 0.8578 | 0.5455 | 0.8440 |
| GRIK1 | 0.4840 | 0.6026 | 0.4019 | 0.4287 | 0.5935 | 0.4137 | 0.4867 | 0.5694 | 0.4454 | 0.5327 | 0.5572 | 0.4778 | 0.5473 |
| HIVPR | 0.7766 | 0.8760 | 0.1771 | 0.4281 | 0.8813 | 0.1604 | 0.8281 | 0.8906 | 0.1573 | 0.8854 | 0.8802 | 0.1656 | 0.8896 |
| KPCB | 0.7394 | 0.6858 | 0.5965 | 0.7586 | 0.6880 | 0.6200 | 0.8231 | 0.6802 | 0.6658 | 0.7913 | 0.7246 | 0.5817 | 0.7651 |
| LCK | 0.9105 | 0.8811 | 0.5313 | 0.7705 | 0.8762 | 0.5705 | 0.9139 | 0.9026 | 0.5841 | 0.9188 | 0.8992 | 0.6448 | 0.9207 |
| PTN1 | 0.6649 | 0.7197 | 0.5145 | 0.5724 | 0.7224 | 0.5184 | 0.6044 | 0.7281 | 0.5272 | 0.6496 | 0.7241 | 0.5333 | 0.6851 |
| SRC | 0.5334 | 0.7744 | 0.3477 | 0.5630 | 0.7648 | 0.3743 | 0.7602 | 0.7637 | 0.4219 | 0.7929 | 0.7682 | 0.4289 | 0.7894 |
| THRB | 0.8819 | 0.8206 | 0.5750 | 0.7741 | 0.8126 | 0.5915 | 0.8479 | 0.8231 | 0.5619 | 0.8598 | 0.8354 | 0.5455 | 0.8538 |
| TRY1 | 0.3860 | 0.4059 | 0.5662 | 0.4841 | 0.4370 | 0.5303 | 0.4498 | 0.4043 | 0.5909 | 0.4338 | 0.3971 | 0.6132 | 0.4226 |
| VGFR2 | 0.7983 | 0.7797 | 0.5629 | 0.7862 | 0.7706 | 0.5900 | 0.8853 | 0.7762 | 0.5865 | 0.8617 | 0.7827 | 0.6051 | 0.8330 |
| % (> 0.7) | 44.44 | 61.11 | 5.6 | 38.89 | 61.11 | 5.6 | 61.11 | 61.11 | 5.6 | 66.67 | 66.67 | 5.6 | 61.11 |
| $r_p^2$ to data from ROC AUC, internal dielectric = 1 | | | | | 0.9941 | 0.9860 | 0.1962 | 0.9886 | 0.9539 | 0.0214 | 0.9811 | 0.8944 | 0.0029 |

**Table 11.3:** The ROC AUC results using different internal dielectric values with MM-PBSA virtual screening on test set 2 (ChEMBL test set). Squared Pearson correlation coefficients to results from internal dielectric = 1 are also shown. AM1-BCC was the charging method.

| Gene ID | ROC AUC | ROC AUC, internal dielectric = 1 | | | ROC AUC, internal dielectric = 2 | | | ROC AUC, internal dielectric = 4 | | | ROC AUC, internal dielectric = 6 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | (PLANTS) | non-polar | Ele + polar | total | non-polar | Ele + polar | total | non-polar | Ele + polar | total | non-polar | Ele + polar | total |
| ACES | 0.7934 | 0.7762 | 0.2904 | 0.3824 | 0.7720 | 0.3194 | 0.5533 | 0.7575 | 0.3749 | 0.6651 | 0.7721 | 0.3909 | 0.6929 |
| ADORA1 | 0.7117 | 0.6334 | 0.6180 | 0.7560 | 0.6618 | 0.5896 | 0.7733 | 0.6432 | 0.6008 | 0.7263 | 0.6436 | 0.6068 | 0.7005 |
| AURKA | 0.4995 | 0.7273 | 0.5685 | 0.7724 | 0.7320 | 0.5740 | 0.8558 | 0.7672 | 0.5065 | 0.8371 | 0.7854 | 0.4280 | 0.8242 |
| CA1 | 0.4385 | 0.5865 | 0.5362 | 0.5680 | 0.6247 | 0.5102 | 0.6199 | 0.6481 | 0.4700 | 0.6433 | 0.6594 | 0.4594 | 0.6645 |
| CASP1 | 0.9250 | 0.8580 | 0.9186 | 0.9919 | 0.8557 | 0.9148 | 0.9945 | 0.8860 | 0.9485 | 0.9837 | 0.8848 | 0.9640 | 0.9710 |
| CDK2 | 0.6915 | 0.4792 | 0.5712 | 0.5607 | 0.4714 | 0.5662 | 0.5336 | 0.4491 | 0.5789 | 0.4930 | 0.4424 | 0.5707 | 0.4702 |
| CTSD | 0.9270 | 0.9067 | 0.7076 | 0.8013 | 0.9139 | 0.7190 | 0.9070 | 0.9060 | 0.7555 | 0.9652 | 0.9053 | 0.7462 | 0.9532 |
| EGFR | 0.6780 | 0.7270 | 0.4228 | 0.5832 | 0.7220 | 0.4383 | 0.6837 | 0.7301 | 0.4528 | 0.7280 | 0.7264 | 0.4843 | 0.7418 |
| ELANE | 0.8623 | 0.8230 | 0.4557 | 0.8376 | 0.8361 | 0.4597 | 0.8924 | 0.8340 | 0.5002 | 0.8742 | 0.8347 | 0.5056 | 0.8620 |
| ESR1 | 0.5635 | 0.5437 | 0.7971 | 0.8781 | 0.5381 | 0.8057 | 0.8154 | 0.5204 | 0.8181 | 0.6847 | 0.5174 | 0.8064 | 0.6286 |
| FGFR1 | 0.5270 | 0.8731 | 0.3271 | 0.5169 | 0.8942 | 0.2704 | 0.7862 | 0.8967 | 0.2566 | 0.8895 | 0.8947 | 0.2815 | 0.8986 |
| GSK3B | 0.7151 | 0.7402 | 0.3997 | 0.5497 | 0.7186 | 0.4204 | 0.6370 | 0.7148 | 0.4503 | 0.6698 | 0.7281 | 0.4536 | 0.7160 |
| LCK | 0.5812 | 0.8637 | 0.4596 | 0.8697 | 0.8665 | 0.5046 | 0.9040 | 0.8620 | 0.5496 | 0.8952 | 0.8644 | 0.5487 | 0.8847 |
| MAPK1 | 0.7449 | 0.8244 | 0.2564 | 0.3799 | 0.8250 | 0.2590 | 0.6791 | 0.8680 | 0.2397 | 0.8285 | 0.8666 | 0.2706 | 0.8577 |
| MMP1 | 0.8578 | 0.8119 | 0.4775 | 0.5836 | 0.8002 | 0.4787 | 0.6514 | 0.7851 | 0.4692 | 0.7132 | 0.7804 | 0.4130 | 0.7204 |
| NTRK1 | 0.2840 | 0.7916 | 0.3353 | 0.5206 | 0.7918 | 0.3437 | 0.7710 | 0.8120 | 0.3245 | 0.8258 | 0.8144 | 0.3374 | 0.8269 |
| PPARG | 0.9202 | 0.8073 | 0.5774 | 0.7129 | 0.8315 | 0.5726 | 0.8823 | 0.8081 | 0.6169 | 0.9089 | 0.8089 | 0.6234 | 0.8984 |
| SRC | 0.5956 | 0.7819 | 0.4383 | 0.6790 | 0.8358 | 0.3750 | 0.8657 | 0.8275 | 0.4527 | 0.8561 | 0.8214 | 0.5039 | 0.8444 |
| VGFR2 | 0.6354 | 0.6200 | 0.6411 | 0.7365 | 0.6759 | 0.5810 | 0.7436 | 0.7167 | 0.4916 | 0.7377 | 0.7167 | 0.4929 | 0.7320 |
| % (> 0.7) | 47.37 | 73.68 | 15.79 | 47.37 | 73.68 | 15.79 | 63.16 | 78.95 | 15.79 | 73.68 | 78.95 | 15.79 | 78.95 |
| $r_p^2$ to data from ROC AUC, internal dielectric = 1 | | | | | 0.9838 | 0.9853 | 0.8025 | 0.9627 | 0.9446 | 0.4606 | 0.9599 | 0.9165 | 0.3299 |

in overall improves. This observation indicates that the variation is likely due to the change in electrostatic and polar energy estimations. However, the universal poor performance of electrostatic-polar energies in distinguishing actives from inactive ligands is demonstrated by the almost perfect linear correlations of ROC AUC values between different internal dielectric values. Hence, the change in electrostatic and polar energy estimations are less likely to provide qualitative improvement that contributing to the better performance in total MM-PBSA score when increasing the

**Table 11.4:** The ROC AUC results using different internal dielectric values with MM-PBSA virtual screening on test set 3 (decoy test set). Squared Pearson correlation coefficients to results from internal dielectric = 1 are also shown. AM1-BCC was the charging method.

| Gene ID | ROC AUC | ROC AUC, internal dielectric = 1 | | | ROC AUC, internal dielectric = 2 | | | ROC AUC, internal dielectric = 4 | | | ROC AUC, internal dielectric = 6 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | (PLANTS) | non-polar | Ele + polar | total | non-polar | Ele + polar | total | non-polar | Ele + polar | total | non-polar | Ele + polar | total |
| ACES | 0.7934 | 0.8414 | 0.3996 | 0.5611 | 0.8339 | 0.4319 | 0.7610 | 0.8278 | 0.4728 | 0.8420 | 0.8339 | 0.5035 | 0.8497 |
| CDK2 | 0.7117 | 0.5148 | 0.6910 | 0.7605 | 0.4955 | 0.7058 | 0.7353 | 0.4927 | 0.6993 | 0.6241 | 0.4849 | 0.7051 | 0.5751 |
| DPP4 | 0.4995 | 0.5278 | 0.7542 | 0.7939 | 0.5369 | 0.7655 | 0.8080 | 0.5305 | 0.7802 | 0.7605 | 0.5362 | 0.7825 | 0.7140 |
| EGFR | 0.4385 | 0.5576 | 0.6932 | 0.8026 | 0.5633 | 0.6865 | 0.7797 | 0.5780 | 0.6692 | 0.6935 | 0.5776 | 0.6692 | 0.6550 |
| ESR1 | 0.9250 | 0.6303 | 0.7971 | 0.8687 | 0.6282 | 0.8145 | 0.8231 | 0.6243 | 0.8423 | 0.7502 | 0.6244 | 0.8506 | 0.7181 |
| FA10 | 0.6915 | 0.9466 | 0.3157 | 0.7850 | 0.9311 | 0.4109 | 0.9269 | 0.9260 | 0.5122 | 0.9467 | 0.9995 | 0.4063 | 0.9990 |
| FGFR1 | 0.9270 | 0.7190 | 0.4526 | 0.6305 | 0.7261 | 0.4678 | 0.7597 | 0.7604 | 0.4469 | 0.7907 | 0.7561 | 0.4822 | 0.7929 |
| FNTA | 0.6780 | 0.7940 | 0.3391 | 0.5799 | 0.8146 | 0.3389 | 0.7393 | 0.8177 | 0.3525 | 0.8001 | 0.8244 | 0.3718 | 0.8135 |
| GRIA2 | 0.8623 | 0.3792 | 0.7900 | 0.7579 | 0.3743 | 0.7894 | 0.6703 | 0.3754 | 0.7771 | 0.5794 | 0.3818 | 0.7699 | 0.5470 |
| GRIK1 | 0.5635 | 0.1130 | 0.9584 | 0.9460 | 0.1145 | 0.9569 | 0.7680 | 0.1017 | 0.9578 | 0.3480 | 0.0975 | 0.9577 | 0.2391 |
| HIVPR | 0.5270 | 0.8914 | 0.2496 | 0.6165 | 0.8875 | 0.2613 | 0.8163 | 0.8922 | 0.2719 | 0.8890 | 0.8978 | 0.2705 | 0.8978 |
| KPCB | 0.7151 | 0.7330 | 0.7240 | 0.8577 | 0.7247 | 0.7541 | 0.8743 | 0.7308 | 0.7530 | 0.8368 | 0.7266 | 0.7545 | 0.8094 |
| LCK | 0.5812 | 0.8281 | 0.5679 | 0.8323 | 0.8368 | 0.5912 | 0.8939 | 0.8345 | 0.6746 | 0.8915 | 0.8342 | 0.7336 | 0.8809 |
| PTN1 | 0.7449 | 0.8117 | 0.7114 | 0.7898 | 0.8171 | 0.7166 | 0.8574 | 0.8103 | 0.7311 | 0.9079 | 0.8076 | 0.7537 | 0.9314 |
| SRC | 0.8578 | 0.7942 | 0.4562 | 0.7124 | 0.7950 | 0.4908 | 0.8458 | 0.7954 | 0.5579 | 0.8471 | 0.8007 | 0.5847 | 0.8407 |
| THRB | 0.2840 | 0.9001 | 0.5860 | 0.8539 | 0.8974 | 0.6280 | 0.9361 | 0.8939 | 0.6771 | 0.9394 | 0.9009 | 0.6742 | 0.9330 |
| TRY1 | 0.9202 | 0.8716 | 0.3527 | 0.6916 | 0.8603 | 0.3962 | 0.8837 | 0.8628 | 0.4206 | 0.8926 | 0.8681 | 0.4268 | 0.8889 |
| VGFR2 | 0.5956 | 0.8206 | 0.5467 | 0.7738 | 0.8158 | 0.5620 | 0.8775 | 0.8282 | 0.5617 | 0.8801 | 0.8297 | 0.5760 | 0.8694 |
| % (> 0.7) | 77.78 | 66.67 | 33.33 | 72.22 | 66.67 | 38.89 | 94.44 | 66.67 | 33.33 | 77.78 | 66.67 | 44.44 | 77.78 |
| $r_p^2$ to data from ROC AUC, internal dielectric = 1 | | | | | 0.9980 | 0.9881 | 0.1120 | 0.9948 | 0.9281 | 0.1169 | 0.9936 | 0.9387 | 0.1528 |

internal dielectric values.

Test set 3 was adopted as an additional set to confirm that the above observations are also true when decoys are included in a library. The electrostatic and polar estimations perform better than the other two test sets, resulted in 33.33 - 44.44% above-0.7-ROC-AUC rate across all the internal dielectrics applied(Table 11.4). This is likely due to the more distinct chemical properties of the actives to decoys. Nonetheless, the non-polar energies still give better outcomes; the ROC AUCs calculated above 0.7 still give a rate at 66.67%. The data further confirmed the observations in the other two test sets.

### 11.3.3 Rescoring using the MM-GBSA, and an alternative charging method

To determine whether the observations in the previous section was singular to for the MM-PBSA method or not, the MM-GBSA method was also tested on the test set 1. Similar results were shown in this assessment,(Tables 11.5 to 11.7) which further confirmed that the different behaviours of the decomposed energies are universal among the MM-PB/GBSA models. Both non-polar and polar energies gave close to one ROC AUC $r_p^2$ across the tested internal dielectric values in test set 1 VS. Once

**Table 11.5:** ROC AUC correlation data for MM-GBSA energy compositions with different internal dielectrics (2, 4, and 6) to results using internal dielectric = 1. The AM1-BCC charging method was applied.

| MM-GBSA energy compositions | Internal dielectric = 2 | | | Internal dielectric = 4 | | | Internal dielectric = 6 | | |
|---|---|---|---|---|---|---|---|---|---|
| | $r_p^{2*}$ | $m^{**}$ | $b^{**}$ | $r_p^{2*}$ | $m^{**}$ | $b^{**}$ | $r_p^{2*}$ | $m^{**}$ | $b^{**}$ |
| Ele + GB polar | 0.9996 | 1.0019 | 0.0014 | 0.9951 | 1.0040 | 0.0011 | 0.9870 | 1.0154 | 0.0020 |
| VDW + GB non-polar | 0.9981 | 1.0168 | -0.0141 | 0.9924 | 1.0111 | -0.0090 | 0.9857 | 1.0140 | -0.0124 |
| Total MM-GBSA | 0.9061 | 0.8523 | 0.1214 | 0.3562 | 0.4572 | 0.4060 | 0.0685 | 0.2127 | 0.5775 |

,

* Pearson correlation coefficient squared

** $m$ and $b$ are the fitted parameter to a linear equation of the form $y = mx + b$

**Table 11.6:** The ROC AUC results using MM-PB/GBSA methods and different partial charges when internal dielectric equals to 1.

| Gene ID | ROC AUC (PLANTS) | ROC AUC (MM-PBSA, AM1-BCC) | | | ROC AUC (MM-GBSA, AM1-BCC) | | | ROC AUC (MM-PBSA, Gasteiger) | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | non-polar | Ele + polar | total | non-polar | Ele + polar | total | non-polar | Ele + polar | total |
| ACES | 0.8656 | 0.9209 | 0.2008 | 0.4384 | 0.9088 | 0.4445 | 0.6322 | 0.9190 | 0.1799 | 0.4434 |
| CDK2 | 0.3931 | 0.5136 | 0.5068 | 0.5707 | 0.4762 | 0.4400 | 0.4429 | 0.5074 | 0.3507 | 0.3546 |
| DPP4 | 0.6807 | 0.7756 | 0.4684 | 0.5704 | 0.7654 | 0.5380 | 0.5806 | 0.7590 | 0.5019 | 0.5904 |
| EGFR | 0.6883 | 0.6361 | 0.5403 | 0.7052 | 0.6402 | 0.5270 | 0.5609 | 0.6485 | 0.5883 | 0.7523 |
| ESR1 | 0.5498 | 0.4050 | 0.9282 | 0.9180 | 0.4293 | 0.8399 | 0.8066 | 0.4163 | 0.8201 | 0.8360 |
| FA10 | 0.6149 | 0.8185 | 0.5128 | 0.7867 | 0.7995 | 0.7360 | 0.7819 | 0.8046 | 0.4439 | 0.7262 |
| FGFR1 | 0.8009 | 0.9364 | 0.1487 | 0.4281 | 0.9286 | 0.0861 | 0.1248 | 0.9320 | 0.1781 | 0.4604 |
| FNTA | 0.6049 | 0.6603 | 0.3380 | 0.3945 | 0.6445 | 0.6020 | 0.6171 | 0.6404 | 0.4073 | 0.4714 |
| GRIA2 | 0.7128 | 0.8498 | 0.3940 | 0.6073 | 0.8645 | 0.6690 | 0.7095 | 0.8608 | 0.3443 | 0.5197 |
| GRIK1 | 0.4840 | 0.6026 | 0.4019 | 0.4287 | 0.5896 | 0.5684 | 0.6026 | 0.5792 | 0.4300 | 0.4420 |
| HIVPR | 0.7766 | 0.8760 | 0.1771 | 0.4281 | 0.8396 | 0.7906 | 0.8479 | 0.8844 | 0.2021 | 0.5625 |
| KPCB | 0.7394 | 0.6858 | 0.5965 | 0.7586 | 0.7194 | 0.4854 | 0.5403 | 0.6719 | 0.5769 | 0.6850 |
| LCK | 0.9105 | 0.8811 | 0.5313 | 0.7705 | 0.8704 | 0.3728 | 0.4638 | 0.8792 | 0.4428 | 0.6531 |
| PTN1 | 0.6649 | 0.7197 | 0.5145 | 0.5724 | 0.7399 | 0.6211 | 0.6289 | 0.7197 | 0.5465 | 0.5934 |
| SRC | 0.5334 | 0.7744 | 0.3477 | 0.5630 | 0.7642 | 0.3295 | 0.4004 | 0.7753 | 0.3500 | 0.5591 |
| THRB | 0.8819 | 0.8206 | 0.5750 | 0.7741 | 0.8201 | 0.5039 | 0.6330 | 0.8232 | 0.5110 | 0.7791 |
| TRY1 | 0.3860 | 0.4059 | 0.5662 | 0.4841 | 0.4027 | 0.3931 | 0.3541 | 0.4330 | 0.4984 | 0.3533 |
| VGFR2 | 0.7983 | 0.7797 | 0.5629 | 0.7862 | 0.7842 | 0.5900 | 0.6569 | 0.7882 | 0.5226 | 0.7470 |
| % (> 0.7) | 44.44 | 61.11 | 5.56 | 38.89 | 66.67 | 16.67 | 22.22 | 61.11 | 5.56 | 27.7 |
| $r_p^2$ to data from ROC AUC, internal dielectric = 1 | | | | | 0.9932 | 0.3891 | 0.3231 | 0.9968 | 0.9451 | 0.8555 |

**Table 11.7:** The ROC AUC results using different internal dielectric values with MM-GBSA virtual screening on test set 1. Squared Pearson correlation coefficients to results from internal dielectric = 1 are also shown. AM1-BCC was the charging method.

| Gene ID | ROC AUC (PLANTS) | ROC AUC, internal dielectric = 1 | | | ROC AUC, internal dielectric = 2 | | | ROC AUC, internal dielectric = 4 | | | ROC AUC, internal dielectric = 6 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | non-polar | Ele + polar | total | non-polar | Ele + polar | total | non-polar | Ele + polar | total | non-polar | Ele + polar | total |
| ACES | 0.8656 | 0.9088 | 0.4445 | 0.6322 | 0.9117 | 0.4488 | 0.7528 | 0.9265 | 0.4183 | 0.8491 | 0.9274 | 0.4392 | 0.8842 |
| CDK2 | 0.3931 | 0.4762 | 0.4400 | 0.4429 | 0.4695 | 0.4429 | 0.4570 | 0.4893 | 0.4140 | 0.4406 | 0.5085 | 0.3784 | 0.4395 |
| DPP4 | 0.6807 | 0.7654 | 0.5380 | 0.5806 | 0.7700 | 0.5335 | 0.6284 | 0.7643 | 0.5384 | 0.6958 | 0.7741 | 0.5192 | 0.7417 |
| EGFR | 0.6883 | 0.6402 | 0.5270 | 0.5609 | 0.6350 | 0.5295 | 0.5940 | 0.6437 | 0.5299 | 0.6369 | 0.6460 | 0.5314 | 0.6552 |
| ESR1 | 0.5498 | 0.4293 | 0.8399 | 0.8066 | 0.4265 | 0.8456 | 0.6895 | 0.4253 | 0.8439 | 0.5837 | 0.4219 | 0.8569 | 0.5402 |
| FA10 | 0.6149 | 0.7995 | 0.7360 | 0.7819 | 0.7965 | 0.7404 | 0.8032 | 0.7816 | 0.7512 | 0.8212 | 0.7728 | 0.7630 | 0.8134 |
| FGFR1 | 0.8009 | 0.9286 | 0.0861 | 0.1248 | 0.9295 | 0.0837 | 0.2554 | 0.9281 | 0.0890 | 0.6228 | 0.9291 | 0.0890 | 0.8288 |
| FNTA | 0.6049 | 0.6445 | 0.6020 | 0.6171 | 0.6288 | 0.6049 | 0.6294 | 0.6206 | 0.6078 | 0.6317 | 0.6101 | 0.6107 | 0.6311 |
| GRIA2 | 0.7128 | 0.8645 | 0.6690 | 0.7095 | 0.8640 | 0.6682 | 0.7405 | 0.8590 | 0.6700 | 0.7855 | 0.8500 | 0.6720 | 0.8085 |
| GRIK1 | 0.4840 | 0.5896 | 0.5684 | 0.6026 | 0.5750 | 0.5712 | 0.5996 | 0.5726 | 0.5748 | 0.5962 | 0.5605 | 0.5865 | 0.5853 |
| HIVPR | 0.7766 | 0.8396 | 0.7906 | 0.8479 | 0.8396 | 0.7958 | 0.8833 | 0.8583 | 0.7917 | 0.9073 | 0.8719 | 0.7896 | 0.9115 |
| KPCB | 0.7394 | 0.7194 | 0.4854 | 0.5403 | 0.7085 | 0.4911 | 0.5943 | 0.7011 | 0.5046 | 0.6941 | 0.6858 | 0.5268 | 0.7359 |
| LCK | 0.9105 | 0.8704 | 0.3728 | 0.4638 | 0.8713 | 0.3801 | 0.5479 | 0.8669 | 0.3977 | 0.6800 | 0.8792 | 0.3958 | 0.7573 |
| PTN1 | 0.6649 | 0.7399 | 0.6211 | 0.6289 | 0.7601 | 0.6145 | 0.6430 | 0.7579 | 0.6189 | 0.6623 | 0.7482 | 0.6224 | 0.6890 |
| SRC | 0.5334 | 0.7642 | 0.3295 | 0.4004 | 0.7632 | 0.3304 | 0.4759 | 0.7536 | 0.3463 | 0.5926 | 0.7530 | 0.3547 | 0.6633 |
| THRB | 0.8819 | 0.8201 | 0.5039 | 0.6330 | 0.8256 | 0.5050 | 0.6974 | 0.8482 | 0.5108 | 0.7733 | 0.8411 | 0.5255 | 0.8073 |
| TRY1 | 0.3860 | 0.4027 | 0.3931 | 0.3541 | 0.4019 | 0.3995 | 0.3278 | 0.4131 | 0.3828 | 0.3341 | 0.4107 | 0.3780 | 0.3541 |
| VGFR2 | 0.7983 | 0.7842 | 0.5900 | 0.6569 | 0.7736 | 0.5981 | 0.7369 | 0.7671 | 0.6061 | 0.8003 | 0.7651 | 0.6117 | 0.8350 |
| % (> 0.7) | 44.44 | 66.67 | 16.67 | 22.22 | 66.67 | 16.67 | 27.78 | 66.67 | 16.67 | 33.33 | 61.11 | 16.67 | 55.56 |
| $r_p^2$ to data from ROC AUC, internal dielectric = 1 | | | | | 0.9988 | 0.9998 | 0.9477 | 0.9955 | 0.9971 | 0.5971 | 0.9914 | 0.9923 | 0.2934 |

combining the two energy components, the total MM-GBSA scores provide worsened correlation between different solute dielectrics.

**Table 11.8:** ROC AUC correlation data for MM-PBSA energy compositions with different internal dielectrics (2, 4, and 6) to results using internal dielectric = 1. The Gasteiger charging method was applied.

| MM-GBSA energy compositions | Internal dielectric = 2 | | | Internal dielectric = 4 | | | Internal dielectric = 6 | | |
|---|---|---|---|---|---|---|---|---|---|
| | $r_p^{2*}$ | $m^{**}$ | $b^{**}$ | $r_p^{2*}$ | $m^{**}$ | $b^{**}$ | $r_p^{2*}$ | $m^{**}$ | $b^{**}$ |
| Ele + PB polar | 0.9870 | 0.9954 | 0.0180 | 0.9386 | 0.9598 | 0.0501 | 0.8742 | 0.9152 | 0.0769 |
| VDW + PB non-polar | 0.9952 | 0.9982 | -0.0033 | 0.9869 | 1.0098 | -0.0078 | 0.9854 | 1.0194 | -0.0114 |
| Total MM-PBSA | 0.3728 | 0.6109 | 0.3293 | 0.0628 | 0.2598 | 0.5636 | 0.0228 | 0.1595 | 0.6261 |

$^{*}$ Pearson correlation coefficient squared

$^{**}$ $m$ and $b$ are the fitted parameter to a linear equation of the form $y = mx + b$

**Table 11.9:** The ROC AUC results using different internal dielectric values with MM-PBSA virtual screening on test set 1. Squared Pearson correlation coefficients to results from internal dielectric = 1 are also shown. Gasteiger charges were applied.

| Gene ID | ROC AUC (PLANTS) | ROC AUC, internal dielectric = 1 | | | ROC AUC, internal dielectric = 2 | | | ROC AUC, internal dielectric = 4 | | | ROC AUC, internal dielectric = 6 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | non-polar | Ele + polar | total | non-polar | Ele + polar | total | non-polar | Ele + polar | total | non-polar | Ele + polar | total |
| ACES | 0.8656 | 0.919 | 0.1799 | 0.4434 | 0.9182 | 0.2205 | 0.8398 | 0.9265 | 0.2929 | 0.9178 | 0.9283 | 0.3568 | 0.9309 |
| CDK2 | 0.3931 | 0.5074 | 0.3507 | 0.3546 | 0.5068 | 0.336 | 0.3614 | 0.5034 | 0.3569 | 0.4084 | 0.4949 | 0.3818 | 0.4282 |
| DPP4 | 0.6807 | 0.759 | 0.5019 | 0.5904 | 0.7632 | 0.4989 | 0.6657 | 0.7669 | 0.4838 | 0.7233 | 0.7869 | 0.436 | 0.7549 |
| EGFR | 0.6883 | 0.6485 | 0.5883 | 0.7523 | 0.6537 | 0.586 | 0.7475 | 0.6631 | 0.5665 | 0.7308 | 0.6594 | 0.5656 | 0.7083 |
| ESR1 | 0.5498 | 0.4163 | 0.8201 | 0.836 | 0.414 | 0.8529 | 0.6239 | 0.4191 | 0.9095 | 0.5345 | 0.4197 | 0.9304 | 0.5074 |
| FA10 | 0.6149 | 0.8046 | 0.4439 | 0.7262 | 0.7667 | 0.4959 | 0.8205 | 0.7677 | 0.5118 | 0.8012 | 0.7654 | 0.5338 | 0.7904 |
| FGFR1 | 0.8009 | 0.932 | 0.1781 | 0.4604 | 0.9242 | 0.1947 | 0.7652 | 0.9271 | 0.2114 | 0.9286 | 0.9261 | 0.2383 | 0.9344 |
| FNTA | 0.6049 | 0.6404 | 0.4073 | 0.4714 | 0.6463 | 0.3875 | 0.5321 | 0.6457 | 0.3829 | 0.6171 | 0.6439 | 0.3625 | 0.6305 |
| GRIA2 | 0.7128 | 0.8608 | 0.3443 | 0.5197 | 0.8505 | 0.3808 | 0.682 | 0.8455 | 0.4458 | 0.7762 | 0.8615 | 0.4722 | 0.8155 |
| GRIK1 | 0.4840 | 0.5792 | 0.43 | 0.442 | 0.5758 | 0.4441 | 0.4665 | 0.556 | 0.4635 | 0.4988 | 0.5537 | 0.4722 | 0.5166 |
| HIVPR | 0.7766 | 0.8844 | 0.2021 | 0.5625 | 0.8833 | 0.199 | 0.8552 | 0.8885 | 0.1844 | 0.8875 | 0.8792 | 0.1875 | 0.8948 |
| KPCB | 0.7394 | 0.6719 | 0.5769 | 0.685 | 0.6806 | 0.5996 | 0.7695 | 0.7142 | 0.5542 | 0.7717 | 0.7237 | 0.556 | 0.7638 |
| LCK | 0.9105 | 0.8792 | 0.4428 | 0.6531 | 0.888 | 0.454 | 0.8777 | 0.9056 | 0.4667 | 0.9139 | 0.9017 | 0.5391 | 0.9173 |
| PTN1 | 0.6649 | 0.7197 | 0.5465 | 0.5934 | 0.7285 | 0.5333 | 0.6132 | 0.7351 | 0.5412 | 0.6702 | 0.7281 | 0.5579 | 0.7022 |
| SRC | 0.5334 | 0.7753 | 0.35 | 0.5591 | 0.7537 | 0.3841 | 0.7453 | 0.7638 | 0.4153 | 0.7836 | 0.7651 | 0.4316 | 0.7853 |
| THRB | 0.8819 | 0.8232 | 0.511 | 0.7791 | 0.8112 | 0.5445 | 0.84 | 0.813 | 0.5467 | 0.8446 | 0.8477 | 0.4675 | 0.8474 |
| TRY1 | 0.3860 | 0.433 | 0.4984 | 0.3533 | 0.4155 | 0.5335 | 0.3852 | 0.3947 | 0.5845 | 0.4083 | 0.4035 | 0.5686 | 0.4043 |
| VGFR2 | 0.7983 | 0.7882 | 0.5226 | 0.747 | 0.7842 | 0.5458 | 0.8194 | 0.7782 | 0.5775 | 0.8456 | 0.7857 | 0.5372 | 0.833 |
| % (> 0.7) | 44.44 | 61.11 | 5.56 | 27.78 | 61.11 | 5.56 | 55.56 | 66.67 | 5.56 | 66.67 | 66.67 | 5.56 | 72.22 |
| $r_p^2$ to data from ROC AUC, internal dielectric = 1 | | | | | 0.9971 | 0.9912 | 0.5962 | 0.9925 | 0.9587 | 0.3615 | 0.9917 | 0.9099 | 0.2979 |

Different to the MM-PBSA model, the total MM-GBSA rescoring VS with internal dielectric equals 2 still provide a good correlation to the internal dielectric 1. This good correlation, however, did not extend to the other VS runs using solute dielectric values equal to 4 and 6. More importantly, this seemly improved robustness does not indicate that the MM-GBSA method performs better than MM-PBSA in ranking ligands. It only shows that the performance of the total MM-GBSA score when slightly increase the internal dielectric constant to 2 is as bad as when the value is set to 1.

The MM-GBSA rescoring also performs marginally better than MM-PBSA when comparing the ROC AUCs provided by using the electrostatic and polar estimations; 1 more library (3 out of 18) give an electrostatic-polar ROC AUC above 0.7 across the solute dielectrics tested. Moreover, poor ROC AUC correlations to the MM-PBSA results in electrostatic-polar estimations reflect the differences between the GB/PB

solver.

An alternative charging method was also tested for potential improvement or worsening of results. Gasteiger partial charges are used in the Autodock package[18] and also a popular and efficient method applied in virtual screening. When incorporating this type of partial charge into the MM-PBSA rescoring, the ROC AUC outcomes and the correlations between different internal dielectric choices were similar to the results using the AM1-BCC method (Table 11.8 and 11.9). Hence, it seems reasonable to suggest that the universal poor performance of electrostatic and polar energies is less likely related to the charging method but the calculating method of choice.

### 11.3.4 KDE probability density of different energy compositions

The results so far demonstrated that the MM-PB/GBSA methods give unbalanced VS rescoring performance between non-polar and electrostatic-polar calculations. The poor discrimination power of electrostatic-polar estimations is obviously affecting the total MM-PB/GBSA scores. Increasing internal dielectric values also weakens the effect on the total estimations. To further investigate the underlying reason for such behaviour kernel density estimation (KDE) on the probability density of non-polar, electrostatic-polar, and total estimated energies. In test set 1, there are 13 systems show trends of decreasing in deteriorating effects when increasing the internal dielectric values. However, the enrichment of each energy components remain in the same range(Figures 11.3 and 11.4). Same trends were observed for test set 2 (13 systems, Figures 11.5 and 11.6) and test set 3 (11 systems, Figures 11.7 and 11.8).

Moreover, the KDE analysis outcomes also confirm that the quantitative shifting in the densities of the electrostatic-polar interactions when increasing the internal dielectric values might provide a facade of beneficiary outcome in the final MM-PBSA evaluation. However, the quality of the electrostatic/polar estimations is not improved. More interestingly, the electrostatic-polar energies can be further separated into the gas phase and the PB/GB parts. Further analysis of the gas phase electrostatics revealed that this part performs better than the total electrostatic-polar

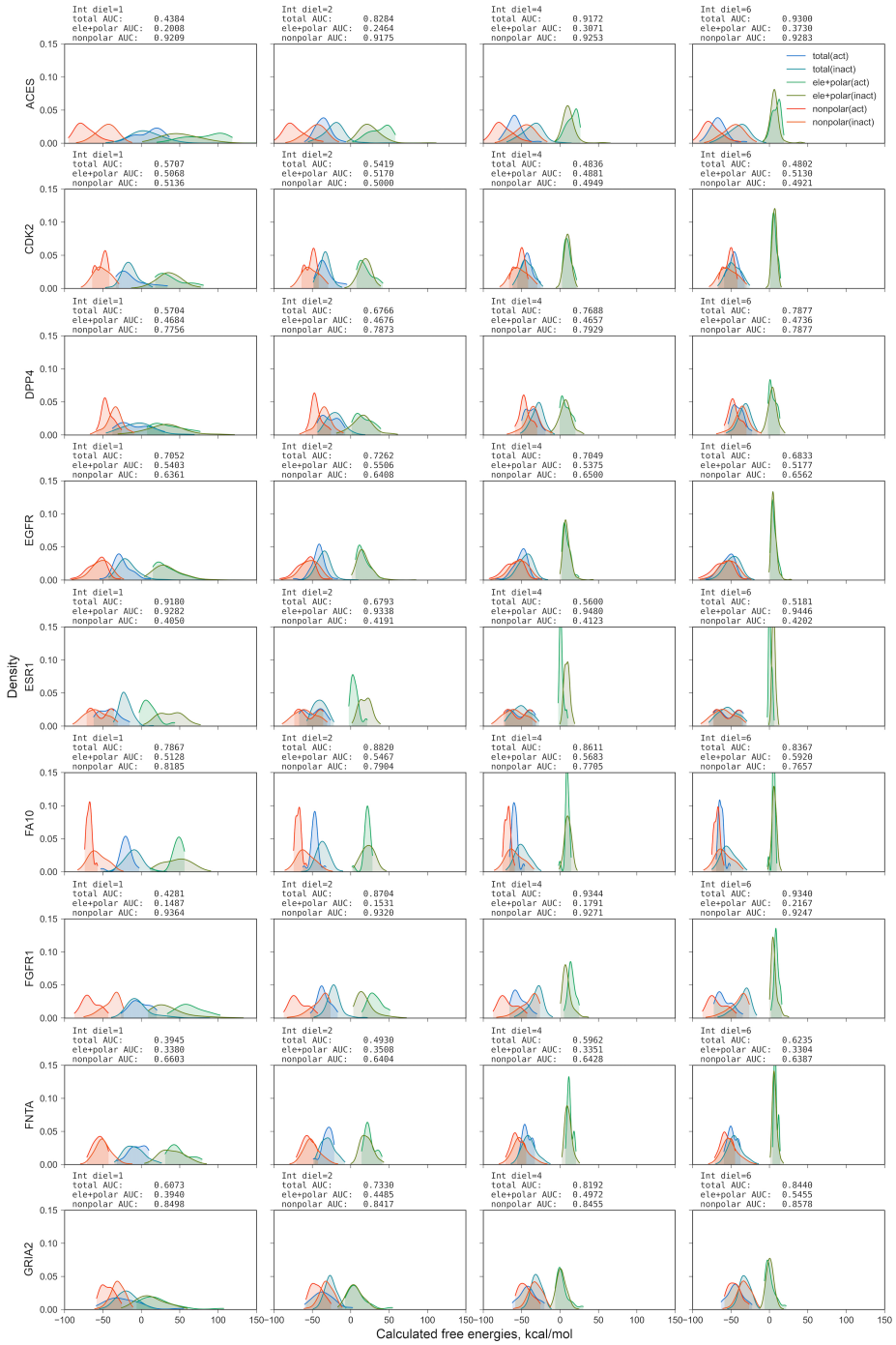



**Figure 11.3:** The KDE plots of different energy compositions of the first 9 receptors from test set 1 (DUDE test set) with internal dielectric values of 1, 2, 4, and 6 applied. ACES, DPP4, EGFR, FA10, FGFR1, FNTA, and GRIA2 were demonstrated with better VS enrichment using non-polar energies than electrostatic/polar energies among the tested libraries. A relatively over-smoothed bandwidth (0.5) was chosen for demonstration purpose as the plots were not intended for the accurate estimation of densities.

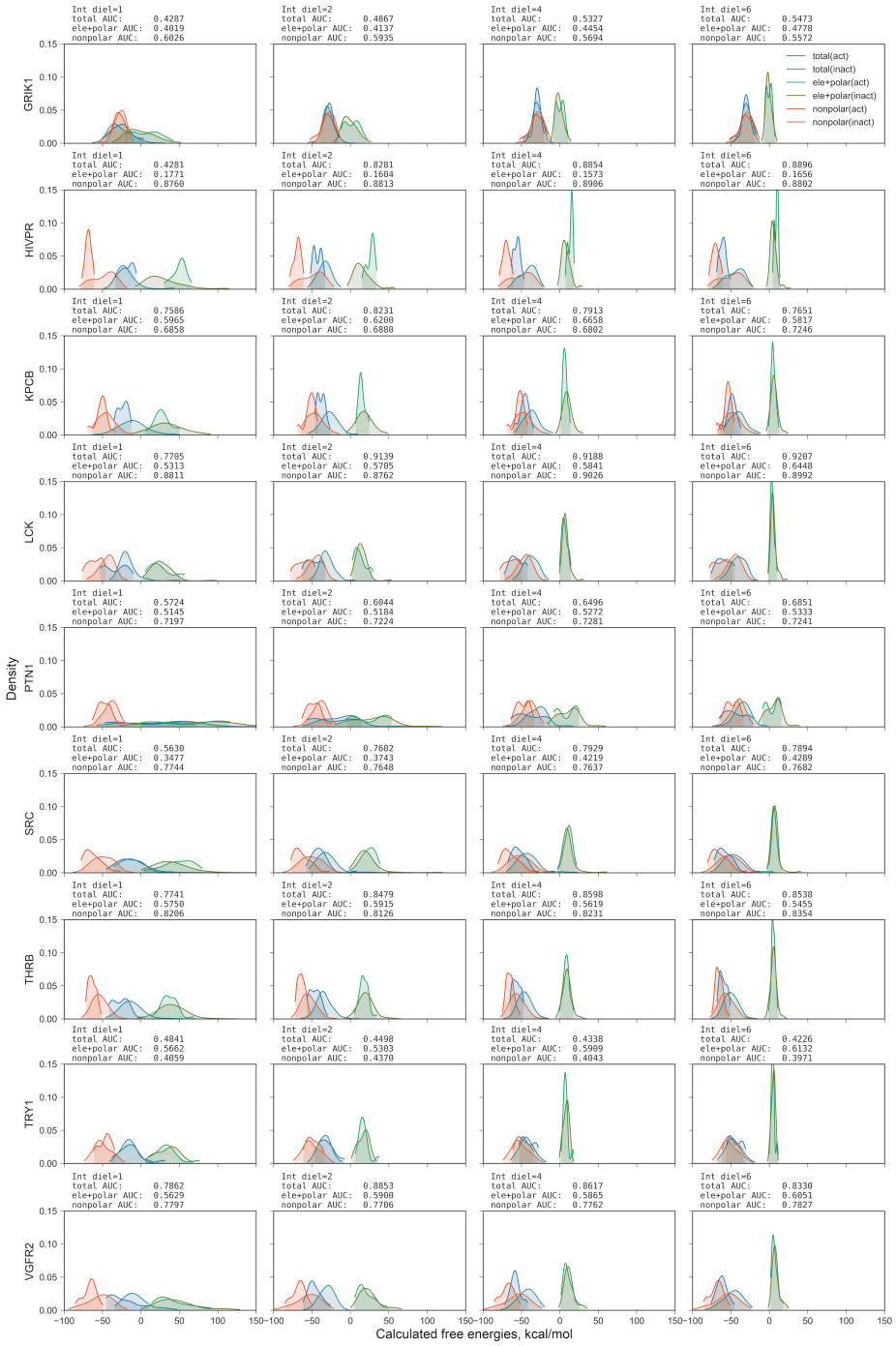



**Figure 11.4:** The KDE plots of different energy compositions of the last 9 receptors from test set 1 (DUDE test set) with internal dielectric values of 1, 2, 4, and 6 applied. All except TRY1 were demonstrated with better VS enrichment using non-polar energies than electrostatic/polar energies among the tested libraries. A relatively over-smoothed bandwidth (0.5) was chosen for demonstration purpose as the plots were not intended for the accurate estimation of densities.

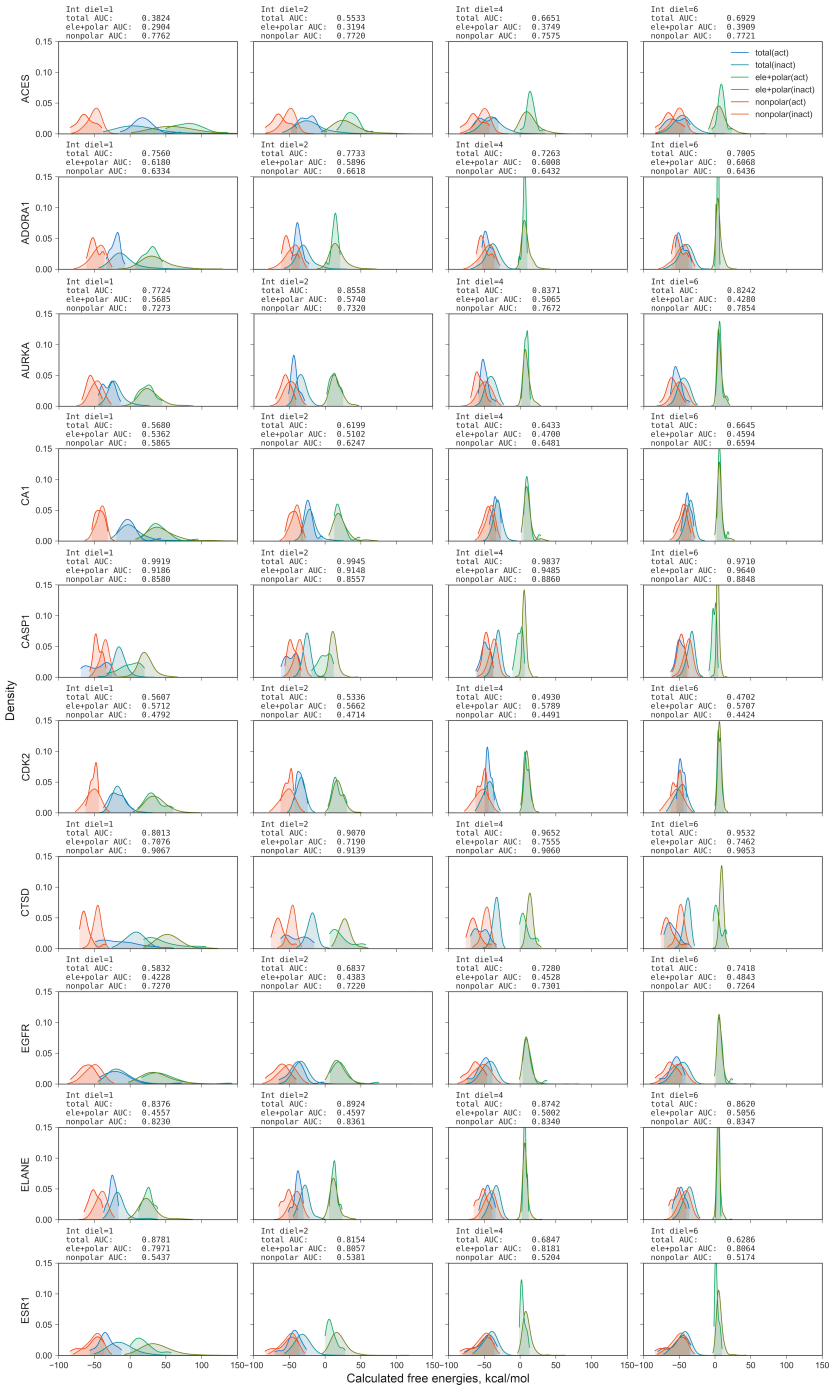



**Figure 11.5:** The KDE plots of different energy compositions of the first 10 receptors from test set 2 (CHEMBL test set) with internal dielectric values of 1, 2, 4, and 6 applied. ACES, ADORA1, AURKA, CA1, CTSD, EGFR, and ELANE were demonstrated with better VS enrichment using non-polar energies than electrostatic/polar energies among the tested libraries. A relatively over-smoothed bandwidth (0.5) was chosen for demonstration purpose as the plots were not intended for the accurate estimation of densities.

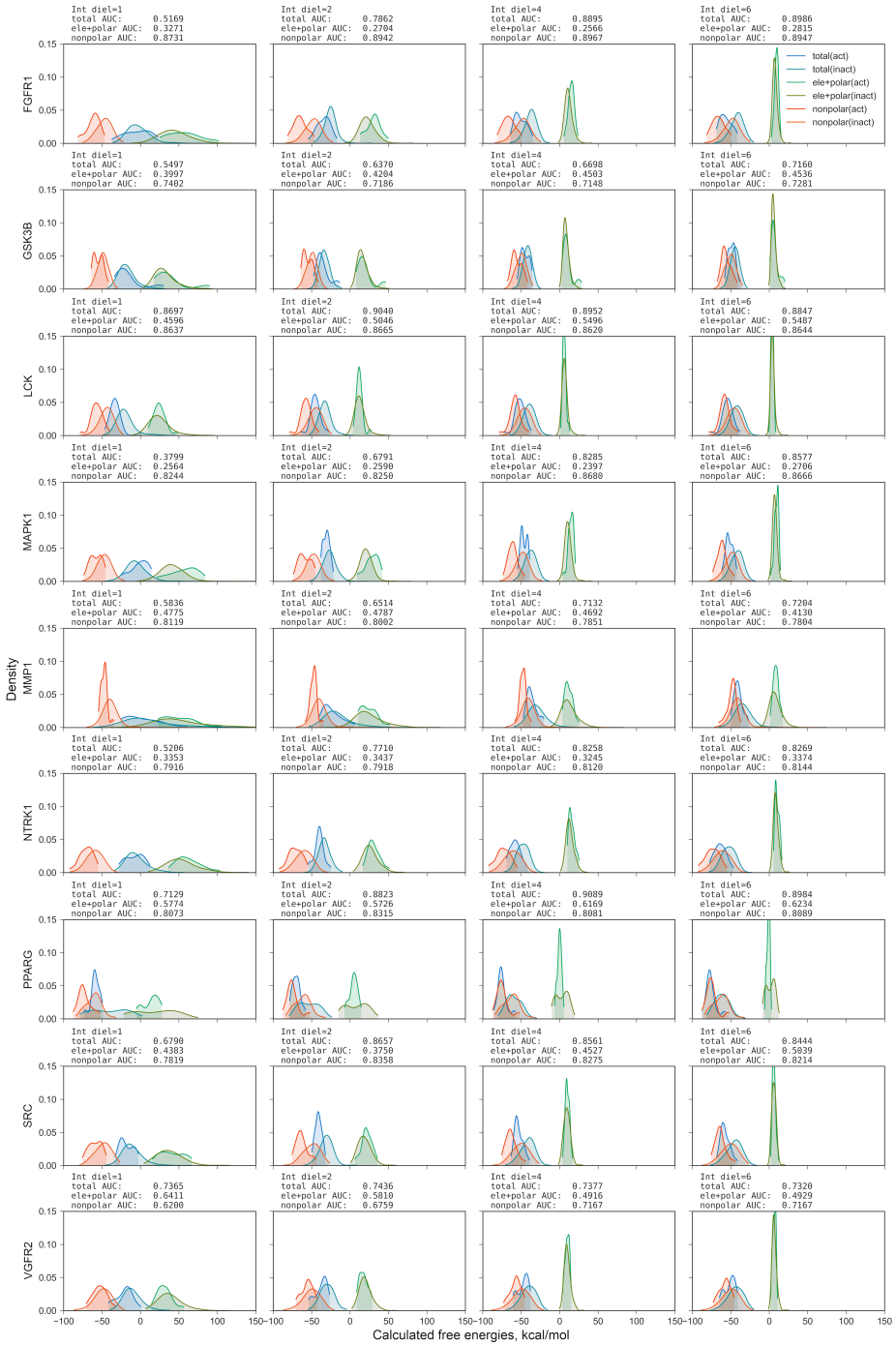



**Figure 11.6:** The KDE plots of different energy compositions of the last 9 receptors from test set 2 (CHEMBL test set) with internal dielectric values of 1, 2, 4, and 6 applied. All receptors were demonstrated with better VS enrichment using non-polar energies than electrostatic/polar energies among the tested libraries. A relatively over-smoothed bandwidth (0.5) was chosen for demonstration purpose as the plots were not intended for the accurate estimation of densities.

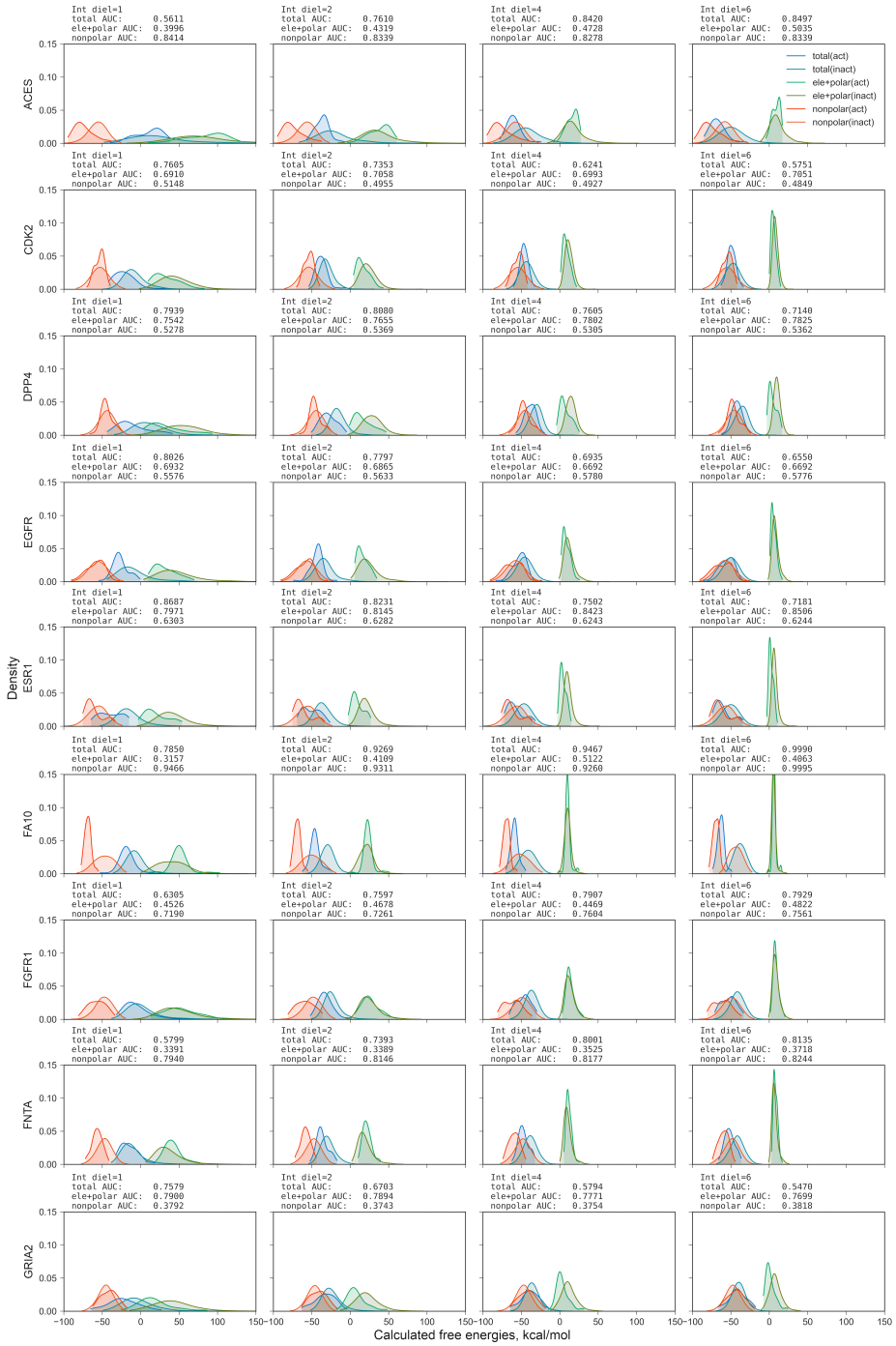



**Figure 11.7:** The KDE plots of different energy compositions of the first 9 receptors from test set 3 (decoys test set) with internal dielectric values of 1, 2, 4, and 6 applied. ACES, FA10, FGFR1, and FNTA were demonstrated with better VS enrichment using non-polar energies than electrostatic/polar energies among the tested libraries. A relatively over-smoothed bandwidth (0.5) was chosen for demonstration purpose as the plots were not intended for the accurate estimation of densities.

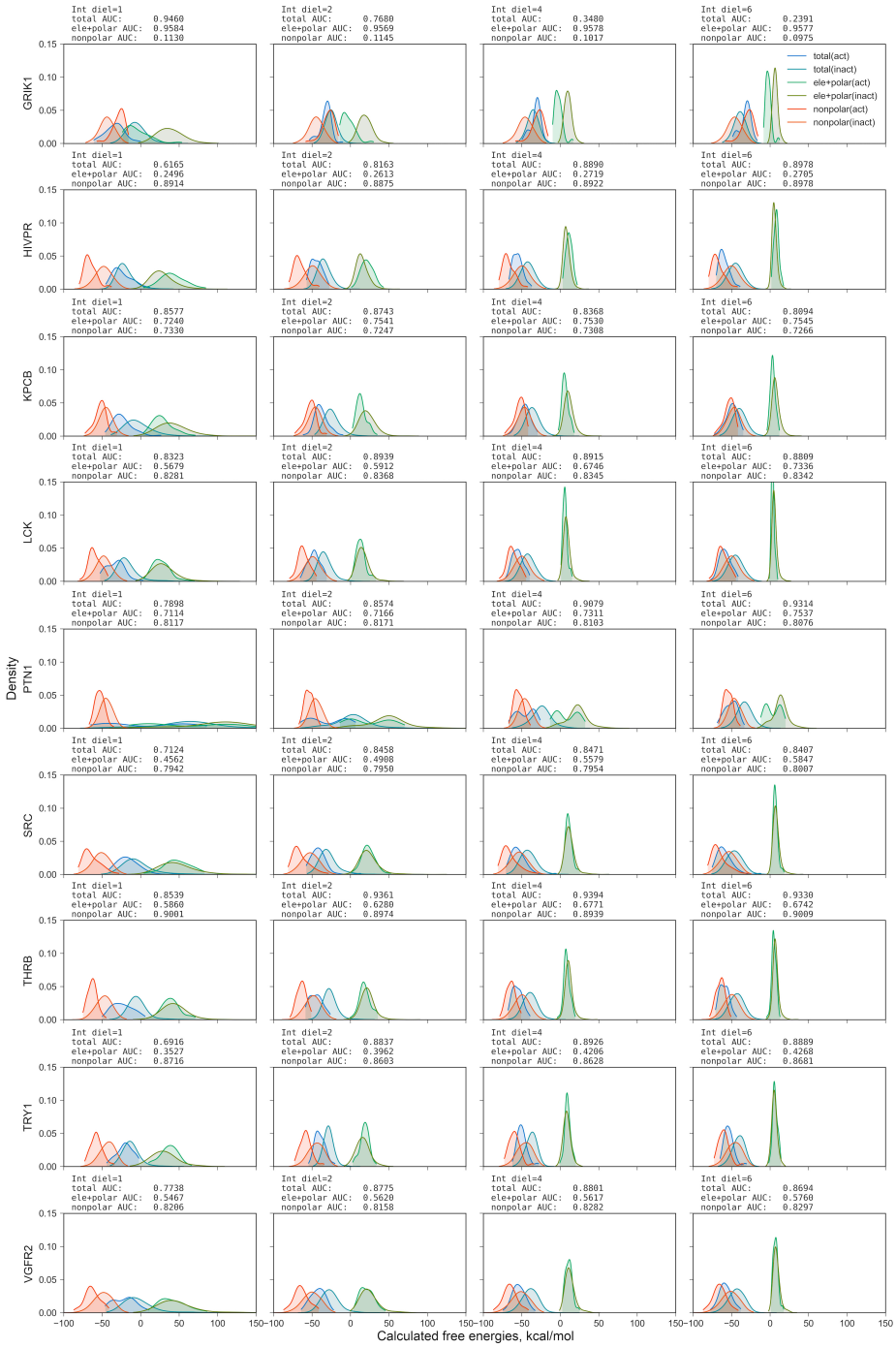



**Figure 11.8:** The KDE plots of different energy compositions of the last 9 receptors from test set 3 (decoys test set) with internal dielectric values of 1, 2, 4, and 6 applied. HIVPR, LCK, PTN1, SRC, THRB, TRY1, and VGFR2 were demonstrated with better VS enrichment using non-polar energies than electrostatic/polar energies among the tested libraries. A relatively over-smoothed bandwidth (0.5) was chosen for demonstration purpose as the plots were not intended for the accurate estimation of densities.

estimations in terms of ROC AUC values (Table 11.10). The outcome implies that it is likely the PB model of choice contributing to the worsening effect on the MM-PBSA enrichment most significantly at low solute dielectric values for the libraries tested.

### 11.3.5 Considering effects from molecular weight

The distribution of the MW within the screening library is closely related to the final enrichment for most of the scoring methods. In general, larger molecules tend to have higher scores. Most libraries in test set 1-3 provide good MW overlay between actives and inactives/decoys. However, ROC AUC analysis has revealed to be a more direct assessment to determine biases from MW (Figure 11.9). Thus the ROC AUC analysis was performed and compared to the other ROC AUCs calculated in the previous sections.

Within the accessed databases (both DUD-E and ChEMBL) for the receptors se-

**Table 11.10:** The ROC AUC outcomes of virtual screening using different gas phase energies. Internal dielectric equals 1.

Panel A: Test set 1

| Gene ID | ROC AUC (PLANTS) | Gas Phase ROC AUC | | |
|---|---|---|---|---|
| | | VDW | ELE | total MM |
| ACES | 0.8656 | 0.9191 | 0.6558 | 0.8325 |
| CDK2 | 0.3931 | 0.5181 | 0.5017 | 0.5147 |
| DPP4 | 0.6807 | 0.7715 | 0.4740 | 0.4876 |
| EGFR | 0.6883 | 0.6409 | 0.6751 | 0.6953 |
| ESR1 | 0.5498 | 0.4027 | 0.7421 | 0.7313 |
| FA10 | 0.6149 | 0.8208 | 0.6109 | 0.7072 |
| FGFR1 | 0.8009 | 0.9364 | 0.8973 | 0.9320 |
| FNTA | 0.6049 | 0.6702 | 0.5361 | 0.5804 |
| GRIA2 | 0.7128 | 0.8500 | 0.5947 | 0.6150 |
| GRIK1 | 0.4840 | 0.6108 | 0.4661 | 0.4762 |
| HIVPR | 0.7766 | 0.8760 | 0.4698 | 0.7073 |
| KPCB | 0.7394 | 0.6946 | 0.6266 | 0.6392 |
| LCK | 0.9105 | 0.8772 | 0.9496 | 0.9555 |
| PTN1 | 0.6649 | 0.7088 | 0.5623 | 0.5706 |
| SRC | 0.5334 | 0.7693 | 0.7442 | 0.7837 |
| THRB | 0.8819 | 0.8264 | 0.5559 | 0.6169 |
| TRY1 | 0.3860 | 0.4139 | 0.5622 | 0.5646 |
| VGFR2 | 0.7983 | 0.7736 | 0.6781 | 0.7641 |
| % (> 0.7) | 44.44 | 61.11 | 22.22 | 44.44 |

Panel B: Test set 2

| Gene ID | ROC AUC (PLANTS) | Gas Phase ROC AUC | | |
|---|---|---|---|---|
| | | VDW | ELE | total MM |
| ACES | 0.7934 | 0.7718 | 0.5114 | 0.568 |
| ADORA1 | 0.7117 | 0.6408 | 0.4400 | 0.5225 |
| AURKA | 0.4995 | 0.7226 | 0.5974 | 0.6444 |
| CA1 | 0.4385 | 0.5850 | 0.5020 | 0.5035 |
| CASP1 | 0.9250 | 0.8436 | 0.9477 | 0.9563 |
| CDK2 | 0.6915 | 0.4887 | 0.2885 | 0.2727 |
| CTSD | 0.9270 | 0.9019 | 0.8106 | 0.8905 |
| EGFR | 0.6780 | 0.7239 | 0.6919 | 0.7109 |
| ELANE | 0.8623 | 0.8202 | 0.5554 | 0.6394 |
| ESR1 | 0.5635 | 0.5394 | 0.5735 | 0.5703 |
| FGFR1 | 0.5270 | 0.8675 | 0.7420 | 0.8134 |
| GSK3B | 0.7151 | 0.7420 | 0.5391 | 0.6009 |
| LCK | 0.5812 | 0.8617 | 0.6330 | 0.6718 |
| MAPK1 | 0.7449 | 0.8148 | 0.7590 | 0.7989 |
| MMP1 | 0.8578 | 0.8032 | 0.4841 | 0.5087 |
| NTRK1 | 0.2840 | 0.7852 | 0.6508 | 0.7092 |
| PPARG | 0.9202 | 0.7992 | 0.6968 | 0.8379 |
| SRC | 0.5956 | 0.7782 | 0.6972 | 0.7287 |
| VGFR2 | 0.6354 | 0.6111 | 0.5970 | 0.6251 |
| % (> 0.7) | 47.37 | 73.68 | 21.05 | 42.11 |

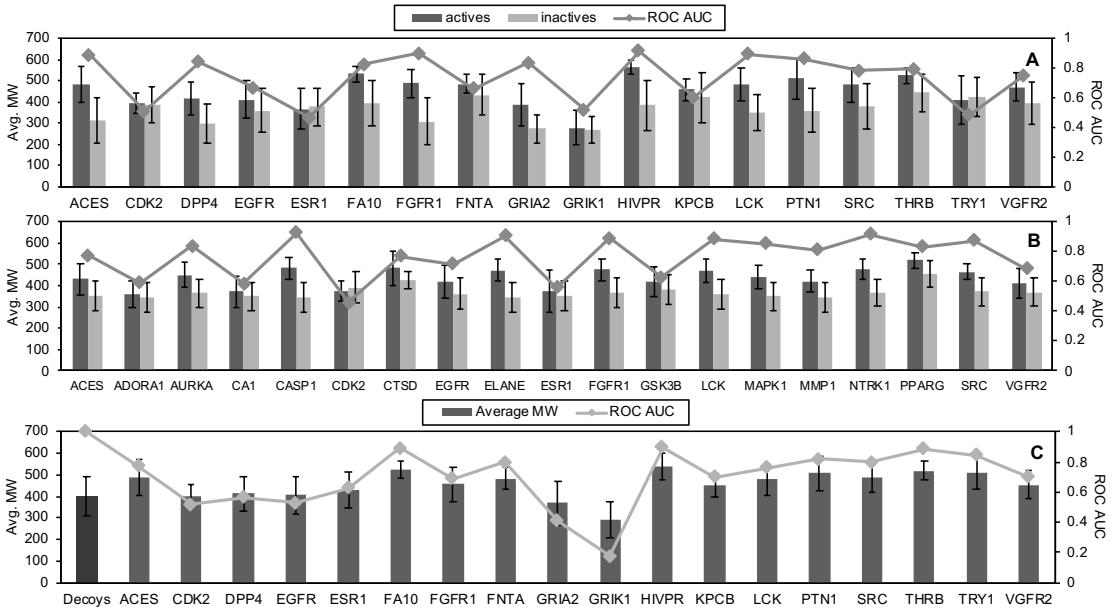



**Figure 11.9:** Comparison between average MW and ROC AUC using MW rankings to determine MW biases. ROC AUCs are the more direct reflection of MW distribution among a library. Libraries assessed are A: test set 1, B: test set2 and C: test set 3.

lected, the results show that good correlations do exist between ROC AUCs calculated from MW and non-polar interactions; the $r_p{}^2$ are 0.939, 0.877, and 0.984 for the test set 1, 2, and 3, respectively(Table 11.11). On the other hand, negative correlations were observed between the electrostatic-polar and the MW results for all the test sets assessed (-0.583, -0.264, and -0.803 for the test set 1, 2, and 3). This indicates that the different active and inactive MW distributions could be causing the out-performance of non-polar interaction in the original test sets.

To expose the relation between electrostatic-polar and non-polar calculated energies among molecules of similar sizes, it is essential to remove the MW bias in the original testing libraries. Hence, the ligand libraries from test set 2 and 3 were processed to lower the ROC AUCs of MW scores. If non-polar interactions are closely connected to the molecular size, this step was expected to lower also the ROC AUC of non-polar estimations. The effects of polar-electrostatic interactions can thus be unmasked under MW-unbiased circumstances.

The ROC AUCs of all the evaluated ranking methods were plotted (Figure 11.10).

**Table 11.11:** Comparing ROC AUC outcomes rescoring using MW and MM-PBSA calculated energy compositions. Internal dielectric equals 1.

Panel A: Test set 1

| Gene ID | ROC AUC | MM-PBSA, internal dielectric = 1 | | |
|---|---|---|---|---|
| | MW | non-polar | Ele + polar | total |
| ACES | 0.8876 | 0.9209 | 0.2008 | 0.4384 |
| CDK2 | 0.5051 | 0.5136 | 0.5068 | 0.5707 |
| DPP4 | 0.8387 | 0.7756 | 0.4684 | 0.5704 |
| EGFR | 0.6645 | 0.6361 | 0.5403 | 0.7052 |
| ESR1 | 0.4632 | 0.4050 | 0.9282 | 0.9180 |
| FA10 | 0.8252 | 0.8185 | 0.5128 | 0.7867 |
| FGFR1 | 0.8943 | 0.9364 | 0.1487 | 0.4281 |
| FNTA | 0.6585 | 0.6603 | 0.3380 | 0.3945 |
| GRIA2 | 0.8348 | 0.8498 | 0.3940 | 0.6073 |
| GRIK1 | 0.5183 | 0.6026 | 0.4019 | 0.4287 |
| HIVPR | 0.9146 | 0.8760 | 0.1771 | 0.4281 |
| KPCB | 0.6004 | 0.6858 | 0.5965 | 0.7586 |
| LCK | 0.8924 | 0.8811 | 0.5313 | 0.7705 |
| PTN1 | 0.8584 | 0.7197 | 0.5145 | 0.5724 |
| SRC | 0.7784 | 0.7744 | 0.3477 | 0.5630 |
| THRB | 0.7929 | 0.8206 | 0.5750 | 0.7741 |
| TRY1 | 0.4802 | 0.4059 | 0.5662 | 0.4841 |
| VGFR2 | 0.75 | 0.7797 | 0.5629 | 0.7862 |
| $r_p^2$ to MW ROC AUC | | 0.9393 | -0.5833 | -0.1372 |

Panel B: Test set 2

| Gene ID | ROC AUC | MM-PBSA, internal dielectric = 1 | | |
|---|---|---|---|---|
| | MW | non-polar | Ele + polar | total |
| ACES | 0.7765 | 0.7762 | 0.2904 | 0.3824 |
| ADORA1 | 0.5881 | 0.6334 | 0.6180 | 0.7560 |
| AURKA | 0.8357 | 0.7273 | 0.5685 | 0.7724 |
| CA1 | 0.5879 | 0.5865 | 0.5362 | 0.5680 |
| CASP1 | 0.925 | 0.8580 | 0.9186 | 0.9919 |
| CDK2 | 0.4569 | 0.4792 | 0.5712 | 0.5607 |
| CTSD | 0.7693 | 0.9067 | 0.7076 | 0.8013 |
| EGFR | 0.7186 | 0.7270 | 0.4228 | 0.5832 |
| ELANE | 0.9072 | 0.8230 | 0.4557 | 0.8376 |
| ESR1 | 0.5587 | 0.5437 | 0.7971 | 0.8781 |
| FGFR1 | 0.8883 | 0.8731 | 0.3271 | 0.5169 |
| GSK3B | 0.6293 | 0.7402 | 0.3997 | 0.5497 |
| LCK | 0.8835 | 0.8637 | 0.4596 | 0.8697 |
| MAPK1 | 0.8551 | 0.8244 | 0.2564 | 0.3799 |
| MMP1 | 0.8102 | 0.8119 | 0.4775 | 0.5836 |
| NTRK1 | 0.917 | 0.7916 | 0.3353 | 0.5206 |
| PPARG | 0.8339 | 0.8073 | 0.5774 | 0.7129 |
| SRC | 0.8765 | 0.7819 | 0.4383 | 0.6790 |
| VGFR2 | 0.6857 | 0.6200 | 0.6411 | 0.7365 |
| $r_p^2$ to MW ROC AUC | | 0.8767 | -0.2639 | 0.1029 |

Panel C: Test set 3

| Gene ID | ROC AUC | MM-PBSA, internal dielectric = 1 | | |
|---|---|---|---|---|
| | MW | non-polar | Ele + polar | total |
| ACES | 0.7738 | 0.8414 | 0.3996 | 0.5611 |
| CDK2 | 0.5169 | 0.5148 | 0.691 | 0.7605 |
| DPP4 | 0.5637 | 0.5278 | 0.7542 | 0.7939 |
| EGFR | 0.5286 | 0.5576 | 0.6932 | 0.8026 |
| ESR1 | 0.6269 | 0.6303 | 0.7971 | 0.8687 |
| FA10 | 0.8923 | 0.9466 | 0.3157 | 0.7850 |
| FGFR1 | 0.6890 | 0.7190 | 0.4526 | 0.6305 |
| FNTA | 0.7987 | 0.7940 | 0.3391 | 0.5799 |
| GRIA2 | 0.4114 | 0.3792 | 0.7900 | 0.7579 |
| GRIK1 | 0.1760 | 0.1130 | 0.9584 | 0.946 |
| HIVPR | 0.8962 | 0.8914 | 0.2496 | 0.6165 |
| KPCB | 0.6989 | 0.7330 | 0.7240 | 0.8577 |
| LCK | 0.7590 | 0.8281 | 0.5679 | 0.8323 |
| PTN1 | 0.8173 | 0.8117 | 0.7114 | 0.7898 |
| SRC | 0.7976 | 0.7942 | 0.4562 | 0.7124 |
| THRB | 0.8839 | 0.9001 | 0.5860 | 0.8539 |
| TRY1 | 0.8426 | 0.8716 | 0.3527 | 0.6916 |
| VGFR2 | 0.6988 | 0.8206 | 0.5467 | 0.7738 |
| $r_p^2$ to MW ROC AUC | | 0.9843 | -0.8026 | -0.4722 |

These include the adjusted and the original MW, non-polar, electrostatic-polar, and total MM-PBSA enrichments for the selected receptors for this part of the analysis. Firstly, the decrease in the MW ROC AUCs does impair the performance of calculated non-polar energies, further confirm the close relationship between the two. Nonetheless, for most of the receptors, the non-polar estimations still provide better enrichment than using MW due to the additional ”fitness” factor that is considered in non-polar calculations.

In addition, the performance of the calculated electrostatic-polar energies improved for the majority of the receptors, comparing to the original test sets. The change in trends can be better observed in the differential ROC AUC plot (Figure 11.11). The impaired non-polar performance due to the change in molecular size composition is compensated by the enrichment improvement of the electrostatic-polar estimation. This compensation safeguards the total MM-PBSA score from severe deteriorating (Figure 11.11A and C). Under this situation, increasing the solute dielectric values

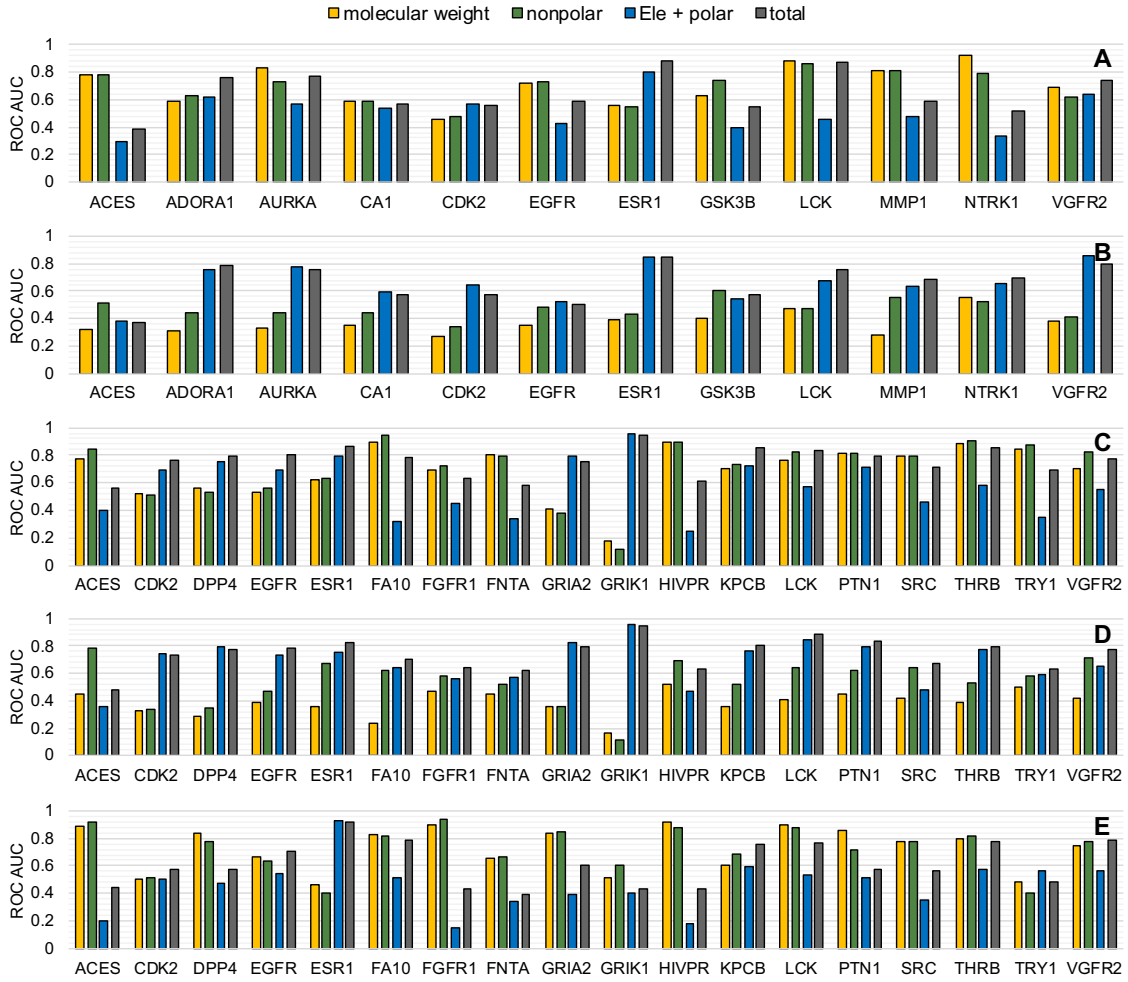



**Figure 11.10:** The bar charts illustrating ROC AUC results for both the virtual screening using the original libraries and the libraries with adjusted molecular weight. A, C, and E are results for the original libraries, while B and D are for the molecular-weight-adjusted libraries. A and B are data generated from the test set 2 (ChEMBL test set). C and D are data generated from the test set 3 (decoy test set). For comparison purpose, the result from the test set 1 (DUD-E test set, shown as E) is also included. Internal dielectric constant = 1.

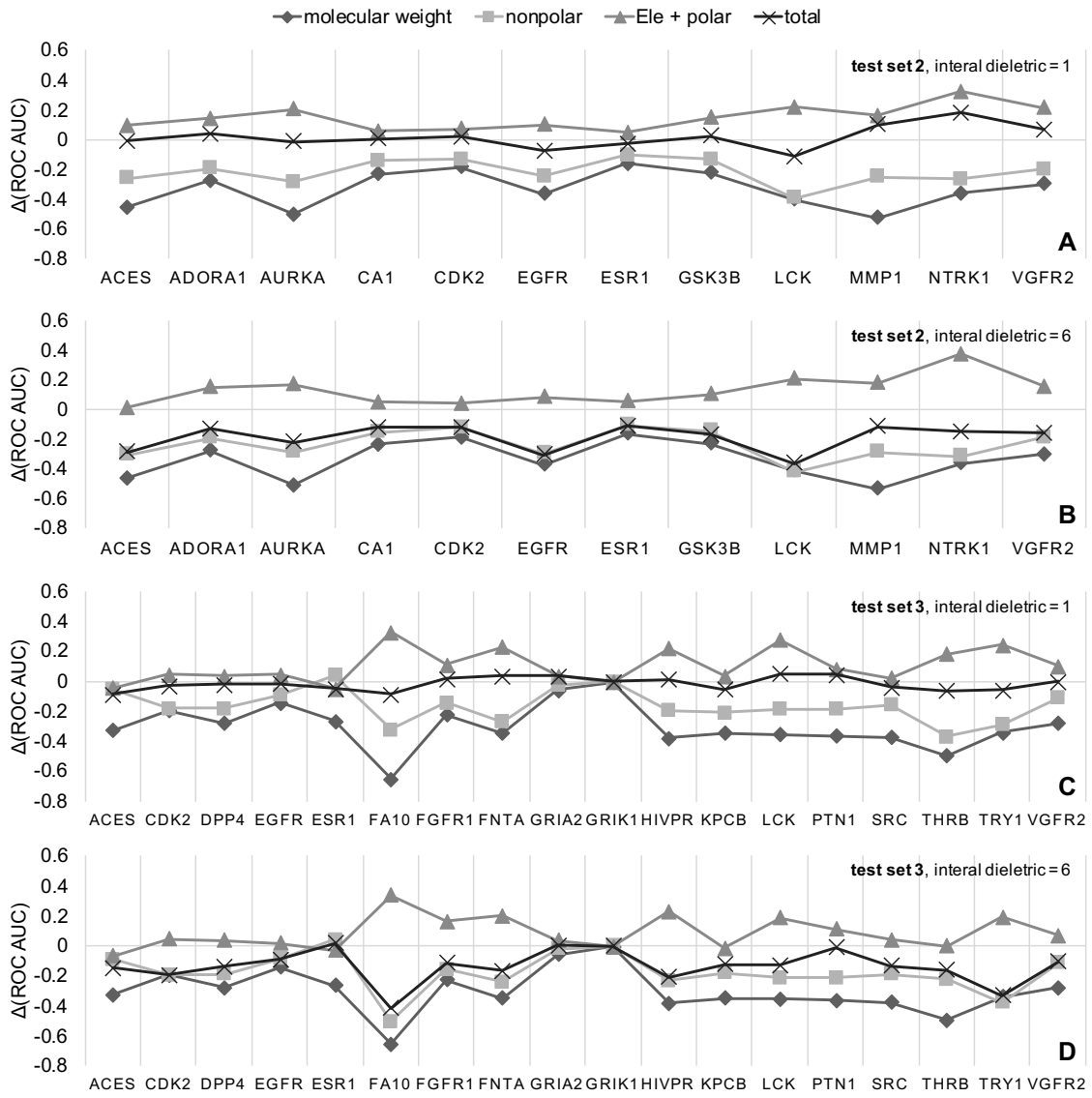



**Figure 11.11:** The differential ROC AUC of the MW-adjusted virtual screening to the original libraries. Different rescoring method are plotted using different greyscale colours and markers as illustrated. A and B are the outcomes calculated for receptors from test set 2 (ChEMBL test set) using internal dielectric values of 1 and 6, respectively. C and D are the same plots for receptors from test set 3 (decoy test set). Lower internal dielectrics show a better safeguarding effects from electrostatic-polar interactions when impairments to the non-polar energies are caused by the change in MW distribution within the tested library.

would quantitatively weaken the safeguarding effects provided by the electrostatic and polar estimations (Figure 11.11B and D). Thus, it seems a harmful approach to increase internal dielectric values with MW-unbiased libraries.

A previous study evaluated the performance of MM-GBSA among proteins from different categories.[13] The study suggested a target-dependent performance of MM-GBSA model.[13] However, in this thesis by using ligand libraries of different nature, the same receptors have given very different outcomes. Comparing the outcomes of test set 1 and 3 in which the same receptor structures were applied, CDK2, DPP4, EGFR, GRIA2, GRIK1, and TRY1 have shown inverted behaviours of non-polar and polar-electrostatic calculated scores (Table 11.2 and 11.4), likely due to the differences in active/inactive ligand molecular size distributions (Figure 11.9). Furthermore, in test set 3, more receptors were observed with alteration in trends of the outcomes for rescoring assessment (FA10, FGFR1, FNTA, KPCB, LCK, PTN1, THRB, and TRY1). This provides evidence for the library-dependent performance of the MM-PBSA method.

Taken together, the analyses indicate that in MM-PBSA rescoring, the relationship between the non-polar and electrostatic-polar interactions is to some extent compensating. When the receptor prefers large molecules as binders, the calculated non-polar energies ensure that these ligands are ranked higher during the scoring process. However, the electrostatic and polar part of the MM-PBSA score is more crucial when the library consists of ligands of similar sizes.

And yet the compensating effect between the two parts of the MM-PBSA score is limited. It has been shown in this thesis that the compensating effect can also switch to cancelling of the beneficiary scores that each part is contributing. The cancelling effect is particularly severe when using an MW-biased library with a low internal dielectric constant or an MW-unbiased library with high internal dielectric values. In the worst scenario, the MM-PBSA rescoring can even end up much worse than a simpler scoring function such as PLANTS.

This work further confirmed that increasing internal dielectric value should not be the universal solution for improved MM-PBSA rescoring. From the results reported here, some improvement caused by increased solute dielectrics was fundamentally a quantitatively weakening of the worsening effect that the calculated electrostatic-polar en-

ergies are contributing to the MM-PBSA total score. The quality of the electrostatic-polar estimations, however, are not improved. Moreover, blindly increasing internal dielectric constants could be dangerous. In a library where molecules are of similar sizes, this can lower the beneficiary performance provided by the electrostatic-polar part of MM-PBSA score.

## 11.4 Conclusions

In this chapter, the in-depth effects by changing the solute dielectric constants virtual screening rescoring using MM-PB/GBSA model were investigated. The performance was assessed under different virtual screening scenarios using larger test sets containing various receptors and ligand libraries with different molecular weight compositions. The results have shown some characteristic behaviours of MM-PB/GBSA model when used as a rescoring method in virtual screening.

Firstly, the initial test sets were mostly biased by MW distributions of the ligand libraries. The active ligands, though have some overlay to the inactive or decoy molecules, are on average larger. Under this scenario, the calculated non-polar interaction would play a dominant role for enrichment within the total MM-PB/GBSA score. Increasing the solute dielectric energy, i.e. lowering the effect from electrostatic-polar energies, would improve the overall performance of the MM-PB/GBSA rescoring.

Secondly, after modulating the MW distributions to a more balanced level, i.e. that active compounds are of similar sizes to the inactives or decoys, the electrostatic and polar interactions become very important in preventing the smaller actives from ranked lower due to worsened non-polar energy contributions. Hence, lowering the contribution of electrostatics and polar energy estimations by increasing internal dielectric constants can impair the enrichment of the MM-PB/GBSA model.

Furthermore, using the alternative Gasteiger charges did not seem to affect the overall trends of MM-PBSA in this study. More interestingly, when only considers the MM

contributions of MM-PB/GBSA calculations, the electrostatic contributions seem less destructive in the original test sets.

The investigations carried out in this work indicate that increasing internal dielectric values could be dangerous. the dependency on library MW distributions suggests that in scenarios where molecules are pre-processed with similar molecular weight, increasing solute dielectric value is not recommended. It will impair the selectivity of active molecules contributed by the electrostatic-polar interactions, likely resulted in an unsuccessful virtual screening experiment.

However, future work should still be considered for using other PB solvers to confirm the universal behaviour of MM-PB/GBSA rescoring. It is still possible that using a different solver might result in a different outcome. However, the comparison between PB and GB involved calculations imply that the underlying theory might be the essential determinant of the observation. Variations between the two solvers still exist, however, the distinction behaviours of electrostatic-polar interactions were similar.

## Chapter references

[1] J. Srinivasan, T.E. Cheatham, P. Cieplak, P.A. Kollman, and D.A. Case. Continuum solvent studies of the stability of dna, rna, and phosphoramidate - dna helices. *J. Am. Chem. Soc.*, 120(37):9401–9409, 1998.

[2] P.A. Kollman, I. Massova, C. Reyes, B. Kuhn, S. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, O. Donini, P. Cieplak, J. Srinivasan, D.A. Case, and T.E. Cheatham. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc. Chem. Res.*, 33(12):889–897, 2000.

[3] D.C. Thompson, C. Humblet, and D. Joseph-McCarthy. Investigation of mm-pbsa rescoring of docking poses. *J. Chem. Inf. Model.*, 48(5):1081–1091, 2008.

[4] G. Rastelli, A.D. Rio, G. Degliesposti, and M. Sgobba. Fast and accurate predictions of binding free energies using mm-pbsa and mm-gbsa. *J. Comput. Chem.*, 31(4):797–810, 2009.

[5] D. Jiao, J. Zhang, R.E. Duke, G. Li, M.J. Schnieders, and P. Ren. Trypsin-ligand binding free energies from explicit and implicit solvent simulations with polarizable potential. *J. Comput. Chem.*, 30(11):1701–1711, 2009.

[6] A. Lindström, L. Edvinsson, A. Johansson, C.D. Andersson, I.E. Andersson, F. Raubacher, and A. Linusson. Postprocessing of docked protein-ligand complexes using implicit solvation models. *J. Chem. Inf. Model.*, 51(2):267–282, 2011.

[7] T. Hou, J. Wang, Y. Li, and W. Wang. Assessing the performance of the mm/pbsa and mm/gbsa methods: Ii. the accuracy of ranking poses generated from docking. *J. Comput. Chem.*, 32(5):866–877, 2011.

[8] T. Yang, J.C. Wu, C. Yan, Y. Wang, R. Luo, M.B. Gonzales, K.N. Dalby, and P. Ren. Virtual screening using molecular simulations. *Proteins Struct. Funct. Bioinf.*, 79(6):1940–1951, 2011.

[9] S. Genheden. Mm/gbsa and lie estimates of host-guest affinities: dependence on charges and solvation model. *J. Comput.-Aided Mol. Des.*, 25(11):1085–1093, 2011.

[10] D.P. Oehme, R.T.C. Brownlee, and D.J.D. Wilson. Effect of atomic charge, solvation, entropy, and ligand protonation state on mm-pb(gb)sa binding energies of hiv protease. *J. Comput. Chem.*, 33(32):2566–2580, 2012.

[11] M. Sgobba, F. Caporuscio, A. Anighoro, C. Portioli, and G. Rastelli. Application of a post-docking procedure based on mm-pbsa and mm-gbsa on single and multiple protein conformations. *Eur. J. Med. Chem.*, 58:431–440, 2012.

[12] L. Xu, H. Sun, Y. Li, J. Wang, and T. Hou. Assessing the performance of mm/pbsa and mm/gbsa methods. 3. the impact of force fields and ligand charge models. *J. Phys. Chem. B*, 117(28):8408–8421, 2013.

[13] X. Zhang, S.E. Wong, and F.C. Lightstone. Toward fully automated high performance computing drug discovery: a massively parallel virtual screening pipeline for docking and molecular mechanics/generalized born surface area rescoring to improve enrichment. *J. Chem. Inf. Model.*, 54(1):324–337, 2014.

[14] H. Sun, Y. Li, S. Tian, L. Xu, and T. Hou. Assessing the performance of mm/pbsa and mm/gbsa methods. 4. accuracies of mm/pbsa and mm/gbsa methodologies evaluated by various simulation protocols using pdbbind data set. *PCCP*, 16(31):16719–16729, 2014.

[15] H. Sun, Y. Li, M. Shen, S. Tian, L. Xu, P. Pan, Y. Guan, and T. Hou. Assessing the performance of mm/pbsa and mm/gbsa methods. 5. improved docking performance using high solute dielectric constant mm/gbsa and mm/pbsa rescoring. *Phys. Chem. Chem. Phys.*, 16 (40):22035–22045, 2014.

[16] I. Maffucci and A. Contini. Improved computation of protein–protein relative binding energies with the nwat-mmgbsa method. *J. Chem. Inf. Model.*, 56(9):1692–1704, 2016.

[17] R. Huey, G.M. Morris, A.J. Olson, and D.S. Goodsell. A semiempirical free energy force field with charge-based desolvation. *J. Comput. Chem.*, 28(6):1145–1152, 2007.

[18] G.M. Morris, R. Huey, W. Lindstrom, M.F. Sanner, R.K. Belew, D.S. Goodsell, and A.J. Olson. Autodock4 and autodocktools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.*, 30(16):2785–2791, 2009.

[19] M.M. Mysinger, M. Carchia, J.J. Irwin, and B.K. Shoichet. Directory of useful decoys, enhanced (dud-e): Better ligands and decoys for better benchmarking. *J. Med. Chem.*, 55(14):6582–6594, 2012.

[20] Yongjun Chu and David R Corey. RNA sequencing: platform selection, experimental design, and data interpretation. *Nucleic acid therapeutics*, 22(4):271–274, 2012.

[21] I. Maffucci, X. Hu, V. Fumagalli, and A. Contini. An efficient implementation of the nwat-mmgbsa method to rescore docking results in medium-throughput virtual screenings. *Front. Chem.*, 6(43), 2018.

[22] J. Willem M. Nissink, Chris Murray, Mike Hartshorn, Marcel L. Verdonk, Jason C. Cole, and Robin Taylor. A new test set for validating predictions of protein-ligand interaction. *Proteins Struct. Funct. Bioinf.*, 49(4):457–471, 2002.

[23] Chemical Computing Group ULC. Moe, molecular operating environment, 2018. 1010 Sherbooke St. West, Suite 910, Montreal, QC, Canada, H3A 2R7.

[24] O. Korb, T. Stützle, and T.E. Exner. An ant colony optimization approach to flexible protein–ligand docking. *Swarm Intell.*, 1(2):115–134, 2007.

[25] O. Korb, T. Stüttzle, and T.E. Exner. Empirical scoring functions for advanced protein-ligand docking with plants. *J. Chem. Inf. Model.*, 49(1):84–96, 2009. doi: 10.1021/ci800298z.

[26] O. Trott and 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(2): 455–461, 2009.

[27] T.A. Halgren, R.B. Murphy, R.A. Friesner, H.S. Beard, L.L. Frye, W.T. Pollard, and J.L. Banks. Glide: a new approach for rapid, accurate docking and scoring. 2. enrichment factors in database screening. *J. Med. Chem*, 47(7):1750–1759, 2004.

[28] R.B. Murphy, M.P. Repasky, J.R. Greenwood, I. Tubert-Brohman, S. Jerome, R. Annabhimoju, N.A. Boyles, C.D. Schmitz, R. Abel, R. Farid, and R.A. Friesner. Wscore: A flexible and accurate treatment of explicit water molecules in ligand–receptor docking. *J. Med. Chem*, 59 (9):4364–4384, 2016.

[29] A. Gaulton, A. Hersey, M. Nowotka, A.P. Bento, J. Chambers, D. Mendez, P. Mutowo, F. Atkinson, L.J. Bellis, E. Cibrián-Uhalte, M. Davies, N. Dedman, A. Karlsson, M.P. Magariños, J.P. Overington, G. Papadatos, I. Smit, and A.R. Leach. The chembl database in 2017. *Nucleic Acids Res.*, 45(D1):D945–D954, 2017.

[30] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The Protein Data Bank. *Nucleic Acids Res.*, 28(1):235–242, 2000.

[31] D. A Case, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, D. Ghoreishi, M.K. Gilson, H. Gohlke, A.W. Goetz, D. Greene, R Harris, N. Homeyer, S. Izadi, A. Kovalenko, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, D.J. Mermelstein, K.M. Merz, Y. Miao, G. Monard, C. Nguyen, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R. Qi, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C.L. Simmerling, J. Smith, R. Salomon-Ferrer, J. Swails, R.C. Walker, J. Wang, H. Wei, R.M. Wolf, X. Wu, L. Xiao, D.M. York, and P.A. Kollman. Amber 2018, 2018.

[32] Felice C. Lightstone, Xiaohua Zhang, and Sergio E. Wong. Vinalc: Parallel molecular docking program. [Computer Software] `https://dx.doi.org/10.11578/dc.20180320.8`, 2014.

# CHAPTER 12

# WSELECT - a program for hydration site selection

## 12.1 Project background

Docking or virtual screening with explicit hydration has become more widely applicable within the field of drug design and discovery. The positions of the water molecules can be either derived from experimental data, i.e. X-ray crystallography or NMR, or computational methods that have become prevalent in recent years. Especially, *in silico* method can provide efficient prediction of water positions with acceptable accuracy using multiple methods, such as 3D-RISM,[1–3] WaterMap,[4,5] JAWS,[6] and SZMAP.[7] However, water selection according to the importance is inevitably required after positioning. This importance is in general determined by the bridging ability of the water molecules through visual inspection.[8] Inevitably, human bias can be easily introduced in this step, which could possibly affect later performance.

Alternatively, the water molecules can be selected through a distance cut-off to both receptor and a known active ligand.[9] This distance cut-off is in general narrow (<3

Å) which is required to exclude water been selected on the solvent side of the ligand. Additionally, AMBER *cpptraj* package[10] provides an option selecting molecules according to ranking in closeness to the targeted molecule(s) or residues. Such selections, in general, come with a sharp cut-off on one or several static structures, completely excluding the minor displacements of water in solution. Furthermore, these methods require a known active molecule in place to facilitate the selection. This could potentially bias the following virtual screening, since the selected hydration centres are characteristic for one or few active ligands, and potentially share certain chemical or structural properties. These can lead to an increase in the number of false negatives, which do not necessarily share these properties.

In this thesis, a simple and quick command-line based method, namely WSELECT, was developed to perform the selection step automatically. It selects water positions in terms of surrounding solvation importance and can also be performed on *apo* structures. Softness is also considered at the cut-off interface considering water positional adaptability.

## 12.2 Implementation

WSELECT is written in C++ with C++11 standard. The pseudocode of the essential algorithms are elucidated as in Algorithm 1 and 2 - separated based on *holo* or *apo* structures. The program uses PDB input files with the AMBER naming standard for amino acids, for both complex or receptor (e.g. complex.pdb or receptor.pdb). An additional file providing the positions of water oxygens (e.g. water_O.pdb) is also required. Multiple input parameters are user adjustable, which also come with default values. The selection performs on either *holo* or *apo* structures depending on user specifications.

For *holo* complexes, the name of the ligand within the *complex.pdb* needs to be provided. Otherwise, the program automatically determines the selection to be performed on an *apo* structure. As for the latter, the centre of the docking site shall be

**Algorithm 1** Water selection with *holo* structures. Output selected water position into $WS$ vector.

**Require:** $C_{pdb}$, complex file. $W_{pdb}$, file for water centre oxygens. $cutoff_{dis}$, distance cut-off (default 4.0 Å). $cutoff_{hydp}$, cut-off for hydropathic value (default -3.9). Recommend: $cutoff_{dis} \geq 3$, $cutoff_{hydp} < 0$ or default.

$C_{input} \leftarrow read\_pdb(C_{pdb})$ ▷ read pdb file into vector of atom object
$W \leftarrow read\_pdb(W_{pdb})$
$L \leftarrow lig(C_{input})$
$R \leftarrow rec(C_{input})$
**for** $w \leftarrow W(first\ to\ last)$ **do**
  **for** $r \leftarrow R(first\ to\ last)$ **do**
    **for** $l \leftarrow L(first\ to\ last)$ **do**
      $l_{dis} \leftarrow min\_distance(l\ to\ w)$ ▷ minimum distance from water centre to ligand atom
      $r_{dis} \leftarrow min\_distance(r\ to\ w)$ ▷ minimum distance from water centre to receptor atom
      **if** $l_{dis} \leq cutoff_{dis} and r_{dis} \leq cutoff_{dis}$ **then**
        **if** $w$ is close to backbone *and* interaction angle > cutoff **then**
          $WS.append(w)$
        **else**
          **if** $r_{dis} \leq 3$ **then**
            $w.hydp \leftarrow w.hydp + r.hydp$
          **else if** $3 < r_{dis} \leq cutoff_{dis}$ **then**
            $w.hydp \leftarrow w.hydp + GAUSS\_MOD(r.hydp, r_{dis})$
          **else**
            $skip()$
          **end if**
        **end if**
      **end if**
    **end for**
  **end for**
**end for**
**function** GAUSS_MOD(hydp, dis)
  $frac \leftarrow -exp((dis-3)^2/(2 \times 0.667^2))$
  **return** $hydp \times frac$
**end function**

specified. In the case of ligand *in situ*, WSELECT first screen out water positions based on distances (default: 4.0 Å) from oxygen to both ligand and protein. The calculated distances should be smaller than the provided cut-off for the water oxygen to be preserved for further screening.

When the site of interest is vacant, only the distances to the receptor and a volume

**Algorithm 2** Water selection with *apo* structures. Output selected water position into $WS$ vector.

**Require:** $C_{pdb}$, complex file. $W_{pdb}$, file for water centre oxygens. $cutoff_{dis}$, distance cut-off (default 4.0 Å). $cutoff_{hydp}$, cut-off for hydropathic value (default -3.9). $rad$, radius from the centre of the binding site (default 7.0 Å). $center$, centre coordinates of the binding site in the form of $x\ y\ z$. Recommend: $cutoff_{dis} \geq 3$, $cutoff_{hydp} < 0$ or default.

$C_{input} \leftarrow read\_pdb(C_{pdb})$ ▷ read pdb file into vector of atom object
$W \leftarrow read\_pdb(W_{pdb})$
$L \leftarrow lig(C_{input})$
$R \leftarrow rec(C_{input})$
$rad \leftarrow 7.0$
$center \leftarrow x\ y\ z$
$min_rad \leftarrow find\_min\_rad()$ ▷ detailed in implementation section
**for** $w \leftarrow W(first\ to\ last)$ **do**
    **for** $r \leftarrow R(first\ to\ last)$ **do**
        **for** $l \leftarrow L(first\ to\ last)$ **do**
            $c_{dis} \leftarrow min\_distance(l\ to\ center)$ ▷ minimum distance from water centre to ligand atom
            $r_{dis} \leftarrow min\_distance(r\ to\ w)$ ▷ minimum distance from water centre to receptor atom
            **if** $min\_rad < c_{dis} \leq radandr_{dis} \leq cutoff_{dis}$ **then** ▷ hollow spherical region
                **if** $w$ is close to backbone *and* interaction angle > cutoff **then**
                $WS.append(w)$
                **else**
                **if** $r_{dis} \leq 3$ **then**
                $w.hydp \leftarrow w.hydp + r.hydp$
                **else if** $3 < r_{dis} \leq cutoff_{dis}$ **then**
                $w.hydp \leftarrow w.hydp + GAUSS\_MOD(r.hydp, r_{dis})$
                **else**
                $skip()$
                **end if**
                **end if**
            **end if**
        **end for**
    **end for**
**end for**

around the centre are considered for first step selection. Since a certain volume of space at the specified site centre is more likely to be occupied by ligand, water molecules presented within such space shall be considered highly displaceable. Thus we introduce the selection space to be hollow-spherical according to further criteria (illustrated in Figure 12.1), and all water centres within this region are passed onto

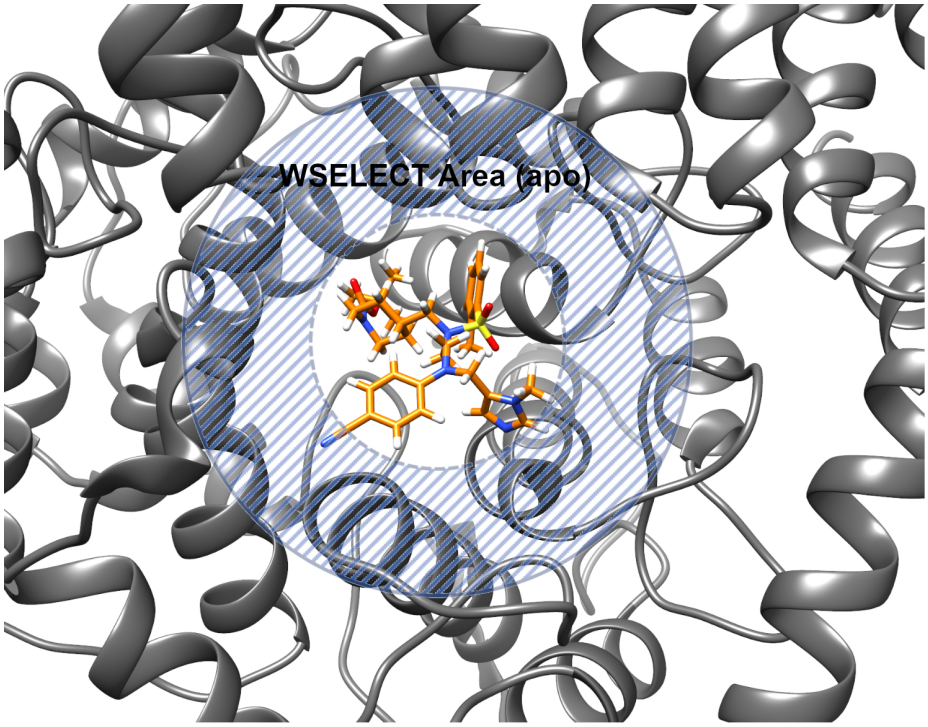



**Figure 12.1:** A representation of the concept of the hollow-spherical space for WSELECT performed on *apo* structures. The space is defined by two radii and the hollow space is reserved for ligand binding.

later selections. Such treatment not only reserves space for the ligand to occupy but also improves computational efficiency during the docking process. Two radii around the site centre define this hollow-spherical space: a maximum and a minimum. The maximum is the input radius for the search space (default: 7 Å). The minimum radius cut-off is determined with the consideration of binding site geometric properties. A part of the receptor close to the binding site is considered. This part consists of receptor atoms within the radius 2 Å larger than the maximum radius. The middle points between every two atoms are determined within the selected region. The distance from each middle point to the binding site centre is calculated and the mid-point furthest away from the centre is determined. Additionally, the minimum distance from protein to the site centre is also calculated. Thus, the minimum radius cut-off is calculated by averaging the values of the minimum protein-to-centre distance and the furthest mid-point distance to the centre. In this way, both the input radius and the rough shape of the binding site are considered to generate the hollow-spherical

region. Using the default radius (7 Å), the minimum radii vary from 4.6-5.48 Å in the *apo* test set (specified in later sections) using the above method, allowing variation between different receptors.

After the initial distance-based selection, the environment of the water centre is then considered. If the water oxygen is in proximity to a backbone peptide bond entity with a good angular directionality the water is prioritised for selection. The decent angular directionality is based on the O$\cdots$H-N or O$\cdots$O=C angles bigger than 120° and 115° respectively, with the latter smaller angle cut-off taken the potential water hydrogen into account.

For water centres left out after the backbone cycle of selection, the effects of sidechain are then evaluated. We adopted a series of values, namely hydrophatic characters,[11] for different amino acids to further score the importance of water centres based on their sidechain environments. The hydrophatic characters are derived from the water-vapour transfer free energies and the interior-exterior distribution of amino acid side-chains and have previously demonstrated with the capability to distinguish buried and solvent-exposed regions of proteins.[11] It was also adapted to select surface water molecules from 3D-RISM calculations.[12] These studies have provided inspiration in this work to apply such values to the WSELECT program. In addition to consider hydropathic values of the surrounding amino acids, a modulation according to distance on top of these values was also introduced, to allow certain softness in the cut-off. The modulation is implemented in the manner of a Gaussian function to the hydropathic contribution from the proximate amino acid side chains. This strategy can be represented as below:

$$h_{a.a.\rightarrow water} = \gamma \times h_{a.a.} \tag{12.1}$$

where

$$\gamma = \begin{cases} 1, & \text{if } r \leq 3 \\ e^{-\frac{(r-3)^2}{2\sigma^2}}, & \text{if } 3 < r \leq r_{max} \end{cases}$$

The $h_{a.a.\rightarrow water}$ is the hydropathic contribution from each amino acid to the designated water centre and the $h_{a.a.}$ is the original hydropathic character of the amino acid. The $\gamma$ is the correcting factor if the distance to side-chain ($r$) falls within the defined range. The $\sigma$ is set to 0.667 for providing the optimised results in the final selection. The hydropathic value for each water centre is calculated in an accumulative manner. The default hydropathic value cut-off is set to -3.9 but can be also adjusted by the user, the same for distance cut-off.

### 12.2.1 Input files

**Complex or receptor PDB file** - The PDB files with amber residue naming custom are preferred, though normal PDB residue names are also accepted. However, non-standard residues have no supported hydrophathic values. The values can be added through modifying the std::map included in the wsdelcare.cpp source file, though this is not recommended. Metal ions are not considered in the case of hydrophatic evaluation. The complex/receptor input should not include any water molecules.

**PDB file for water centres** - The file should strictly include water oxygen centres only.

**Reference PDB file** - WSELECT can optionally output comparison information to a reference file, most likely crystal reference, under the condition that a pdb file is inputted with *-cpdb* option. The output file of comparison lists the distances from the selected water centre to the closest reference. The RMSD to the reference, as well as the count of numbers water centre close to the reference ($< 2$ Å) are also included.

### 12.2.2 Output files

**Prefix** - A prefix of the output files can be defined using –o option. Default: WS.

**Selected water molecules** - The program will generate a PDB file containing selected water molecules. The two columns representing occupancy and temperature factor are replaced by the criteria that the water molecule has passed during the selection. "1.00 0.00" suggesting that the water molecule was determined to be close to the backbone. "0.00 x.xx" with x.xx a value suggesting the selected water molecule satisfies the hydropathic cut-off at a value of x.xx.

**Miscellaneous files** (*WS* can be replaced by -o input)
*WS_cmd_input* - Since the program was initially developed for the following application of the PLANTS docking software, a partial PLANTS configure input file is also automatically generated for the water in the required format for docking.
*WS_selected_info.dat* - Only generated if reference *-cpdb* option is on. This file includes information of input options for selection, distances of each water centre to the reference structure, RMSD to the reference structure, and the number of water centre selected that are close to the reference (distance $< 2$ Å).

## 12.3 Test set preparations

The same test set as the test set 0 in Chapter 11 was applied as the *holo* test set in this work. The preparations of the receptor and ligand were carried out in the same procedures. For 3D-RISM calculation included in AMBER simulation package,[10] the ligands were prepared with antechamber using AM1-BCC partial charges.[13] The *ff03*[14] and *gaff*[13] force fields were applied for protein and ligand, respectively. The topology files were generated using *tleap*.[10]

The *apo* test set was generated based on the *holo* test set (Table 12.1). It consisted of *apo* structures published along with the complexes presented within the *holo* test

**Table 12.1:** Test sets applied for the WSELECT testing

| Class | Size | Protein contained (PDB ID) |
|---|---|---|
| *holo* test set | 73 (97 binding sites) | Same as in test set 0, Table 11.1, Chapter 11 |
| *apo* test set | 12 (31 binding sites) | 1bya, 1c5b, 1c5v, 1htd, 1kem, 1mri, 1ttb, 1ttc, 1xib, 2ctb, 2lao, 2qwa |

set. Since not all structures were published with an *apo* complement, the size of this test set is significantly smaller. However, each binding site presented with the structures were considered separately. Thus, the test set resulted in 31 binding sites. The structure preparation was the same as the *holo* test set only that the ligand-concerned procedures were excluded.

## 12.4 Testing methodology

WSELECT was tested both on experimentally determined and *in silico* calculated water positions. The standard procedures of 3D-RISM calculation in AMBER[10,15] was followed using TIP3P water, $Na^+$, and $Cl^-$ with concentrations of 55.5M, 0.15M, and 0.15M, respectively. The Kovalenko-Hirata closure relation[16] was also applied.

Between the water positions selected from crystal structures or calculated results, the performance of WSELECT was challenged by counting correlations, true positive rate (TPR), and docking experiments. Additionally, WSELECT was also compared to the selection by using only distance cut-offs at 3 and 4 Å. The water selected through distance cut-off was defined as the selected water molecules are closer than the cut-off to both of the ligand and the receptor in the *holo* test set or to only the receptor in the *apo* test set. The docking procedures were performed as same as depicted in Chapter 10 except in the PLANTS configure file, where the information of the water coordinates are added accordingly.

## 12.5 Results and discussions

The correlation of selected water position counts between WSELECT selections using crystal structures and 3D-RISM-placevent calculated data was performed for the *holo* test set. A poorer Pearson correlation coefficient (0.59) was observed if the total selected water positions were considered. However, if only consider the ones that are closed to the crystal structures (distance $< 2$ Å), the Pearson correlation

coefficient improved to 0.75. The average counts of water positions selected from crystal structures (2.49) is also similar to those reproduced positions selected from 3D-RISM results (2.82). The same correlation behaviour was also observed for the *holo* test set; Pearson correlation coefficient improved from 0.50 to a good 0.78 when considering the reproduced water positions between the two WSELECT testing on

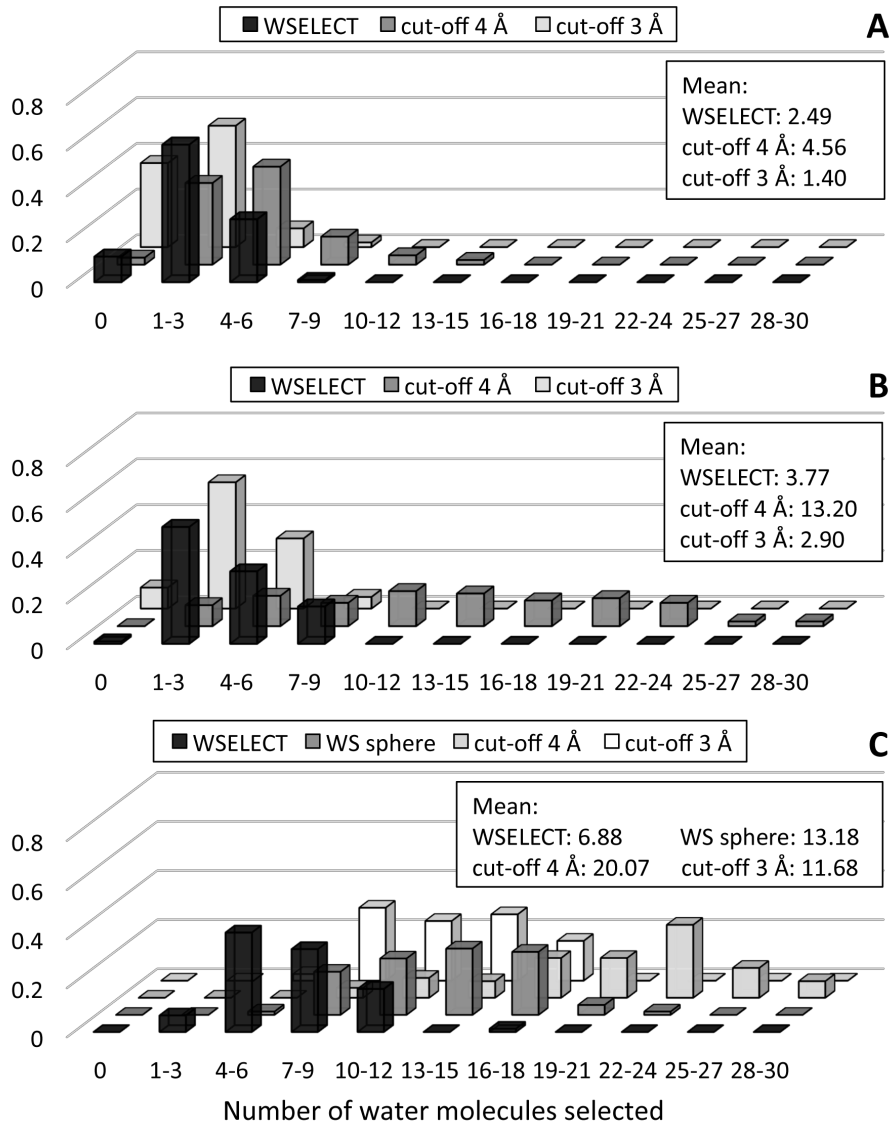



**Figure 12.2:** Bar chart histograms of numbers of water molecules selected using different methods and on different test sets. A: selections performed on the crystal structures of the *holo* test set. B: selections performed on the 3D-RISM calculated water positions from *holo* test set. C: selections performed on the 3D-RISM calculated *apo* test set. Means of selected waters using each method are shown in boxes.

**Table 12.2:** The information including TPR regarding selected water positions in the *holo* and *apo* test set

Panel A: the *holo* test set

| | WSELECT | distance cut-off 3 Å | distance cut-off 4 Å |
|---|---|---|---|
| Total Selected | 369 | 284 | 1294 |
| Total Reproduced* | 276 | 182 | 686 |
| TPR | 0.748 | 0.641 | 0.530 |

Panel B: the *apo* test set

| | WSELECT | WS sphere** | distance cut-off 3 Å | distance cut-off 4 Å |
|---|---|---|---|---|
| Total Selected | 172 | 355 | 421 | 649 |
| Total Reproduced* | 137 | 288 | 327 | 457 |
| TPR | 0.797 | 0.811 | 0.777 | 0.704 |

* The selected water positions close to crystal water centres
** The WSELECT method on an *apo* structure in a spherical space in the binding site

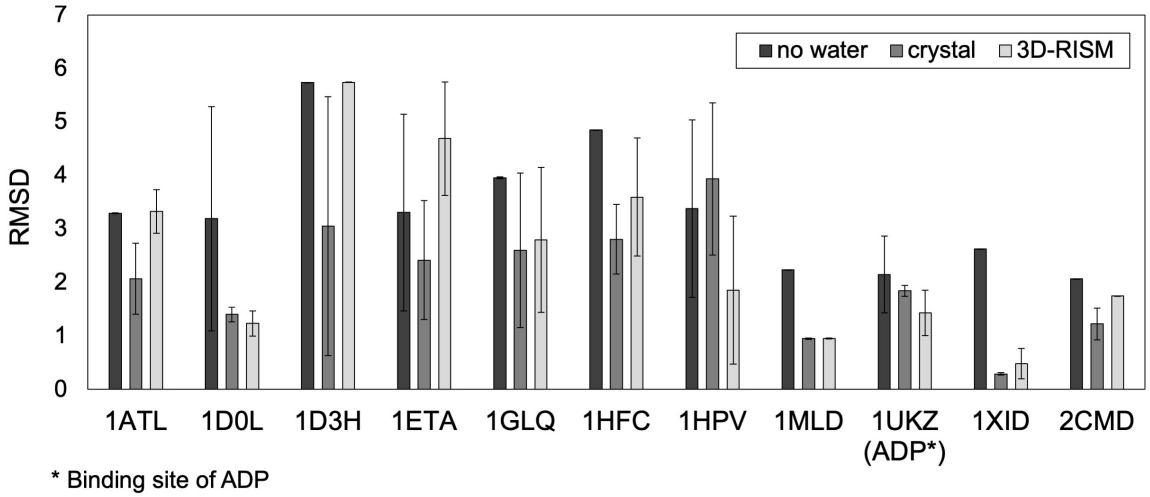



**Figure 12.3:** RMSD bar charts of complexes with improved crystal pose reproducibility when explicit hydration was included. Both experimental-derived and calculated water positions were tested.

crystal structures and calculated position respectively. These suggest that the water positions selected using WSELECT program provide some false positives, but the majority of the experimental water positions within the binding sites were reproduced by 3D-RISM calculation and WSELECT selection.

Moreover, performance comparison was carried out between WSELECT to distance cut-off water selections on the 3D-RISM-placevent calculated water positions in the *holo* test set. WSELECT tends to provide a lower number of selected positions from the count histogram analysis (Figure 12.2), especially when comparing to the selection using distance cut-off of 4 Å. Furthermore, the TPR analysis (Table 12.2 Panel A) reflects that significant higher false positives were selected when using distance cut-off methods.

As for the *apo* test set, an additional comparison was performed with a spherical selection space around the binding site centre (WS-sphere), to validate the effectiveness of the hollow sphere strategy adopted by WSELECT. From histogram analyses, only the WSELECT method provides a condensed distribution of numbers of selected water molecules around 4-12 water molecules. This observation suggests the robustness of WSELECT when performing selections on both *holo* and *apo* structures. However, the TPR analysis showed a different trend as in the *holo* test set; the WSELECT failed to provide a much better TPR than other method applied (Table 12.2 Panel B). It is likely that without a ligand occupying a big part of the binding site, the 3D-RISM-placevent calculation provided a very condensed water density within the binding sites. Multiple water centres resulted in close proximity to each crystal resolved water centre, leading to higher TPR values.

Initial tests on docking performance also shown some promising results. Multiple systems within the *holo* test set provided binding poses closer to the crystal pose once explicit waters are included (Figure 12.3). Interestingly, one among them is 1hpv, the structure of HIV-1 protease complexed with its inhibitor, VX-478.[17] It was suggested that one water is important to bridging the ligand to the receptor. This water was selected by WSELECT suggesting that WSELECT is likely able to determine the important water within the binding site. However, this is yet to be confirmed with a larger test set consisting of receptors with known important bridging waters upon ligand binding.

## 12.6 Conclusion

WSELECT provides an effective and efficient water selection on a single-frame basis. It considers not only the chemical environment of each water molecule presented but also the geometric property of the binding site when an *apo* structure is considered. It demonstrated good improvements in two decent test sets under both *holo* and *apo* scenario to traditional distance cut-offs. The algorithm, as demonstrated, is simple

and straightforward, and is not computationally expensive. With implementation in C++11 standard, depending on the size of the system, the selection time is within the range of seconds. Additionally, it can also be easily incorporated into other software packages as intended.

However, the method still requires further testing and adjustment. The selection performance should be compared to more available software and more docking validations should also be carried out. It also worths to assess the WSELECT when incorporated into a virtual screening workflow. Technically, the water selection on an *apo* structure should be further optimised. It remains challenging for the WSELECT program to work in a significantly larger binding site due to the smaller default radii applied. One approach can be outlining the surface of the receptor to introduce more geometric specificity of each receptor. Nonetheless, WSELECT can be a useful tool when a relatively enclosed binding site is considered, or in a ligand-bound complex structure.

## 12.7 Example commands

*holo* structure with regular ligand (ligand name in pdb is LIG):

```
wselect -pdb complex.pdb -opdb water_O.pdb -t l LIG
wselect -pdb complex.pdb -opdb water_O.pdb -cpdb crystal.pdb -t l LIG
```

*holo* structure with peptide (peptide sequence from residue 1 to 10):

```
wselect -pdb complex_pep.pdb -opdb water_O_pep.pdb -t p 1 10
wselect -pdb complex_pep.pdb -opdb water_O_pep.pdb -cpdb crystal_pep.pdb -t p 1 10
```

*apo* structure:

```
wselect -pdb complex_apo.pdb -opdb water_O_apo.pdb -center x y z
wselect -pdb complex_apo.pdb -opdb water_O_apo.pdb -cpdb crystal_apo.pdb -center x y z
```

# Chapter references


[1] A. Kovalenko and F. Hirata. Three-dimensional density profiles of water in contact with a solute of arbitrary shape: a RISM approach. *Chemical Physics Letters*, 290(1):237–244, 1998.

[2] A. Kovalenko and F. Hirata. Self-consistent description of a metal–water interface by the Kohn–Sham density functional theory and the three-dimensional reference interaction site model. *The Journal of Chemical Physics*, 110(20):10095–10112, 1999.

[3] A. Kovalenko. Multiscale modeling of solvation in chemical and biological nanosystems and in nanoporous materials. *Pure and Applied Chemistry*, 85(1):159–199, 2013.

[4] Tom Young, Robert Abel, Byungchan Kim, Bruce J Berne, and Richard A Friesner. Motifs for molecular recognition exploiting hydrophobic enclosure in protein–ligand binding. *Proceedings of the National Academy of Sciences*, 104(3):808–813, 2007.

[5] R. Abel, T. Young, R. Farid, B.J. Berne, and R.A. Friesner. Role of the active-site solvent in the thermodynamics of factor Xa ligand binding. *Journal of the American Chemical Society*, 130(9):2817–2831, 2008.

[6] J. Michel, J. Tirado-Rives, and W.L. Jorgensen. Prediction of the water content in protein binding sites. *Journal of Physical Chemistry B*, 113(40):13337–13346, 2009.

[7] OpenEye. SZMAP 1.2.1.4, 2015. URL `http://www.eyesopen.com`.

[8] R. Thilagavathi and R.L. Mancera. Ligand-protein cross-docking with water molecules. *J. Chem. Inf. Model.*, 50(3):415–421, 2010.

[9] B.C. Roberts and R.L. Mancera. Ligand−Protein Docking with Water Molecules. *J. Chem. Inf. Model.*, 48(2):397–408, 2008.

[10] D. A Case, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, D. Ghoreishi, M.K. Gilson, H. Gohlke, A.W. Goetz, D. Greene, R Harris, N. Homeyer, S. Izadi, A. Kovalenko, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, D.J. Mermelstein, K.M. Merz, Y. Miao, G. Monard, C. Nguyen, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R. Qi, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C.L. Simmerling, J. Smith, R. Salomon-Ferrer, J. Swails, R.C. Walker, J. Wang, H. Wei, R.M. Wolf, X. Wu, L. Xiao, D.M. York, and P.A. Kollman. Amber 2018, 2018.

[11] J. Kyte and R.F. Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology*, 157(1):105–132, 1982.

[12] W.J. Huang, N. Blinov, D.S. Wishart, and A. Kovalenko. Role of water in ligand binding to maltose-binding protein: Insight from a new docking protocol based on the 3D-RISM-KH molecular theory of solvation. *Journal of Chemical Information and Modeling*, 55(2):317–328, 2015.

[13] J. Wang, R.M. Wolf, J.W. Caldwell, P.A. Kollman, and D.A. Case. Development and testing of a general amber force field. *J. Comput. Chem.*, 25(9):1157–1174, 2004.

[14] Y. Duan, C. Wu, S. Chowdhury, M.C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, J. Caldwell, J. Wang, and P. Kollman. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *Journal of computational chemistry*, 24(16):1999–2012, 2003.

[15] Sindhikara D. Dan sindhikara's tutorials - using 3d-rism and placevent. `http://dansindhikara.com/Tutorials/Entries/2012/1/1_Using_3D-RISM_and_PLACEVENT.html`, 2012. Accessed: Oct 2016.

[16] A. Kovalenko and F. Hirata. Self-consistent, Kohn-Sham DFT and three-dimensional RISM description of a metal-molecular liquid interface. *Journal of Molecular Liquids*, 90(1):215–224, 2001.

[17] E.E. Kim, C.T. Baker, M.D. Dwyer, M.A. Murcko, B.G. Rao, R.D. Tung, and M.A. Navia. Crystal structure of HIV-1 protease in complex with VX-478, a potent and orally bioavailable inhibitor of the enzyme. *Journal of the American Chemical Society*, 117(3):1181–1182, 1995.

# CHAPTER 13

# Workflow for network analysis

## 13.1 RNA-sequencing and differential gene expression

RNA-sequencing (RNA-seq) method has become an essential tool to identify and quantify gene expression in a high-throughput manner. The technique applies the next-generation sequencing to detect and quantify RNA of a biological sample at a given state.[1?] The complete process has been well reviewed elsewhere.[1–3] Supported by the deep sequencing techniques, it allows accurate sequencing and counting of a library of fragmented RNA or cDNA. The reads are then aligned to a reference genome or transcriptome, or undergo a *de novo* assembling process to construct a transcription map.

Several essential steps should be taken into consideration for a complete RNA-seq pipeline to ensure successful assessments. The pre-analysis steps include experimental design, sequencing design, and quality control.[3] The highly abundant ribosomal RNA (rRNA) should be removed prior to analysis to ensure only the more important

messenger RNAs (mRNA) are sequenced. The sequenced RNA library size is related to the precision of the quantification and multiple replicates should also be performed for determining the statistical significance.

The analysis steps start with the quality control of the sequencing data. The raw reads are checked for multiple properties including sequence quality, GC content, and the presence of adaptors using popular toolkits such as FastQC,[4] NGSQC[5], and FASTX-Toolkit[6]. The alignment quality of the RNA fragments should also be assessed. A uniformed coverage on the exons the RNA fragments mapped to are expected for valid RNA-seq analyses.

In general, the reads are mapped to a reference genome or transcriptome. The results can also be used for novel transcripts discovery or even assembled *de novo* into a new transcriptome. The quantifications of transcripts are determined by counts accompanied by appropriate sample normalisation methods, such as reads per kilobase of exon model per million reads (RPKM) and fragments per kilobase of exon model per million mapped reads (FKM). These normalisation methods generally involve the correction for gene length since longer genes tend to have more fragments, i.e., more reads. However, it is not always necessary for comparing differences in gene expression among the same genes, for example, the application in differential gene expression (DGE) analysis.

In DGE analysis, the gene expression values are compared between samples. The analysis relies on test statistics to determine which genes have statistically significant changes in the expression. Non-parametric methods, in theory, are applicable accounting for gene expression probability distributions. However, this would require a larger number of replicates to reach acceptable detection sensitivity. Hence, most of the DGE analyses involve the application of discrete probability distributions, for instances, the Poisson or negative binomial,[7,8] to compute differential expression. The popular tool, edgeR, takes in the raw count data and integrated normalisation modelling negative binomial (NB) distributions along with the DGE analysis.[9]

In this work, the standardised edgeR differential expression analysis pipeline has

been characterised for the current DGE analysis between drug-responsive and non-responsive gastric cancer cell lines. Multiple certain steps were automated for comparisons between various samples and some quality controls are also included. Moreover, multiple output files were also generated automatically for further data analysis and inspections. The workflow is also applicable for smallRNA-sequencing analysis and can be pipelined to network analysis.

## 13.2 Network between gene expression and cellular phenotypes

Genes rarely work alone. The observable characteristics, namely phenotypes, are determined by a network of co-regulated genes. Though in the past years a large amount of information regarding individual genes and their ontology relationships, it remains a challenge to establish an integrated and accurate network between cellular phenotypes and gene expressions. Recent years multiple approaches have been developed to facilitate producing gene regulatory networks (GRNs) with genes as vertices connected by edges defined by different criteria. Multiple approaches have been proposed such as Bayesian Networks, Relevance Networks, and Graphical Gaussian Models.[10–12]

One recent approach, ARACNE (Algorithm for the Reconstruction of Accurate Cellular Networks), was introduced as an information-theoretic algorithm with the edges defined as irreducible statistical dependencies between gene expression profiles.[13] Essentially similar to Relevance Networks[10], the algorithm infers the dependencies (edges) by removing the indirect candidate interactions based on the data processing inequality (DPI) theory. It has been shown with a generalised application in combination with mutual information (MI) ranking and provides good resilience to estimation errors.[13,14] Multiple studies have shown promising applications of the method, such as GRNs and miRNA cross-talks in multiple cancer cell lines.[15–19] Moreover, Turan et al. have explored the usefulness of ARACNE in investigating the statistical relationship between whole-body systemic variables to expression profiles and have

derived a relatively accurate network of muscles from the chronic obstructive pulmonary disease on a systemic level.[20] All have proven that ARACNE in combination with MI is an efficient and powerful tool for inferring large networks.

With the objective of investigating the connections between cellular phenotype measurements to changes in genomic expression, the initial steps of interaction network inference and analysis have been assessed using DGE and differential microRNA expression (DME) data of gastric cancer cell lines. The method was expected to be able to identify biomolecular markers for drug-responsive or non-responsive cell lines. The workflow can be extended to include phenotype measurements once available and is expected to provide complete and unbiased information regarding cross-talk between genome and phenotype properties.

## 13.3 Methodologies

### 13.3.1 The briefer overview

The workflow consists of two separate stages: the DGE/DME analysis step followed by the network inference. The DGE/DME analysis takes in the raw counts from RNA-seq or small RNA-seq experiments and automatically output results for each comparison (for example between drug-treated and negative control for one specific cell line). The output results in this step include one .cvs file including all the statistical data and a summary of DGE information. Additionally, three more figures were also generated for data visualisations for each comparison. Following this, the DGE/DME statistical data were used as inputs to the second stage of the workflow. The logFC outcomes were combined, after steps of filtering and transformation, for network inferences. Only genes or miRNAs with significant differential expressions in certain comparisons were included in the analysis. Data regarding the degree of connectivity of each node (representing genes) are saved into an output file. Each network was also plotted for visualisation and a preliminary gene ontology analysis was also included in the workflow. A brief outline of the overall workflow is illustrated

in Figure 13.1 and the detailed information regarding each step inside the workflow are elucidated in the following sections.

### 13.3.2 DGE and differential microRNA analysis

The RNA-seq and smallRNA-seq raw count data of gastric cancer cell lines were provided by Biomax Informatics AG, Munich, Germany. The cell lines included are Hs746T (gastric adenocarcinoma), MKN1 (gastric adenosquamous carcinoma), MKN7 (gastric tubular adenocarcinoma), and NCI-N87 (gastric tubular adenocarcinoma). The cell lines were treated by cancer drugs in a series of experiments. The drugs included are Afatinib (tyrosine kinase inhibitor), Cetuximab (EGFR inhibitor, a monoclonal antibody), and Trastuzumab (a monoclonal antibody for breast cancer). The responsive and non-responsive cell lines were determined through certain criteria and RNA-seq and smallRNA-seq information were also collected.

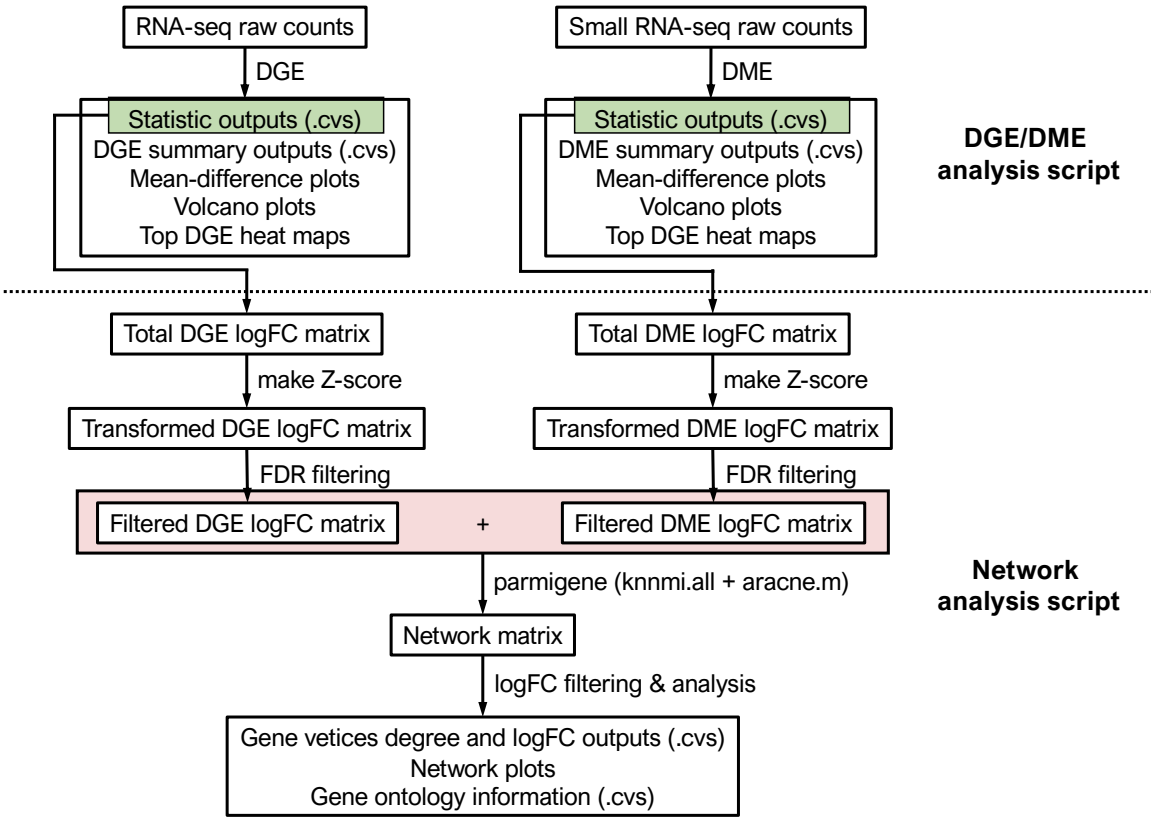
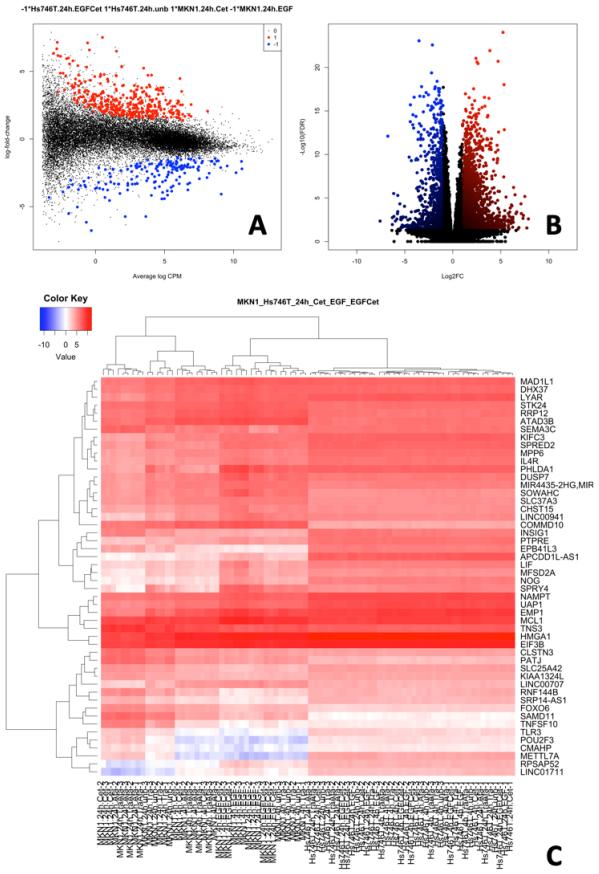



**Figure 13.1:** The flow chat for the DGE/DME-network workflow.

The DGE and DME analyses were performed using edgeR package[9] in R.[21] The raw count data include three replicates of each experiment. The column name of the experimental data are characterised by the cell line name, treatment time length, type of treatment, and repeat index. For example, "Hs746T_24h_Afa_1" represents that the HS746T cells were treated with Afatinib for 24 hrs and this is the first replicate of the experiment. Hence, initial steps within the DGE/DME analysing script are data managements by grouping the three replicates and recording with the same experiment name, as this grouping information is also required for edgeR analysis. For example, for the three columns named "Hs746T_24h_Afa_1", "Hs746T_24h_Afa_2", and "Hs746T_24h_Afa_3", after data processing, will be grouped under the name "Hs746T.24h.Afa" in the group vector as one of the edgeR inputs. An additional design matrix (a template matrix for contrast matrix construction) is also generated automatically at this step.

One additional feature of the script is that it can automatically generate the list of contrast for drug treatment against control according to column name pattern of the data. Contrast matrix is an additional parameter that is required for gene-wise statistical tests in edgeR.[9] It defines the relation of experiments to be retrieved for statistical comparisons. The auto-generated contrast list only considers the comparisons to controls. However, the user can also define a customised list for more complicated comparison. One useful feature included in edgeR is that it has incorporated the possibility of performing the non-pairwise comparison, as long as the corrected contrast matrix is provided. For example, one auto-generated contrast can be comparing 24hrs Afatinib treatment of Hs746T to control (i.e. the RNA-seq experiment performed without treatments). In this work, however, it is also interesting to investigate the DGE between Hs746T (non-responsive) and NCI-N87 (responsive) under Afatinib treatment. Hence, the contrast can be depicted in the way that the treatment of the individual cell line to control will be considered plus the comparison between different cell lines. The contrast list is also saved in an additional file for data verifications.




**D**

| Type of DGE | No. of genes |
|---|---|
| Down | 165 |
| NotSig | 22023 |
| Up | 342 |

**E**

| SYMBOL | logFC | logCPM | F | PValue | FDR |
|---|---|---|---|---|---|
| LINC00941 | -3.9497078 | 4.85570685 | 234.727608 | 2.28E-27 | 5.13E-23 |
| METTL7A | 5.23409799 | 3.48444478 | 195.590853 | 9.31E-25 | 1.05E-20 |
| NOG | -3.5180594 | 4.62439929 | 182.2936 | 8.66E-24 | 6.50E-20 |
| SOWAHC | -2.1226066 | 6.37886733 | 175.953313 | 2.61E-23 | 1.47E-19 |
| SAMD11 | 3.86266878 | 3.94843797 | 167.517465 | 1.18E-22 | 5.30E-19 |
| CLSTN3 | 2.41313028 | 5.75197188 | 156.587098 | 8.93E-22 | 3.35E-18 |
| TNFSF10 | 4.71937373 | 4.05489536 | 152.702528 | 1.97E-21 | 6.34E-18 |
| SEMA3C | 2.52225504 | 6.29353584 | 151.434306 | 2.39E-21 | 6.74E-18 |
| PATJ | 2.62534562 | 5.92050134 | 149.381718 | 3.57E-21 | 8.93E-18 |
| KIFC3 | -2.2115168 | 6.2999866 | 137.023589 | 4.24E-20 | 9.56E-17 |
| COMMD10 | -1.6365593 | 5.84554119 | 126.492699 | 3.90E-19 | 7.98E-16 |
| POU2F3 | 5.336369 | 1.89008523 | 122.428492 | 9.44E-19 | 1.77E-15 |
| PHLDA1 | -3.2521912 | 6.82159034 | 120.891416 | 1.61E-18 | 2.74E-15 |
| MIR4435-2HG | -1.8301424 | 5.82370682 | 119.760133 | 1.70E-18 | 2.74E-15 |
| STK24 | -0.9904071 | 6.88487272 | 119.163446 | 1.94E-18 | 2.92E-15 |
| ATAD3B | -1.4884384 | 6.90874347 | 117.922162 | 2.57E-18 | 3.38E-15 |
| DUSP7 | -1.8770775 | 5.84055578 | 117.580016 | 2.77E-18 | 3.38E-15 |
| MPP6 | -1.369225 | 5.92920624 | 117.497596 | 2.82E-18 | 3.38E-15 |
| INSIG1 | 2.17221535 | 5.51252684 | 117.45759 | 2.85E-18 | 3.38E-15 |
| MAD1L1 | -1.4653282 | 6.74174921 | 114.64736 | 5.38E-18 | 6.06E-15 |
| UAP1 | -1.4133094 | 7.53269404 | 113.744543 | 6.62E-18 | 7.10E-15 |
| LINC00707 | -2.779213 | 4.27737869 | 113.012857 | 7.83E-18 | 8.01E-15 |
| KIAA1324L | 1.31817409 | 4.50175454 | 111.649527 | 1.07E-17 | 1.05E-14 |
| EMP1 | -2.5338364 | 8.23321195 | 110.51927 | 1.39E-17 | 1.31E-14 |
| MCL1 | -1.4366622 | 8.66587742 | 109.988475 | 1.58E-17 | 1.42E-14 |
| RPSAP52 | -3.4337895 | 2.33475809 | 108.981559 | 1.99E-17 | 1.73E-14 |
| APCDD1L-AS1 | -3.0312832 | 5.86182338 | 108.878389 | 2.19E-17 | 1.83E-14 |
| LINC01711 | -4.026778 | 1.76471144 | 107.809516 | 2.63E-17 | 2.11E-14 |
| EPB41L3 | 3.29895024 | 4.54578808 | 107.752851 | 2.76E-17 | 2.14E-14 |

**Figure 13.2:** The output files for each contrast experiment using the DGE/DME analysis script. A: MD plot. B: Volcano plot. C: Heat map of top differentiated genes of the contrast. D: the summary file of up- or down-expressed genes using the contrast. E: The file including the detailed statistic data for each gene, ranked according to FDR.

The complete raw data were read from a .csv file, underwent the above-mentioned data processing, and were converted to a DGEList object in edgeR. Genes with a minimum count of 10 and a minimum total count of 15 for each group are initially filtered out. Then the normalisation factors for scaling the raw library sizes were calculated using the weighted trimmed mean of M-values.[22]

The DGE/DME analysis was done on a per-contrast basis. The data retrieved for each contrast were undergone gene-wise statistical tests by fitting to a quasi-likelihood negative binomial generalised log-linear model included in the edgeR package. The empirical Bayes quasi-likelihood F-tests was applied. A file summarising the logFC, logCPM (count-per-million), F-test, p-value, and false discovery rate (FDR) of all

the genes after count filtering was generated. Moreover, an additional summary file of the numbers of up- or down-regulated genes was also among the outputs. For data visualisation, three types of plots are also included inside the script. A mean-difference plot was generated for each contrast with the up- and down-regulated genes highlighted in red and blue respectively. The same type of colouring was also adopted in a volcano plot. Furthermore, a heat map illustrating the expression levels of the top FDR-ranked genes across all the related experiments (i.e. experiments performed on the same cell line(s)) of the contrast is also included in the outputs. A demonstration of the output files for each contrast is shown in Figure 13.2.

### 13.3.3 Network inferences combining DGE and DME

Network analysis was performed on the DGE and DME logFC data generated from the previous step. Since the RNA-seq and small RNA-seq experiments were performed with the same set of drug treatments, this resulted in the sets of output files from DGE and DME analysis have the same name patterns. This provided convenience in file management prior to network inferences. However, other analysis can also be included within the file with proper adaptations, though within the scope of this thesis data other than (small) RNA-seq experiments were not considered.

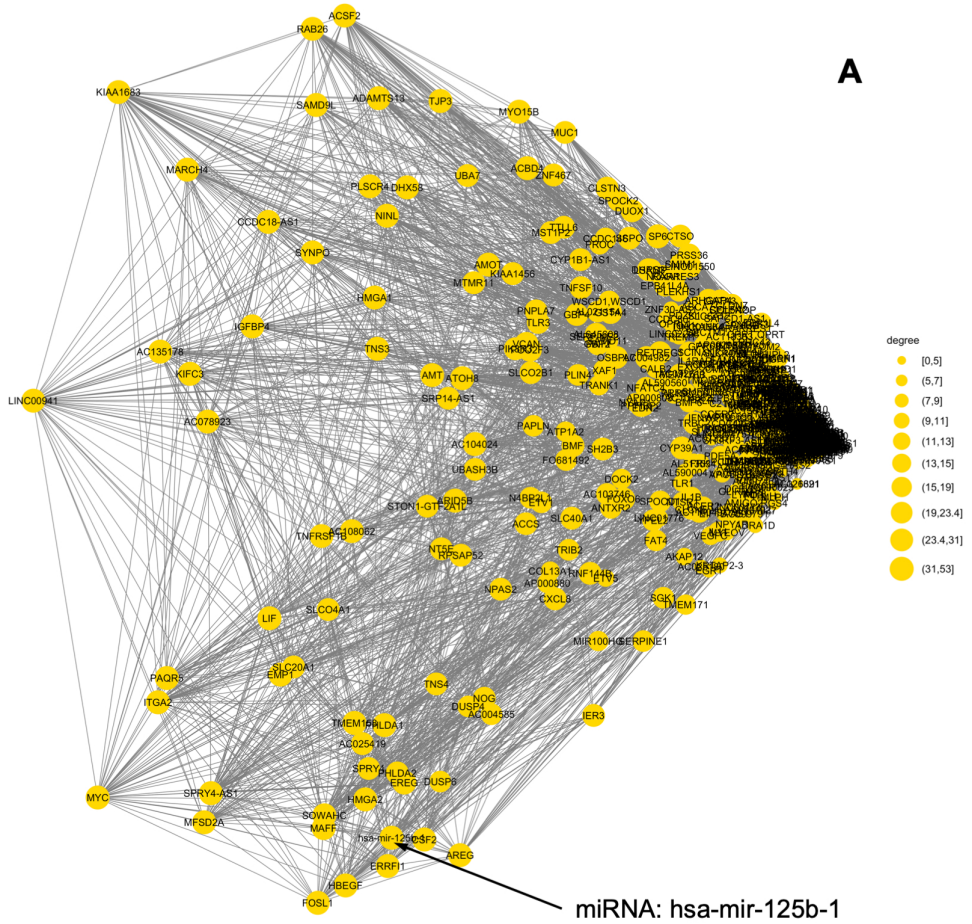
Some initial testings were performed on the complete DGE and DME including all the genes and miRNAs. However, this leads to the universal identification of the category of chaperone proteins as the most connected nodes among all the experiments. Moreover, these identified genes were not statistically differentially expressed, i.e. the FDR values of these tested genes were all larger than 0.05. This is a reasonable outcome since chaperones are responsible for protein folding, which would likely have the best connectivity to other genes in network analysis. However, they probably have the same expression level across all gastric cancer cell lines. Moreover, the high connectivities of chaperones or other generally highly recruited proteins would mask the more significantly differentially expressed genes (but less connected genes) between drug treating experiments. Hence, it is essential to include FDR filtering

prior to network inference for more accurate analysis. The DGE results were firstly filtered with a more strict 0.01 FDR cut-off; any genes with an FDR larger than 0.01 were not included in the later analysis. Due to the nature of the small RNA-seq data provided, none of the miRNA tested has shown any significant differential expression in any contrast experiment analysed. Hence, for testing purpose, only the 20 miRNA with the lowest FDR were included in the network inference.

The network analysis R script consists of two loops. The first loop works to record the list of DGE contrasts that have more than 1000 genes with an FDR less than 0.01 as the looping criteria for the second loop. Furthermore, the loop also collects all the logFC information from each DGE and DME statistics files and forming two matrices for DGE and DME, respectively. Combining two datasets from different types of experiments (RNA-seq and small RNA-seq in this case), it requires them to follow the same statistical characteristics, i.e. they should be fitted to the same probability density function to allow further data comparison. Hence, both DGE and DME logFC matrices underwent Z-score transformation (the data mean was adjusted to 0 and the standard deviation to 1). This step forms the transformed matrices for DGE and DME analysis and the second loop starts.

The second loop goes through the list of comparisons (contrasts) recorded in loop one and performs network inference and analysis for each contrast. For each contrast, the significantly differentially expressed genes (FDR $< 0.01$) or miRNAs (the 20 miRNA with the lowest FDR values) were first identified and selected for further analysis. A new matrix was formed by retrieving information of the selected genes and miRNAs from the transformed logFC matrices for DGE and DME analysis. This matrix then underwent parallel mutual information estimation between all matrix rows using the *knnmi.all* function with default settings included within the *parmigene* package.[14] Following this, the gene interaction network was reconstructed using the ARACNE algorithm implemented by *parmigene* package[14] using the multiplicative model. The recommended settings specified in Ref. 14 were applied ($\tau = 0.15$). The network was then further filtered with an absolute logFC cut-off at 2; this ensures that not only the

differential expression of the selected genes are statistically significant, but also the fold change in expression levels are more than 4-fold differences. This cut-off value can be lowered down if required in other analysis. The connectivity degrees were




**B**

| ENSEMBLE_ID | SYMBOL | degree | logFC |
|---|---|---|---|
| ENSG00000137309 | HMGA1 | 332 | -2.2079331 |
| ENSG00000028137 | TNFRSF1B | 331 | -2.0095414 |
| ENSG00000134531 | EMP1 | 316 | -2.5338364 |
| ENSG00000141753 | IGFBP4 | 316 | -2.4910823 |
| ENSG00000198142 | SOWAHC | 283 | -2.1226066 |
| ENSG00000140859 | KIFC3 | 278 | -2.2115168 |
| ENSG00000235884 | LINC00941 | 260 | -3.9497078 |
| ENSG00000260328 | AC104024 | 241 | -2.1389807 |
| ENSG00000265749 | AC135178 | 239 | 2.81449857 |
| ⋮ | | | |

**C**

| GO_ID | Freq | Term |
|---|---|---|
| GO:0007165 | 49 | signal transduction |
| GO:0007275 | 39 | multicellular organism development |
| GO:0030154 | 33 | cell differentiation |
| GO:0045944 | 31 | positive regulation of transcription by RNA polymerase II |
| GO:0007155 | 30 | cell adhesion |
| GO:0006355 | 27 | regulation of transcription, DNA-templated |
| GO:0008284 | 26 | positive regulation of cell proliferation |
| GO:0010628 | 25 | positive regulation of gene expression |
| GO:0019221 | 24 | cytokine-mediated signaling pathway |
| ⋮ | | |

**Figure 13.3:** The output files for each contrast experiment using the neetwork analysis script. A: gene interaction network plot. B: data for gene (node) connectivity degree and logFC. C: preliminary gene ontology information.

then summarised along with logFC of each selected gene or miRNA and written into a .cvs file. The network was plotted using the *ggnet2* function included in the *ggplot2* package[23] with the *eigen* mode. A preliminary consensus gene ontology analysis was also included, which requires the loading of a file mapping each gene to its gene ontology at the start of the network analysis. The data were retrieved from *GO.db* package.[24] An example of the output files from network analysis script is shown in Figure 13.3.

## 13.4 Discussions and future prospects

Turan et al. have demonstrated the usefulness of ARACNE in connecting phenotypes to a list of targeted genes. This application provided a much clearer interaction network with a much well-defined pathway. In this study, the genes or miRNAs of interest were left to be identified by the network analysis. This complicates the connectivity to a much greater extent thus leading to a series of filters being applied in the process.

Much work is still needed for the improved performance of the workflow. For example, the network analysis stage should be adapted to use other experimental data (such as motility or protein phosphorylation assays) as input. Moreover, on the technical term, it is more practical and less error-prone if an S3 or S4 class object can be designed to hold the network related data and functions. The visualisation of the network requires more testings in customisation for getting the most suitable plotting method. Gene ontology analysis should incorporate improved automatisation in data reading and enrichment testing. Furthermore, issues have remained within the initial raw count data, especially for small RNA-seq experiments. No significantly differentially expressed miRNA were detected in any of the contrast. Some cell lines were determined responsive to one drug, however, were not responsive on the gene level. All these imply that the data might require better quality controls or more experimental repeats should be performed to confirm.

Despite all the required improvements and issue related to the raw data, promising outcomes were still observed by applying the workflow. The interaction network analysis using *parmigene* package[14] was capable of identify multiple cancer/tumor-involved genes as important nodes. Moreover, though some low-quality DME data were included, several miRNAs were identified as well connected nodes. For example, hsa-mir-125b-1 miRNA, identified in Cetuximab treatments comparing MKN1 (responsive) and Hs746T (non-responsive) cell lines, (Figure 13.3A) is one of the important miRNAs related to cancer (entries for "MicroRNAs in cancer" on PathCards[25]). It was ranked the 38th most connected nodes in overall 579 selected genes/miRNAs. Though more detailed analyses should be performed in the future after improving the performance of the workflow, it has demonstrated good indications in constructing accurate interaction networks.

## Chapter references


[1] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10:57, 2009.

[2] Kimberly R Kukurba and Stephen B Montgomery. RNA Sequencing and Analysis. *Cold Spring Harbor protocols*, 2015(11):951–969, 2015.

[3] Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michal Wojciech Szczesniak, Daniel J Gaffney, Laura L Elo, Xuegong Zhang, and Ali Mortazavi. A survey of best practices for RNA-seq data analysis. *Genome biology*, 17: 13, 2016.

[4] Andrews S. Fastqc. a quality control tool for high throughput sequence data. `http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`, 2018. Accessed: Dec 2018.

[5] M. Dai, R.C. Thompson, C. Maher, R. Contreras-Galindo, M.H. Kaplan, D.M. Markovitz, G. Omenn, and F. Meng. NGSQC: cross-platform quality analysis pipeline for deep sequencing data. *BMC genomics*, 11 Suppl 4:S7, 2010.

[6] Fastx-toolkit. `http://hannonlab.cshl.edu/fastx_toolkit/.`, 2014. Accessed: Dec 2018.

[7] M.D. Robinson and G.K. Smyth. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics (Oxford, England)*, 23(21):2881–2887, 2007.

[8] S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome biology*, 11(10):R106, 2010.

[9] M.D. Robinson, D.J. McCarthy, and G.K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010. ISSN 1367-4803.

[10] A.J. Butte and I.S. Kohane. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac. Symp. Biocomput.*, pages 418–429, 2000.

[11] Nir Friedman. Inferring cellular networks using probabilistic graphical models. *Science (New York, N.Y.)*, 303(5659):799–805, 2004.

[12] J. Schafer and K. Strimmer. An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics (Oxford, England)*, 21(6):754–764, 2005.

[13] A.A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R.D. Favera, and A. Califano. ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinformatics*, 7(1):S7, 2006.

[14] G. Sales and C. Romualdi. parmigene–a parallel R package for mutual information estimation and gene network reconstruction. *Bioinformatics (Oxford, England)*, 27(13):1876–1877, 2011.

[15] L. Hua, L. Li, and P. Zhou. Identifying breast cancer subtype related miRNAs from two constructed miRNAs interaction networks in silico method. *BioMed research international*, 2013:798912, 2013.

[16] A. Colaprico, C. Cava, G. Bertoli, G. Bontempi, and I. Castiglioni. Integrative analysis with Monte Carlo cross-validation reveals miRNAs regulating pathways cross-talk in aggressive breast cancer. *BioMed research international*, 2015:17, 2015. ISSN 2314-6133.

[17] A. Remo, I. Simeone, M. Pancione, P. Parcesepe, P. Finetti, L. Cerulo, H. Bensmail, D. Birnbaum, S.J. Van Laere, V. Colantuoni, F. Bonetti, F. Bertucci, E. Manfrin, and M. Ceccarelli. Systems biology analysis reveals NFAT5 as a novel biomarker and master regulator of inflammatory breast cancer. *Journal of Translational Medicine*, 13(1):138, 2015.

[18] Y. Şenbabaoğlu, S.O. Sümer, F. Sánchez-Vega, D. Bemis, G. Ciriello, N. Schultz, and C. Sander. A Multi-Method Approach for Proteomic Network Inference in 11 Human Cancers. *PLOS Computational Biology*, 12(2):e1004765, 2016.

[19] P. Martini, L. Paracchini, G. Caratti, M. Mello-Grand, R. Fruscio, L. Beltrame, E. Calura, G. Sales, A. Ravaggi, E. Bignotti, F.E. Odicino, E. Sartori, P. Perego, D. Katsaros, I. Craparotta, G. Chiorino, S. Cagnin, L. Mannarino, L. Ceppi, C. Mangioni, C. Ghimenti, M. D'Incalci, S. Marchini, and C. Romualdi. lncRNAs as Novel Indicators of Patients' Prognosis in Stage I Epithelial Ovarian Cancer: A Retrospective and Multicentric Study. *Clinical Cancer Research*, 23(9):2356 – 2366, 2017.

[20] N. Turan, S. Kalko, A. Stincone, K. Clarke, A. Sabah, K. Howlett, S.J. Curnow, D.A. Rodriguez, M. Cascante, L. O'Neill, S. Egginton, J. Roca, and F. Falciani. A Systems Biology Approach Identifies Molecular Networks Defining Skeletal Muscle Abnormalities in Chronic Obstructive Pulmonary Disease. *PLOS Computational Biology*, 7(9):e1002129, 2011.

[21] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. URL `https://www.R-project.org/`.

[22] M.D. Robinson and A. Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3):R25, 2010.

[23] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL `http://ggplot2.org`.

[24] M. Carlson. *GO.db: A set of annotation maps describing the entire Gene Ontology*, 2018. R package version 3.7.0.

[25] F. Belinky, N. Nativ, G. Stelzer, S. Zimmerman, T. Iny Stein, M. Safran, and D. Lancet. PathCards: multi-source consolidation of human biological pathways. *Database : the journal of biological databases and curation*, 2015:bav006, 2015.

# CHAPTER 14

# Final remarks

The second part of the work involved multiple designs and developments that can be implemented within a CADD project. The MM-PBSA rescoring workflow provides a complete package for carrying out virtual screening using a better-parameterised scoring method. The MM-PBSA rescoring assessment covered the common validation steps of developing a VS method from generating the test sets, procedure optimisation, and final data analysis and interpretation. The design and development of the WSELECT program suggested that even a small additional step within the VS process requires detailed validations and performance comparisons. Lastly, the network analysis workflow, though yet to be further improved, indicated bioinformatics as a powerful tool in making connections between experimental data.

The future prospects and improvements are better summarised by the end of each part of the works. Though in general much further investigations and adjustments are still required, each study has achieved some initial success or in-depth knowledge. This thesis still left an enormous CADD space unexplored. However, it took peeks at

the space from multiple perspectives and is expected to provide insights into future expeditions into the field.

# Part IV:

# Appendix

# Abbreviations

3PO, 3-(3-pyridinyl)-1-(4-pyridinyl)-2-propen-1-one

AMPK, AMP-activated protein kinase

apoB, apolipoprotein B

CADD, computer-aided drug design

CVDs, cardiovascular diseases

DGE, differential gene expression

DME, differential microRNA expression

DPI, data processing inequality

ECs, endothelial cells

F1,6BP, fructose-1,6-bisphosphate

F1,6BPase, fructose-1,6-bisphosphatase

F2,6BP, fructose-2,6-bisphosphate

F6P, fructose-6-phosphate

FDR, false discovery rate

FKM, fragments per kilobase of exon model per million mapped reads

GRNs, gene regulatory networks

HFD, high-fat diet

HIF, hypoxia inducible factor

HREs, hypoxia-response elements

HTS, high-throughput screening

IGF, insulin/insulin-like growth factor

KDE,kernel density estimation

KLF2, Krüppel-like factor 2

LBDD, ligand-based drug design

LDL, low density lipoprotein

MD, molecular dynamic

MI, mutual information

MM, molecular mechanics

MM-GBSA, molecular mechanics generalized Born surface area

MM-PBSA , molecular mechanics Poisson-Boltzmann surface area

mRNA, messenger RNA

MW, molecular weight

NB, negative binomial

NMR, nuclear magnetic resonance

PFK1, 6-phosphofructo-1-kinase

PFKFB, 6-phosphofructo-2-kinase/fructose-2,6-biphosphatase

PKC, protein kinase C

QSAR, quantitative structure–activity relationship

RMSD, root-mean-squared displacement

RNA-seq, RNA-sequencing

ROC AUC, receiver operating characteristic area under curve

RPKM, reads per kilobase of exon model per million reads

rRNA, ribosomal RNA

SBDD, structure-based drug design

TECs, tumour endothelial cells

VEGF, vascular endothelial growth factor

vina, Autodock Vina

VS, virtual screening

# Bibliography

## Bibliography (alphabetic order)

(2014). Fastx-toolkit. `http://hannonlab.cshl.edu/fastx_toolkit/`. Accessed: Dec 2018.

Abel, R., Young, T., Farid, R., Berne, B., and Friesner, R. (2008). Role of the active-site solvent in the thermodynamics of factor Xa ligand binding. *Journal of the American Chemical Society*, 130(9):2817–2831.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome biology*, 11(10):R106.

Assad, T. and Hemnes, A. (2015). Metabolic Dysfunction in Pulmonary Arterial Hypertension. *Curr. Hypertens. Rep.*, 17(3):20.

Atsumi, T., Chesney, J., Metz, C., Leng, L., Donnelly, S., Makita, Z., Mitchell, R., and Bucala, R. (2002). High Expression of Inducible 6-Phosphofructo-2-Kinase/Fructose-2,6-Bisphosphatase (iPFK-2; PFKFB3) in Human Cancers. *Cancer Res.*, 62(20):5881–5887.

Atsumi, T., Nishio, T., Niwa, H., Takeuchi, J., Bando, H., Shimizu, C., Yoshioka, N., Bucala, R., and Koike, T. (2005). Expression of Inducible 6-Phosphofructo-2-Kinase/Fructose-2,6-Bisphosphatase/PFKFB3 Isoforms in Adipocytes and Their Potential Role in Glycolytic Regulation. *Diabetes*, 54(12):3349–3357.

Austin, E. and Loyd, J. (2014). The genetics of pulmonary arterial hypertension. *Circ. Res.*, 115(1):189–202.

Baek, K., Li, R., Jen, N., Choi, H., Kaboodrangi, A., Ping, P., Liem, D., Beebe, T., and Hsiai, T. (2017). Flow-Responsive Vascular Endothelial Growth Factor Receptor-Protein Kinase C Isoform Epsilon Signaling Mediates Glycolytic Metabolites for Vascular Repair. *Antioxid. Redox Signal.*, 28(1):31–43.

Bando, H., Atsumi, T., Nishio, T., Niwa, H., Mishima, S., Shimizu, C., Yoshioka, N., Bucala, R., and Koike, T. (2005). Phosphorylation of the 6-Phosphofructo-2-Kinase/Fructose 2,6-

Bisphosphatase/PFKFB3 Family of Glycolytic Regulators in Human Cancer. *Clin. Cancer Res.*, 11(16):5784–5792.

Barabasi, A.-L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nature Rev. Genet.*, 12(1):56–68.

Bartrons, R. and Caro, J. (2007). Hypoxia, glucose metabolism and the Warburg's effect. *J. Bioenerg. Biomembr.*, 39(3):223–229.

Batsanov, S. (2001). Van der Waals radii of elements. *Inorg. Mater.*, 37(9):871–885.

Bekkering, S., van den Munckhof, I., Nielen, T., Lamfers, E., Dinarello, C., Rutten, J., de Graaf, J., Joosten, L., Netea, M., Gomes, M., and Riksen, N. (2016). Innate immune cell activation and epigenetic remodeling in symptomatic and asymptomatic atherosclerosis in humans in vivo. *Atherosclerosis*, 254:228–236.

Belinky, F., Nativ, N., Stelzer, G., Zimmerman, S., Iny Stein, T., Safran, M., and Lancet, D. (2015). PathCards: multi-source consolidation of human biological pathways. *Database : the journal of biological databases and curation*, 2015:bav006.

Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I., and Bourne, P. (2000). The Protein Data Bank. *Nucleic Acids Res.*, 28(1):235–242.

Bhattacharya, R., SenBanerjee, S., Lin, Z., Mir, S., Hamik, A., Wang, P., Mukherjee, P., Mukhopadhyay, D., and Jain, M. (2005). Inhibition of Vascular Permeability Factor/Vascular Endothelial Growth Factor-mediated Angiogenesis by the Kruppel-like Factor KLF2. *Journal of Biological Chemistry*, 280(32):28848–28851.

Bissantz, C., Bernard, P., Hibert, M., and Rognan, D. (2003). Protein-based virtual screening of chemical databases. II. Are homology models of G-Protein Coupled Receptors suitable targets? *Proteins*, 50(1):5–25.

Bobarykina, A., Minchenko, D., Opentanova, I., Moenner, M., Caro, J., Esumi, H., and Minchenko, O. (2006). Hypoxic regulation of PFKFB-3 and PFKFB-4 gene expression in gastric and pancreatic cancer cell lines and expression of PFKFB genes in gastric cancers. *Acta Biochim. Pol.*, 53(4):789–799.

Bondi, A. (1964). Van der Waals volumes and radii. *J. Phys. Chem.*, 68(3):441–451.

Boutard, N., Białas, A., Sabiniarz, A., Guzik, P., Banaszak, K., Biela, A., Bień, M., Buda, A., Bugaj, B., Cieluch, E., Cierpich, A., Dudek, Ł., Eggenweiler, H.-M., Fogt, J., Gaik, M., Gondela, A., Jakubiec, K., Jurzak, M., Kitlińska, A., Kowalczyk, P., Kujawa, M., Kwiecińska, K., Leś, M., Lindemann, R., Maciuszek, M., Mikulski, M., Niedziejko, P., Obara, A., Pawlik, H., Rzymski, T., Sieprawska-Lupa, M., Sowińska, M., Szeremeta-Spisak, J., Stachowicz, A., Tomczyk, M., Wiklik, K., Włoszczak, Ł., Ziemiańska, S., Zarebski, A., Brzózka, K., Nowak, M., and Fabritius, C.-H. (2018). Discovery and Structure–Activity Relationships of N-Aryl 6-Aminoquinoxalines as Potent PFKFB3 Kinase Inhibitors. *ChemMedChem*, (Early View).

Boyd, S., Brookfield, J., Critchlow, S., Cumming, I., Curtis, N., Debreczeni, J., Degorce, S., Donald, C., Evans, N., Groombridge, S., Hopcroft, P., Jones, N., Kettle, J., Lamont, S., Lewis, H., MacFaull, P., McLoughlin, S., Rigoreau, L., Smith, J., St-Gallay, S., Stock, J., Turnbull, A., Wheatley, E., Winter, J., and Wingfield, J. (2015). Structure-Based Design of Potent and Selective Inhibitors of the Metabolic Kinase PFKFB3. *J. Med. Chem.*, 58(8):3611–3625.

Boyle, J., Wilson, B., Bicknell, R., Harrower, S., Weissberg, P., and Fan, T. (2000). Expression of angiogenic factor thymidine phosphorylase and angiogenesis in human atherosclerosis. *J. Pathol.*, 192(2):234–242.

Brahimi-Horn, M., Chiche, J., and Pouysségur, J. (2007). Hypoxia and cancer. *J. Mol. Med.*, 85(12):1301–1307.

Brooke, D., van Dam, E., Watts, C., Khoury, A., Dziadek, M., Brooks, H., Graham, L.-J., Flanagan, J., and Denny, W. (2014). Targeting the Warburg Effect in cancer; relationships for 2-arylpyridazinones as inhibitors of the key glycolytic enzyme 6-phosphofructo-2-kinase/2,6-bisphosphatase 3 (PFKFB3). *Bioorganic Med. Chem*, 22(3):1029–1039.

Bui, T. and Thompson, C. (2006). Cancer's sweet tooth. *Cancer Cell*, 9(6):419–420.

Butte, A. and Kohane, I. (2000). Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac. Symp. Biocomput.*, pages 418–429.

Calvo, M., Bartrons, R., Castaño, E., Perales, J., Navarro-Sabaté, A., and Manzano, A. (2006). Pfkfb3 gene silencing decreases glycolysis, induces cell-cycle delay and inhibits anchorage-independent growth in hela cells. *FEBS Lett.*, 580(13):3308–3314.

Camejo, G. (1982). The Interaction of Lipids and Lipoproteins with the Intercellular Matrix of Arterial Tissue: Its Possible Role in Atherogenesis. *Adv. Lipid Res.*, 19:1–53.

Camejo, G., Fager, G., Rosengren, B., Hurt-Camejo, E., and Bondjers, G. (1993). Binding of low density lipoproteins by proteoglycans synthesized by proliferating and quiescent human arterial smooth muscle cells. *J. Biol. Chem.*, 268(19):14131–14137.

Camejo, G., Hurt, E., Wiklund, O., Rosengren, B., López, F., and Bondjers, G. (1991). Modifications of low-density lipoprotein induced by arterial proteoglycans and chondroitin-6-sulfate. *Biochim. Biophys. Acta*, 1096(3):253–261.

Cantelmo, A., Conradi, L.-C., Brajic, A., Goveia, J., Kalucka, J., Pircher, A., Chaturvedi, P., Hol, J., Thienpont, B., Teuwen, L.-A., Schoors, S., Boeckx, B., Vriens, J., Kuchnio, A., Veys, K., Cruys, B., Finotto, L., Treps, L., Stav-Noraas, T., Bifari, F., Stapor, P., Decimo, I., Kampen, K., De Bock, K., Haraldsen, G., Schoonjans, L., Rabelink, T., Eelen, G., Ghesquière, B., Rehman, J., Lambrechts, D., Malik, A., Dewerchin, M., and Carmeliet, P. (2016). Inhibition of the Glycolytic Activator PFKFB3 in Endothelium Induces Tumor Vessel Normalization, Impairs Metastasis, and Improves Chemotherapy. *Cancer Cell*, 30(6):968–985.

Cao, R. and Wang, Y. (2016). Predicting Molecular Targets for Small-Molecule Drugs with a Ligand-Based Interaction Fingerprint Approach. *ChemMedChem*, 11(12):1352–1361.

Carlson, M. (2018). *GO.db: A set of annotation maps describing the entire Gene Ontology.* R package version 3.7.0.

Caro, J. (2001). Hypoxia Regulation of Gene Transcription. *High Altitude Med. Biol.*, 2(2):145–154.

Case, D. A., Ben-Shalom, I., Brozell, S., Cerutti, D., Cheatham, T., Cruzeiro, V., Darden, T., Duke, R., Ghoreishi, D., Gilson, M., Gohlke, H., Goetz, A., Greene, D., Harris, R., Homeyer, N., Izadi, S., Kovalenko, A., Kurtzman, T., Lee, T., LeGrand, S., Li, P., Lin, C., Liu, J., Luchko, T., Luo, R., Mermelstein, D., Merz, K., Miao, Y., Monard, G., Nguyen, C., Nguyen, H., Omelyan, I., Onufriev, A., Pan, F., Qi, R., Roe, D., Roitberg, A., Sagui, C., Schott-Verdugo, S., Shen, J., Simmerling, C., Smith, J., Salomon-Ferrer, R., Swails, J., Walker, R., Wang, J., Wei, H., Wolf, R., Wu, X., Xiao, L., York, D., and Kollman, P. (2018). Amber 2018.

Cavalier, M., Kim, S.-G., Neau, D., and Lee, Y.-H. (2011). Molecular basis of the fructose-2,6-bisphosphatase reaction of PFKFB3: Transition state and the C-terminal function. *Proteins*, 80(4):1143–1153.

Celletti, F., Waugh, J., Amabile, P., Brendolan, A., Hilfiker, P., and Dake, M. (2001). Vascular endothelial growth factor enhances atherosclerotic plaque progression. *Nat. Med.*, 7:425.

Chang, M., Potter-Perigo, S., Tsoi, C., Chait, A., and Wight, T. (2000). Oxidized Low Density Lipoproteins Regulate Synthesis of Monkey Aortic Smooth Muscle Cell Proteoglycans That Have Enhanced Native Low Density Lipoprotein Binding Properties. *J. Biol. Chem.*, 275(7):4766–4773.

Charifson, P., Corkery, J., Murcko, M., and Walters, W. (1999). Consensus Scoring: A Method for Obtaining Improved Hit Rates from Docking Databases of Three-Dimensional Structures into Proteins. *J. Med. Chem.*, 42(25):5100–5109.

Cheng, T., Li, Q., Wang, Y., and Bryant, S. (2011). Identifying compound-target associations by combining bioactivity profile similarity search and public databases mining. *J. Chem. Inf. Model.*, 51(9):2440–2448.

Cheng, T., Li, X., Li, Y., Liu, Z., and Wang, R. (2009). Comparative Assessment of Scoring Functions on a Diverse Test Set. *J. Chem. Inf. Model.*, 49(4):1079–1093.

Chesney, J., Clark, J., Klarer, A., Imbert-Fernandez, Y., Lane, A., and Telang, S. (2014). Fructose-2,6-bisphosphate synthesis by 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase 4 (PFKFB4) is required for the glycolytic response to hypoxia and tumor growth. *Oncotarget*, 5(16):6670–6686.

Chesney, J., Mitchell, R., Benigni, F., Bacher, M., Spiegel, L., Al-Abed, Y., Han, J., Metz, C., and Bucala, R. (1999). An inducible gene product for 6-phosphofructo-2-kinase with an au-rich instability element: Role in tumor cell glycolysis and the warburg effect. *Proc. Natl. Acad. Sci. U.S.A.*, 96(6):3047–3052.

Chu, Y. and Corey, D. R. (2012). RNA sequencing: platform selection, experimental design, and data interpretation. *Nucleic acid therapeutics*, 22(4):271–274.

Chung, B., Tallis, G., Yalamoori, V., Anantharamaiah, G., and Segrest, J. (1994). Liposome-like particles isolated from human atherosclerotic plaques are structurally and compositionally similar to surface remnants of triglyceride-rich lipoproteins. *Arterioscler. Thromb. Vasc. Biol.*, 14(4):622–635.

Clem, B., Telang, S., Clem, A., Yalcin, A., Meier, J., Simmons, A., Rasku, M., Arumugam, S., Dean, W., Eaton, J., Lane, A., Trent, J., and Chesney, J. (2008). Small-molecule inhibition of 6-phosphofructo-2-kinase activity suppresses glycolytic flux and tumor growth. *Mol. Canc. Therapeut.*, 7(1):110–120.

Clem, B. F., O'Neal, J., Tapolsky, G., Clem, A. L., Imbert-Fernandez, Y., Kerr, D. A., Klarer, A. C., Redman, R., Miller, D. M., Trent, J. O., Telang, S., and Chesney, J. (2013). Targeting 6-Phosphofructo-2-Kinase (PFKFB3) as a Therapeutic Strategy against Cancer. *Mol. Cancer Ther.*, 12(8):1461–1470.

Clinic, M. (2018). Arteriosclerosis / atherosclerosis - diagnosis & treatment. `https://www.mayoclinic.org/diseases-conditions/arteriosclerosis-atherosclerosis/diagnosis-treatment/drc-20350575`. Accessed: Dec 2018.

Colaprico, A., Cava, C., Bertoli, G., Bontempi, G., and Castiglioni, I. (2015). Integrative analysis with Monte Carlo cross-validation reveals miRNAs regulating pathways cross-talk in aggressive breast cancer. *BioMed research international*, 2015:17.

Colomer, D., Vives-Corrons, J., Pujades, A., and Bartrons, R. (1987). Control of phosphofructokinase by fructose 2,6-bisphosphate in b-lymphocytes and b-chronic lymphocytic leukemia cells. *Cancer Res.*, 47(7):1859–1862.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szczesniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., and Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome biology*, 17:13.

Corporation, A. (2018). Asinex - screening libraries. https://http://www.asinex.com/. Accessed: Apr 2018.

D., S. (2012). Dan sindhikara's tutorials - using 3d-rism and placevent. http://dansindhikara.com/Tutorials/Entries/2012/1/1_Using_3D-RISM_and_PLACEVENT.html. Accessed: Oct 2016.

Dai, M., Thompson, R., Maher, C., Contreras-Galindo, R., Kaplan, M., Markovitz, D., Omenn, G., and Meng, F. (2010). NGSQC: cross-platform quality analysis pipeline for deep sequencing data. *BMC genomics*, 11 Suppl 4:S7.

Daiber, A., Steven, S., Weber, A., Shuvaev, V. V., Muzykantov, V. R., Laher, I., Li, H., Lamas, S., and Munzel, T. (2017). Targeting vascular (endothelial) dysfunction. *Br. J. Pharmacol.*, 174(12):1591–1619.

Dang, C. and Semenza, G. (1999). Oncogenic alterations of metabolism. *Trends Biochem. Sci.*, 24(2):68–72.

Davidson, S. and Duchen, M. (2007). Endothelial Mitochondria. *Circ. Res.*, 100(8):1128–1141.

de Beer, S., Vermeulen, N., and Oostenbrink, C. (2010). The Role of Water Molecules in Computational Drug Design. *Curr. Top. Med. Chem.*, 10(1):55–66.

Deng, Z., Morse, J., Slager, S., Cuervo, N., Moore, K., Venetos, G., Kalachikov, S., Cayanis, E., Fischer, S., Barst, R., Hodge, S., and Knowles, J. (2000). Familial Primary Pulmonary Hypertension (Gene PPH1) Is Caused by Mutations in the Bone Morphogenetic Protein Receptor–II Gene. *Am. J. Hum. Genet.*, 67(3):737–744.

De Bock, K., Georgiadou, M., and Carmeliet, P. (2013a). Role of Endothelial Cell Metabolism in Vessel Sprouting. *Cell Metab.*, 18(5):634–647.

De Bock, K., Georgiadou, M., Schoors, S., Kuchnio, A., Wong, B., Cantelmo, A., Quaegebeur, A., Ghesquière, B., Cauwenberghs, S., Eelen, G., Phng, L.-K., Betz, I., Tembuyser, B., Brepoels, K., Welti, J., Geudens, I., Segura, I., Cruys, B., Bifari, F., Decimo, I., Blanco, R., Wyns, S., Vangindertael, J., Rocha, S., Collins, R., Munck, S., Daelemans, D., Imamura, H., Devlieger, R., Rider, M., Van Veldhoven, P., Schuit, F., Bartrons, R., Hofkens, J., Fraisl, P., Telang, S., DeBerardinis, R., Schoonjans, L., Vinckier, S., Chesney, J., Gerhardt, H., Dewerchin, M., and Carmeliet, P. (2013b). Role of PFKFB3-Driven Glycolysis in Vessel Sprouting. *Cell*, 154(3):651–663.

DiMasi, J., Hansen, R., and Grabowski, H. (2003). The price of innovation: new estimates of drug development costs. *J. Health. Econ.*, 22(2):151–185.

Doddaballapur, A., Michalik, K., Manavski, Y., Lucas, T., Houtkooper, R., You, X., Chen, W., Zeiher, A., Potente, M., Dimmeler, S., and Boon, R. (2015). Laminar shear stress inhibits endothelial cell metabolism via klf2-mediated repression of pfkfb3. *Arter. Thromb. Vasc. Biol.*, 35(1):137–145.

Draoui, N., de Zeeuw, P., and Carmeliet, P. (2017). Angiogenesis revisited from a metabolic perspective: role and therapeutic implications of endothelial cell metabolism. *Open Biol.*, 7(12).

Duan, Y., Wu, C., Chowdhury, S., Lee, M., Xiong, G., Zhang, W., Yang, R., Cieplak, P., Luo, R., Lee, T., Caldwell, J., Wang, J., and Kollman, P. (2003). A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *Journal of computational chemistry*, 24(16):1999–2012.

Dupradeau, F.-Y., Cezard, C., Lelong, R., Stanislawiak, E., Pecher, J., Delepine, J. C., and Cieplak, P. (2008). R.E.DD.B.: a database for RESP and ESP atomic charges, and force field libraries. *Nucleic acids research*, 36(Database issue):D360–7.

Eberini, I., Daniele, S., Parravicini, C., Sensi, C., Trincavelli, M., Martini, C., and Abbracchio, M. (2011). In silico identification of new ligands for GPR17: a promising therapeutic target for neurodegenerative diseases. *J. Comput. Aided. Mol. Des.*, 25(8):743–752.

el Maghrabi, M., Claus, T., Pilkis, J., and Pilkis, S. (1982). Regulation of 6-phosphofructo-2-kinase activity by cyclic amp-dependent phosphorylation. *Proc. Natl. Acad. Sci. U.S.A.*, 79(2):315–319.

El-Maghrabi, M., Fox, E., Pilkis, J., and Pilkis, S. (1982). Cyclic AMP-dependent phosphorylation of rat liver 6-phosphofructo 2-kinase/fructose 2,6-bisphosphatase. *Biochem. Biophys. Res. Commun.*, 106(3):794–802.

Engin, H., Gursoy, A., Nussinov, R., and Keskin, O. (2014). Network-based strategies can help mono- and poly-pharmacology drug discovery: a systems biology view. *Curr. Pharm. Des.*, 20(8):1201–1207.

Evers, A. and Klabunde, T. (2005). Structure-based drug discovery using GPCR homology modeling: successful virtual screening for antagonists of the alpha1A adrenergic receptor. *J. Med. Chem.*, 48(4):1088–1097.

Evers, A. and Klebe, G. (2004). Successful virtual screening for a submicromolar antagonist of the neurokinin-1 receptor based on a ligand-supported homology model. *J. Med. Chem.*, 47(22):5381–5392.

Fogel, D. and B., G. (2009). Computational Intelligence Methods for Docking Scores. *Curr. Comput. Aided Drug Des.*, 5(1):56–68.

Fogelstrand, P. and Borén, J. (2012). Retention of atherogenic lipoproteins in the artery wall and its role in atherogenesis. *Nutr. Metab. Cardiovasc. Dis.*, 22(1):1–7.

Friedman, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science (New York, N.Y.)*, 303(5659):799–805.

Fryer, J., Myers, P., and Appleberg, M. (1987). Carotid intraplaque hemorrhage: the significance of neovascularity. *J. Vasc. Surg.*, 6(4):341–349.

Fukasawa, M., Tsuchiya, T., Takayama, E., Shinomiya, N., Uyeda, K., Sakakibara, R., and Seki, S. (2004). Identification and Characterization of the Hypoxia-Responsive Element of the Human Placental 6-Phosphofructo-2-Kinase/Fructose-2,6-Bisphosphatase Gene. *J. Biochem.*, 136(3):273–277.

G., S. and R., U. (2015). The mm/pbsa and mm/gbsa methods to estimate ligand-binding affinities. *Expert Opin. Drug Discovery*, 10(5):449–461.

Galie, N., Humbert, M., Vachiery, J.-L., Gibbs, S., Lang, I., Torbicki, A., Simonneau, G., Peacock, A., Vonk Noordegraaf, A., Beghetti, M., Ghofrani, A., Gomez Sanchez, M., Hansmann, G., Klepetko, W., Lancellotti, P., Matucci, M., McDonagh, T., Pierard, L., Trindade, P., Zompatori, M., and Hoeper, M. (2016). 2015 ESC/ERS Guidelines for the diagnosis and treatment of pulmonary hypertension: The Joint Task Force for the Diagnosis and Treatment of Pulmonary Hypertension of the European Society of Cardiology (ESC) and the European Respiratory Society (ERS): Endorsed by: Association for European Paediatric and Congenital Cardiology (AEPC), International Society for Heart and Lung Transplantation (ISHLT). *European heart journal*, 37(1):67–119.

Galili, O., Sattler, K., Herrmann, J., Woodrum, J., Olson, M., Lerman, L., and Lerman, A. (2005). Experimental hypercholesterolemia differentially affects adventitial vasa vasorum and vessel structure of the left internal thoracic and coronary arteries. *J. Thorac. Cardiovasc. Surg.*, 129(4):767–772.

Gaulton, A., Hersey, A., Nowotka, M., Bento, A., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L., Cibrián-Uhalte, E., Davies, M., Dedman, N., Karlsson, A., Magariños, M., Overington, J., Papadatos, G., Smit, I., and Leach, A. (2017). The chembl database in 2017. *Nucleic Acids Res.*, 45(D1):D945–D954.

Genheden, S. (2011). Mm/gbsa and lie estimates of host-guest affinities: dependence on charges and solvation model. *J. Comput.-Aided Mol. Des.*, 25(11):1085–1093.

Gijsen, F., Wentzel, J., Thury, A., Mastik, F., Schaar, J., Schuurbiers, J., Slager, C., van der Giessen, W., de Feyter, P., van der Steen, A., and Serruys, P. (2008). Strain distribution over plaques in human coronary arteries relates to shear stress. *Am. J. Physiol. Heart Circ. Physiol.*, 295(4):H1608–14.

Glass, C. and Witztum, J. (2001). Atherosclerosis: The Road Ahead. *Cell*, 104(4):503–516.

Gopalan, D., Delcroix, M., and Held, M. (2017). Diagnosis of chronic thromboembolic pulmonary hypertension. *Eur. Respir. Rev.*, 26(143):160108.

Gu, M., Li, L., Zhang, Z., Chen, J., Zhang, W., Zhang, J., Han, L., Tang, M., You, B., Zhang, Q., and You, Y. (2017). PFKFB3 promotes proliferation, migration and angiogenesis in nasopharyngeal carcinoma. *J. Cancer*, 8(18):3887–3896.

Gunther, J., Bergner, A., Hendlich, M., and Klebe, G. (2003). Utilising structural knowledge in drug design strategies: applications using Relibase. *J. Mol. Biol.*, 326(2):621–636.

Gustafsson, M., Levin, M., Skalen, K., Perman, J., Friden, V., Jirholt, P., Olofsson, S.-O., Fazio, S., Linton, M., Semenkovich, C., Olivecrona, G., and Boren, J. (2007). Retention of low-density lipoprotein in atherosclerotic lesions of the mouse: evidence for a role of lipoprotein lipase. *Circ. Res.*, 101(8):777–783.

Gustafsson, N., Färnegårdh, K., Bonagas, N., Ninou, A., Groth, P., Wiita, E., Jönsson, M., Hallberg, K., Lehto, J., Pennisi, R., Martinsson, J., Norström, C., Hollers, J., Schultz, J., Andersson, M., Markova, N., Marttila, P., Kim, B., Norin, M., Olin, T., and Helleday, T. (2018). Targeting PFKFB3 radiosensitizes cancer cells and suppresses homologous recombination. *Nat. Commun.*, 9(1):3872.

Halgren, T., Murphy, R., Friesner, R., Beard, H., Frye, L., Pollard, W., and Banks, J. (2004). Glide: a new approach for rapid, accurate docking and scoring. 2. enrichment factors in database screening. *J. Med. Chem*, 47(7):1750–1759.

Hamilton, J., Callaghan, M., Sutherland, R., and Watts, C. (1997). Identification of prg1, a novel progestin-responsive gene with sequence homology to 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase. *Mol. Endocrinol.*, 11(4):490–502.

Hasemann, C., Istvan, E., Uyeda, K., and Deisenhofer, J. (1996). The crystal structure of the bifunctional enzyme 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase reveals distinct domain homologies. *Structure*, 4(9):1017–1029.

Heinke, R., Spannhoff, A., Meier, R., Trojer, P., Bauer, I., Jung, M., and Sippl, W. (2009). Virtual screening and biological characterization of novel histone arginine methyltransferase PRMT1 inhibitors. *ChemMedChem*, 4(1):69–77.

Hendlich, M., Bergner, A., Günther, J., and Klebe, G. (2003). Relibase: Design and development of a database for comprehensive analysis of protein-ligand interactions. *J. Mol. Biol.*, 326(2):607–620.

Herrmann, J., Lerman, L., Mukhopadhyay, D., Napoli, C., and Lerman, A. (2006). Angiogenesis in atherogenesis. *Arter. Thromb. Vasc. Biol.*, 26(9):1948–1957.

Hers, H. (1983). Fructose 2,6-bisphosphate. *Biochem Soc Trans.*, 11(3):250–251.

Hers, H. (1984). The discovery and the biological role of fructose 2,6-bisphosphate. *Biochem Soc Trans.*, 12(5):729–735.

Hoeks, A., Reesink, K., Hermeling, E., and Reneman, R. (2008). Local Blood Pressure Rather Than Shear Stress Should Be Blamed for Plaque Rupture. *J. Am. Coll. Cardiol.*, 52(13):1107–1108.

Hong, S.-P., Leiper, F., Woods, A., Carling, D., and Carlson, M. (2003). Activation of yeast Snf1 and mammalian AMP-activated protein kinase by upstream kinases. *Proc. Natl. Acad. Sci. U.S.A.*, 100(15):8839–8843.

Hoshiga, M., Alpers, C., Smith, L., Giachelli, C., and Schwartz, S. (1995). Alpha-v beta-3 integrin expression in normal and atherosclerotic artery. *Circ. Res.*, 77(6):1129–1135.

Hou, T., Wang, J., Li, Y., and Wang, W. (2011). Assessing the performance of the mm/pbsa and mm/gbsa methods: Ii. the accuracy of ranking poses generated from docking. *J. Comput. Chem.*, 32(5):866–877.

Hu, X., Maffucci, I., and Contini, A. (2018). Advances in the Treatment of Explicit Water Molecules in Docking and Binding Free Energy Calculations. *Curr. Med. Chem.*, 25:1–23.

Hua, L., Li, L., and Zhou, P. (2013). Identifying breast cancer subtype related miRNAs from two constructed miRNAs interaction networks in silico method. *BioMed research international*, 2013:798912.

Huang, S.-Y., Grinter, S., and Zou, X. (2010). Scoring functions and their evaluation methods for protein-ligand docking: recent advances and future directions. *Phys. Chem. Chem. Phys.*, 12(40):12899–12908.

Huang, S.-Y. and Zou, X. (2010). Inclusion of Solvation and Entropy in the Knowledge-Based Scoring Function for Protein−Ligand Interactions. *J. Chem. Inf. Model.*, 50(2):262–273.

Huang, W., Blinov, N., Wishart, D., and Kovalenko, A. (2015). Role of water in ligand binding to maltose-binding protein: Insight from a new docking protocol based on the 3D-RISM-KH molecular theory of solvation. *Journal of Chemical Information and Modeling*, 55(2):317–328.

Huey, R., Morris, G., Olson, A., and Goodsell, D. (2007). A semiempirical free energy force field with charge-based desolvation. *J. Comput. Chem.*, 28(6):1145–1152.

Humbert, M. (2010). Pulmonary arterial hypertension and chronic thromboembolic pulmonary hypertension: pathophysiology. *European respiratory review : an official journal of the European Respiratory Society*, 19(115):59–63.

Huo, Y., Guo, X., Li, H., Wang, H., Zhang, W., Wang, Y., Zhou, H., Gao, Z., Telang, S., Chesney, J., Chen, Y., Ye, J., Chapkin, R., and Wu, C. (2010). Disruption of Inducible 6-Phosphofructo-2-kinase Ameliorates Diet-induced Adiposity but Exacerbates Systemic Insulin Resistance and Adipose Tissue Inflammatory Response. *J. Biol. Chem.*, 285(6):3713–3721.

Huo, Y., Guo, X., Li, H., Xu, H., Halim, V., Zhang, W., Wang, H., Fan, Y.-Y., Ong, K., Woo, S.-L., Chapkin, R., Mashek, D., Chen, Y., Dong, H., Lu, F., Wei, L., and Wu, C. (2012). Targeted Overexpression of Inducible 6-Phosphofructo-2-kinase in Adipose Tissue Increases Fat Deposition but Protects against Diet-induced Insulin Resistance and Inflammatory Responses. *J. Biol. Chem.*, 287(25):21492–21500.

Hurt, E., Bondjers, G., and Camejo, G. (1990). Interaction of LDL with human arterial proteoglycans stimulates its uptake by human monocyte-derived macrophages. *J. Lipid Res.*, 31(3):443–454.

Hurt-Camejo, E., Camejo, G., Rosengren, B., Lopez, F., Ahlstrom, C., Fager, G., and Bondjers, G. (1992). Effect of arterial proteoglycans and glycosaminoglycans on low density lipoprotein

oxidation and its uptake by human macrophages and arterial smooth muscle cells. *Arterioscler. Thromb. Vasc. Biol.*, 12(5):569–583.

Ihling, C., Szombathy, T., Bohrmann, B., Brockhaus, M., Schaefer, H., and Loeffler, B. (2001). Coexpression of endothelin-converting enzyme-1 and endothelin-1 in different stages of human atherosclerosis. *Circulation*, 104(8):864–869.

Inc., E. B. (2010). Encyclopaedia britannica - atherosclerosis. `https://www.britannica.com/science/atherosclerosis/media/40908/95216`. Accessed: January 4, 2019.

Irwin, J., Sterling, T., Mysinger, M., Bolstad, E., and Coleman, R. (2012). ZINC: A Free Tool to Discover Chemistry for Biology. *Journal of Chemical Information and Modeling*, 52(7):1757–1768.

Jain, R. (2014). Antiangiogenesis Strategies Revisited: From Starving Tumors to Alleviating Hypoxia. *Cancer Cell*, 26(5):605–622.

Jakalian, A., Jack, D., and Bayly, C. (2002). Fast, efficient generation of high-quality atomic charges. am1-bcc model: Ii. parameterization and validation. *J. Comput. Chem.*, 23(16):1623–1641.

Jhoti, H., Rees, S., and Solari, R. (2013). High-throughput screening and structure-based approaches to hit discovery: is there a clear winner?

Jiao, D., Zhang, J., Duke, R., Li, G., Schnieders, M., and Ren, P. (2009). Trypsin-ligand binding free energies from explicit and implicit solvent simulations with polarizable potential. *J. Comput. Chem.*, 30(11):1701–1711.

Johnson, J. and Newby, A. (2009). Macrophage heterogeneity in atherosclerotic plaques. *Curr. Opin. Lipidol.*, 20(5):370–378.

Jung, K., Zheng, H.-M., Jeong, Y., Choi, M.-J., Lee, H., Hong, S.-W., Lee, H.-S., Son, M., Lee, S., Hong, S., and Hong, S.-S. (2013). Suppression of tumor proliferation and angiogenesis of hepatocellular carcinoma by HS-104, a novel phosphoinositide 3-kinase inhibitor. *Cancer Lett.*, 328(1):176–187.

Kalyaanamoorthy, S. and Chen, Y.-P. (2011). Structure-based drug design to augment hit discovery. *Drug Discov. Today*, 16(17):831–839.

Kawanami, D., Mahabeleshwar, G., Lin, Z., Atkins, G., Hamik, A., Haldar, S., Maemura, K., LaManna, J., and Jain, M. (2009). Kruppel-like Factor 2 Inhibits Hypoxia-inducible Factor $1\alpha$ Expression and Function in the Endothelium. *Journal of Biological Chemistry*, 284(31):20522–20530.

Keiser, M., Setola, V., Irwin, J., Laggner, C., Abbas, A., Hufeisen, S., Jensen, N., Kuijer, M., Matos, R., Tran, T., Whaley, R., Glennon, R., Hert, J., Thomas, K., Edwards, D., Shoichet, B., and Roth, B. (2009). Predicting new molecular targets for known drugs. *Nature*, 462:175.

Kim, E., Baker, C., Dwyer, M., Murcko, M., Rao, B., Tung, R., and Navia, M. (1995). Crystal structure of HIV-1 protease in complex with VX-478, a potent and orally bioavailable inhibitor of the enzyme. *Journal of the American Chemical Society*, 117(3):1181–1182.

Kim, S.-G., Cavalier, M., El-Maghrabi, M., and Lee, Y.-H. (2007). A Direct Substrate–Substrate Interaction Found in the Kinase Domain of the Bifunctional Enzyme, 6-Phosphofructo-2-kinase/Fructose-2,6-bisphosphatase. *J. Mol. Biol.*, 370(1):14–26.

Kim, S.-G., Manes, N., El-Maghrabi, M., and Lee, Y.-H. (2006). Crystal structure of the hypoxia-inducible form of 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase (pfkfb3): A possible new target for cancer therapy. *J. Biol. Chem.*, 281(5):2939–2944.

Kinkade, R., Dasgupta, P., Carie, A., Pernazza, D., Carless, M., Pillai, S., Lawrence, N., Sebti, S., and Chellappan, S. (2008). A Small Molecule Disruptor of Rb/Raf-1 Interaction Inhibits Cell

Proliferation, Angiogenesis, and Growth of Human Tumor Xenografts in Nude Mice. *Cancer Res.*, 68(10):3810–3818.

Kireev, D., Wigle, T., Norris-Drouin, J., Herold, J., Janzen, W., and Frye, S. (2010). Identification of non-peptide malignant brain tumor (MBT) repeat antagonists by virtual screening of commercially available compounds. *J. Med. Chem.*, 53(21):7625–7631.

Kitchen, D. B., Decornez, H., Furr, J. R., and Bajorath, J. (2004). Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov.*, 3(11):935–949.

Kluge, R., Barthel, H., Pankau, H., Seese, A., Schauer, J., Wirtz, H., Seyfarth, H.-J., Steinbach, J., Sabri, O., and Winkler, J. (2005). Different mechanisms for changes in glucose uptake of the right and left ventricular myocardium in pulmonary hypertension. *J. Nucl. Med.*, 46(1):25–31.

Kollman, P., Massova, I., Reyes, C., Kuhn, B., Huo, S., Chong, L., Lee, M., Lee, T., Duan, Y., Wang, W., Donini, O., Cieplak, P., Srinivasan, J., Case, D., and Cheatham, T. (2000). Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc. Chem. Res.*, 33(12):889–897.

Korb, O., Stüttzle, T., and Exner, T. (2009). Empirical scoring functions for advanced protein-ligand docking with plants. *J. Chem. Inf. Model.*, 49(1):84–96.

Korb, O., Stützle, T., and Exner, T. (2007). An ant colony optimization approach to flexible protein–ligand docking. *Swarm Intell.*, 1(2):115–134.

Kovalenko, A. (2013). Multiscale modeling of solvation in chemical and biological nanosystems and in nanoporous materials. *Pure and Applied Chemistry*, 85(1):159–199.

Kovalenko, A. and Hirata, F. (1998). Three-dimensional density profiles of water in contact with a solute of arbitrary shape: a RISM approach. *Chemical Physics Letters*, 290(1):237–244.

Kovalenko, A. and Hirata, F. (1999). Self-consistent description of a metal–water interface by the Kohn–Sham density functional theory and the three-dimensional reference interaction site model. *The Journal of Chemical Physics*, 110(20):10095–10112.

Kovalenko, A. and Hirata, F. (2001). Self-consistent, Kohn-Sham DFT and three-dimensional RISM description of a metal-molecular liquid interface. *Journal of Molecular Liquids*, 90(1):215–224.

Kukurba, K. R. and Montgomery, S. B. (2015). RNA Sequencing and Analysis. *Cold Spring Harbor protocols*, 2015(11):951–969.

Kumar, A., Voet, A., and Zhang, K. (2012). Fragment based drug design: from experimental to computational approaches. *Curr. Med. Chem.*, 19(30):5128–5147.

Kurland, I., Chapman, B., and El-maghrabi, M. (2000). N- and C-termini modulate the effects of pH and phosphorylation on hepatic 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase. *Biochem. J.*, 347(2):459–467.

Kwon, H., Sangiorgi, G., Ritman, E., McKenna, C., Holmes, D., Schwartz, R., and Lerman, A. (1998). Enhanced coronary vasa vasorum neovascularization in experimental hypercholesterolemia. *J. Clin. Invest.*, 101(8):1551–1556.

Kyte, J. and Doolittle, R. (1982). A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology*, 157(1):105–132.

Ladbury, J. (1996). Just add water! The effect of water on the specificity of protein-ligand binding sites and its potential application to drug design. *Chem. Biol.*, 3(12):973–980.

Lane, K., Machado, R., Pauciulo, M., Thomson, J., Phillips, J., Loyd, J., Nichols, W., and Trembath, R. (2000). Heterozygous germline mutations in BMPR2, encoding a TGF-$\beta$ receptor, cause familial primary pulmonary hypertension. *Nat. Genet.*, 26:81.

Lang, I. (2015). Chronic thromboembolic pulmonary hypertension: a distinct disease entity. *European respiratory review : an official journal of the European Respiratory Society*, 24(136):246–252.

Lavecchia, A. and Di Giovanni, C. (2013). Virtual screening strategies in drug discovery: a critical review. *Curr. Med. Chem.*, 20(23):2839–2860.

Lee, K., Jeong, K.-W., Lee, Y., Song, J., Kim, M., Lee, G., and Kim, Y. (2010). Pharmacophore modeling and virtual screening studies for new VEGFR-2 kinase inhibitors. *Eur. J. Med. Chem.*, 45(11):5420–5427.

Lee, Y.-H., Li, Y., Uyeda, K., and Hasemann, C. (2003). Tissue-specific Structure/Function Differentiation of the Liver Isoform of 6-Phosphofructo-2-kinase/Fructose-2,6-bisphosphatase. *J. Biol. Chem.*, 278(1):523–530.

Lightstone, F. C., Zhang, X., and Wong, S. E. (2014). Vinalc: Parallel molecular docking program. [Computer Software] `https://dx.doi.org/10.11578/dc.20180320.8`.

Lindström, A., Edvinsson, L., Johansson, A., Andersson, C., Andersson, I., Raubacher, F., and Linusson, A. (2011). Postprocessing of docked protein-ligand complexes using implicit solvation models. *J. Chem. Inf. Model.*, 51(2):267–282.

Lionta, E., Spyrou, G., Vassilatis, D., and Cournia, Z. (2014). Structure-based virtual screening for drug discovery: principles, applications and recent advances. *Curr. Top. Med. Chem.*, 14(16):1923–1938.

Lipinski, C. (2004). Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discov. Today Technol.*, 1(4):337–341.

Lipinski, C., Lombardo, F., Dominy, B., and Feeney, P. (2001). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.*, 46(1):3–26.

Lu, Y., Wang, R., Yang, C., and Wang, S. (2007). Analysis of ligand-bound water molecules in high-resolution crystal structures of protein-ligand complexes. *J. Chem. Inf. Model.*, 47(2):668–675.

Lundgrin, E., Park, M., Sharp, J., Tang, W., Thomas, J., Asosingh, K., Comhair, S., DiFilippo, F., Neumann, D., Davis, L., Graham, B., Tuder, R., Dostanic, I., and Erzurum, S. (2013). Fasting 2-deoxy-2-[18F]fluoro-D-glucose positron emission tomography to detect metabolic changes in pulmonary arterial hypertension hearts over 1 year. *Ann. Am. Thorac. Soc.*, 10(1):1–9.

Lung, T. N. H. and (NHLBI), B. I. (2018). Atherosclerosis. `https://www.nhlbi.nih.gov/health-topics/atherosclerosis`. Accessed: Dec 2018.

Luttun, A., Tjwa, M., Moons, L., Wu, Y., Angelillo-Scherrer, A., Liao, F., Nagy, J., Hooper, A., Priller, J., De Klerck, B., Compernolle, V., Daci, E., Bohlen, P., Dewerchin, M., Herbert, J.-M., Fava, R., Matthys, P., Carmeliet, G., Collen, D., Dvorak, H., Hicklin, D., and Carmeliet, P. (2002). Revascularization of ischemic tissues by PlGF treatment, and inhibition of tumor angiogenesis, arthritis and atherosclerosis by anti-Flt1. *Nat. Med.*, 8:831.

Macarron, R., Banks, M., Bojanic, D., Burns, D., Cirovic, D., Garyantes, T., Green, D., Hertzberg, R., Janzen, W., Paslay, J., Schopfer, U., and Sittampalam, G. (2011). Impact of high-throughput screening in biomedical research. *Nat. Rev. Drug Discov.*, 10:188.

Macut, H. (2016-2018). personal communication.

Maffucci, I. and Contini, A. (2016). Improved computation of protein–protein relative binding energies with the nwat-mmgbsa method. *J. Chem. Inf. Model.*, 56(9):1692–1704.

Maffucci, I., Hu, X., Fumagalli, V., and Contini, A. (2018). An efficient implementation of the nwat-mmgbsa method to rescore docking results in medium-throughput virtual screenings. *Front. Chem.*, 6(43).

Maier, J., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K., and Simmerling, C. (2015). ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.*, 11(8):3696–3713.

Manes, N. and El-Maghrabi, M. (2005). The kinase activity of human brain 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase is regulated via inhibition by phosphoenolpyruvate. *Arch. Biochem. Biophys.*, 438(2):125–136.

Manzano, A., Rosa, J., Ventura, F., Perez, J., Nadal, M., Estivill, X., Ambrosio, S., Gil, J., and Bartrons, R. (1998). Molecular cloning, expression, and chromosomal localization of a ubiquitously expressed human 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase gene (PFKFB3). *Cytogenet. Cell Genet.*, 83(3-4):214–217.

Margolin, A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R., and Califano, A. (2006). ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinformatics*, 7(1):S7.

Marotta, L., Almendro, V., Marusyk, A., Shipitsin, M., Schemme, J., Walker, S., Bloushtain-Qimron, N., Kim, J., Choudhury, S., Maruyama, R., Wu, Z., Gönen, M., Mulvey, L., Bessarabova, M., Huh, S., Silver, S., Kim, S., Park, S., Lee, H., Anderson, K. S., Richardson, A. L., Nikolskaya, T., Nikolsky, Y., Liu, X. S., Root, D. E., H., W. C., Frank, D., and Polyak, K. (2011). The JAK2/STAT3 signaling pathway is required for growth of CD44+CD24– stem cell–like breast cancer cells in human tumors. *J. Clin. Investig*, 121(7):2723–2735.

Marsboom, G., Wietholt, C., Haney, C., Toth, P., Ryan, J., Morrow, E., Thenappan, T., Bache-Wiig, P., Piao, L., Paul, J., Chen, C.-T., and Archer, S. (2012). Lung $^{18}$F-Fluorodeoxyglucose Positron Emission Tomography for Diagnosis and Monitoring of Pulmonary Arterial Hypertension. *Am. J. Respir. Crit. Care Med.*, 185(6):670–679.

Marshall, M., Goldberg, D., Neal, F., and Millar, D. (1978). Enzymes of glucose metabolism in carcinoma of the cervix and endometrium of the human uterus. *Br. J. Cancer*, 37(6):990–1001.

Marsin, A.-S., Bouzin, C., Bertrand, L., and Hue, L. (2002). The Stimulation of Glycolysis by Hypoxia in Activated Monocytes Is Mediated by AMP-activated Protein Kinase and Inducible 6-Phosphofructo-2-kinase. *J. Biol. Chem.*, 277(34):30778–30783.

Martini, P., Paracchini, L., Caratti, G., Mello-Grand, M., Fruscio, R., Beltrame, L., Calura, E., Sales, G., Ravaggi, A., Bignotti, E., Odicino, F., Sartori, E., Perego, P., Katsaros, D., Craparotta, I., Chiorino, G., Cagnin, S., Mannarino, L., Ceppi, L., Mangioni, C., Ghimenti, C., D'Incalci, M., Marchini, S., and Romualdi, C. (2017). lncRNAs as Novel Indicators of Patients' Prognosis in Stage I Epithelial Ovarian Cancer: A Retrospective and Multicentric Study. *Clinical Cancer Research*, 23(9):2356 – 2366.

Massova, I. and Kollman, P. (2000). Combined molecular mechanical and continuum solvent approach (mm-pbsa/gbsa) to predict ligand binding. *Perspect. Drug Discovery Des.*, 18(1):113–135.

Mateu, L., Ávila, E., Camejo, G., León, V., and Liscano, N. (1984). The structural stability of low-density lipoprotein: A kinetic X-ray scattering study of its interaction with arterial proteoglycans. *Biochim. Biophys. Acta*, 795(3):525–534.

Maxwell, P., Pugh, C., and Ratcliffe, P. (2001). Activation of the HIF pathway in cancer. *Curr. Opin. Genet. Dev.*, 11(3):293–299.

Meagher, K., Redman, L., and Carlson, H. (2003). Development of polyphosphate parameters for use with the AMBER force field. *Journal of computational chemistry*, 24(9):1016–1025.

Melo-Filho, C., Braga, R., and Andrade, C. (2014). 3D-QSAR approaches in drug design: perspectives to generate reliable CoMFA models. *Curr. Comput. Aided Drug Des.*, 10(2):148–159.

Merchan, J., Kovács, K., Railsback, J., Kurtoglu, M., Jing, Y., Piña, Y., Gao, N., Murray, T., Lehrman, M., and Lampidis, T. (2010). Antiangiogenic Activity of 2-Deoxy-D-Glucose. *PLoS One*, 5(10):e13699.

Mestas, J. and Ley, K. (2008). Monocyte-Endothelial Cell Interactions in the Development of Atherosclerosis. *Trends Cardiovasc. Med.*, 18(6):228–232.

Michel, J., Tirado-Rives, J., and Jorgensen, W. (2009). Prediction of the water content in protein binding sites. *Journal of Physical Chemistry B*, 113(40):13337–13346.

Minchenko, A., Leshchinsky, I., Opentanova, I., Sang, N., Srinivas, V., Armstead, V., and Caro, J. (2002). Hypoxia-inducible Factor-1-mediated Expression of the 6-Phosphofructo-2-kinase/fructose-2,6-bisphosphatase-3 (PFKFB3) Gene: ITS POSSIBLE ROLE IN THE WARBURG EFFECT. *J. Biol. Chem.*, 277(8):6183–6187.

Minchenko, O., Opentanova, I., and Caro, J. (2003). Hypoxic regulation of the 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase gene family (pfkfb-1–4) expression in vivo. *FEBS Lett.*, 554(3):264–270.

Moitessier, N., Englebienne, P., Lee, D., Lawandi, J., and Corbeil, C. (2008). Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. *Br. J. Pharmacol.*, 153 Suppl 1:S7–26.

Moos, M., John, N., Gräbner, R., Noßmann, S., Günther, B., Vollandt, R., Funk, C., Kaiser, B., and Habenicht, A. (2005). The Lamina Adventitia Is the Major Site of Immune Cell Accumulation in Standard Chow-Fed Apolipoprotein E–Deficient Mice. *Arter. Thromb. Vasc. Biol.*, 25(11):2386–2391.

Morris, G., Huey, R., Lindstrom, W., Sanner, M., Belew, R., Goodsell, D., and Olson, A. (2009). Autodock4 and autodocktools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.*, 30(16):2785–2791.

Mukai, H., Muramatsu, A., Mashud, R., Kubouchi, K., Tsujimoto, S., Hongu, T., Kanaho, Y., Tsubaki, M., Nishida, S., Shioi, G., Danno, S., Mehruba, M., Satoh, R., and Sugiura, R. (2016). PKN3 is the major regulator of angiogenesis and tumor metastasis in mice. *Sci. Rep.*, 6:18979.

Murphy, R., Repasky, M., Greenwood, J., Tubert-Brohman, I., Jerome, S., Annabhimoju, R., Boyles, N., Schmitz, C., Abel, R., Farid, R., and Friesner, R. (2016). WScore: A Flexible and Accurate Treatment of Explicit Water Molecules in Ligand-Receptor Docking. *J. Med. Chem.*, 59(9).

Mysinger, M., Carchia, M., Irwin, J., and Shoichet, B. (2012). Directory of useful decoys, enhanced (dud-e): Better ligands and decoys for better benchmarking. *J. Med. Chem.*, 55(14):6582–6594.

Nakano, T., Nakashima, Y., Yonemitsu, Y., Sumiyoshi, S., Chen, Y.-X., Akishima, Y., Ishii, T., Iida, M., and Sueishi, K. (2005). Angiogenesis and lymphangiogenesis and expression of lymphangiogenic factors in the atherosclerotic intima of human coronary arteries. *Hum. Pathol.*, 36(4):330–340.

Nakano, T., Ninomiya, T., Sumiyoshi, S., Onimaru, M., Fujii, H., Itabe, H., Nakashima, Y., Sueishi, K., Tsuruya, K., Oda, Y., Kitazono, T., and Kiyohara, Y. (2013). Chronic kidney disease is associated with neovascularization and intraplaque hemorrhage in coronary atherosclerosis in elders: results from the Hisayama Study. *Kidney Int.*, 84(2):373–380.

Nakashima, Y., Fujii, H., Sumiyoshi, S., Wight, T., and Sueishi, K. (2007). Early Human Atherosclerosis. *Arter. Thromb. Vasc. Biol.*, 27(5):1159–1165.

Nakashima, Y., Wight, T., and Sueishi, K. (2008). Early atherosclerosis in humans: role of diffuse intimal thickening and extracellular matrix proteoglycans. *Cardiovasc. Res.*, 79(1):14–23.

Nguyen, H., Roe, D., and Simmerling, C. (2013). Improved generalized born solvent model parameters for protein simulations. *J. Chem. Theory Comput.*, 9(4):2020–2034.

Nidhi, Glick, M., Davies, J., and Jenkins, J. (2006). Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. *J. Chem. Inf. Model.*, 46(3):1124–1133.

Nissink, J. W. M., Murray, C., Hartshorn, M., Verdonk, M. L., Cole, J. C., and Taylor, R. (2002). A new test set for validating predictions of protein-ligand interaction. *Proteins Struct. Funct. Bioinf.*, 49(4):457–471.

Novellasdemunt, L., Obach, M., Millán-Ariño, L., Manzano, A., Ventura, F., Rosa, J., Jordan, A., Navarro-Sabate, À., and Bartrons, R. (2012). Progestins activate 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase 3 (PFKFB3) in breast cancer cells. *Biochem. J*, 442(2):345–356.

Obach, M., Navarro-Sabaté, À., Caro, J., Kong, X., Duran, J., Gómez, M., Perales, J., Ventura, F., Rosa, J., and Bartrons, R. (2004). 6-Phosphofructo-2-kinase (pfkfb3) Gene Promoter Contains Hypoxia-inducible Factor-1 Binding Sites Necessary for Transactivation in Response to Hypoxia. *J. Biol. Chem.*, 279(51):53562–53570.

O'Boyle, N., Banck, M., James, C., Morley, C., Vandermeersch, T., and Hutchison, G. (2011). Open Babel: An open chemical toolbox. *J. Cheminformatics*, 3(1):33.

Oehme, D., Brownlee, R., and Wilson, D. (2012). Effect of atomic charge, solvation, entropy, and ligand protonation state on mm-pb(gb)sa binding energies of hiv protease. *J. Comput. Chem.*, 33(32):2566–2580.

Ohayon, J., Finet, G., Gharib, A., Herzka, D., Tracqui, P., Heroux, J., Rioufol, G., Kotys, M., Elagha, A., and Pettigrew, R. (2008). Necrotic core thickness and positive arterial remodeling index: emergent biomechanical factors for evaluating the risk of plaque rupture. *Am. J. Physiol. Heart Circ. Physiol.*, 295(2):H717–27.

Okamoto, M., Takayama, K., Shimizu, T., Ishida, K., Takahashi, O., and Furuya, T. (2009). Identification of death-associated protein kinases inhibitors using structure-based virtual screening. *J. Med. Chem.*, 52(22):7323–7327.

Okar, D. A. and Lange, A. J. (1999). Fructose-2,6-bisphosphate and control of carbohydrate metabolism in eukaryotes. *BioFactors*, 10(1):1–14.

Okar, D. A., Manzano, A., Navarro-Sabate, A., Riera, L., Bartrons, R., and Lange, A. J. (2001). PFK-2/FBPase-2: maker and breaker of the essential biofactor fructose-2,6-bisphosphate. *Trends Biochem. Sci.*, 26(1):30–35.

Okar, D. A., Wu, C., and Lange, A. J. (2004). Regulation of the regulatory enzyme, 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase. *Adv. Enzyme Regul.*, 44(1):123 – 154.

Onufriev, A., Bashford, D., and Case, D. (2004). Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins Struct. Funct. Bioinf.*, 55(2):383–394.

OpenEye (2015). SZMAP 1.2.1.4.

Pasterkamp, G., Schoneveld, A., Hijnen, D., de Kleijn, D., Teepen, H., van der Wal, A., and Borst, C. (2000). Atherosclerotic arterial remodeling and the localization of macrophages and matrix metalloproteases 1, 2 and 9 in the human coronary artery. *Atherosclerosis*, 150(2):245–253.

Paterson, J. (1936). Vascularization and hemorrhage of the intima of arteriosclerotic arteries. *Arch. Pathol.*, 22:312–324.

Paul, S., Mytelka, D., Dunwiddie, C., Persinger, C., Munos, B., Lindborg, S., and Schacht, A. (2010). How to improve R&D productivity: the pharmaceutical industry's grand challenge.

Peng, F., Li, Q., Sun, J.-Y., Luo, Y., Chen, M., and Bao, Y. (2018). PFKFB3 is involved in breast cancer proliferation, migration, invasion and angiogenesis. *Int. J. Oncol.*, 52(3):945–954.

Perez-Pineiro, R., Burgos, A., Jones, D., Andrew, L., Rodriguez, H., Suarez, M., Fairlamb, A., and Wishart, D. (2009). Development of a Novel Virtual Screening Cascade Protocol to Identify Potential Trypanothione Reductase Inhibitors. *J. Med. Chem.*, 52(6):1670–1680.

Quintero, M., Colombo, S., Godfrey, A., and Moncada, S. (2006). Mitochondria as signaling organelles in the vascular endothelium. *Proc. Natl. Acad. Sci. U.S.A.*, 103(14):5379–5384.

R., B. (2018). Amber parameter database. `http://research.bmh.manchester.ac.uk/bryce/amber/`. Accessed: April 2016.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raha, K. and Merz, K. (2005). Large-Scale Validation of a Quantum Mechanics Based Scoring Function: Predicting the Binding Affinity and the Binding Mode of a Diverse Set of Protein−Ligand Complexes. *J. Med. Chem.*, 48(14):4558–4575.

Rastelli, G., Rio, A., Degliesposti, G., and Sgobba, M. (2009). Fast and accurate predictions of binding free energies using mm-pbsa and mm-gbsa. *J. Comput. Chem.*, 31(4):797–810.

Regazzoni, L. (2018). personal communication.

Rehman, J. and Archer, S. (2010). A proposed mitochondrial–metabolic mechanism for initiation and maintenance of pulmonary arterial hypertension in fawn-hooded rats: The warburg model of pulmonary arterial hypertension. In Yuan, J.-J. and Ward, J., editors, *Membrane Receptors, Channels and Transporters in Pulmonary Circulation*, pages 171–185, Totowa, NJ. Humana Press.

Remo, A., Simeone, I., Pancione, M., Parcesepe, P., Finetti, P., Cerulo, L., Bensmail, H., Birnbaum, D., Van Laere, S., Colantuoni, V., Bonetti, F., Bertucci, F., Manfrin, E., and Ceccarelli, M. (2015). Systems biology analysis reveals NFAT5 as a novel biomarker and master regulator of inflammatory breast cancer. *Journal of Translational Medicine*, 13(1):138.

Richardson, P., Davies, M., and Born, G. (1989). Influence of plaque configuration and stress distribution on fissuring of coronary atherosclerotic plaques. *Lancet*, 334(8669):941–944.

Rider, M. H., Bertrand, L., Vertommen, D., Michels, P. A., Rousseau, G. G., and Hue, L. (2004). 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase: head-to-head with a bifunctional enzyme that controls glycolysis. *Biochem. J*, 381(3):561–579.

Riera, L., Manzano, A., Navarro-Sabaté, A., Perales, J., and Bartrons, R. (2002). Insulin induces PFKFB3 gene expression in HT29 human colon adenocarcinoma cells. *Biochim. Biophys. Acta, Mol. Cell. Res.*, 1589(2):89–92.

Roberts, B. and Mancera, R. (2008). Ligand−Protein Docking with Water Molecules. *J. Chem. Inf. Model.*, 48(2):397–408.

Robinson, M., McCarthy, D., and Smyth, G. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140.

Robinson, M. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3):R25.

Robinson, M. and Smyth, G. (2007). Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics (Oxford, England)*, 23(21):2881–2887.

Rueda, M., Bottegoni, G., and Abagyan, R. (2009). Consistent improvement of cross-docking results using binding site ensembles generated with elastic network normal modes. *J. Chem. Inf. Model.*, 49(3):716–725.

Rueda, M., Bottegoni, G., and Abagyan, R. (2010). Recipes for the Selection of Experimental Protein Conformations for Virtual Screening. *J. Chem. Inf. Model.*, 50(1):186–193.

Ruparelia, N., Chai, J. T., Fisher, E. A., and Choudhury, R. P. (2017). Inflammatory processes in cardiovascular disease: a route to targeted therapies. *Nat. Rev. Cardiol.*, 14(3):133–144.

S., A. (2018). Fastqc. a quality control tool for high throughput sequence data. `http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`. Accessed: Dec 2018.

Sakai, A., Kato, M., Fukasawa, M., Ishiguro, M., Furuya, E., and Sakakibara, R. (1996). Cloning of cdna encoding for a novel isozyme of fructose 6-phosphate,2-kinase/fructose 2,6-bisphosphatase from human placenta1. *J. Biochem.*, 119(3):506–511.

Sakakibara, R., Kato, M., Okamura, N., Nakagawa, T., Komada, Y., Tominaga, N., Shimojo, M., and Fukasawa, M. (1997). Characterization of a human placental fructose-6-phosphate, 2-kinase/fructose- 2,6 - bisphosphatase1. *J. Biochem.*, 122(1):122–128.

Sales, G. and Romualdi, C. (2011). parmigene–a parallel R package for mutual information estimation and gene network reconstruction. *Bioinformatics (Oxford, England)*, 27(13):1876–1877.

Schafer, J. and Strimmer, K. (2005). An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics (Oxford, England)*, 21(6):754–764.

Semenza, G. (2006). Development of novel therapeutic strategies that target HIF-1. *Expert Opin. Ther. Targets*, 10(2):267–280.

Seo, M., Kim, J.-D., Neau, D., Sehgal, I., and Lee, Y.-H. (2011). Structure-Based Development of Small Molecule PFKFB3 Inhibitors: A Framework for Potential Cancer Therapeutic Agents Targeting the Warburg Effect. *PLoS One*, 6(9):e24179.

Sgobba, M., Caporuscio, F., Anighoro, A., Portioli, C., and Rastelli, G. (2012). Application of a post-docking procedure based on mm-pbsa and mm-gbsa on single and multiple protein conformations. *Eur. J. Med. Chem.*, 58:431–440.

Sheridan, R. and Kearsley, S. (2002). Why do we need so many chemical similarity search methods? *Drug Discov. Today*, 7(17):903–911.

Simonneau, G., Torbicki, A., Dorfmüller, P., and Kim, N. (2017). The pathophysiology of chronic thromboembolic pulmonary hypertension. *Eur. Respir. Rev.*, 26(143):160112.

Sindhikara, D., Yoshida, N., and Hirata, F. (2012). Placevent: An algorithm for prediction of explicit solvent atom distribution-Application to HIV-1 protease and F-ATP synthase. *Journal of Computational Chemistry*, 33(18):1536–1543.

Sluimer, J. and Daemen, M. (2009). Novel concepts in atherogenesis: angiogenesis and hypoxia in atherosclerosis. *J. Pathol.*, 218(1):7–29.

Sluimer, J., Gasc, J.-M., van Wanroij, J., Kisters, N., Groeneweg, M., Sollewijn Gelpke, M., Cleutjens, J., van den Akker, L., Corvol, P., Wouters, B., Daemen, M., and Bijnens, A.-P. (2008). Hypoxia, Hypoxia-Inducible Transcription Factor, and Macrophages in Human Atherosclerotic Plaques Are Correlated With Intraplaque Angiogenesis. *J. Am. Coll. Cardiol.*, 51(13):1258–1265.

Sluimer, J., Kolodgie, F., Bijnens, A., Maxfield, K., Pacheco, E., Kutys, B., Duimel, H., Frederik, P., van Hinsbergh, V., Virmani, R., and Daemen, M. (2009). Thin-Walled Microvessels in Human Coronary Atherosclerotic Plaques Show Incomplete Endothelial Junctions: Relevance of Compromised Structural Integrity for Intraplaque Microvascular Leakage. *J. Am. Coll. Cardiol.*, 53(17):1517–1527.

Smolders, V. (2018). personal communication.

Sobrino, F. and Gualberto, A. (1985). Hormonal regulation of fructose 2,6-bisphosphate levels in epididymal adipose tissue of rat. *FEBS Lett.*, 182(2):327–330.

Sommer, K., Friedrich, N.-O., Bietz, S., Hilbig, M., Inhester, T., and Rarey, M. (2016). Unicon: A powerful and easy-to-use compound library converter. *J. Chem. Inf. Model.*, 56(6):1105–1111.

Song, C., Lim, S., and Tong, J. (2009). Recent advances in computer-aided drug design. *Brief. Bioinformatics*, 10(5):579–591.

Spyrakis, F. and Cavasotto, C. (2015). Open challenges in structure-based virtual screening: Receptor modeling, target flexibility consideration and active site water molecules description. *Arch. Biochem. Biophys.*, 583:105–119.

Srinivasan, J., Cheatham, T., Cieplak, P., Kollman, P., and Case, D. (1998). Continuum solvent studies of the stability of dna, rna, and phosphoramidate - dna helices. *J. Am. Chem. Soc.*, 120(37):9401–9409.

St-Gallay, S., Bennett, N., Critchlow, S., Curtis, N., Davies, G., Debreczeni, J., Evans, N., Hardern, I., Holdgate, G., Jones, N., Leach, L., Maman, S., McLoughlin, S., Preston, M., Rigoreau, L., Thomas, A., Turnbull, A., Walker, G., Walsh, J., Ward, R., Wheatley, E., and Winter-Holt, J. (2017). A High-Throughput Screening Triage Workflow to Authenticate a Novel Series of PFKFB3 Inhibitors. *SLAS Discov.*, 23(1):11–22.

States), C. A. U. (2018). Web of science. Accessed: 2018-12-04.

Suarez, S. and Ballmer-Hofer, K. (2001). VEGF transiently disrupts gap junctional communication in endothelial cells. *J. Cell Sci.*, 114(6):1229 LP – 1235.

Sun, H., Li, Y., Shen, M., Tian, S., Xu, L., Pan, P., Guan, Y., and Hou, T. (2014a). Assessing the performance of mm/pbsa and mm/gbsa methods. 5. improved docking performance using high solute dielectric constant mm/gbsa and mm/pbsa rescoring. *Phys. Chem. Chem. Phys.*, 16(40):22035–22045.

Sun, H., Li, Y., Tian, S., Xu, L., and Hou, T. (2014b). Assessing the performance of mm/pbsa and mm/gbsa methods. 4. accuracies of mm/pbsa and mm/gbsa methodologies evaluated by various simulation protocols using pdbbind data set. *PCCP*, 16(31):16719–16729.

Szulcek, D. (2018). personal communication.

Tabas, I. (2010). Macrophage death and defective inflammation resolution in atherosclerosis. *Nat. Rev. Immunol.*, 10(1):36–46.

Tange, O. (2018). GNU Parallel 2018.

Tawakol, A., Singh, P., Mojena, M., Pimentel-Santillana, M., Emami, H., MacNabb, M., Rudd, J., Narula, J., Enriquez, J., Través, P., Fernández-Velasco, M., Bartrons, R., Martín-Sanz, P., Fayad, Z., Tejedor, A., and Boscá, L. (2015). Hif-1α and pfkfb3 mediate a tight relationship

between proinflammatory activation and anerobic metabolism in atherosclerotic macrophages. *Arter. Thromb. Vasc. Biol.*, 35(6):1463–1471.

Teichert-Kuliszewska, K., Kutryk, M., Kuliszewski, M., Karoubi, G., Courtman, D., Zucco, L., Granton, J., and Stewart, D. (2006). Bone Morphogenetic Protein Receptor-2 Signaling Promotes Pulmonary Arterial Endothelial Cell Survival. *Circ. Res.*, 98(2):209–217.

Telang, S., Yalcin, A., Clem, A., Bucala, R., Lane, A., Eaton, J., and Chesney, J. (2006). Ras transformation requires metabolic control by 6-phosphofructo-2-kinase. *Oncogene*, 25:7225.

Thilagavathi, R. and Mancera, R. (2010). Ligand-protein cross-docking with water molecules. *J. Chem. Inf. Model.*, 50(3):415–421.

Thompson, D., Humblet, C., and Joseph-McCarthy, D. (2008). Investigation of mm-pbsa rescoring of docking poses. *J. Chem. Inf. Model.*, 48(5):1081–1091.

Trefely, S., Khoo, P.-S., Krycer, J., Chaudhuri, R., Fazakerley, D., Parker, B., Sultani, G., Lee, J., Stephan, J.-P., Torres, E., Jung, K., Kuijl, C., James, D., Junutula, J., and Stöckli, J. (2015). Kinome Screen Identifies PFKFB3 and Glucose Metabolism as Important Regulators of the Insulin/Insulin-like Growth Factor (IGF)-1 Signaling Pathway. *J. Biol. Chem.*, 290(43):25834–25846.

Trenti, A., Tedesco, S., Boscaro, C., Ferri, N., Cignarella, A., Trevisi, L., and Bolego, C. (2017). The Glycolytic Enzyme PFKFB3 Is Involved in Estrogen-Mediated Angiogenesis via GPER1. *Journal of Pharmacology and Experimental Therapeutics*, 361(3):398–407.

Trott, O. and Olson, A. (2009). Autodock vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.*, 31(2):455–461.

Tura, O. (2018). personal communication.

Turan, N., Kalko, S., Stincone, A., Clarke, K., Sabah, A., Howlett, K., Curnow, S., Rodriguez, D., Cascante, M., O'Neill, L., Egginton, S., Roca, J., and Falciani, F. (2011). A Systems Biology Approach Identifies Molecular Networks Defining Skeletal Muscle Abnormalities in Chronic Obstructive Pulmonary Disease. *PLOS Computational Biology*, 7(9):e1002129.

Twickler, T., Dallinga-Thie, G., Chapman, M., and Cohn, J. (2005). Remnant lipoproteins and atherosclerosis. *Curr. Atheroscler. Rep.*, 7(2):140–147.

ULC, C. C. G. (2018). Moe, molecular operating environment. 1010 Sherbooke St. West, Suite 910, Montreal, QC, Canada, H3A 2R7.

Upritchard, J. and Sutherland, W. (1999). Oxidation of heparin-treated low density lipoprotein by peroxidases. *Atherosclerosis*, 146(2):211–219.

Usman, A., Ribatti, D., Sadat, U., and Gillard, J. (2015). From Lipid Retention to Immune-Mediate Inflammation and Associated Angiogenesis in the Pathogenesis of Atherosclerosis. *J. Atheroscler. Thromb.*, 22(8):739–749.

van de Veerdonk, M., Bogaard, H., and Voelkel, N. (2016). The right ventricle and pulmonary hypertension. *Heart Fail. Rev.*, 21(3):259–271.

Van Schaftingen, E., Hue, L., and Hers, H. (1980a). Control of the fructose 6-phosphate/fructose 1,6-bisphosphate cycle in isolated hepatocytes by glucose and glucagon. role of a low-molecular-weight stimulator of phosphofructokinase. *Biochem. J*, 192(3):887–895.

Van Schaftingen, E., Hue, L., and Hers, H. (1980b). Fructose 2,6-bisphosphate, the probably structure of the glucose- and glucagon-sensitive stimulator of phosphofructokinase. *Biochem. J*, 192(3):897–901.

Virmani, R., Kolodgie, F., Burke, A., Finn, A., Gold, H., Tulenko, T., Wrenn, S., and Narula, J. (2005). Atherosclerotic plaque progression and vulnerability to rupture: angiogenesis as a source of intraplaque hemorrhage. *Arter. Thromb. Vasc. Biol.*, 25(10):2054–2061.

Vuorinen, A. and Schuster, D. (2015). Methods for generating and applying pharmacophore models as virtual screening filters and for bioactivity profiling. *Methods*, 71:113–134.

Wang, J., Morin, P., Wang, W., and Kollman, P. (2001). Use of MM-PBSA in reproducing the binding free energies to HIV-1 RT of TIBO derivatives and predicting the binding mode to HIV-1 RT of efavirenz by docking and MM-PBSA. *J. Am. Chem. Soc.*, 123(22):5221–5230.

Wang, J., Wang, W., Kollman, P., and Case, D. (2006). Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of Molecular Graphics and Modelling*, 25(2):247–260.

Wang, J., Wolf, R., Caldwell, J., Kollman, P., and Case, D. (2004a). Development and testing of a general amber force field. *Journal of computational chemistry*, 25(9):1157–1174.

Wang, J., Wolf, R., Caldwell, J., Kollman, P., and Case, D. (2004b). Development and testing of a general amber force field. *J. Comput. Chem.*, 25(9):1157–1174.

Wang, R. and Wang, S. (2001). How Does Consensus Scoring Work for Virtual Library Screening? An Idealized Computer Experiment. *J. Chem. Inf. Comput. Sci.*, 41(5):1422–1426.

Wang, X., Wei, X., Thijssen, B., Das, J., Lipkin, S., and Yu, H. (2012). Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nature Biotechnol.*, 30(2):159–164.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10:57.

Warburg, O. (1956). On the origin of cancer cells. *Science*, 123(3191):309–314.

Waskom, M., Botvinnik, O., Hobson, P., Cole, J., Halchenko, Y., Hoyer, S., Miles, A., Augspurger, T., Yarkoni, T., Megies, T., Coelho, L., Wehner, D., Cynddl, Ziegler, E., Diego0020, Zaytsev, Y., Hoppe, T., Seabold, S., Cloud, P., Koskinen, M., Meyer, K., Qalieh, A., and Allan, D. (2014). seaborn: v0.5.0 (November 2014).

Weber, G. (1977). Enzymology of cancer cells. *N. Engl. J. Med.*, 296(10):541–551.

(WHO), W. H. O. (2017). Fact sheets - cardiovascular diseases. `https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)`. Accessed: Dec 2018.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Williams, J., Armstrong, M., and Heistad, D. (1988). Vasa vasorum in atherosclerotic coronary arteries: responses to vasoactive stimuli and regression of atherosclerosis. *Circ. Res.*, 62(3):515–523.

Williams, K. and Tabas, I. (1995). The response-to-retention hypothesis of early atherogenesis. *Arter. Thromb. Vasc. Biol.*, 15(5):551–561.

Williams, T., C., K., and many others (2018). Gnuplot 5.2: an interactive plotting program. `http://gnuplot.sourceforge.net/`.

Wong, S., Amaro, R., and McCammon, J. (2009). MM-PBSA Captures Key Role of Intercalating Water Molecules at a Protein-Protein Interface. *J. Chem. Theory Comput.*, 5(2):422–429.

Wong, S. and Lightstone, F. (2011). Accounting for water molecules in drug design. *Expert Opin. Drug Discov.*, 6(1):65–74.

Wu, C., Khan, S. A., Peng, L.-J., and Lange, A. (2006). Roles for fructose-2,6-bisphosphate in the control of fuel metabolism: Beyond its allosteric effects on glycolytic and gluconeogenic enzymes. *Adv. Enzyme Regul.*, 46(1):72–88.

Xu, L., Sun, H., Li, Y., Wang, J., and Hou, T. (2013). Assessing the performance of mm/pbsa and mm/gbsa methods. 3. the impact of force fields and ligand charge models. *J. Phys. Chem. B*, 117(28):8408–8421.

Xu, Y., An, X., Guo, X., Habtetsion, T., Wang, Y., Xu, X., Kandala, S., Li, Q., Li, H., Zhang, C., Caldwell, R., Fulton, D., Su, Y., Hoda, M., Zhou, G., Wu, C., and Huo, Y. (2014). Endothelial PFKFB3 Plays a Critical Role in Angiogenesis. *Arter. Thromb. Vasc. Biol.*, 34(6):1231–1239.

Yang, T., Wu, J., Yan, C., Wang, Y., Luo, R., Gonzales, M., Dalby, K., and Ren, P. (2011a). Virtual screening using molecular simulations. *Proteins Struct. Funct. Bioinf.*, 79(6):1940–1951.

Yang, X., Long, L., Reynolds, P., and Morrell, N. (2011b). Expression of Mutant BMPR-II in Pulmonary Endothelial Cells Promotes Apoptosis and a Release of Factors that Stimulate Proliferation of Pulmonary Arterial Smooth Muscle Cells. *Pulm. Circ.*, 1(1):103–110.

Young, T., Abel, R., Kim, B., Berne, B. J., and Friesner, R. A. (2007). Motifs for molecular recognition exploiting hydrophobic enclosure in protein–ligand binding. *Proceedings of the National Academy of Sciences*, 104(3):808–813.

Zhang, X., Wong, S., and Lightstone, F. (2014). Toward fully automated high performance computing drug discovery: a massively parallel virtual screening pipeline for docking and molecular mechanics/generalized born surface area rescoring to improve enrichment. *J. Chem. Inf. Model.*, 54(1):324–337.

Zhao, S. and Li, S. (2010). Network-Based Relating Pharmacological and Genomic Spaces for Drug Target Identification. *PLoS ONE*, 5(7):e11764.

Şenbabaoğlu, Y., Sümer, S., Sánchez-Vega, F., Bemis, D., Ciriello, G., Schultz, N., and Sander, C. (2016). A Multi-Method Approach for Proteomic Network Inference in 11 Human Cancers. *PLOS Computational Biology*, 12(2):e1004765.