Graduate School of Frontier Sciences, The University of Tokyo

Department of Computational Biology

2012-2014

Master's Thesis

Development of a virtual screening tool and its application in the discovery of G9a inhibitors

Submitted: August, 2014

Adviser: Professor Dr. Kam Y.J. Zhang

47-126921 Lee Xiao Yin

2012-2014 Master's Thesis

Development of a virtual screening tool and its application in the discovery of G9a inhibitors

Lee Xiao Yin

Acknowledgements

My special thanks go to my supervisor Prof. Dr. Kam Y. J. Zhang for his great guidance and help throughout my Master's degree studies. His knowledge and commitment massively motivated me in my research. I am very thankful for his encouragement and patience during my Master studies. He is the friendliest supervisor that I ever had.

I am grateful to Dr. Arnout Voet who had mentored me during my research projects. I have learned all my computational drug design skills from him. He has taught me how to perform and analyze my results. I would like to thank him for his patience and guidance throughout my Master's project. Without his advice and assistance, I would not have completed my project.

Moreover, I would like to thank Francois Berenger for his assistance on solving all my programming problems. I am thankful for his help throughout my project. Without his advice and support, I would not be able to run my experiment so smoothly.

Besides, I want to express my thanks to my lab members in RIKEN that giving me a lot of advices during my research. They are David Simoncini, Ashutosh Kumar, Rojan Shrestha, Muhammad Muddassar, Taeho Jo and Kamlesh Sahu. Because of them, the laboratory is always full of laughter. Thanks for being with me as nice colleagues throughout the days in my Master's research. Furthermore, I would like to thank the collaborators of my project, Dr. Yoshida and Dr. Ito for their support in testing my compounds.

I would like to thank my friends who have given me encouragement and help when I was facing some problems in my research. I really appreciated for their kindness and support. Moreover, I would like to thank all the professors in The University of Tokyo. I have always felt welcome to ask for any possible help and support that I needed. In this way I would like to express special regards to. I am also grateful to the Ajinomoto Scholarship Foundation for grants that have funded part of this study.

At this point, I warmly thank my family members who have given me their full support both emotionally and physically throughout the years. Their encouragement had motivated me to finish my research work on time. Without their support and tolerance, I would not have succeeded in the completion of this work.

Table of Contents

List of Figures

List of Tables

Aims of the Thesis

1. G9a

In our research, we are investigating the histone lysine methylation by G9a, which is reported to be an "on and off" switch that is crucial for many molecular biological processes in human cells. Hence, G9a has become an interesting target for the drug discovery and design field. The major challenge of the research on the histone H3 at lysine 9 (H3K9) of G9a is to find a novel drug that can specifically inhibit the methylation of the histone by G9a. The methylation of histone H3 at lysine 9 (H3K9) is related to diverse molecular biology processes like DNA methylation, heterochromatin formation and gene transcriptional silencing. Those epigenetic changes will result in the severe changes of the cellular regulatory system and may result in the uncontrolled growth and cell division of cells, more commonly known as cancer. Currently, there are few reported inhibitors targeting the G9a protein, typically acting on the substrate site, and all belong to the same chemotype. For others the exact mechanism of action still remains unknown. For my Master's thesis the primary goal was to identify novel inhibitors of G9a, preferably having a different mechanism of action than the Epidithiodiketopiperazine (ETP) class. An introduction to G9a can be found in Chapter 1. To reach this goal computational drug design methods were used. A general introduction to computational drug design methods is given in Chapter 2. During the virtual process for the novel G9a inhibitors we developed a novel computational drug design algorithm called EleKit2 and applied it to enhance the virtual screening performance. The results can be found in Chapter 4.

2. EleKit2

Complementarity at the receptor-ligand interface is of major importance for molecular recognition and is therefore critical for the rational design of receptor-ligand complexes. Next to complementarity in shape, it is generally accepted that the long range electrostatic interactions should also exhibit complementarity for optimal recognition. The second goal of my thesis was to develop a tool (named EleKit2) that measures the electrostatic complementarity between a protein receptor and a small molecule ligand, benchmark it on a database of experimentally determined receptor ligand complexes and to evaluate the improvement when applied in a virtual screening experiment. The results can be found in Chapter 3.

Chapter 1

G9a as a drug target

1.1 Epigenetics

The word "epigenetics" was first introduced by Conrad H. Waddington in 1942 and is defined as "the branch of biology that studies the causal interactions between genes and their products, which bring the phenotype into being" [1]. The epigenetic term is derived from the Greek word of "epigenesis" which translates as the developmental processes of an individual [2]. However, nowadays epigenetic is commonly defined as a genetic alteration that involves an inheritable phenotype without a single change in the underlying DNA sequence [3]. Epigenetic mechanisms act as a gatekeeper in the human genome that can turn the genes on and off. Epigenetic regulation is a crucial mechanism in the human body and keeps the gene activity on the right track. Moreover, epigenetic regulation is essential in an organism's growth by maintaining the cell system, such as the cell growth, cell regeneration and cell death [4]. Through the influence of the external environment effects, with "aging" as one of the most common examples, will increase the likelihood of epigenetic changes to happen. Different environment factors, such as living lifestyle, prenatal stress, physiopathological situations and pharmacological interventions will also contribute to epigenetic changes [5]. These epigenetic changes will cause a major disruption in the gene expression which may lead to cancer and other severe diseases in human bodies [6-8]. The common examples of the epigenetic defects are non-coding RNAs gene silencing [9, 10], DNA methylation [6, 11, 12] and the histone modification [13-16].

1.1.1 Histone modification

Histones are the proteins of chromatin that associate with DNA to form chromosomes. Histones act as a spool around which DNA can wind, see Figure 1. The alteration in histones will affect the normal arrangement of chromatin which can cause or reverse the formation of heterochromatin (compact form of the chromatin) and inactivate the process of DNA transcription [14]. Histone modification is generally described as post-translational changes of the histone and chromatin which is important in governing the gene regulation in

epigenetic states [17-19]. These post-translational histone modifications will alter packing of chromatin and will result in the alterations of the genetic signaling process. The post-translational modifications usually happen at the N-terminal histone tails and may consist of acetylation [20], methylation [21], ubiquitination [22], phosphorylation [23], sumoylation [24], ADP ribosylation [25], deamination [26] and/or proline isomerization [27]. Those changes are the key to cause the chromatin remodeling in epigenetic states [28]. Many malignant alterations are related to cancers like prostate cancer, leukemia, pancreatic cancer, breast cancer, colorectal cancer and so on. As the epigenetic change creates a mark or error on the gene, therefore they have also been recognized as epigenetic biomarkers, which are commonly used in cancer diagnosis for early stage cancer detection [8].

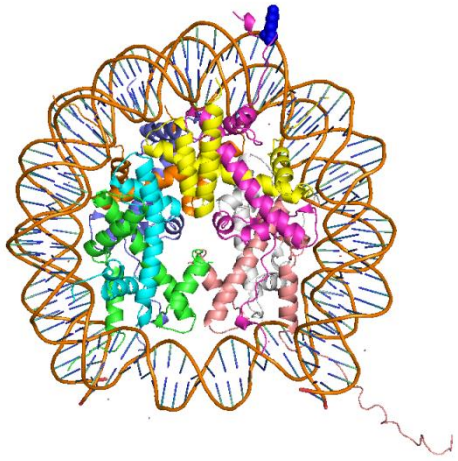


Figure 1 Structure representation of core components of nucleosome structure. The nucleosome (PDB 1AOI, cartoon representation) is composed of eight histone proteins of H3, H4, H2A and H2B pairs which depicts here in different colored cartoons. DNA (around 147 base pairs per nucleosome) is wrapped around the octomer core proteins (in orange). The N-terminal tail of histone 3 is extended out and the position of lysine 9 is indicated with dark colored spheres.

The occurrence of epigenetic errors in one's gene will disrupt the cell's development system, which may transform normal cells to cancer cells [29]. One of the contributions for the uncontrolled development of the cells would be the histone methylation. The methylation on the lysine tail at different loci of histone will modulate the transcriptional activation and transcriptional repression of the gene that results in the abnormal expression of the corresponding protein. Studies have shown that the histone methylation that happens at the lysine tail of histone H3 at lysine 4 marking the error on the active genes [4, 30]. However, when methylation of lysine (K9) on a specific histone H3 (H3K9me) will cause the gene

repression [16, 30]. The methylation of the H3K9me is usually mediated by the histone methyltransferases (HKMTs) and it serves an important role in the studies of the methylation mechanism in H3K9me. Many of the HKMTs containing a conserved SET domain and many studies investigate this domain in order to have a deeper understanding on the histone methylation as well as to validate it as a drug target.

1.2 What is G9a?

G9a (also known as KMT1C or EHMT2) is one of the identified lysine methyltransferase from the Suv39h protein family that is involved in the mono-, di- and tri-methylation of the euchromatic histone lysine, mostly histone 3 lysine 9 (H3K9) [31]. The methylation of H3K9 by G9a occurs at the ε-amino group of lysine residues in euchromatic region. For the structure of the protein, see Figure 2. The G9a encoded gene (roughly 1000 amino acids, depending on alternate splicing) is located in the class III region of the human major histocompatibility complex (MHC) [32]. Furthermore, G9a contains a polyglutamic acid stretch, ankyrin (ANK) repeats and SET domain at its amino terminus which was initially found in the Notch protein of Drosophila [21, 33].

The formation di- and tri- methylated H3K9 (annotated as H3K9me2 and HK39me3) are common in non-active genes and seldom in active genes [34, 35]. The methylation of histone H3 Lys 9 (H3K9me) by G9a initiates the formation of the pericentric heterochromatin and also the condensation of mitotic chromosome [36]. H3K9me acts as a specific tag to recruit HP1 protein through binding of the HP1 chromodomain to the methylated lysine tail [35, 36]. This particularly happens on the H3K9 and H3K27 [37]. HP1 protein is a heterochomatic adapter molecules [37, 38] that contains a specialized chromodomain which recognizes and binds the methylated lysine 9 on H3 during histone methylation and mediates the epigenetic transcriptional repression of the active gene [39-41].

Additionally, GLP is a protein related to G9a, and is also known as EuHMTase/ KMT1D. It also plays an important role in gene silencing. GLP possesses a high similar specificity of substrate on histones with the G9a and they are found especially in the H3K9me1 and H3K9me2 [42-44]. The formation of a complex of GLP with G9a is mediated via heterodimerization at the SET domains and the importance was indicated during *in vivo* studies investigating H3K9 methylation [42, 45].

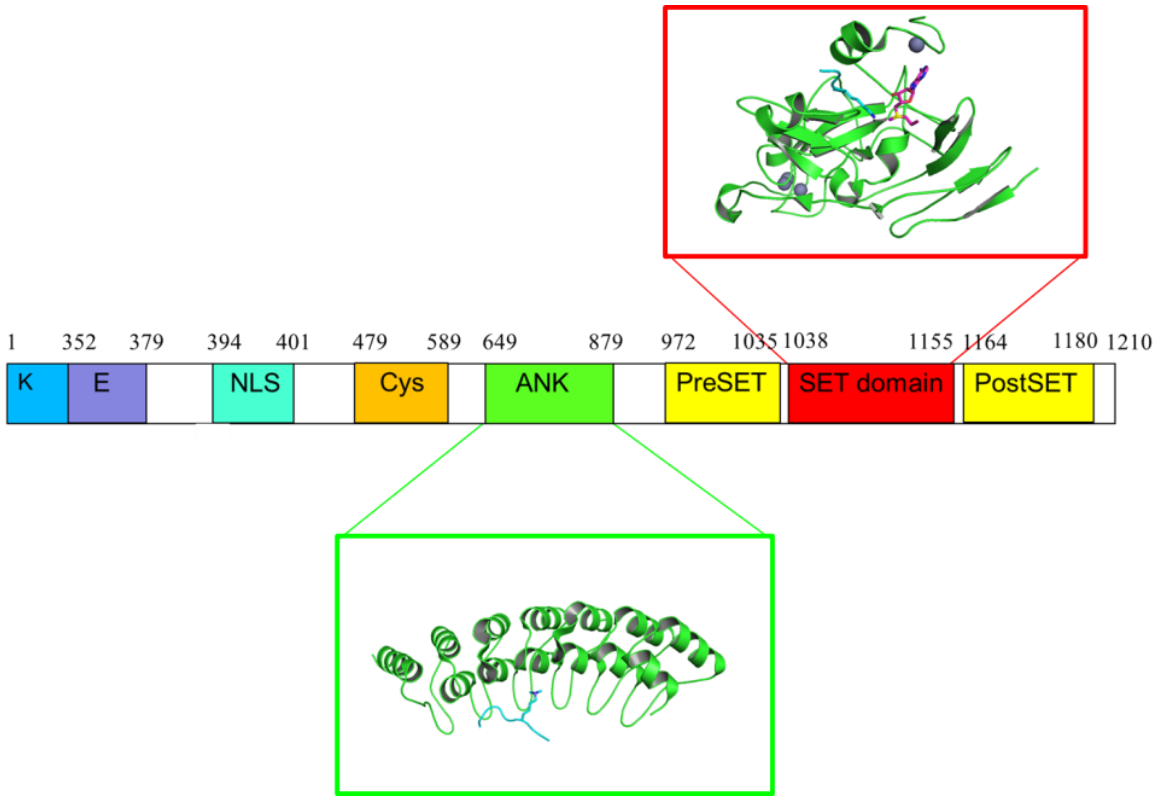


Figure 2 Schematic diagram of the G9a domains structure. G9a is composed of 1298 amino acid residues, with different conserved domains. The G9a protein contains an N-terminal lysine tail (K), a glutamate rich domain (E), a Nuclear Localization Sequence (NLS), a cysteine- rich domain (Cys), a ankyrin repeat domain (ANK) , a PreSET domain (PreSET), SET domain and a postSET domain (Post SET).

1.2.1 SET domain

The SET domain is a conserved protein domain, consisting of around 130 amino acid residues, and typically found in chromatin regulator proteins. The SET domain was originally found in *Drosophila* from a genetic screen of position effect variegation (PEV) at the suppressor protein SU(VAR)3–9 [46], the Polycomb-group protein Enhancer of zeste [47] and the trithorax-group protein Trithorax [48]. As such, the SET domain name is the abbreviation of SU(VAR)3-9, Enhancer-of-zeste, Trithorax domain. The SET domain consists of both cysteine-rich preSET and postSET motifs that have been identified as essential for the enzymatic activity in protein methyltransferase [21]. SET-domain containing enzymes are mainly found in histone lysine methyltransferases (HKMTs), which act as a catalyst for histone methylation. The SET domain harbors the enzyme functionality, which specifically transfers methyl groups onto histone lysine. Moreover, SET domain proteins also take part in the modulation of their own gene activity [49]. There are several HKMTs that

contain the SET domain, such as human SUV39H1 [50], SUV39H2 [50, 51] and the discussed protein G9a [45]. *Strahl, et al.* performed the *in vivo* studies on G9a and discovered that the methylation happened specifically at lysine 9 and lysine 27 of histone H3 [52]. Besides, Tachibana, *et al.* also found that the existence of the specific activity in G9a was 10 to 20 fold higher than to the other Su39h protein, such as Suv39h1 [45]. Besides, the SET domain is also important for mediating the protein-protein interaction while interacting with the dual-specificity phosphatases (dsPTPases) proteins [53]. These findings have proven that G9a is a potential target for the studies on histone methyltransferase.

1.2.2 G9a catalytic mechanism

S-adenosyl-L-methionine (SAM) is an essential cofactor for G9a and other protein lysine methyltransferase (PKMTs), in which they transfer the methyl group from SAM to the ε-amino group of lysine residues of the target protein [30]. The catalytic mechanism is started when both the SAM molecule and the lysine tail are bound simultaneously at the catalytic active site of SET domain. After the binding, the Tyrosine (Tyr) 287 residue, located near the active site, will deprotonate the ε-amino of the lysine tail [54]. The chemical reaction happens after the deprotonatation of ε-amino of lysine tail by the Tyr 287 when the deprotonated lysine tail makes a nucleophilic attack on the methyl group, which is carried on the sulfur atom of the SAM molecule. This causes the methyl group to transfer to the lysine tail [54]. For a schematic depiction, see Figure 3. The transfer process of the methyl group from the SAM molecule to the lysine tail is called histone lysine methylation. The methylation of the lysine group on histone will result in the modification of histone and generates the binding site that attracts the repressor or activator proteins with specific domains, such as chromodomains to bind [55]. The proteins that bind to the histone will determine the differences of the transcriptional changes of the chromatin structures. The association of the repressor proteins will induce transcriptional repression, while the activator proteins will cause the transcriptional expression to happen. After the histone methylation, the chromatin structure will cause a deformation and leads either to a condensed or decondensed state. The condensation of the normal form of chromatin into a compact heterochromatin structure will cause the inaccessibility of the transcriptional machineries which leads to the gene silencing [56]. However, gene activation may occur when the heterochromatin reverts into chromatin [57].

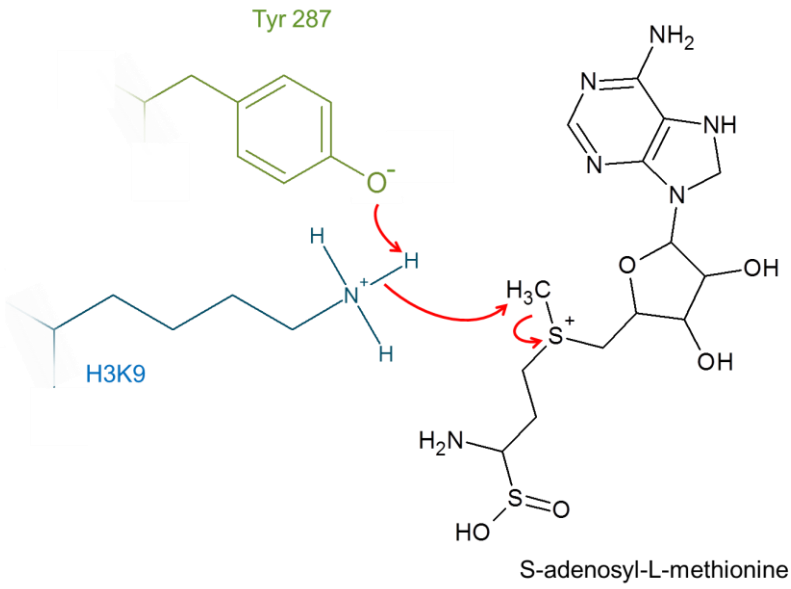


Figure 3 G9a catalytic mechanism. The schematic diagram shows the methyl transfer reaction from the SAM molecule to the lysine side chain of histone 3. The red arrows have shown the electron transfer from the lysine tail to the sulfur atom at the SAM molecule**.**

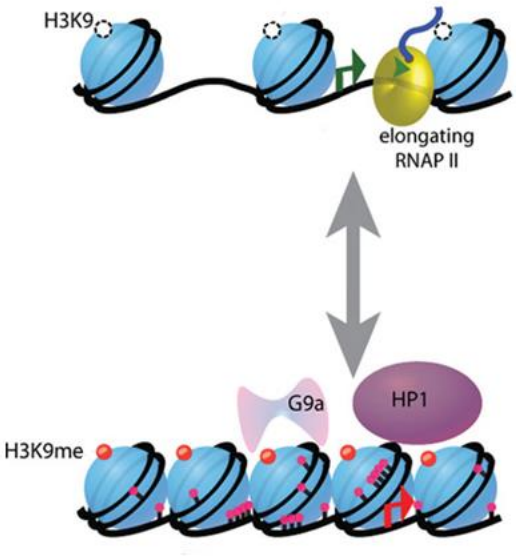


Figure 4 Schematic diagram of G9a methylation. Histone methylation by G9a on the histone H3 (H3K9 me) will cause a reversible effect between the normal chromatin packing (shown on the top) and heterochromatin (shown on the bottom). Transcription repression and activation of the genes contained within this chromatin region is due to the regulatory proteins that bind the lysine tail of histone during histone methylation. The figure shows the binding of HP1 protein on the binding site which disrupts the transcription expression resulting in gene silencing. The schematic diagram is adapted from the Sha K, *et al.* 2009.

1.3 The role and relation of G9a and human health

G9a has an essential role for the epigenetic regulation of gene expression in euchromatin. G9a is usually identified as a corepressor that is linked to gene silencing because of the methylation of Lys 9 on the histone 3 that marks transcriptional repression in genes. However, it also acts as a coactivator depending on the location of the chromosome [58]. As a coactivator, G9a will recruit the coactivator-associated proteins that help in the activation of transcription genes by sending out an activation signal to the transcriptional machinery. On the other side, G9a can also act as a corepressor by associating with transcriptional repressor proteins that repress the transcription signal. In short, G9a is functioning as a gatekeeper that turns the genes "on" and "off". In order to have a good understanding of the G9a, a lot of research has been carried out to elucidate the functional roles of the G9a in the cancer, immune system, drug design/drug discovery, embryogenesis, neuron studies, stem cell and cell development. Often, mice samples are used for G9a studies as it possesses a highly similar biological system compared to the one of humans.

1.3.1 G9a and embryogenesis

In embryogenesis studies, G9a was shown to be an important enzyme in the regulation of the cell differentiation in embryonic stem cells (ESC) [59]. In G9a knockout cells, the absence of DNA methylation and the binding of HP1 to euchromatin causes the loss of silencing on the homeobox gene Oct ¾ of ESC which leads to an early embryonic lethality [59] and growth defects [60, 61]. From studies performed on germ-lineage knockout mice, it was observed that G9a has an important role during the regulation of the gametogenesis [67]. In the G9a-deficient cells, the overexpression of the G9a regulated genes disrupted the germ cell development and the meiotic pathway, causing the loss of adult gamete cells in either sex.

1.3.2 G9a and gametogenesis

G9a plays a key role in the regulation of gene expression in the imprinted genes [62-64]. Generally, a fertilized egg will inherit two copies of single gene from their parents. However, in genomic imprinting, one copy is active (imprinted gene) and the other copy will remain silence due to the occurrence of the epigenetic methylation in the histone. Improper imprinting will lead to abnormalities in cell development and division, leading to cancer. The methylation of H3K9 was reported to mark the imprinting control region of the gene and result in the allelic repression of trophoblast genes [62]. However, the understanding about the function of G9a in the imprinted genes is still very limited. The methylation of H3K9 is involved in the silencing of the X chromosome in early development of females [64].

1.3.3 G9a and immunobiology

From an immunological perspective, G9a acts as a coactivator for nuclear hormone receptors [65], while in some cases it can play a role as a corepressor as well [66]. *Lee, et al*. proposed that the promoter context and the combination of different proteins decides the G9a activity as a coactivator or corepressor [65]. Besides, *Thomas, et al*. reported that B-cell-specific G9a knockout mice have shown defects in the differentiation of the mature B cells into plasma cells [67]. These finding claims that that the methylation of H3K9me2 will affect the lymphocyte development and its activation. Besides, the combination of the G9a with the chromatin-modifying enzymes and has been shown to affect DNA methylation which results in transcriptional gene repression [63, 68].

1.3.4 G9a and neurology

Research of the phenotypic behavioral changes of neuron-specific G9a mutant mice has been carried out to investigate the role of G9a on a neurological level [69]. The GLP/G9a histone methyl transferase complex is able to control the cognition and adaptive responses in brain neuron. The neuronal G9a/GLP deficient mice showed de-repression of many non-neuronal and neuron progenitor genes in adult neurons [69] resulting in severe behavioral abnormalities [70, 71]. Similar findings were reported in studies of EHMT/G9a-deficient flies [72, 73]. These findings highlighted the importance of the GLP/G9a in the human mental retardation studies as the transcriptional disruption of neuron-specific deficiency of GLP/G9a is similar with the human 9q34 symptoms from patients [70, 72, 74].

1.3.5 G9a and cancer

G9a acts as a coactivator for cancer cell growth by silencing of tumor-suppressor genes, such as p53. The overexpression of G9a contributes to various kinds of human cancers, such as liver cancer [75], leukemia [76], prostate cancer [76] , lung cancer [77], pancreatic cancer [78] and so on. Furthermore, G9a knockdown can also inhibit the proliferation of cancer cells [9]. The fact that G9a is a potential coactivator was also indicated by Vandel and Trouche in the research of the interaction of the CREB-binding protein (CRB)/p300, the histone acetyltransferase with H3K9 in the reported by Vandel and Trouche [79].

Previous examples indicate that G9a is a coactivator and corepressor for several biological mechanisms in human bodies [58]. These findings highlight the importance of the molecular mechanism of G9a on gene transcription gene regulation and thus it make G9a an interesting research target, especially for drug discovery and design.

1.4 G9a inhibitors

1.4.1 Why do we need to develop new inhibitors?

In our research, we are focusing on the design of novel inhibitors of the G9a protein or the G9a-like protein (GLP) by targeting the cofactor site and protein substrate binding site using computational methods, often referred to as *in silic*o screening. While a number of inhibitors have already been reported, either accurate mechanism of action remains unknown, or the compounds suffer from poor properties for further development. In order to develop better probes to study the G9a protein or to increase the therapeutic potential of G9a inhibitors, we aim to design novel inhibitors belonging to different chemotypes and/or different mechanisms of action compared to those previously reported. In order to identify these inhibitors we relied on existing virtual screening methods incorporating pharmacophore screening [80] and molecular docking simulation [81].

1.4.2 What inhibitors are known?

The majority of reported inhibitors target the G9a protein on the histone tail binding site, also known as the substrate site and therefore these compounds are also known as substrate-competitive inhibitors. Although different inhibitors are reported, they all belong to the same chemotype of epidithioiketopiperazines (ETP). ETP is a toxic secondary metabolites produced specifically by fungi [82]. BIX01294 (diazepin-quinazolin-amine derivative) [83] was the first identified inhibitor for the protein substrate site of G9a and GLP by *Kubicek et al.* [84] It impairs the G9a methyltransferase by decreasing the dimethylation of the H3K9 in ES Lines [85]. Besides, BIX01294 has proven to break the protein-protein interaction [86]. The potency of the BIX01294 on the G9a with IC_{50} values of 1.9 μM and the GLP with the IC50 values of 0.7 μM [86]. Compounds with improved drug potency were further reported with a potency of IC_{50} of 250 nM and IC_{50} 27 nM in respect to the G9a and GLP respectively [87]. UNC0638 [88] exhibits a higher potency, selectivity and less toxicity in inhibiting both the G9a and GLP in comparing to the BIX01294. The UNC0638 resulted in the decline of H3K9me2 level in the mouse ESC. It shows potency of $IC_{50} < 15$ nM ($n = 4$) in G9a and $IC_{50} = 19 \pm 1$ nM ($n = 2$) for GLP by SAHH-coupled assays [89]. UNC0224 is another inhibitor that is an analogue of BIX01294. It also possesses high potency and selectivity for G9a. Its potency (IC_{50}) for G9a or GLP are respectively 15nM and 20nM in a Thioglo assay and 289nM and 58nM in an AlphaScreen assay [90]. UNC0321 is the result of rational design experiment investigating the structure activity relationship (SAR) of ETP derivatives and analyzing the crystal structure of BIX01294 and UNC0224. It has a very low picomolar

potency of (Morrison K_i = 63 pM) in *in vitro* studies [91] , however it suffers from the drawback of bad cell membrane permeability[89]. UNC0646 [92] and UNC0631 [93] inhibitors were developed to improve the cellular permeability and lipophilicity and they are having a IC_{50} <0.6nM in G9a while the IC_{50} < 0.015 µM for GLP in the *in vitro* studies [94].

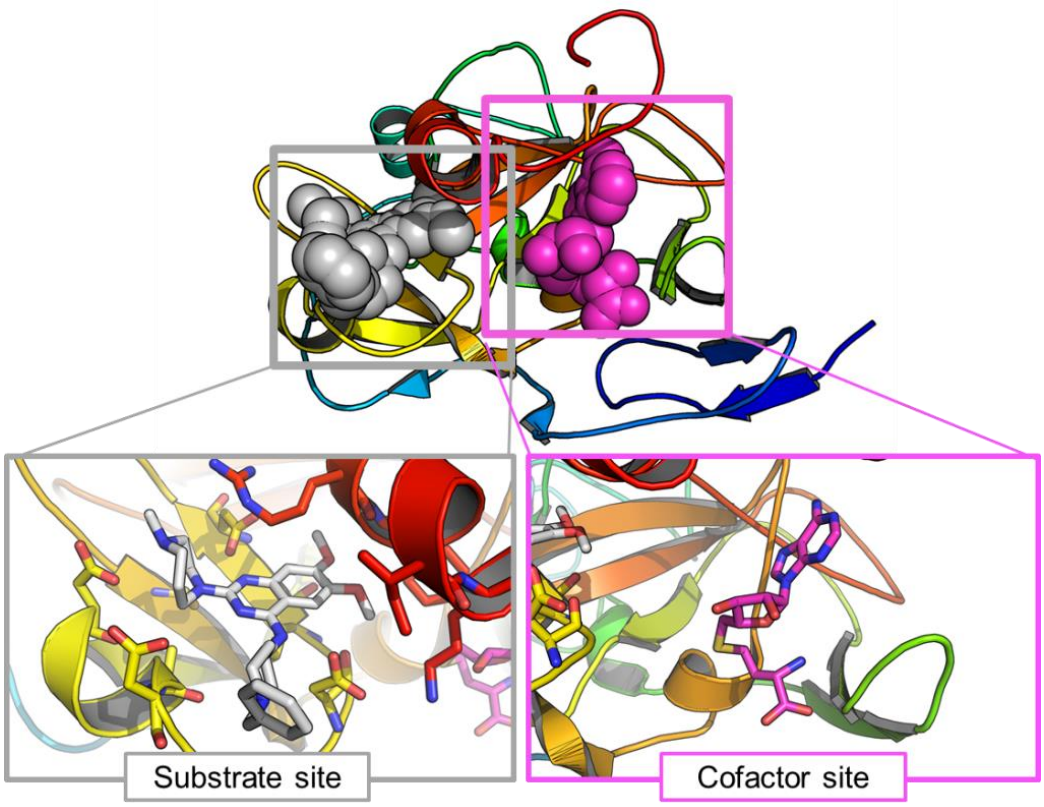


Figure 5 Binding sites of G9a inhibitors. The structure of G9a is depicted as a cartoon and rainbow colored from N to C-terminus. The left box zooms in on the substrate where the ETP class of inhibitors bind, competing with the histone lysine. In this figure BIX01294 is depicted as sticks. In the magenta box on the right side, the SAM cofactor molecule is depicted.

The first SAM cofactor-competitive inhibitor for G9a was believed to be Chaetocin. Chaetocin originates from a natural product from the *Chaetomium* fungi species. Chaetocin exhibits apoptotic effects on tumor cells [95]. Chaetocin was reported to inhibit the G9a at the IC_{50} of 1.1 µm in *in vitro* studies and mouse G9a with IC_{50} = 2.5 µM [96]. Chaetocin is structurally very different from ETP class of HKMT inhibitors in protein binding site of G9a. While it was first thought to be a competitive inhibitors at the SAM cofactor site in the SET domain, recent report opposed this claim [97]. Thus, the exact mechanism of action in G9a still remains unknown [97].

BIX-01338 [98] is another G9a inhibitor. It has a potency of IC_{50} of 4.7 µM in G9a is reported by Cherblanc *et al.* [99]. BIX-01338 is thought to be an aspecific SAM-competitive inhibitor and thus inhibits a variety of histone methyltransferase. The exact mechanism of action in G9a however remains unknown in the cellular and cancer studies [85, 100].

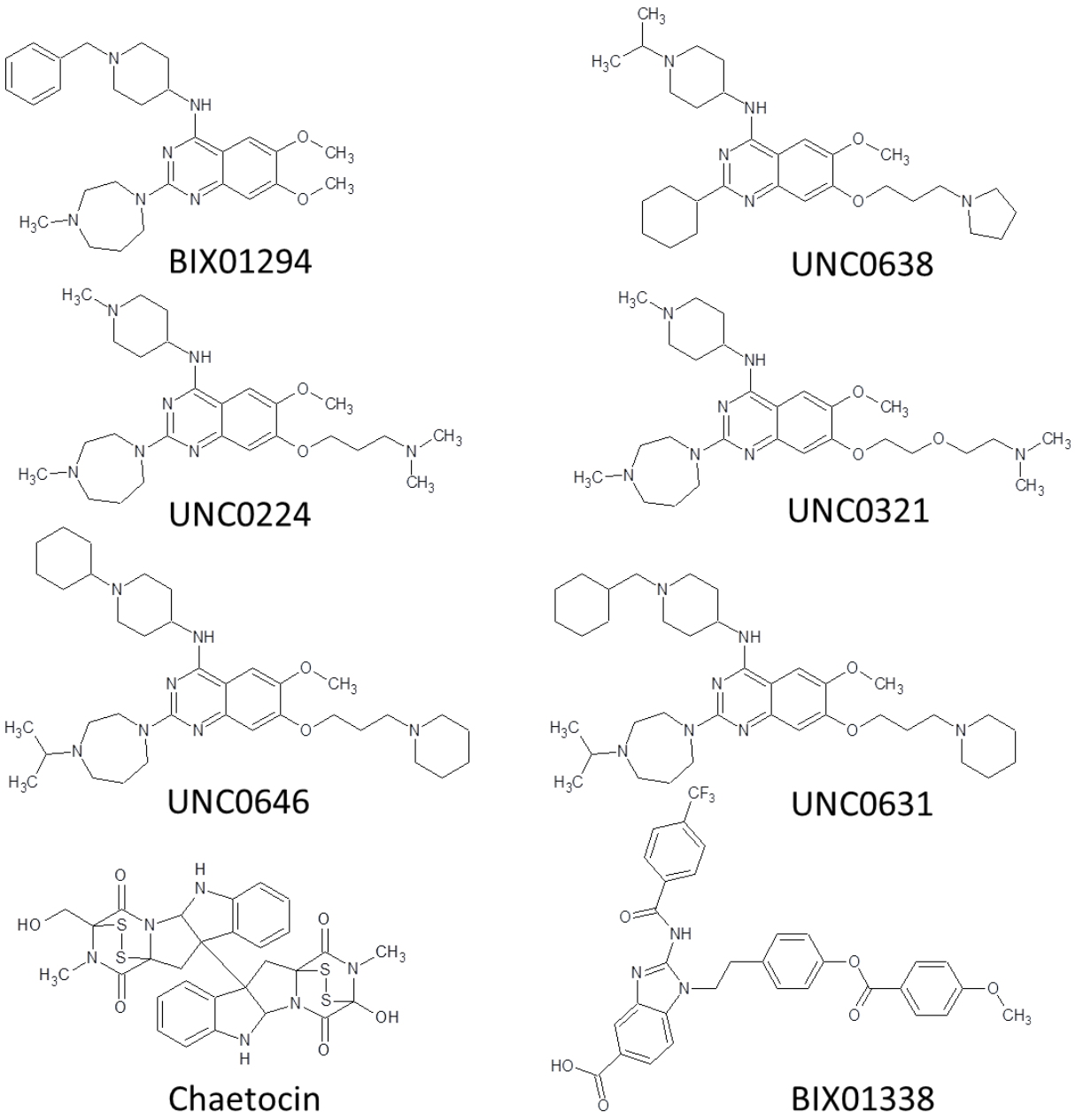


Figure 6 G9a known inhibitor structures. The known G9a inhibitors targeting the substrate site are shown in the top 3 rows. All of them belong to the ETP class of inhibitors. The bottom row G9a inhibitors are thought to have an SAM competitive inhibitory profile but exact mechanism of action remains unclear.

Chapter 2

Computer-aided Drug Design (CADD)

2.1 Introduction

Drug design is the expensive and laborious process of developing new medicine. This process finds its origin in the herbal remedies [101]. Only since the last century drugs have a (semi) synthetic origin [102]. The first hit-compounds often lack the potency and safety to be used and thus optimization is required. While historically optimization was a brute force approach by trial-and-error [103, 104], soon rational strategies towards potency optimization [105, 106] came into place. Since 1980's the computer has become a more prominent and an inevitable tool in the drug discovery process [107]. The cross-over between the computational and the pharmaceutical research field is typically designated as Computer Aided Drug Design (CADD) [108, 109].

CADD covers a broad range of applications spanning the drug discovery pipeline, although highly concentrated at the early phases (see Figure 7). The typical drug design and drug discovery research process, from the concept to the market takes approximately 13 years and cost almost $1 billion [110]. The whole drug discovery process is expensive and time consuming. Thus, CADD can speed up and rationalize the drug design process while reducing the costs [111]. At the earliest phase in drug discovery, the aim is to identify the first hit compounds. The *in silico* counter part of the *in vitro* High Throughput Screening (HTS) is referred to as Virtual Screening (VS) and aims at filtering libraries of molecules using computational methods thereby prioritizing compounds most likely to be active for a given target [112]. Later-on in the drug discovery pipeline, the potency of the hit and lead compounds needs to be improved [113]. New d erivatives are designed whether with or without a different scaffold at the core of the molecule [114]. The ultimate goal is to design highly potent and specific molecules which also have a suitable IP position [115]. This can be achieved by classical medicinal chemistry approaches, where rational design can be based on the observed SAR or rational design based structural information [116]. Computational

method however can also be used to virtually create derivatives, hop chemical scaffolds [117, 118], and then score them for improved potency thereby prioritizing the most promising derivatives [119, 120]. However not only the potency of the compounds are of importance, also their pharmacokinetic behavior: Absorption, Distribution, Metabolization, Excretion and Toxicity which is often referred to as ADMET [121-123]. Next to the battery of *in vitro* and *in vivo* ADMET experiments, virtual experiments have also been developed to profile drug-like compounds early during the development process.

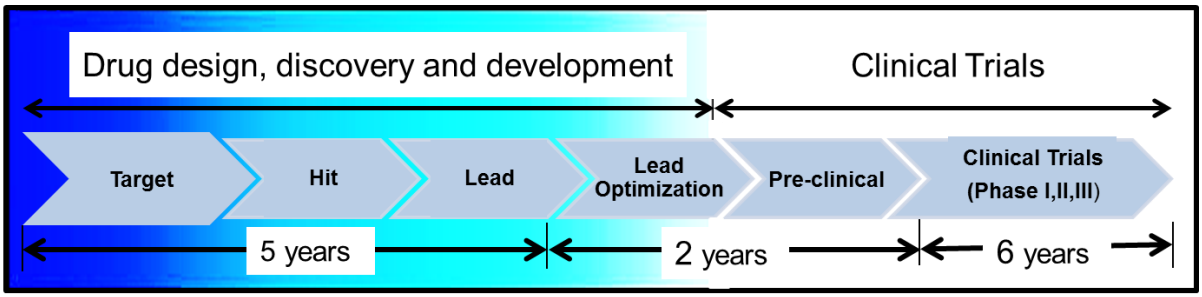


Figure 7 Drug discovery pipeline. The schematic diagram indicates the workflow of drug discovery to the final clinical validation state before the drug is approved by FDA and release to market. The first step of drug discovery is started with the identification and validation of biological target. This is followed by the hit and lead discovery. The lead compounds are optimized for the potency and check for the ability of their clinical efficacy by bioassay. The final step is clinical trials that comprise three phases: Phase I, Phase II and Phase III, with animal and human testing before the drug is send for approval and apply in the market. Besides, the blue gradient indicates the importance of CADD during drug design.

2.2 Chemoinformatics

2.2.1 What is chemoinformatics?

The basis to all CADD methods is chemo-informatics. Chemoinformatics is the technique which exploits the computational informatics in the analysis of the chemistry data in the drug discovery field. Chemoinformatics is also useful in handling the problem of data explosion as it can systematically store, handle, and retrieve the chemical structures and chemical property data [124, 125]. The word Chemoinformatics is often used to describe computational methods for the analysis of potential drug compounds by analyzing the chemical space and the chemical properties which are related to "drug-likeness" and ADMET properties. To others, chemoinformatics in drug discovery is related to "molecular similarity searching", where molecular descriptor and molecular fingerprints are calculated for comparing chemical libraries [126].

In the next two sections, two chemoinformatics subfields which are important for the research performed during my master thesis will be discussed.

2.2.2 Conformation generation

In CADD, 3D structures of small molecules are crucial for rational drug design [110]. From a (2D) chemical structure, the 3D structural information of a molecule can be calculated and analyzed. Typically, the experimental 3D structure of the biological target or small molecules are normally obtained from X-ray crystallography or NMR experiments [127]. However, for the majority of molecules structural information is not always readily available since they are hard to obtain. Hence, the computational approaches to generate the 3D structure of the protein have broadly applied in drug design. Compounds however often can be stable in more than one conformation and thus it is essential to be able to sample many different stable biological relevant conformations. In our research, we have used Omega v2.4.6 [128] with default parameters to construct the 3D conformations of the molecules from DUD dataset.

2.2.3 Similarity searching

The concept that molecules with a similar structure tend to possess similar activity was introduced by Johnson and Maggiora in 1990 [129]. Similarity searching is the discipline where molecules with a similar structure are sought within the large chemical database and it is generally applied in the early phase of lead discovery. A single bioactive molecule is sufficient to begin a search [130]. For similarity searching, molecules are commonly represented using molecular fingerprints. Molecular fingerprints can be classified as 2-dimensional or 3-dimensional fingerprints, and are a vector of elements describing the presence of substructures or chemical properties and composition. The 2D fingerprint is an essential tool in ligand-based virtual screening and it performs better in scaffold hopping when applied to a diverse database [131]. Often, 2D fingerprints are the best for obtaining the active compounds with the better enrichment factor with a high database rank [132]. The 2D fingerprints which are commonly used include Daylight [133], Dragon [134], MACCS [135], TOPOISM [136] and BCI [137].

The most popular similarity metric to calculate the similarity of 2D fingerprint is the Tanimoto coefficient (T_c) [138].

$$\boldsymbol{T_c = \frac{c}{a+b-c}}$$

Where, ***a*** and ***b*** are the number of individual elements of a molecular fingerprint for molecules A and B respectively, and ***c*** of the number of common elements in both fingerprints. The Tanimoto coefficient also known as Jaccard coefficient is used to determine the number of features between both molecules a and b and compared to the number of features that are union, with the result of 1 indicates a perfect matche and 0 the absence of matches.

3D fingerprint are usually applied as a complement to the 2D fingerprint. In 3D fingerprint, the conformation of the molecules is taken into the consideration during similarity searching [139].

The conformation of molecules is important to determine the active molecules that highly correlate with the experimentally defined ligand conformation. The commonly used 3D fingerprints are 3D pharmacophore fingerprints (see section 2.6.3) and 3D shape similarity. Contrary to fingerprint based methods, 3D shape similarity measure the overlap region of the atomic volumes during the alignment of the query to the 3D conformer of the molecule [140]. The shape similarity is able to find the compounds that have the similar shape and chemical properties base on the information of the atom types. The commonly used 3D shape similarity programs are ROCS [141], and Phase-shape [31].

2.3 Virtual screening

As the size of the chemical libraries has drastically increased over the last few years, it causes the active compound screening process to be more costly and time consuming. To solve this problem, these libraries can be enriched for compounds using *in silico* methods. Virtual screening (VS) is a widely used preselecting method [142] for compounds screening. VS is used to identify the compounds that are most likely to be active for a given target from commercial small molecule databases, by eliminating the compounds that are unlikely to be active.

A VS experiment is typically carried out subjectively by the researcher, according to his preference and experiences. They can either use one VS tool, or combine multiple VS tools. The most commonly used virtual screening tools can be classified in different categories,

such as molecular docking [143] , pharmacophore screening [144], and similarity searches [145]. Except for these VS methods, often simple molecular descriptor based filters are used at the start of the virtual screening process to remove the undesired compounds.

Virtual screening can be treated as a filter tool to remove the inactive compounds. Often multiple VS methods are combined. In this case, the VS process is often depicted as a funnel. In the funnel, the large size at the top indicates the many compounds which are used to start the VS experiment, and during the VS process, the compounds unlikely to be active are removed at every step. As such the library size decreases step by step, similar to the shape of a funnel. Typically the computationally least demanding steps are placed on top of the funnel with the more demanding steps to the bottom.

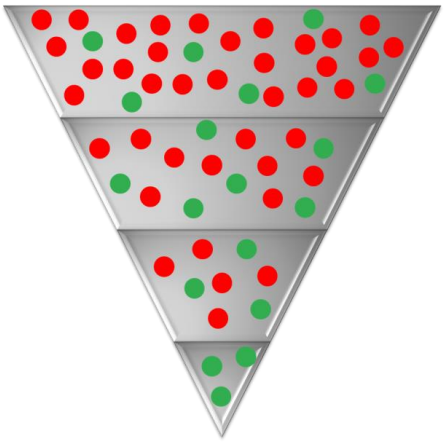


Figure 8 Funnel-based virtual screening. The funnel shows the virtual screening process that discarding at every step the molecules that are unlikely to be active. Across the funnel from the top to the bottom shows the compounds that have been selected each step for further analysis in the next step. The bottom layer shows the finals selected molecules after the experimental testing.

2.4 Docking

2.4.1 What is docking?

Docking is the computational method to predict the binding conformation and affinity of one molecule when bound to the other molecule. Docking was designed to predict the conformations of the ligand that bind to the protein or DNA in its active site. The binding orientation of each ligand to the receptor is calculated mathematically by using the binding energy score (molecular interaction). The first docking program, DOCK was developed by Kuntz, *et al.* in 1982 in predicting the binding orientation of two molecules [146]. Since the development of DOCK, many other tools have been reported including AutoDock [147], GOLD [148], Glide [149], FRED [150], FlexX [151] eHiTS [152] and Surflex [153]. The

differences of these docking tools can be found in the different search algorithms which can be used to predict the binding pose of the compounds as well as the scoring functions to measure the quality of the binding pose as well as to predict the affinity of the compound for the protein target.

2.4.2 Search algorithms

The currently used search algorithms include flexibility of the ligand by generating different position, conformation, position and orientation of ligand to bind in the active site. Less frequent, search algorithms can also include flexibility of the receptor binding site. The search algorithms will try to calculate and analyze all the possible binding poses of the protein and ligand complex. In general, the search algorithms can be divided into two major groups, the deterministic algorithm and the stochastic algorithm. The deterministic algorithm is reproducible, while the stochastic algorithm is designed to randomly pick the molecules to process and it is not reproducible. The most commonly used search algorithms are the genetic algorithm (GOLD [147] and AutoDock [147]), fragment-based incremental construction methods (FlexX [151] and DOCK [154]), Monte Carlo simulated annealing (QXP(Flo)[155], LigandFit [156], ICM [157] and MCDOCK [158]) , systematic searches (FRED [150] and Glide [149]). These search algorithms are different approaches that are used to find the better solution for the binding of the ligand to the protein target.

2.4.2.1 Genetic Algorithm

The Genetic Algorithm (GA) was introduced by John Holland, University of Michigan, United States in the early 1970's. The GA is a heuristic searcher that can mimic the evolution process and genetic changes of the chromosome, with the changes in fitness of an individual [159]. The GA uses genetic mutations to a population which can enhance the whole performance of the search. GA was developed as a machine leaning method with the concept of representing all the variables as genetic information stored in a chromosome, which represents one individual in a population. The population consists of many individuals that all carry their own individual genetic information (variables) that represents a possible solution, evaluated by a fitness function. Following the rules of evolution (mutation, reproduction and crossover), the population will evolve towards individuals with the best fitness, corresponding to genes with the best solution.

In docking, the variables that are encoded by the genes represent the variables of the docking solution: the X, Y, Z coordinates of the central atom of the ligand as well as the rotational

angles to determine the orientation with respect to the receptor structure, the torsion angle of the rotatable bonds, and the rotamers of the amino acid of the receptor protein (if flexible side chains are allowed) and the orientation of bounded water molecules (if these are taken explicitly into account). The value of these genes corresponds to a docking solution which is evaluated by a scoring function.

The genetic algorithm starts with creating a starting population in a random manner. Then the population is scored and individuals with a good fitness are allowed to reproduce. Offspring is created according to crossing over, and point mutations can be introduced by searching the optimum value of a gene. Next, the new population is evaluated again: bad individuals will die, thereby eliminating the bad solutions. Individuals with a top fitness score however will be allowed to reproduce again. This process continues until all individuals have similar genes (the solution has converged) or another end criteria is met (typically a maximum number of generations is defined).

In the end the solution corresponding to the genes with a top fitness are considered the correct solution for this search. As this process starts with a random population with random gene values, it is a stochastic process and multiple runs will result in multiple results. Often however these are very similar.

GOLD is one of the docking simulation software that is based on the genetic algorithm. (see section 2.7.4).

2.4.3 Scoring functions

A docking scoring function is used to assess the quality of the predicted binding poses of the ligands that binds to the binding site. The second function of scoring function is to calculate the free binding energy of the ligand. There are three general groups of scoring function that are normally applied in the docking simulation.

2.4.3.1 Force field scoring function

In this case a force field is used to calculate the binding affinities between the interactions of two binding molecules. The force field scoring function usually include only the non-bonded interaction terms of a force field: the van der Waals interactions, the electrostatic interactions, and sometimes the hydrogen bond interaction of the putative complex. Force field scoring function were first implemented in the early version of DOCK docking software.

The force field scoring function that calculates the molecular mechanics interaction energies is represented by the following equation:

$$E = \sum_{i=1}^{lig} \sum_{j=1}^{rec} \left(\frac{A_{ij}}{r_{ij}^{12}} \right) - \left(\frac{B_{ij}}{r_{ij}^{6}} \right) + 332 \frac{q_i q_j}{D r_{ij}}$$

Where, Aij and Bij are the van der Waals parameters to check the intermolecular energy of the interaction between two molecules. And, qi and qj are the atomic charges, rij describes the distance between the atomic charges of the protein atom, i and ligand atom, j. While, Drij is a dielectric constant in Coulomb energy. The 332 value is used to change the energy into kcal/mol.

2.4.3.2 Empirical scoring function

The empirical scoring function is a scoring function that sums up the energy terms, which are determined by empirically derived parameters, to predict the binding affinities of two molecules [160]. Empirical scoring functions usually take the form as shown below:

$$\begin{aligned} \Delta G = {} & \Delta G_0 + \Delta G_0 \times N_{rot} \\ & + \Delta G_{hb} \sum_{hb} f\,(\Delta R, \Delta \alpha) \\ & + \Delta G_{ion} \sum_{ion} f\,(\Delta R, \Delta \alpha) \\ & + \Delta G_{aro} \sum_{iaro} f\,(\Delta R, \Delta \alpha) \\ & + \Delta G_{lip} \sum_{lip} f^{*}\,(\Delta R) \end{aligned}$$

Where, ΔG indicates the stable free energy which combines different energy terms. ΔG_{hb} describes the hydrogen bond energy, ΔG_{ion} represent the salt bridges, $\Delta G_{aro}$is the aromatic energy and ΔG_{lip} is the lipophilic energy which approximate all the pairwise atom contacts. *f is* a penalty function for the ideal geometry by considering the deviant from the ideal value ΔR (distance) while Δα (angle) in an interaction. For lipophilicity, the penalization function

f * is calculated differently. Empirical scoring function are a simple and fast method and are the most commonly used method to estimate the free energy of binding for the protein-ligand putative complex with the known 3D structure [161]. Furthermore, empirical scoring functions are generally applied in some existing docking software, such as PLP [162-164], Glide [149], ICM [157], ChemScore [161], FlexX [151], Surflex [153] , MedusaScore [165] ,SFCscore [166] and LUDI [167].

In our studies, we have employed the Piecewise Linear Potential Scoring Function (PLP) [164] as a scoring function for docking. PLP is a simple scoring function that can accurately predict the intermolecular interaction energy of protein and ligand complex [162, 168].

2.4.3.3 Knowledge-based scoring function

Knowledge-based scoring functions are derived using statistical method from the structural information of the crystallographically determined structures and converts the statistical information into energy terms for protein-ligand atom pairs [169, 170]. The Knowledge-based scoring function is based on the statistics of occurrence (the histogram of the interatomic distances) of a set of atom pairs by using the inverse Boltzmann distribution and it is often used to rank the protein and ligand problem [171-173]. Knowledge-based scoring function is a simple scoring function and can quickly screening through a large size database of compounds [169]. A functional form of the knowledge-based scoring function is shown as following:

$$W(r) = -k_B T ln\,[g(r)], g(r) = p(r)/p^*\,(r)$$

Where, $W(r)$ is the ligand interaction free binding energy. And, $-k_B$ is the Boltzmann constant. T indicates the absolute temperature. r depicts the distance of the protein and ligand atom pairs. $g(r)$ is the normalized mean radial pair distribution for a distance r. $p(r)$ is the atom pairs distribution in the distance r. $p^*\,(r)$ is the reference state of 0 where the interaction of the radial atom pairs distribution occur. Moreover, knowledge-based scoring function is a robust approach that can accurately predicts the protein structure and the binding geometry of the ligand to the protein [174, 175]. Knowledge-based scoring functions have been used in the following docking software: GOLD [176], DrugScore [177, 178], ITScore [179], PMF [180], BLEEP [181], KScore [182], DFIRE [183] and SMoG [184].

2.5 Docking post filtering

Molecular docking simulations are without any doubt the most commonly applied VS method (if a receptor structure is available). However, there are some drawbacks for using docking simulations. For example, docking cannot accurately rank the molecules by their binding affinities. This will affect the docking result as binding affinity is relatively important for ligands that bind tightly to the active site. Moreover, docking also has been reported to have failed in predicting the right binding pose with the right ligands into the active site. Combining these limitations in the analysis of hundreds to millions of compounds, many of them likely to be inactive, it is logical that is a very difficult task to identify the active compounds at the top ranking, and remove the false positives.

To compensate the limitations of docking simulations, post-filtering the docking results are used in order to obtain a better ranking result. There are many post-filtering methods and programs are proposed, such as Cartesian Genetic Programming [185], SIFT (Structural Interaction Fingerprints) or PLIF (Protein Ligand Interaction Fingerprints) [186], ChemBioServer [187] and so on.

Another frequently used method is to post filter the docking simulations using a pharmacophore model. The pharmacophore model is designed based on the non-bonded interactions of the ligand to the receptor. This model will eliminate the docked molecules that do not fulfill the required interaction with the binding receptor. The combination of docking simulations and pharmacophore post filtering have shown to improve the performance of virtual screening [188]. The docking simulations take cares of analyzing the possible poses of the molecules that bind to active site; while the pharmacophore model helps in filtering the docked poses and ensures that only the molecules with the correct binding pose are retained.

In my research, I have also investigated the possibility of exploiting the electrostatic complementarity between the receptor protein and the docked solution as a docking post-filter. For the background and results, I refer to Chapter 3.

2.6 Pharmacophore modelling

2.6.1 Introduction

The first person to use the term pharmacophore was Lemont B. Kier, during the early 1970's. He was the first person that identified the pharmacophore for the muscarinic receptor by analyzing the common elements in the small molecule muscarinic receptor modulators (which he defines as 'muscarinic pharmacophore').

In 1977, Peter Gund defined the pharmacophore as *"a set of structural features in a molecule that is recognized at a receptor site and its responsible for the molecule's biological activity"* [189]. The formal definition of the pharmacophore was later launched by IUPAC as "*A pharmacophore is the ensemble steric and electric features necessary to ensure the optimal supramolecular interactions with a specific biological target and to trigger (or block) its biological response*" [190]. A pharmacophore is a simple concept that considers the molecular interaction between the target receptor with the chemical molecules nearby. Pharmacophores have been widely used in CADD for molecular recognition in predicting the biological activities. In computational chemistry, a pharmacophore is presented as a three-dimensional feature based model with specific location and geometric constraint by taking account the distances between the pharmacophore features. Alternatively, all the pharmacophore features of a small molecule can be combined in groups of 3 or 4 and can be used to construct a pharmacophore fingerprint.

2.6.2 Pharmacophore model

A pharmacophore model, also known as a pharmacophore query when it is used in virtual screening, is a collection of pharmacophore features. Pharmacophore features describe the key elements of molecular recognition that are required to recognize and bind to the target receptor. The pharmacophore features are usually represented as spheres, with the radius determining the tolerable deviation of the ideal position. The features typically represent hydrogen bond donor, hydrogen bond acceptor, hydrophobic, aromatic, anionic and cationic and any logical combination. A pharmacophore model is represented in a 3D form and allows clear visualization of the key elements of recognition between the ligand and the receptor.

The graphical representation of the pharmacophore features is shown in the Figure 9. The inclusion of these bonded interaction of the ligand-protein, has made pharmacophore query a useful molecular recognition tool to investigate the protein–ligand complex.

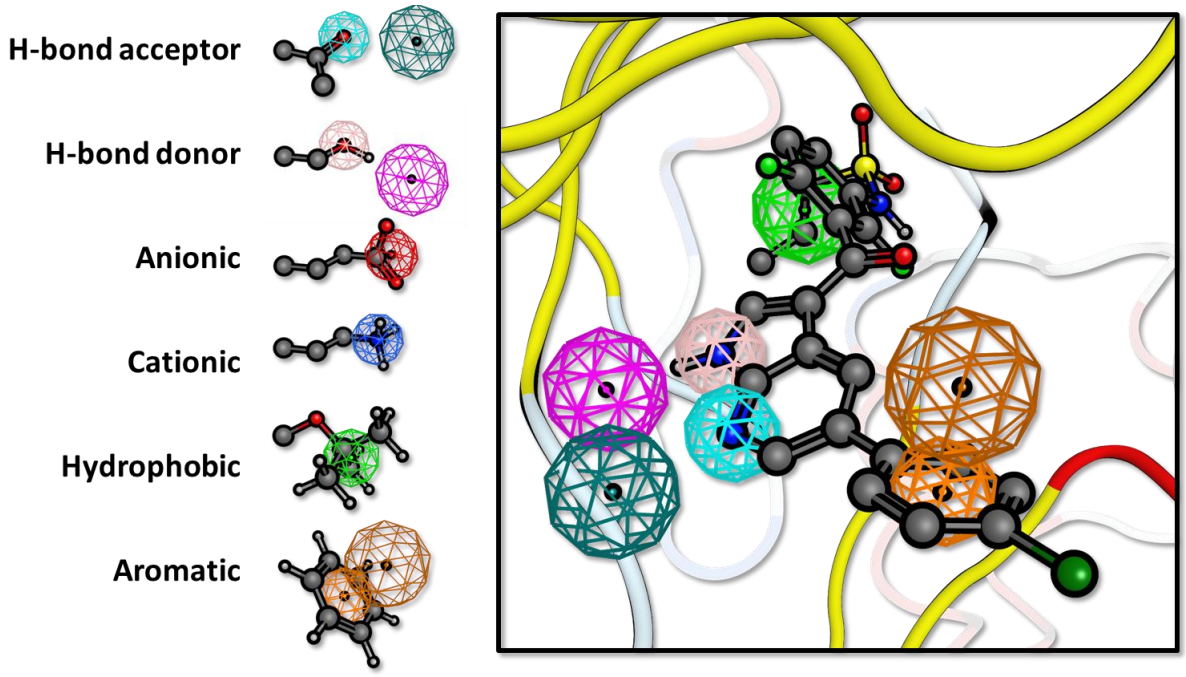


Figure 9 Pharmacophore query. A Pharmacophore query is comprised of different features. The features represent molecular recognition motifs such as hydrogen bond acceptors or donors, anionic, cationic, hydrophobic and aromatic groups. The radius of the sphere determines the strictness of the geometric constraint. For features where the correct orientation of the interaction is important such as hydrogen bonds and the aromatic plane, a second feature can be used indicating the vector of the interaction (or the normal of the plane). A pharmacophore query can combine any of these features, with different radius and logic operations such as "AND", "OR" and "NOT". On the left a hypothetical pharmacophore query for BRAF kinase is given.

A pharmacophore model can be applied for various purposes such as analyzing the SAR of a class of molecules and the rational design of novel compounds. The most common application of a pharmacophore model however is as virtual screening query to identify novel hit compounds, or new derivatives. In this case, a small molecule ligand database should contain all low energy conformations of the small molecule, which are attempted to fit into the query. The hits are molecules that fulfill the query and possess the correct chemical functionalities in the correct chemical organization to be able to recognize and bind to the receptor protein.

2.6.3 Construction of the pharmacophore model

The approach to construct a pharmacophore model is dependent on the prior knowledge of the drug discovery process. There are four different possibilities (see Figure 10):

1) Absence of both the ligand and protein information

There is no option for using pharmacophore modelling for VS. At best a pharmacophore fingerprint can be used to create a diverse set for HTS.

2) Active ligands are known, but no structural information of the receptor is available

The active ligands can be aligned according to the common molecular recognition properties. These common properties can be assigned to a feature and the radius is chosen such that the atoms of the different active ligands are included. Inactive ligands can be used to validate the pharmacophore features or to assign forbidden features.

3) Active ligands as well as structural information is available

Preferably the structural information is crystallographical, but molecular modelling can also be used to dock the active ligands into the receptor. The alignment of the active molecules can be accurately performed by aligning the protein structures. The common interactions with the receptors can be used to create the pharmacophore query.

4) Only the receptor structure is available

In this case, different molecular modelling methods are possible to analyze the binding site. The identification of the possible interactions in the binding site can be carried out by using receptor pocket analysis algorithms, such as GRID and FTMAP [191]. The key interactions are projected into the binding activity and can be assigned to pharmacophore features that represent the most likely molecular recognition motif for this given area within the pocket.

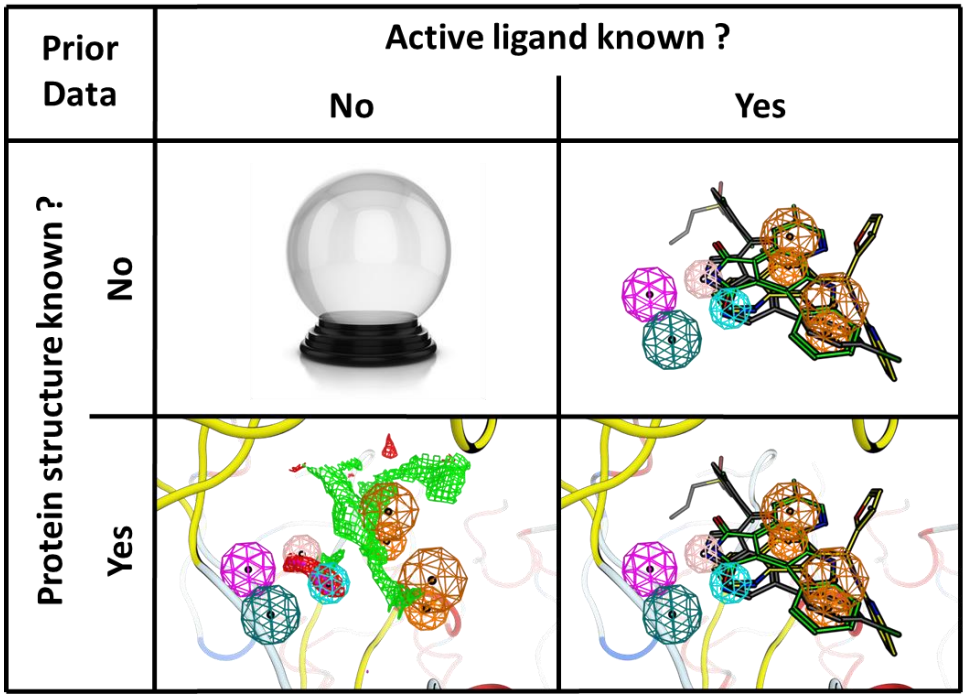


Figure 10 Four different situations for the pharmacophore search. The figure shows the four different situations that may be encountered when starting a VS. The situations include the absence of both the ligand and protein structure information, where experimental screening is the only option. The second option is the presence of active ligands, but the protein structure is unknown, where pharmacophores can be used for ligand based virtual screening. The best situation is when binding ligand and structural information is present. The most challenging option is when only a protein structure is available.

2.7 Software used during research

2.7.1 PyMOL

PyMOL is an open source molecular visualization tool that is distributed by Schrodinger. It was first developed by Warren L Delano as a freely available molecular visualization tool for drug design. PyMOL provides a graphical interface system for the visualization of the molecules. PyMOL also provides the setting to produce a high quality representation of the molecules which is suitable for publication by the ray tracing engine. The image output is generated in the Portable Network Graphics (PNG) format. PyMOL is freely available and it can be downloaded from http://www.pymol.org/. PyMOL is operational on Windows, Linux, MAC OSX and GNU/UNIX-based systems.

2.7.2 Molecular Operating Environment (MOE)

The Molecular Operating Environment (MOE) is a graphical molecular modeling tool that is popular among chemoinformatics, bioinformatics and CADD researchers. MOE is developed by The Chemical Computation Group Inc. MOE is a commercial software package that integrates a wide variety of functions. MOE supports the molecular visualization, molecular modeling, molecular simulation, data processing, protein structure analysis, protein modeling, high throughput virtual screening and development of novel algorithms. For the later, MOE employs the SVL (Scientific Vector Language) scripting language which provides flexibility to the user for customization. MOE supports the installation on Windows, Linux, OSX, SUN Solaris and HPC-clusters. Furthermore, MOE can perform all operations graphically, via a command line toolbox, or complete in command line modules using the MOE/batch application.

2.7.3 OMEGA

OMEGA [128, 192] is a robust knowledge-based conformer generator developed by OpenEye Scientific Software, USA. It is developed to deal with the large size libraries. Omega reassembles 3D fragments, from the fragments library into 3D conformations. While assembling the fragments, Omega iterates over the rotatable torsion angles using knowledge of the active conformations of similar molecules [192]. OMEGA not only can rapidly produce multi-conformer small molecule databases, it is also able to reproduce the 3D conformations of the crystallographic-generated conformers [128, 192]. Thus OMEGA provides great reliability and performance of generating multi-conformers when applied on large compound databases [193]. During my research, OMEGA was used for the generation of all 3D small molecule libraries using basic parameters unless otherwise specified.

2.7.4 GOLD

GOLD (Genetic Optimisation for Ligand Docking) is a molecular docking simulation software that uses the stochastic genetic algorithm [176]. The genetic algorithm of GOLD treats the ligands and receptor protein as partially flexible. GOLD was developed by the CCDC (Cambridge Crystallographic Data Centre) in Cambridge, UK. GOLD provides a graphical interface for structure visualization. Besides, GOLD provides an easy parameters and docking setup which are user friendly and can be easy to use by non-experts. During the development and evaluation of GOLD, 305 protein-ligand complexes from the PDB used as a

validation set and the result showed that GOLD can accurately predict the top-ranked solutions in 78% percent of the cases [148].
Furthermore, GOLD allows a great flexibility for users to set up the simulation parameters, and allows users to modify the parameters of the genetic search algorithm as well as to define backbone flexibility, protein site chains flexibility, ligand ring conformational flexibility, inclusion of the water molecules, specific metal coordination geometries and so on.

For scoring, GOLD is implemented with multiple scoring functions such as GoldScore, ChemScore, Astex Statistical Potential (ASP) and Piecewise Linear Potential (PLP) scoring functions. For this research, the PLP scoring function was utilized.

Chapter 3

EleKit2 as a virtual screening post filter

3.1 Introduction

EleKit2 is a modification of the previously reported EleKit algorithm [194]. In the previous EleKit algorithm, the similarity between the electrostatic potential fields of a protein ligand and a small molecule ligand was studied. This is useful for the rational design of small molecule inhibitors of protein-protein interactions, and its utility was demonstrated in the virtual screening of HIV integrase inhibitors in the SAMPL4 challenge [195].

However, EleKit2 is a novel Poisson Boltzmann (PB) electrostatics toolkit that serves a different purpose by investigating the electrostatic complementarity at the binding interface between the small molecules and the protein. Electrostatic complementarity is very important for the molecular recognition of rational drug design and therefore the inclusion of the electrostatic complementary is necessary. EleKit2 builds upon PDB2PQR [196] and APBS [197] to compute the partial charges and PB electrostatic potentials. In this chapter, I assess the ability of EleKit2 in measuring the electrostatic complementarity of protein-small molecule ligand interactions as a post-filter for molecular docking simulations packages.

3.2 Electrostatic Complementarity in the binding of protein-small molecule ligands

3.2.1 Introduction

All biological processes rely on the specific interaction between multiple molecules. For correct functioning it is therefore critical for these molecules to specifically recognize and interact with each other. Molecular recognition is dependent on the complementarity in shape and various specific interactions between the receptor and the ligand molecules [198-200]. Next to complementarity in shape, a key contributor to the long range recognition of molecules is considered to be in many cases the electrostatic potential [201]. The electrostatic potential originates from the electronegativity, dipole moment and the distribution of atomic partial charges within molecules. Electrostatic interactions between macromolecules will

influence the binding affinity, enzyme catalytic properties and the specificity of the molecular complexes. Complementarity of the binding of a receptor and a ligand is similar to the precision of a lock and a key, which not only requires a perfect match in shape but also requires electrostatic complementarity at the contacting surfaces [202]. Coulomb's law is commonly applied for the calculation of the electrostatic potential energy of the interaction of the two different charged particles.

The coulomb's law equation:

$$U_{coul} = k\frac{q_1 q_2}{r} = \frac{1}{4\pi\epsilon_0}\frac{q_1 q_2}{r}$$

Where

k is the Coulomb's constant (8.9876×109 N m² C−2)
r is the radius
ε_0 is the permittivity of free space
q1, q2 are the partial charges of the interacting molecules

In molecular graphics software, the electrostatic potential is illustrated in different color shades, which depict the polarity of the macromolecule. Generally, the polar areas are colored in red or blue, representing negative or positive electrostatic potentials respectively, while the non-polar areas are colored in lighter/white color.

Electrostatic complementarity has been widely studied starting from the early nineties on the interaction between two protein molecules. McCoy, *et al.* have reported the existence of anti-correlated surface electrostatic potentials, indicating electrostatic complementarity, while investigating the electrostatic potential at protein-protein interfaces [203]. The importance of the electrostatic complementarity in stabilizing the binding of the protein-protein complexes was reported by Karshikov, *et al.* in their research on the thrombin and hirudin molecules [204]. Moreover, the existence of the electrostatic complementarity on the protein-protein interface were also reported by Braden, *et al.*, Demchuk, *et al.*, Hendsch, *et al.*,Lescar, *et al.*, Karshikov, *et al.* and Novotny, *et al.* [205-210]. However, electrostatic complementarity does not only exist at the protein-protein interaction level but also in the small molecule and protein complexes. In 1994, Chau and Dean investigated complementarity using a simplified electrostatics model on a limited set of protein-ligand complexes [211-213]. Complementarity was calculated between the surface potential points of 34 receptor-ligand

complexes from the Brookhaven Protein Databank [214]. The partial charges of these surface points derived from the protein and ligand respectively were computed using the CNDO/2 (ligand) [215], ECEPP(protein) [216] and Mulliken, VSS program (van der Waals surface of ligand) [217] respectively. Their work has demonstrated the existence of complementarity at the protein-ligand interface by only analyzing the points surrounding the ligand surface at the interface region. However their method has some drawbacks including missing out the points that are lying in between the contacting surface of the receptor with the ligand which is equally important in contributing to the calculation of the complementarity.

As there is currently no method readily available to calculate the protein-ligand electrostatic complementarity, we set out to create a novel toolkit EleKit2 employing a more accurate electrostatics algorithm. EleKit2 has implemented the PB which is an efficient way in checking the complementarity between the protein and small molecule when compared to the conventional method applied in the Chau and Dean studies.

3.2.2 What is Poisson-Boltzmann electrostatics?

Poisson-Boltzmann (PB) theory is a commonly used theory in molecular biophysics to calculate the electrostatic potential of macromolecular systems in ionic solution. PB plays a major role in the biochemical reaction as it describes the electrostatic interactions and the binding energy of molecules in the solution condition. In the early 20^{th} century, the PB equation was first introduced by Gouy [218] and Chapman [219] , by equating the forces at the small volumes and the electric potential for the equilibrium distribution of the charge ion between two charged electrodes on the solute. Furthermore, the PB definition was later generalized by the Debye–Hückel [220]. The most recent definition was formulated by Grochowski and Trylska as: "The Poisson-Boltzmann equation is a solvent model for electrostatic continuum studies of the protein, nucleic acid, ion and water molecules" [221]. PB can be utilized for the calculation of the binding free energy, the electrostatic energy and the solvation energy in the solution. Moreover, the PB equation can be formulated as a linearized or a non-linearized equation which describes the interface problems between the biomolecules. The linearized PB equation was developed by Debye–Hückel, *et al.* [220] in 1923. They have developed a simple linearized PB equation to facilitate the calculation the binding free energy. The details on calculation of the PB using a non-linearized equation were reported by Gronwall *et al.* [222] in 1928. Because of the continuous expansion of protein structural data and the increasing number of solved protein complexes, researchers

experienced increasing difficulties for the calculation of the electrostatic properties. Therefore, accurate computational PB solvers have been developed to numerically solve the PB equation that comprise of the complexity of biomolecular system.

The Poisson Boltzmann (PB) equation:

$$\nabla \cdot \varepsilon(r)\nabla\varphi(r) - \varepsilon(r)k(r^2)\sinh[\varphi(r)] + 4\pi\rho(r)/kT = 0$$

Where

φ is the electrostatic potential.
r is the position vector
k is the Boltzmann's constant
T is the absolute temperature of the solutes
ρ is the molecular fixed charged constant.

In the last few decades, different algorithms have been developed to solve the Poisson–Boltzmann equation numerically. For example, the finite element algorithm [30, 223, 224], which is a fast computed adaptive nonlinear algorithm that reduces the computation time in solving the equation. In this method, a geometrically accurate mesh is created as a finite element. The finite element provides more flexibility in the local mesh refinement [225, 226] and meticulous convergence analysis on the non-linearized PB equation [226]. The multilevel finite element method that was used in solving the non-linearized PB equation for biomolecules system was later reported by Host *et al.* [227]. Improvements to shorten the solution time by using the finite element algorithm were later reported by Cortis *et al.* [223].

Another way to solve the PB equation is to use a finite difference algorithm [228, 229] that provides another numerical solution. Later Electronic Population Analysis on LCAO–MO Molecular Wave Functions that incorporates a multigrid technique have been applied to improve the efficiency [217]. The first finite difference method was reported by Warwicker *et al.* to solve the PB equation on a regular lattice [230]. He proposed a method to deal with the arbitrary shape or complex charges of the biomolecules by integrating the PB equation in a grid, which discretized the dielectric and molecular charges. The finite difference method is an efficient PB solver which is applicable on both the linearized and the non-linearized PB equation. Multiple PB solver packages that employ the finite difference algorithm are available including Delphi [231] , Grasp [232], MIBPB [233], ZAP [234], MEAD [235], UHBD [236] and PBEQ [237]. Both finite difference and finite element methods require a discretization of three-dimensional space [238].

APBS [197], short for Adaptive PB Solver, is one of the open-source computational programs that is widely used to solve the PB equation numerically by implementing both the final elements and finite difference algorithm. APBS can accurately calculate the electrostatic potential of biomolecules, thus it is suitable and was used to develop our new toolkit, EleKit2.

Another numerical solution is a boundary element algorithm [239, 240] which was inspired by Green's theorem. The boundary element algorithm is a reformulated equation from the original differential equation of the linearized PB equation, and it is known as the boundary integral equation [239, 241]. The generation of the molecular surface is needed in the boundary element algorithm, which was used to discretize the low polarization charges from the high polar solvent for the initial linear algebraic equations. The discretization is crucial in improving the efficiency and the accuracy of the solution [238].

3.3 Material and Methods

3.3.1 Software on which EleKit2 is built

3.3.1.1 PDBPQR

PDB2PQR [242-244] is a Python software package that was developed by Todd Dolinsky at Washington University in St. Louis. PDB2PQR was initially designed using the C++ algorithm by Jens Erik Nielsen which is used for the structure preparation for APBS. PDB2PQR is widely used for the continuum electrostatics calculations for biomolecules system as well as for the preparation of docking simulations using AutoDock [147, 245, 246]. PDB2PQR functions as a converter that converts the input PDB format [247] of protein receptor to the PQR output format by adding charges and solvation parameters to the PDB files. PDB2PQR also performs the optimization of the original PDB file by assigning missing parameters, adding and optimizing the geometry of hydrogens to the heavy atoms and determines the pKa values of the amino acids to ensure correct protonation state. The optimized protonated protein structure is amenable to the continuum electrostatics calculations by numerous computational biology software. In addition, PDB2PQR format is a commonly-used format and acceptable by many molecular modeling software packages, such as AutoDock [246], AMBER [248], MEAD [235], EleKit1 [194] ,VMD [249], PyMOL [250] and PMV [251]. Furthermore, PDB2PQR is used by APBS to automatically prepare the input file for solvation energy calculation and visualizing the electrostatics potentials as calculated by APBS. PDB2PQR provides an easy and simple setup, execution and analysis

of the Poisson-Boltzmann electrostatics calculations of APBS [244]. The latest version PDB2PQR [242] supports numerous force fields, such as CHARMM27 [252], AMBER99 [253], PARSE [254], optimized-Poisson Boltzmann force field [255] and user-defined force fields for proteins. Ligands however are parameterized using the PEOE force field [256]. PDB2PQR is freely available from http://pdb2pqr.sf.net/.

3.3.1.2 APBS

Adaptive Poisson-Boltzmann Solver (APBS) is a finite-volume-based PB solver software package that incorporates the global inexact-Newton method [228] to solve the PB equation numerically by calculating the finite volume mesh [257]. The Finite Element ToolKit has been used by the APBS to solve the elliptic nonlinear partial differential PB equations [258]. Moreover, PMG (Parallel algebraic MultiGrid) [259, 260], the adaptive multilevel finite element algorithm is also implemented in APBS, that allows discretization of the Cartesian mesh resulting in the rapid solution of the partial differential PB equations [228, 261]. APBS is frequently utilized as a continuum electrostatic model for the electrostatic studies of biomolecular system in solutes. APBS investigates the electrostatic interaction of all the macromolecules present in solution, and also plays a vital role in the drug design field by calculating the solvation and binding energy of the protein-protein/protein-ligand interaction. Recently, Unni *et al.* have reported the new Opal-based Web services of the APBS software package that enable easy access through the local web-server [244]. The Opal Toolkit [262, 263] was developed to provide flexibility for the APBS user to have the electrostatics calculations on the different computing platforms. The APBS software package is downloadable from https://www.poissonboltzmann.org/apbs/

3.3.2 EleKit2 algorithms

EleKit2 requires two different molecular 3D structures in the same Cartesian space, one representing the protein receptor in the PDB file format, one the ligand molecule (in the mol2 file format for small molecules). First, the electrostatic potentials around the receptor protein and the ligand molecule are computed separately and stored in 3D grids. Second, since only the binding site is most likely to be relevant for molecular recognition, a bit mask is created for these electrostatic potential grid points. The goal of this mask is to take into account only points in space that are at the receptor-ligand interface. The final mask is the Boolean intersection between the receptor protein's "thick surface" and the ligand molecule's "thick surface". To create this mask, a single distance cut-off parameter called "surface thickness" is needed. A "thick surface" is the volume difference between the union of atoms whose radii

were added half of "surface thickness" (dilatation) and the same atoms whose radii were removed half of "surface_thickness" (erosion). The first parameter is skin width (sw), to define a layer surrounding the receptor (or ligand) surface. The final mask includes the points belonging to the cross section of both parameters. Points inside the receptor molecule are excluded from the mask. For a graphical representation see Figure 11. Finally, the complementarity between electrostatic potentials of the ligand and receptor is assessed by correlating values at grid points within the mask using the Spearman rank-order correlation coefficient ρ. A positive value indicates similarity while a negative value represents complementarity. Additional similarity scores [264], Hodgkin index [265], Pearson's R [266] and a Tanimoto score [267] are also calculated. A graphical overview of the EleKit2 method is shown in Figure 12. EleKit2 is written in OCaml [268] and computations are parallelized with the Parmap library [269].

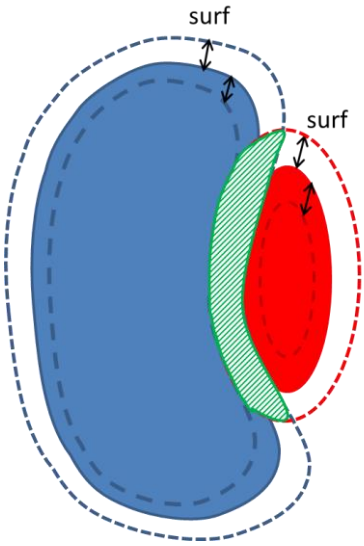


Figure 11 EleKit2 mask region. The schematic diagram displays the mask region that created between the protein-ligand complex. Protein is indicated in blue and the ligand is shown in red. The dashes lines surrounded the protein and ligand demonstrates the shell widths that have created. The overlapping region of both the shell widths from the protein and the ligand respectively are collected as the mask. The electrostatic potential of the grid points within the mask is calculated.

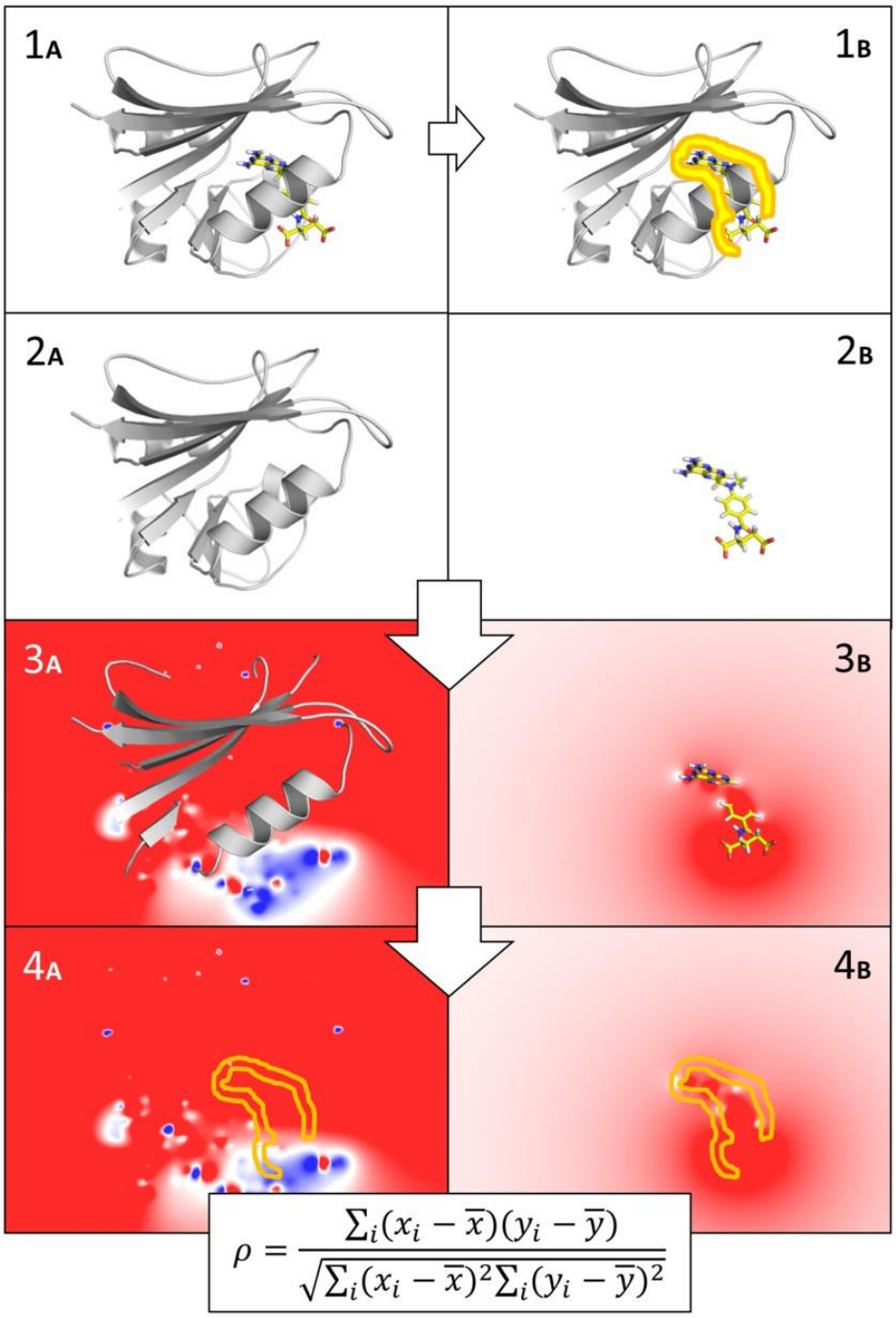


Figure 12 Overview of EleKit2. EleKit2 requires a protein-ligand complex shown in (1A). The receptor protein is shown as a grey cartoon while the ligand is depicted in sticks. As we are only interested in the points at the interface of the protein-ligand complex, a mask is created at the intersection of the protein and ligand (1B). The ligand complex works with an isolated receptor (2A) and ligand (2B) structure in the same Cartesian space. Using APBS the electrostatic potential of the molecules are calculated and stored in distinct grids (3A and 3B). Finally, the complementarity between electrostatic potentials for the points within the mask is calculated using the Spearman rank correlation coefficient (4A and 4B).

3.3.2.1 Parameterization

Five shell widths parameters of 0.4 Å, 0.5 Å, 0.6 Å, 0.7 Å and 0.8 Å were analyzed using the Iridium benchmark dataset (Figure 13) to decide the optimum parameter for further experiments.

3.3.3 Hardware

This work has been carried out using Dell Precision T5400 workstation with a quadcore 2.0 GHz Intel Xeon CPU and the Intel Xeon 5570 based massively parallel PC cluster of the RIKEN Integrated Cluster of Clusters (RICC).

3.3.4 Datasets

3.3.4.1 Iridium dataset

The Iridium dataset [270], is a freely available protein-ligand structure database that was released by OpenEye Scientific Software Inc. The Iridium dataset was initially derived from 728 receptor-ligand structures from the published databases and it was then further refined to determine the quality of the structures. Those structures with no experimentally determined free energy of binding value were discarded and the rest was kept in the dataset. The Iridium dataset consist of a total of 233 protein-ligand dataset, and almost half of the dataset has been manually curated to ensure the quality of the structures. To maintain the quality of the dataset, some aspects are considered during the curation step, including the functional groups, atom charges, atom bond, stereochemistry and tautomer of the protein structures derived from the Protein Data Bank (PDB). The Iridium dataset can be divided into 3 different categories based on the precision of the structures regarding to the experimental data, resulting in an Iridium Highly Trustworthy (Iridium-HT), an Iridium Moderately Trustworthy (Iridium-MT) and an Iridium Not Trustworthy (Iridium-NT) subset. Iridium-HT consists of 121 structures with highly trustworthy structures. Iridium-MT contains 104 structured with a moderate quality. And another 8 structures with serious flaws are classified as Iridium-NT. The main objective for the generating of a high accuracy of protein-ligand structural databases is to validate the molecular docking algorithms. As the dataset comprise of different level of the accuracy and precision of the structures, it facilitates the user to utilize it as a prediction model in validating any of the structure-based drug design tools.

For the evaluation of the EleKit2 algorithm and the study of the presence of electrostatic complementarity in existing receptor-ligand complexes, the Iridium dataset with the “highly trustworthy” and “medium trustworthy” subsets were used for benchmarking. However, because of limitations of PDB2PQR that is a part of EleKit2, EleKit2 is unable to correctly

handle entries with metal ions or cofactors. Therefore, these entries were removed from the dataset during the course the evaluation experiment, delivering a new Iridium subset indicated as "Optimized Iridium". EleKit2 was applied using a desolvated protein environment with a solute dielectric constant of 2.0 and a shell width of 0.5Å. The Iridium dataset is freely available online and can be downloaded from http://www.eyesopen.com/iridium

3.3.4.2 DUD dataset

The Directory of Useful Decoys (DUD) [271] is a benchmarking dataset for virtual screening. The DUD was developed by the Shoichet Laboratory, Department of Pharmaceutical Chemistry at the University of California, San Francisco (UCSF). The DUD dataset is a collection of 40 different protein targets with known active molecules as well as inactive "decoys" per target drawn from the ZINC database. For each target, a crystal structure of the receptor-ligand complex is available. There are around 3000 decoys of ligand per protein target and every ligand has another 30 decoys that are topological different from each other.

For the evaluation of the utility of EleKit2 in virtual screening as a post-filter of docking results, DUD dataset was used. Targets which contain either cofactors or metal ions were excluded resulting in a remaining 28 test cases. This dataset was also used to analyze the relationship between the correctness of the binding mode (RMSD) versus the electrostatic complementarity. DUD is freely available online and can be downloaded at http://blaster.docking.org/dud/

3.3.5 GOLD docking simulation software

The GOLD [272] docking simulation software was used for the virtual screening benchmarking. Using GOLD all the ligands and decoys were docked to their related targets from the DUD dataset. The basic GOLD settings for virtual screening were applied with 10 simulations per ligand molecule. In order to analyze the docking results, the docking scoring function (PLP_f), and the electrostatic complementarity ($EC\rho$) for every binding mode per compound were calculated.

To calculate the relationship between the binding mode quality and the electrostatic complementarity, the reference ligands, for which a crystallographically determined binding mode is available, of the DUD targets, were docked 25 times with the targets by using GOLD.

3.3.6 fconv

fconv is an open source C++ software under GNU General Public License. It is designed as a robust tool for the manipulating the molecular data for computational drug design. fconv functions as a format converter and the acceptable format are PDB(QT), MOL2, SDF, DLG and CIF . Moreover, it can also be used as a substructure search engine to search for the drug molecules. Furthermore, it can also use to detect binding cavities as well as the calculation of molecular descriptors such as molecular weight and number of rotational bonds. Furthermore, fconv is a popular tool for calculating the root-mean-square-deviation (RMSD) of ligand molecules in absolute space or after alignment. During the EleKit2 related research, fconv was applied to determine the difference between docked and experimentally determined binding mode of the 28 DUD targets in absolute space [30]. fconv is freely available and can be downloaded from http://www.agklebe.de.

3.4 Results and Discussion

3.4.1 Validation of EleKit2

3.4.1.1 Shell width parameter

As EleKit2 uses a single parameter, our first experiment was to determine the optimal value (ranging from 0.4 to 0.8) of this parameter using the Iridium optimized benchmark set. The shell width that showed the best complementarity and least similarity was selected (see Figure 3). From this point on, EleKit2 was applied using a half shell width of 0.5Å in the subsequent research experiments.

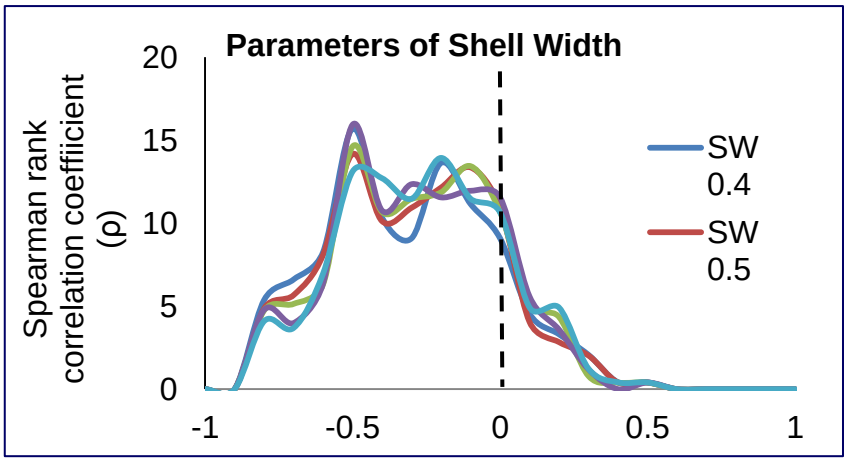


Figure 13 Parameters of shell width. The curves indicate the normalized distribution of the ECρ of the shell widths of 0.4 Å, 0.5 Å, 0.6 Å, 0.7 Å and 0.8 Å. The distribution of the curves towards -1 demonstrates complementarity.

3.4.1.2 Benchmarking by using Iridium dataset

The presence of electrostatic complementarity in existing receptor-ligand complexes was first analyzed using EleKit2. For this purpose we used the Iridium dataset that is based on high resolution crystal structures and is manually curated. Both the highly trustworthy dataset (Iridium-HT) and the mildly trustworthy dataset (Iridium-MT) were used (Figure 4 and 5).

A majority (85%) of the protein-ligand complexes in the Iridium dataset showed negative ECρ values indicating electrostatic complementarity (Figure 4). However, a minority of the complexes exhibited significant similarity. Therefore, all cases where ECρ > 0.1 were visually inspected and it was observed that in all cases the ligand binding to the receptor was either mediated via metal ligation or by binding onto a cofactor/ prosthetic group. The Iridium results are shown in Addendum 1.

When a protein structure contains bound metal ions or prosthetic groups at the protein-ligand interface, APBS cannot calculate the electrostatic field accurately unless further optimization of parameters is performed. Upon removal of all targets containing metal ions or prosthetic groups from the Iridium dataset, the distribution indeed improved to a 91% of complexes showing electrostatic complementarity (Figure 14 and Figure 15).

Analysis of the Iridium data revealed that a few protein-ligand complexes still exhibit ECρ values corresponding to random or weak similarity. This was to be expected as not all interactions in nature are driven by charged interactions; some are driven by hydrophobic or entropic interactions. Visual inspection revealed that in these cases the receptor-ligand interaction was of a highly hydrophobic nature, such as compounds with a steroid core. Furthermore, carbohydrate ligands were also present here as their binding is mediated via a multitude of hydrogen bonds in which the orientation of lone pairs and polar hydrogens are very important. As lone pairs are condensed into the nucleus, assigning partial charges to a mono-polar atom is not accurate enough for these cases.

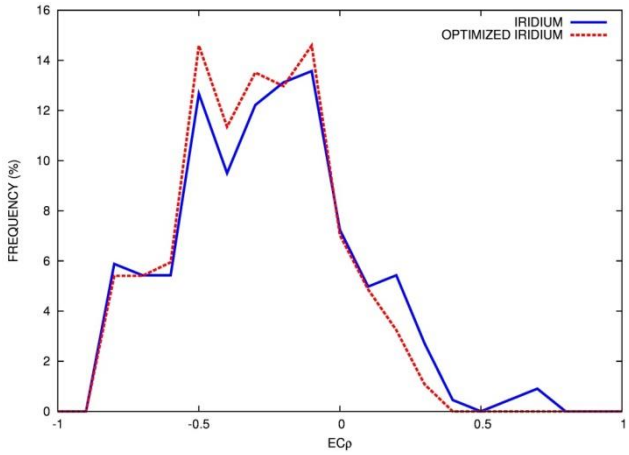


Figure 14 ECρ distribution in the Iridium dataset. The curves show the normalized distribution of ECρ without (blue) and after (red) dataset optimization. The majority of the population resides between -1 and 0 demonstrating complementarity. Optimization of the dataset by removing all the cofactors and metals significantly improves the result by reducing the number of cases without complementarity.

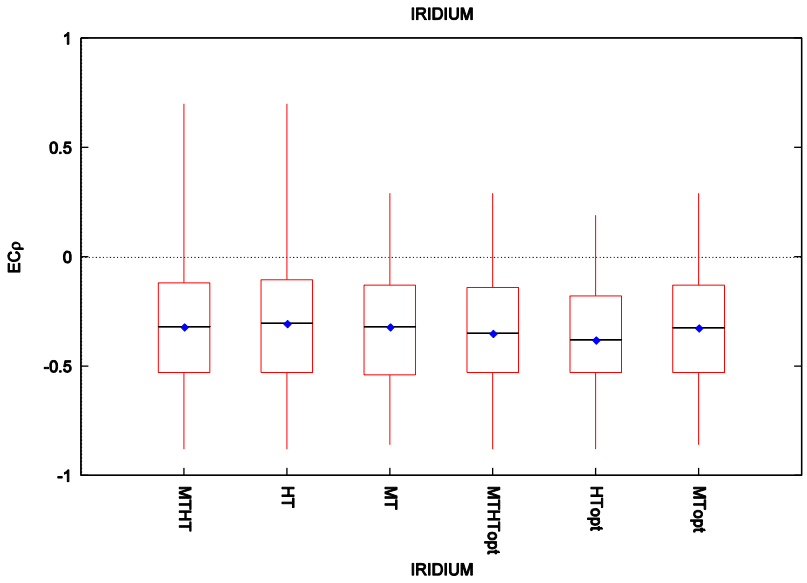


Figure 15 The distribution of ECρ for three Iridium datasets before and after dataset optimization. The box plots represent the distribution of ECρ for the different Iridium datasets. HT, MT, MTHT, HTopt, MTopt, MTHTopt represent respectively the Iridium highly trustworthy, Iridium mildly trustworthy, Iridium highly plus mildly trustworthy, the optimized Iridium highly trustworthy, optimized Iridium mildly trustworthy and optimized Iridium highly plus mildly trustworthy datasets (optimized means all entries with cofactors and metals have been removed).

3.4.1.3 Solute dielectric constant parameter

The influence of the dielectric constant on the electrostatic complementarity was investigated. The previously mentioned experiments were carried out using a solute dielectric constant of 2.0 which corresponds to desolvated molecules, or in fact interactions at the interface between molecules. However, since molecular recognition already happens in the long range and to further explore the influence of the solvated state of the molecules, the experiment was repeated with a dielectric constant of 78. The average electrostatic complementarity (ECρ) using the solvated model is -0.23 as compared to -0.35 using the desolvated model (Addendum 1). Therefore, the desolvated model was used for all the subsequent analyses with the DUD dataset. The analysis of the structures of the Iridium dataset indeed indicate that there is complementarity at the protein-ligand interface in the majority of cases, implying a potential value for computational structured based drug design.

3.4.2 Virtual screening

Since the benchmarking of EleKit2 revealed that our software can indeed calculate a complementarity metric, the next step was to analyse the putative usage of EleKit2 as a virtual screening postfilter. The workflow on the analysis of the utility of the EleKit2 as a post-filter tool for virtual screening has shown in Figure 16.

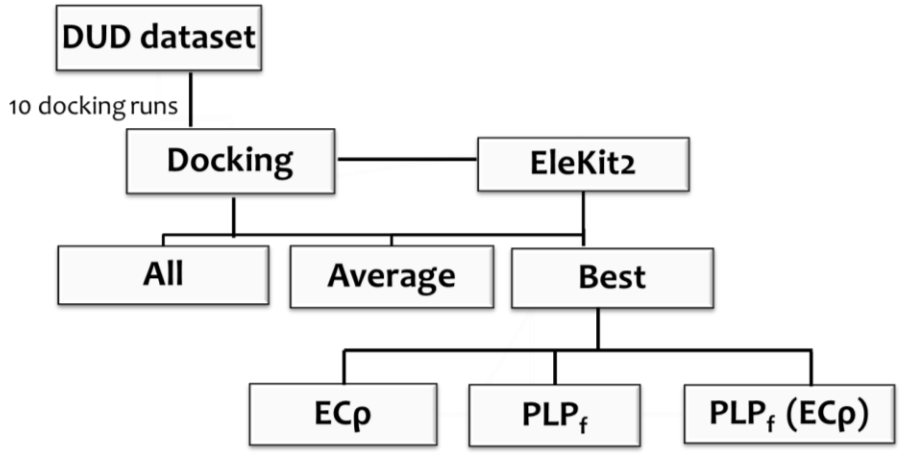
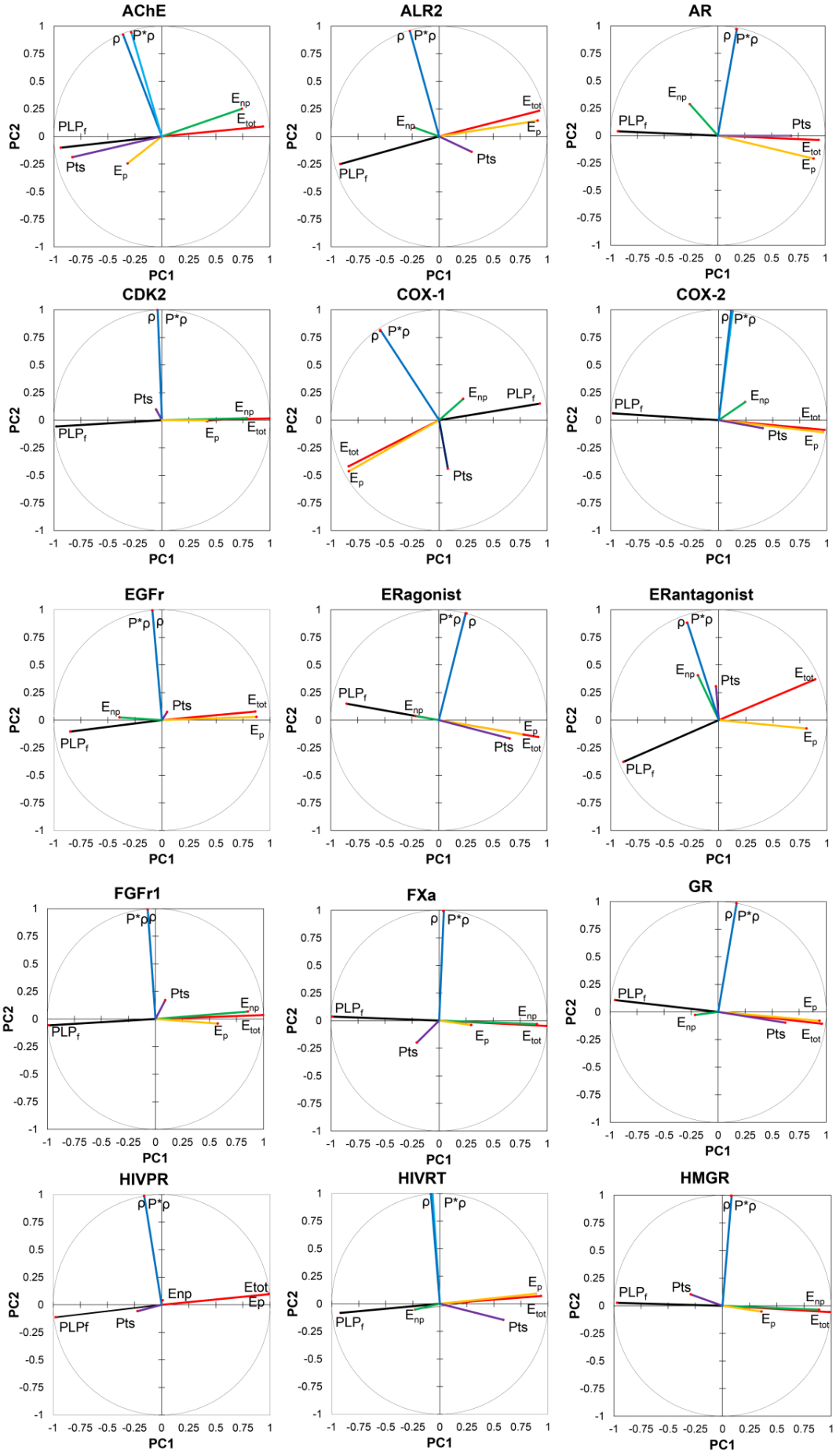


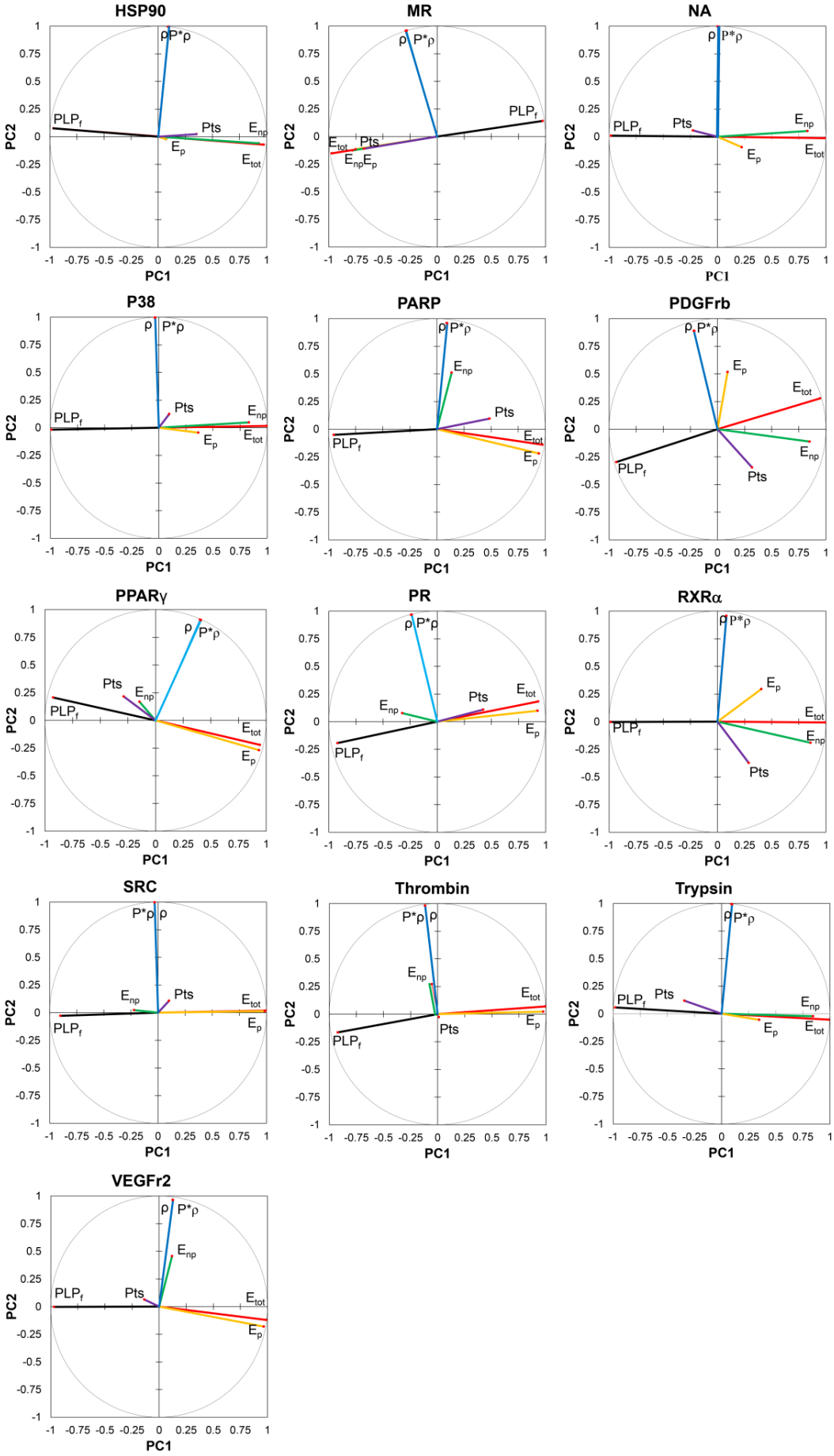
Figure 16 EleKit2 work flow. Schematic representation of the EleKit2 experimental workflow. The major steps are depicted from the top row to the bottom row. The experimental process is starting from the DUD dataset, which all the ligands or decoys molecules are docked 10 different times. The docked result is sent to the EleKit2 toolkit to check for the presence of complementarity. Subsequently, the docked result and EleKit2 result are processed by using three options, by considering: (i) All the 10 docking solutions scores. (ii) The average scores of the 10 docking solution. (iii) The best score of the 10 docking solutions. For this last option there are a further three different options to analyze the ranking: (i) The best complementarity scores (ECρ). (ii) The best docking fitness scores (PLP_f). (iii) The best docking score corresponding to the best complementarity scores PLP_f (ECρ).

3.4.2.1 Is EleKit2 Independent with Docking?

The applicability of EleKit2 in virtual screening by molecular docking simulations was evaluated using the DUD dataset. All compounds (actives and decoys) were docked into their respective receptors using GOLD with the PLP_f scoring function. There was no correlation between the docking scores and the ECρ values observed. The Pearson's ***r*** between PLP_f and ECρ has a mean of -0.04 with a variance of 0.09 for different DUD targets. Thus, the ECρ values produced by EleKit2 are independent from the molecular docking scores although it used the docked poses of ligands. The independence between ECρ values and docking scores is further shown in the Principle Component Analysis biplots (Figure 17) of the scoring function parameters for the DUD targets. It can be seen that the ECρ is unrelated (orthogonal) to the docking scores as well as the polar-contribution of the docking scores. Therefore, we reasoned that the ECρ values can indeed produce additional information regarding to the docking pose scores. The docking scores however are dominated by either hydrophobic or hydrophilic terms depending on the DUD target (polar or nonpolar targets). For many hydrophobic cases, this is not really unexpected since scoring functions predominantly aim to predict the affinity of compounds during docking simulations and binding affinity is often dominated by the hydrophobicity of the compound whereas the polar interactions contribute to the specificity.

Figure 17 The biplots of PCA on DUD dataset (next two pages). The biplots of Principal Component Analysis indicate the independency of the complementarity score towards the GOLD docking score. In the biplots, the variables of ρ, Pts, P*ρ, PLP_f, Etot, Ep, Enp, represent the Spearman score, the number of Points in the EleKit2 mask, the product of Pts and ρ, PLP Fitness score, PLP predicted total free energy of binding, PLP polar contribution, and the PLP non-polar contribution. The biplots show that the EleKit2 scores (ρ and P*ρ) are orthogonal to the GOLD docking scores (PLP_f).


AChE
ALR2
AR
CDK2
COX-1
COX-2
EGFr
ERagonist
ERantagonist
FGFr1
FXa
GR
HIVPR
HIVRT
HMGR
PC1
PC2
ρ
P*ρ
Pts
Enp
Etot
Ep
PLPf
Enp
Etot
Ep
PLPf


HSP90
MR
NA
P38
PARP
PDGFrb
PPARγ
PR
RXRα
SRC
Thrombin
Trypsin
VEGFr2
PC1
PC2
ρ
P*ρ
PLP_f
Pts
E_np
E_p
E_tot

3.4.2.2 Experiments

We have explored several options for the handling of multiple docked poses for each compound as a result of multiple docking runs. The enrichment of actives due to each option were analyzed by calculating the area under curve (AUC) of the receiver operating characteristic (ROC) plot [273]. Since we have run 10 docking simulations for each compound in the DUD dataset, there are 10 top docked poses for each compound generated. Three different ways of handling these multiple docked poses for each compound were evaluated. First, all docked poses were included, with the drawback that many incorrect binding modes are considered to be correct. Secondly, the average score of all docked poses are used in which good and bad information can be averaged out. Thirdly, only the top scoring binding mode is used, with the risk that the correct binding mode is discarded. The first two options were analyzed but failed to deliver any satisfactory results. An overview of the AUC values is given in Table 1 and Table 2.

We have focused on the third option of which the overview is given in Table 4 by taking the best out of all docked poses for further analysis.

There are three strategies to rank compounds when faced with the third option of taking the best out of all docked poses resulted from multiple docking runs:

1) To take the PLP_f docking score of the best scoring docked pose (denoted as PLP_f).
2) To take the ECρ of the docked pose with the highest complementarity (denoted as ECρ).
3) To take the PLP_f docking score of the docked pose with the highest complementarity (denoted as PLP_f (ECρ).

A comparison of the first and second ranking strategies is shown in the Table 3. It shows that the ECρ strategy performed worse than PLP_f. However, the PLP_f (ECρ) ranking strategy clearly outperformed PLP_f (Table 4).

Table 1 An overview of the AUCs of the 28 DUD targets using all the 10 docking poses.

Targets	**PLP_f**	**ECρ**
AChE	0.72	0.49
ALR2	0.45	0.55
AR	0.56	0.68
CDK2	0.50	0.53
COX-1	0.32	0.38
COX-2	0.88	0.57
EGFr	0.52	0.54
ERagonist	0.76	0.56
ERantagonist	0.85	0.70
FGFr1	0.36	0.37
FXa	0.63	0.59
GR	0.45	0.43
HIVPR	0.42	0.56
HIVRT	0.65	0.47
HMGR	0.42	0.51
HSP90	0.66	0.49
MR	0.46	0.68
NA	0.39	0.71
P38	0.69	0.51
PARP	0.75	0.76
PDGFrb	0.45	0.67
PPARγ	0.46	0.55
PR	0.39	0.56
RXRα	0.73	0.70
SRC	0.64	0.55
Thrombin	0.77	0.44
Trypsin	0.64	0.54
VEGFr2	0.61	0.54
Number of best cases	14	14

Table 2 An overview of the AUCs of the 28 targets from DUD using a ranking based on the average values of the 10 docking poses.

Targets	**PLP_f**	**ECρ**
AChE	0.73	0.47
ALR2	0.51	0.59
AR	0.64	0.74
CDK2	0.53	0.56
COX-1	0.31	0.39
COX-2	0.91	0.63
EGFr	0.49	0.54
ERagonist	0.79	0.57
ERantagonist	0.89	0.72
FGFr1	0.39	0.29
FXa	0.67	0.66
GR	0.50	0.45
HIVPR	0.42	0.61
HIVRT	0.63	0.51
HMGR	0.41	0.52
HSP90	0.70	0.41
MR	0.70	0.79
NA	0.37	0.81
P38	0.65	0.54
PARP	0.82	0.79
PDGFrb	0.47	0.71
PPARγ	0.41	0.56
PR	0.41	0.60
RXRα	0.70	0.77
SRC	0.68	0.56
Thrombin	0.87	0.36
Trypsin	0.65	0.57
VEGFr2	0.59	0.59
Number of best cases	14	13

Table 3 Comparing the AUCs for the DUD targets ranked by the best conformation using the PLP_f or the ECρ.

Targets	PLP_f	ECρ
AChE	0.76	0.49
ALR2	0.47	0.49
AR	0.68	0.69
CDK2	0.51	0.57
COX-1	0.28	0.35
COX-2	0.89	0.32
EGFr	0.41	0.40
ERagonist	0.77	0.53
ERantagonist	0.88	0.63
FGFr1	0.35	0.32
FXa	0.67	0.53
GR	0.49	0.40
HIVPR	0.53	0.61
HIVRT	0.57	0.45
HMGR	0.46	0.56
HSP90	0.67	0.36
MR	0.70	0.69
NA	0.40	0.80
P38	0.64	0.50
PARP	0.79	0.64
PDGFrb	0.45	0.63
PPARγ	0.38	0.55
PR	0.39	0.51
RXRα	0.63	0.68
SRC	0.68	0.55
Thrombin	0.88	0.31
Trypsin	0.65	0.47
VEGFr2	0.56	0.59
Number of best cases	16	12

Table 4 Comparing the average AUCs for each DUD targets ranked by the best scoring conformation or the most complementary conformation using the PLP_f scoring function.

Targets	**PLP_f**	**ECρ**	**PLP_f(ECρ)**
AChE	**0.76**	0.49	0.72
ALR2	0.47	0.49	**0.54**
AR	0.68	**0.69**	0.66
CDK2	0.51	**0.57**	0.54
COX-1	0.28	**0.35**	0.32
COX-2	**0.89**	0.32	**0.89**
EGFr	0.41	0.40	**0.52**
ERagonist	0.77	0.53	**0.79**
ERantagonist	**0.88**	0.63	**0.88**
FGFr1	0.35	0.32	**0.42**
FXa	**0.67**	0.53	0.64
GR	0.49	0.40	**0.53**
HIVPR	0.53	**0.61**	0.48
HIVRT	0.57	0.45	**0.64**
HMGR	0.46	**0.56**	0.43
HSP90	0.67	0.36	**0.69**
MR	0.70	0.69	**0.71**
NA	0.40	**0.80**	0.39
P38	**0.64**	0.50	**0.64**
PARP	0.79	0.64	**0.82**
PDGFrb	0.45	**0.63**	0.46
PPARγ	0.38	**0.55**	0.47
PR	0.39	**0.51**	0.41
RXRα	0.63	0.68	**0.71**
SRC	**0.68**	0.55	0.62
Thrombin	**0.88**	0.31	0.73
Trypsin	**0.65**	0.47	0.64
VEGFr2	0.56	0.59	**0.60**
#best PLP_f (ECρ) : 14			
#best PLP_f : 8			
#best ECρ : 9			

Out of a total of 28 cases, PLP_f (ECρ) performed the best 14 times versus 8 times for PLP_f and 9 times for ECρ. PLP_f (ECρ) and PLP_f performed equally well in 3 cases. In 12 out of the 28 cases, ECρ performed better than PLP_f (Table 3). This is especially true in the case of the Neuraminidase (NA) inhibitors. It's worth noting that the first NA inhibitors originated from a rational design strategy incorporating Molecular Interaction Fields including the electrostatic potential [27] and thus explaining the high level of complementarity. As the analysis of the Iridium dataset revealed that complementarity exists, but is widely spread between -1 and 0, it also showed the limitations of the ECρ value as an indicator of compound binding affinity. The ECρ is designed only to measure the electrostatic complementarity and many other factors that are important for binding are not included, and therefore is not suitable for ranking. Consequently we explored ranking the "most complementary" binding modes of each compound according to the docking score, as the docking score is designed to measure binding affinity. This is exactly why we proposed the ranking strategy of PLP_f (ECρ) as a post filtering step when dealing with multiple docked poses. Indeed PLP_f (ECρ) is the best scoring metric compared to PLP_f and ECρ.

For post filtering, another combination method of PLP_f and the ECρ values score was analyzed as well. Cut-offs of -0.2, -0.1, 0, 0.1, 0.2 were used, representing good complementarity to absolutely no complementarity. The compounds that did not meet the desired electrostatic complementarity cut-off score were ranked at the end of the list. The analysis was performed on all 10 docking poses (All Combo), the average docking score/ECρ values (Average Combo), or the docked result with the best docking/ECρ value (Best Combo). The performance to enrich the active compounds towards the top of the ranking (based either on docking scores, EleKit2 scores and/or their combination) were analyzed by the AUC of the ROC curve, based on scoring according to the docking score, the ECρ score and post filtering the docking scores according to complementarity values. An overview of the results is given in Table 5, 6 and 7. Post-filtering by combining the docking score with ECρ-based cut-off values does not improve the results as the AUC values are generally averaged out between the highest and the lowest values. This observation is true for ROC and BEDROC values for all cases (all docked poses, best docked poses and average docking/ ECρ values). Therefore, the combination method is not a suitable post-filtering method to improve the virtual screening performance.

Table 5 Performance using different ECρ cutoffs on all docking scores.

ALL							
Targets	**PLP_f**	**Spear**	**combo_0.0**	**combo_0.1**	**combo_0.2**	**combo_-0.1**	**combo_-0.2**
AChE	0.72	0.49	0.62	0.62	0.64	0.61	0.62
ALR2	0.45	0.55	0.51	0.52	0.49	0.48	0.46
AR	0.56	0.68	0.66	0.64	0.63	0.66	0.62
CDK2	0.50	0.53	0.53	0.53	0.52	0.52	0.51
COX-1	0.32	0.38	0.33	0.32	0.34	0.33	0.33
COX-2	0.88	0.57	0.79	0.83	0.85	0.75	0.73
EGFr	0.52	0.54	0.53	0.54	0.55	0.52	0.50
ERagonist	0.76	0.56	0.69	0.71	0.72	0.69	0.68
ERantagonist	0.85	0.70	0.80	0.83	0.86	0.80	0.82
FGFr1	0.36	0.37	0.33	0.29	0.29	0.37	0.42
FXa	0.63	0.59	0.63	0.65	0.65	0.62	0.61
GR	0.45	0.43	0.45	0.43	0.41	0.42	0.43
HIVPR	0.42	0.56	0.51	0.49	0.47	0.50	0.47
HIVRT	0.65	0.47	0.54	0.58	0.62	0.53	0.58
HMGR	0.42	0.51	0.44	0.44	0.44	0.46	0.47
HSP90	0.66	0.49	0.58	0.54	0.60	0.62	0.64
MR	0.46	0.68	0.61	0.60	0.57	0.59	0.60
NA	0.39	0.71	0.58	**0.53**	0.47	0.61	0.62
P38	0.69	0.51	0.59	0.61	0.66	0.62	0.65
PARP	0.75	0.76	0.81	0.81	0.80	0.79	0.77
PDGFrb	0.45	0.67	0.61	0.56	0.51	0.61	0.60
PPARγ	0.46	0.55	0.48	0.46	0.46	0.50	0.49
PR	0.39	0.56	0.45	0.48	0.46	0.46	0.48
RXRα	0.73	0.70	0.73	0.74	0.74	0.74	0.76
SRC	0.64	0.55	0.59	0.59	0.61	0.61	0.62
Thrombin	0.77	0.44	0.61	0.69	0.74	0.57	0.61
Trypsin	0.64	0.54	0.63	0.62	0.64	0.60	0.58
VEGFr2	0.61	0.54	0.58	0.61	0.62	0.57	0.57

Table 6 Performance using different ECρ cutoffs on the average docking score.

AVERAGE							
Targets	**PLP_f**	**Spear**	**combo_0.0**	**combo_0.1**	**combo_0.2**	**combo_-0.1**	**combo_-0.2**
AChE	0.73	0.47	0.64	0.64	0.65	0.63	0.67
ALR2	0.51	0.59	0.54	0.55	0.52	0.50	0.51
AR	0.64	0.74	0.74	0.71	0.70	0.73	0.68
CDK2	0.53	0.56	0.55	0.54	0.53	0.55	0.56
COX-1	0.31	0.39	0.33	0.31	0.32	0.33	0.33
COX-2	0.91	0.63	0.86	0.88	0.90	0.84	0.85
EGFr	0.49	0.54	0.53	0.53	0.54	0.52	0.48
ERagonist	0.79	0.57	0.72	0.74	0.75	0.71	0.72
ERantagonist	0.89	0.72	0.86	0.89	0.91	0.87	0.89
FGFr1	0.39	0.29	0.26	0.20	0.24	0.34	0.42
FXa	0.67	0.66	0.71	0.73	0.71	0.70	0.69
GR	0.50	0.45	0.47	0.47	0.46	0.43	0.44
HIVPR	0.42	0.61	0.51	0.49	0.46	0.51	0.48
HIVRT	0.63	0.51	0.58	0.59	0.61	0.58	0.64
HMGR	0.41	0.52	0.43	0.44	0.43	0.45	0.45
HSP90	0.70	0.41	0.60	0.49	0.58	0.70	0.72
MR	0.70	0.79	0.76	0.76	0.74	0.74	0.74
NA	0.37	0.81	0.62	0.52	0.44	0.68	0.70
P38	0.65	0.54	0.61	0.62	0.64	0.63	0.63
PARP	0.82	0.79	0.86	0.88	0.86	0.85	0.84
PDGFrb	0.47	0.71	0.65	0.59	0.52	0.65	0.62
PPARγ	0.41	0.56	0.39	0.37	0.38	0.44	0.47
PR	0.41	0.60	0.50	0.51	0.48	0.51	0.54
RXRα	0.70	0.77	0.77	0.75	0.72	0.80	0.79
SRC	0.68	0.56	0.64	0.62	0.63	0.66	0.69
Thrombin	0.87	0.36	0.69	0.79	0.84	0.63	0.71
Trypsin	0.65	0.57	0.67	0.65	0.66	0.63	0.60
VEGFr2	0.59	0.59	0.58	0.61	0.61	0.58	0.60

Table 7 Performance using different ECρ cutoffs of values on the best docking score.

BEST								
Targets	**PLP_f**	**Spear**	**PLP_f (ECρ)**	**combo_0.0**	**combo_0.1**	**combo_0.2**	**combo_-0.1**	**combo_-0.2**
AChE	0.76	0.49	0.72	0.70	0.73	0.73	0.66	0.62
ALR2	0.47	0.49	0.54	0.50	0.49	0.48	0.49	0.51
AR	0.68	0.69	0.66	0.74	0.72	0.71	0.74	0.72
CDK2	0.51	0.57	0.54	0.55	0.52	0.50	0.55	0.58
COX-1	0.28	0.35	0.32	0.29	0.28	0.29	0.29	0.30
COX-2	0.89	0.32	0.89	0.86	0.87	0.89	0.78	0.62
EGFr	0.41	0.40	0.52	0.43	0.44	0.43	0.44	0.39
ERagonist	0.77	0.53	0.79	0.72	0.77	0.76	0.69	0.67
ERantagonist	0.88	0.63	0.88	0.84	0.91	0.89	0.78	0.77
FGFr1	0.35	0.32	0.42	0.19	0.22	0.29	0.21	0.29
FXa	0.67	0.53	0.64	0.67	0.67	0.67	0.67	0.66
GR	0.49	0.40	0.53	0.46	0.46	0.45	0.40	0.42
HIVPR	0.53	0.61	0.48	0.56	0.53	0.54	0.55	0.55
HIVRT	0.57	0.45	0.64	0.53	0.53	0.54	0.49	0.49
HMGR	0.46	0.56	0.43	0.47	0.48	0.47	0.50	0.50
HSP90	0.67	0.36	0.69	0.46	0.42	0.54	0.52	0.59
MR	0.70	0.69	0.71	0.80	0.75	0.74	0.76	0.76
NA	0.40	0.80	0.39	0.46	0.43	0.41	0.48	0.57
P38	0.64	0.50	0.64	0.58	0.63	0.65	0.57	0.57
PARP	0.79	0.64	0.82	0.81	0.83	0.81	0.77	0.73
PDGFrb	0.45	0.63	0.46	0.53	0.51	0.46	0.56	0.59
PPARγ	0.38	0.55	0.47	0.40	0.39	0.39	0.43	0.46
PR	0.39	0.51	0.41	0.44	0.42	0.40	0.43	0.49
RXRα	0.63	0.68	0.71	0.71	0.68	0.65	0.76	0.75
SRC	0.68	0.55	0.62	0.63	0.64	0.66	0.62	0.60
Thrombin	0.88	0.31	0.73	0.86	0.87	0.88	0.75	0.62
Trypsin	0.65	0.47	0.64	0.67	0.65	0.65	0.63	0.60
VEGFr2	0.56	0.59	0.60	0.54	0.56	0.56	0.53	0.56

When analyzing the distribution of the PLP_f or the ECρ scores for the ligands and decoys in the DUD dataset by using the three options (except the Combo), it is revealed that taking all 10 docked poses is randomly distributed around 0 (Figure 18). Furthermore no clear separation is shown between the active ligands and decoys of the individual DUD targets. When taking only the best ECρ per compound, the distribution shifts towards complementarity. However, still no clear tendency towards separation between active ligands and decoys can be observed (Figure 19). A trend for separation is observed between ligands and decoys when the PLP_f docking score is used for the best scoring docking poses (Figure 20). The separation becomes more pronounced when the scores of the most complementary (lowest ECρ) binding mode are used (Figure 21). These results indicate the capability of the ECρ metric to select good docked binding modes in a post-filtering step.

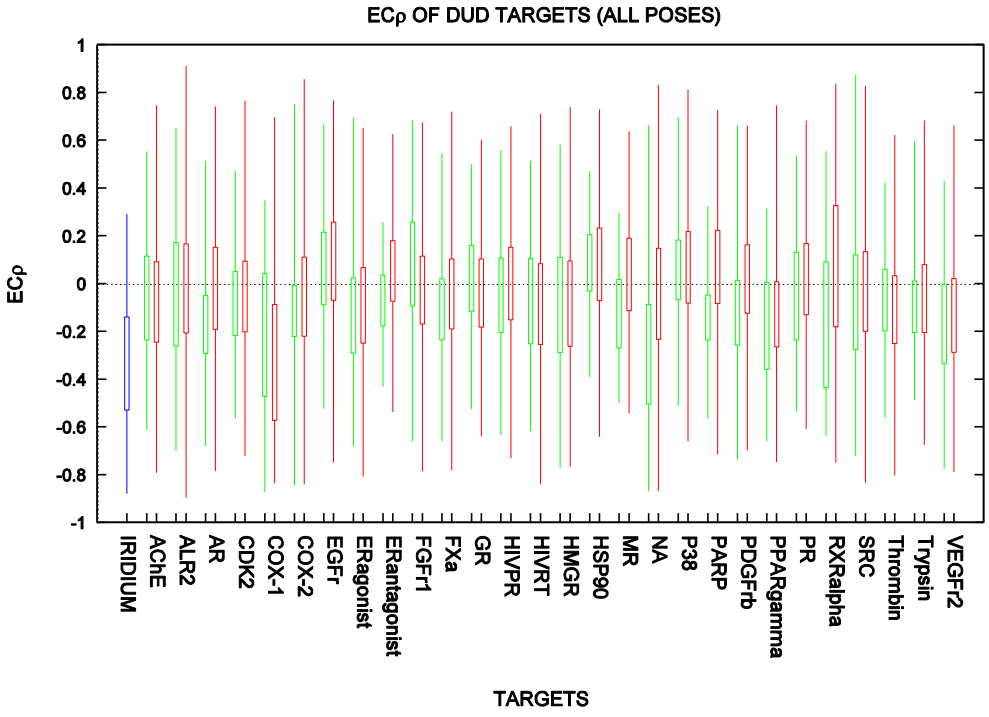


Figure 18 The ECρ distribution of DUD targets including all 10 docked poses per compound. The box-plot represents the distribution of ECρ in the 10 docking solutions for the different DUD targets of ligand (green) versus the decoys (red). The distribution of the ECρ is compared with the distribution of the Iridium dataset (blue). Not only does this show that in the majority of the cases the docking simulations have a random ECρ distribution (around 0).

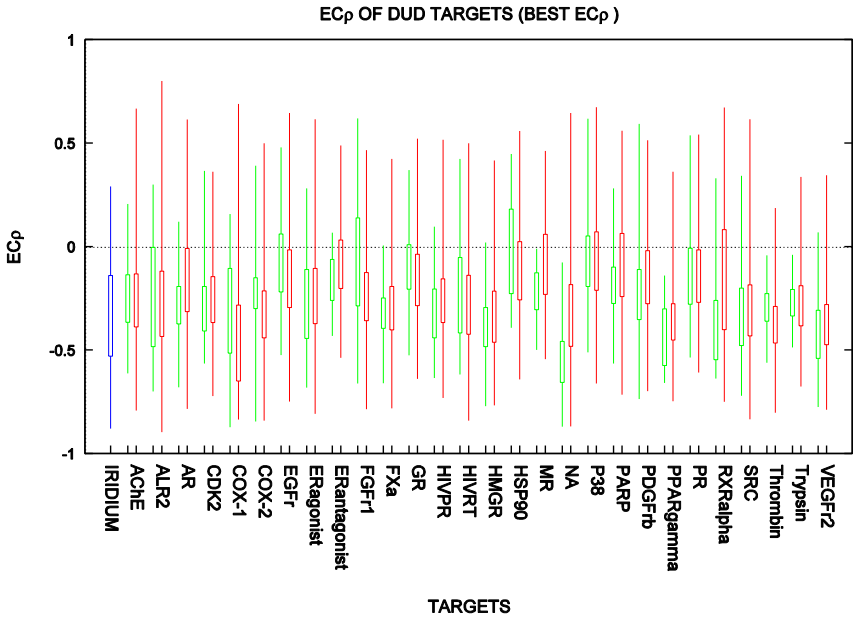


Figure 19 The distribution of ECρ when only the most complementarity docked pose for a ligand is considered for each target of the DUD dataset. The box-plot represents the distribution of ECρ for the different DUD targets of ligand (green) versus the decoys (red) for the lowest ECρ value out of the 10 docked solutions per compound. The distribution of the ECρ is compared with the distribution of the Iridium dataset (blue).

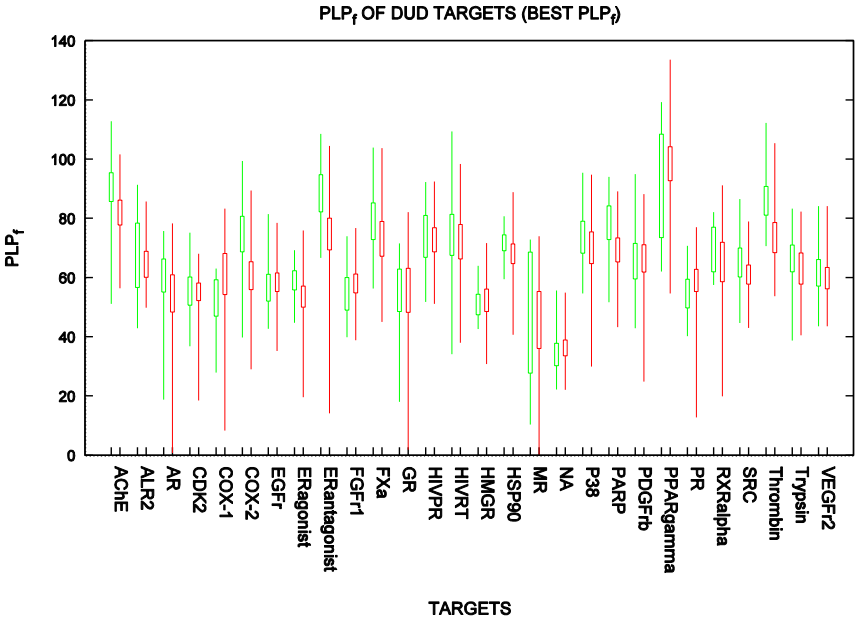


Figure 20 The distribution of the PLP_f scores of the DUD targets based on the docking pose with the best PLP_f score. The box-plot represents the distribution of the PLP_f for the different DUD targets of ligand (green) versus the decoys (red) for the best PLP_f value out of the 10 docked solutions per compound.

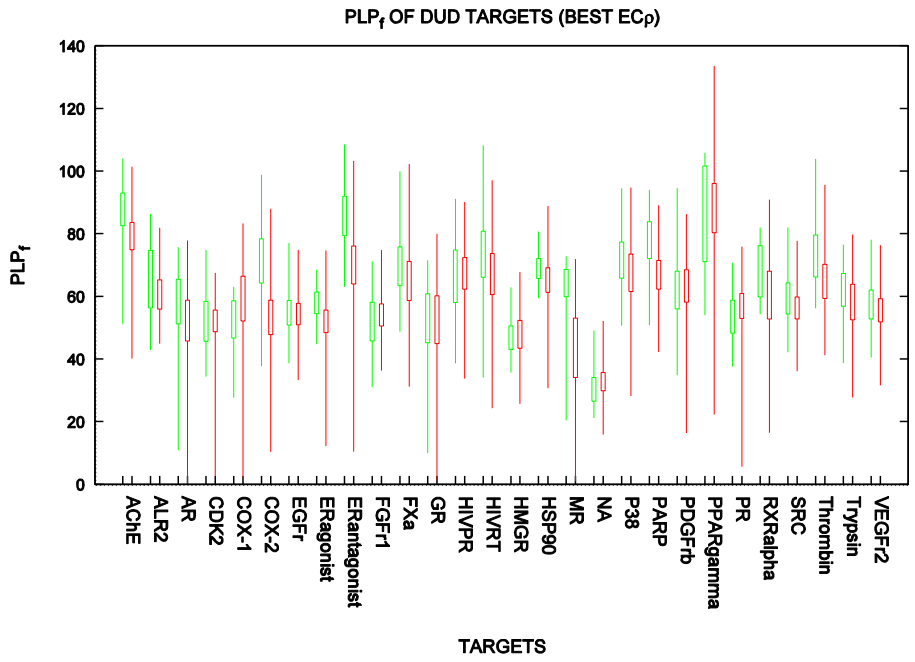


Figure 21 The distribution of the PLP_f scores in the DUD targets based on the most complementary docked poses. The box-plot represents the distribution of the PLP_f for the different DUD targets of ligand (green) versus the decoys (red) for the lowest ECρ value out of the 10 docked solutions per compound.

Next, we analyzed the relationship between successes in pose prediction (RMSD from crystallographic pose) and the ECρ of the corresponding poses. It can be observed that the best docking results (lowest RMSD) typically agree with the best ECρ (Figure 22). The majority of cases show that indeed good docking poses (low RMSD) reside in the area exhibiting electrostatic complementarity. Furthermore a correlation can be observed for these cases where worse binding modes also exhibit worse ECρ values. A few cases however do not exhibit a correlation. For these cases the points are not widely spread as the docking algorithm only returned very accurate binding modes with similar RMSD and ECρ values. This correlation between RMSD and ECρ is in agreement with the observation that selecting the docked poses with the best ECρ gives better performances on the DUD dataset compared to selecting the best docking score. Our results show that EleKit2 can indeed measure the complementarity and provides additional information to the docking simulations. The best approach is to prioritize per compound the docking mode with the highest complementarity.

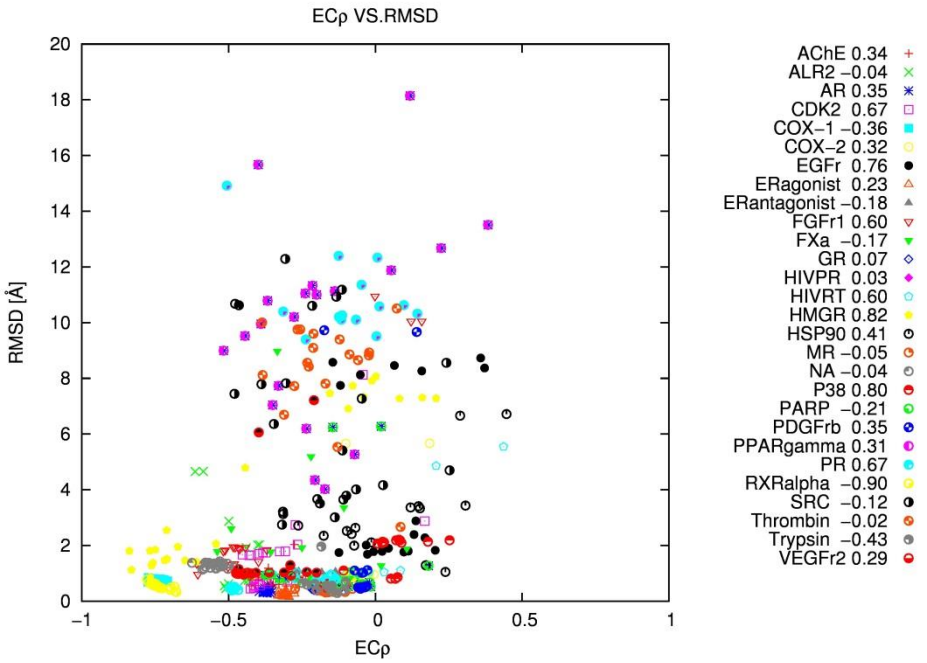


Figure 22 Good binding modes agree with complementarity ECρ values. The RMSD indicating the correctness of the binding mode is plotted against the corresponding ECρ values of one representative ligand for each of the 28 targets in the DUD dataset. The clusters in the lower-left region reveal the tendency that good binding modes agree with electrostatic complementarity. The in general positive values of the spearman rank-correlation coefficients between RMSD and ECρ further indicate this tendency.

However, all statistical analysis (student t-test) revealed that the PLP_f (ECρ) method was not significantly better than the other options (PLP_f or ECρ). Nevertheless, promising trends were observed in the boxplots analysing the differences in populations (Figure 21). Therefore enrichment studies were performed to analyse the increase in chance to identify an active inhibitor and remove decoy docking solutions at a given ECρ cut-off value.

The enrichment factor was calculated for the 28 DUD cases spanning a range of -0.5 to 0.5 ECρ as following:

$$EF = \frac{(A_F/T_F)}{(A/T)}$$

Where, EF is the Enrichment Factor. A is the number of the active ligands before filtering, A_F indicates the number of active ligands after filtering, and T_F is the total number of ligands after filtering, while T depicts the total number ligands before filtering. Furthermore, the fraction of retained active and total ligands were calculated since often enrichment factors can significantly increase while in fact the number of removed actives may be very high as well, indicating useless values. For an overview see Figure 23.

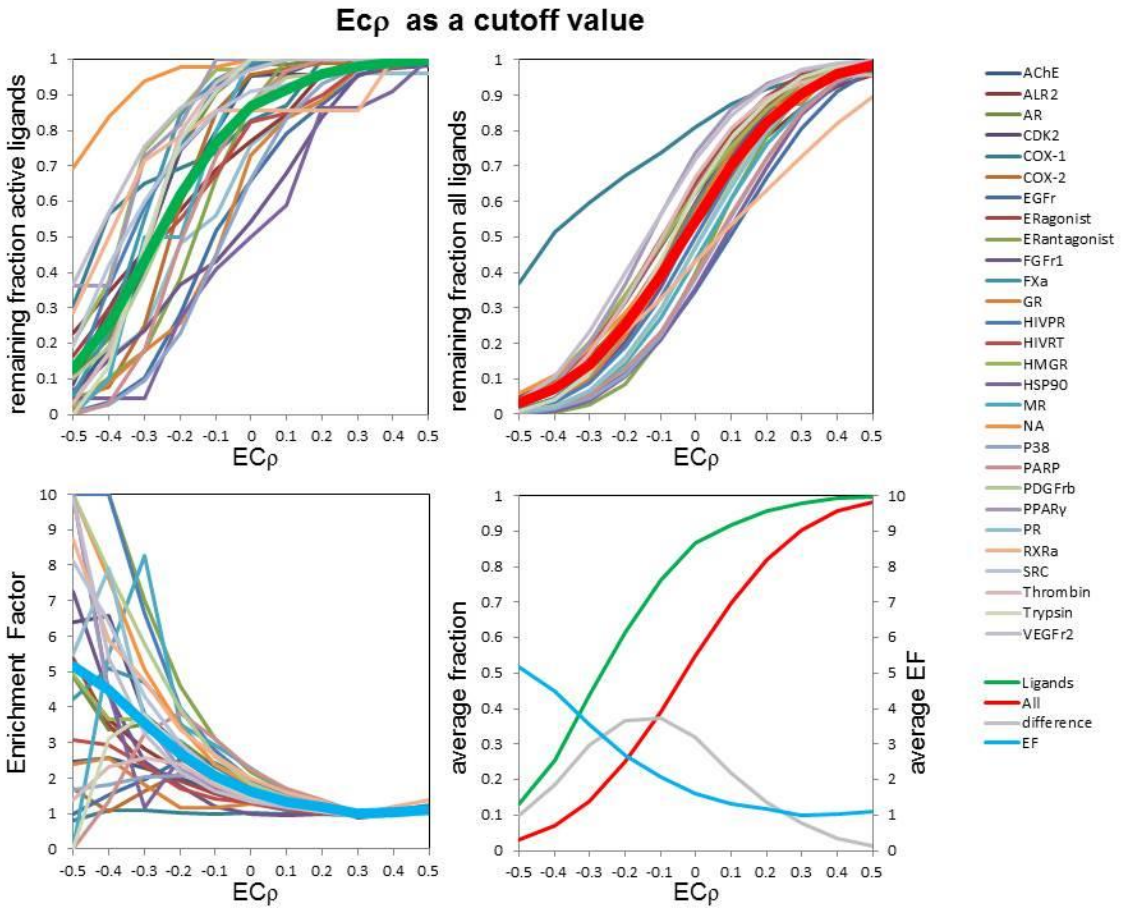


Figure 23 Overview of the enrichment and retainment factors of the DUD targets for different ECρ cut-off values. While at low ECρ values the enrichment factor is high, the number of retained active molecules is very low. Therefore, a cut-off value of -0.1 to 0.0 ECρ appears to be an optimum value to remove a significant amount of decoys without discarding too many active ligands

From this analysis, it is clear that indeed at low (-0.5) ECρ values, the EF is very high. However, only a few compounds remain. At -0.1 ECρ cutoff, almost half of the decoy population is removed without interfering too much with the active ligand population (80-90% remains). While, the EF at -0.1 ECρ cutoff appears to be rather low (EF 2.06), it is nevertheless a significant improvement as half of the docked solutions can be removed and the chance to identify active ligands in the remaining population almost doubles. The fact that no significant improvement could be observed using ROC plot AUC values can be explained by the low quality of the ECρ or PLP_f score as a ranking factor.

3.5 Conclusion

In our research, we have calculated the electrostatic complementarities of the small molecule ligands and receptor by using the popular continuum Poisson-Boltzmann electrostatic model. The Poisson-Boltzmann equation has already been extensively studied to investigate the long

range electrostatic interactions between biomolecules, mainly for protein-protein interactions. However, the research that investigates the protein and a small molecule ligand complex has been neglected. Hence, to check and emphasize the importance of the present of the binding of small molecule ligands in the computer-aided drug design, an open source tool, EleKit2 has been developed.

EleKit2 was built upon the existing computational software package of PDB2PQR [242] and Adaptive Poisson-Boltzmann Solver (APBS) [274]. PDB2PQR and APBS are commonly used continuum models for biomolecular simulation and modeling. By solving the PB equation numerically, EleKit2 is efficient in checking the electrostatic interactions between biomolecules in either solvated and desolvated conditions. EleKit2 calculates the electrostatic complementarity in the contacting surfaces of protein-small molecule ligands complexes in computational drug design. By inspecting the complementarity of the biomolecules complexes, EleKit2 has further demonstrated the existence of the high complementarity in contacting surfaces of biomolecules as previously claimed by Honig and Nicholls [202].

Using EleKit2 toolkit, we have performed the retrospective analysis of the Iridium receptor-ligand benchmark set and observed the presence of electrostatic complementarity in crystal structures. To evaluate the utility of electrostatic complementarity for virtual screening, we analyzed the DUD dataset using EleKit2 as a docking filter. Our results indicate that the best docking pose agrees with the best complementarity but not with the best docking score. The virtual screening performance appeared to have improved by selecting the docking poses based on the complementarity prior to their ranking but prove statistically insignificant. Nevertheless, enrichment studies indicated that using a cutoff value of ECρ -0.1, half of the docked solutions can be removed without influencing the active ligands, doubling the chance to identify active ligands. Therefore, we suggest to use EleKit2 as a docking post filter to remove compounds without electrostatic complementarity.

To conclude, EleKit2 toolkit is also useful for computational drug design. It helps to identify better docking poses, and thereby also improves virtual screening. Furthermore analysis revealed that although complementarity is present in nature it is not taken into account during docking simulations and it may thus be beneficial to include complementarity during the pose prediction to improve docking algorithms.

Chapter 4

Virtual screening targeting G9a SET domain

4.1 Introduction

G9a is a histone lysine methyltransferase that predominantly methylates the lysine K9 on histone H3. G9a plays a key role in transcriptional silencing of the chromatin by H3K9 methylation. The formation of the heterochromatin can trigger the disruption of gene expression and is related to cancer and diseases. G9a contains a SET domain that is responsible for the catalytic activity. The active site in the SET domain consists of a substrate binding site, where the lysine containing histone tails binds, and a cofactor site where the methyl donating SAM cofactor binds. The structure of the SET domain in complex with the substrate, substrate competing inhibitors, and cofactor has been well studied and has contributed to the rational development of inhibitors. However, many of the inhibitors are either not selective in targeting the G9a or toxic. Therefore, in order to develop better probes to study the G9a protein or to increase the therapeutic potential of G9a inhibitors, we aimed to design novel inhibitors belonging to different chemotypes and different mechanisms of action compared to those previously reported inhibitors.

4.2 Material and Methods

4.2.1 Enamine database

The Enamine druglike molecular database is supplied by Enamine Ltd. This library contains approximately 1800 000 commercially available chemical compounds that can be ordered for experimental testing. These compounds have been designed to exhibit favorable ADMET properties. The majority of the compounds have a molecular weight lower than 400 Da and a cLogP below 3 which indicates good solubility as well.

The compounds from the Enamine database are delivered in a 2D database file. As pharmacophore modelling is a 3D method, a 3D database is required. Therefore, we prepared

a 3D conformational dataset of this database by using the OMEGA [275] conformation generator.

4.2.2 Pharmacophore modelling

The first step in our VS approach is based on pharmacophore modeling. The Molecular Operating Environment (MOE) was used to manually generate the pharmacophore models, and use them as a query for searching Enamine database. Pharmacophore based post filtering of the docking results was performed as well using the MOE pharmacophore implementation on the absolute position of the docked ligands.

4.2.3 Docking

All docking simulations were performed using GOLD using the default virtual screening search parameters, with PLP_f as a scoring function. The crystal structure of the both the histone lysine methyltransferase G9a, and GLP (G9a-like protein) were downloaded from the PDB database (entries 3RJW [89] and 3SW9 [276] respectively). The PDB files were prepared for docking using the Protonate_3D algorithm implemented in MOE. Each ligand was docked 10 individual times. The simulation was executed parallelly on the RIKEN Integrated Cluster of Clusters (RICC) supercomputer, RIKEN, Japan.

4.2.4 AlphaLISA protocol

The G9a-inhibitory activity was evaluated using the AlphaLISA assay (PerkinElmer) [277, 278] at 500 μM. The acquired compounds were dissolved to a final concentration of 10mM in Dimethyl Sulphoxide (DMSO). From this stock solution, 1 μL was added in the measuring wells of white mOptiPlate-384. Next, 7 μL of G9a protein in assay buffer was added to the wells. After 10 minutes of incubation at room temperature, 1 μL of both buffered (50 mM Tris-HCl, pH 9.0, 50 mM NaCl, 0.01% Tween-20) b-H3 peptide (first 21 amino acids of histone 3) and SAM (Sigma) were added to the wells (final concentration of 50 nM and 15 μM , respectively). After 1 hour incubation at room temperature, the G9a enzymatic reaction was quenched by adding 7.5 μL acceptor and donor beads mixture in 1-fold epigenetic buffer (final concentration of 10 μg/ ml) to the wells, followed by 1 hour incubation in the dark at room temperature. Finally, the inhibitory signal is measured by using AlphaLISA (PerkinElmer Inc.). The signal is proportional to the level of substrate modification by G9a.

4.2.5 TruHits assay

The TruHits assay was designed as a counter assay AlphaLisa to identify false positives which are often observed during AlphaLISA. In the TruHits assay, the streptavidin-coated

donor beads and biotinylated acceptor beads are directly linked to generate the light signal. This assay can effectively identify compounds that interfere with the light signal produced during the assay. False positive compounds are typically either, colored compounds, insoluble compounds that scatter the produced light signal or singlet oxygen quenching compounds.

4.3 EleKit2

Except for pharmacophore post filtering, EleKit2 was one of the docking post filtering tools that was used during the virtual screening. The application of EleKit2 as a docking post filter is described in detail in Chapter 2. During this virtual screening experiment, all the binding modes of the compounds that are not complementary on the electrostatic level were eliminated. Only the remaining compounds were considered for the next step.

4.4 AlphaLISA assay

The AlphaLISA is a bioluminescent assay that relies on the bioconjugation of donor and acceptor beads resulting in light emission. The assay can be used to investigate the presence of methylated G9a substrate peptide. The principle of the AlphaLISA for G9a inhibitory activity is shown in Figure 24. The donor beads are coated with Streptavidin, which binds biotin that is linked to a peptide containing the H3K9 motif. On the other side, the acceptor beads are coated with an antibody that specifically recognizes the H3K9me3 motif. With the presence of the methylated H3K9 motif, the acceptor beads will bind on the substrate which is linked to the donor beads. When these donor beads are excited with a high energy laser light at 680 nm, singlet oxygen will be released and diffuse to the nearby acceptor beads. The maximal distance the singlet oxygen can travel is around 200 nm, thus the acceptors bead must be very close in order to trigger the excitation of the acceptor bead. The excitation of the acceptor beads produces a chemiluminescent signal at 615 nm. The light emission by the acceptor beads will be then be measured. The amount of the light that produced is proportional to the concentration of methylated substrate peptide. This concentration however is dependent on the activity of the G9a enzyme which is also present in the assay. Therefore, adding inhibitors that inhibit the G9a enzyme will result in lower concentrations of the methylated peptide substrate and lower light production by the acceptor beads compared to the control experiment.

In our research, hit molecules selected from the virtual screening were identified by AlphaLISA assay (in 100 µM or 500 µM) and further evaluated using the TruHits assay to remove false positives.

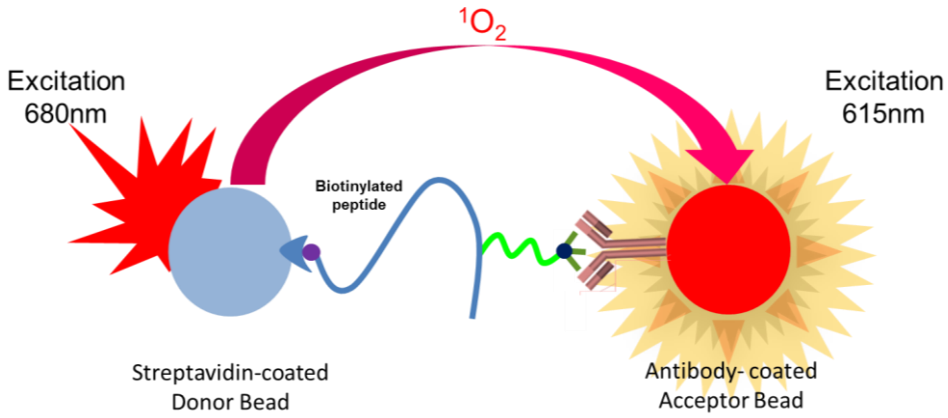


Figure 24 AlphaLISA assay. The figure shows the principle of AlphaLISA assay. The donor bead that coated with streptavidin is shows in blue. The acceptor bead coated with an antibody is shows in red. The Streptavidin-coated donor bead binds to the biotinylated peptide protein. When the methylated peptide protein present, both the donor and acceptor beads are bind to the methylated protein substrate with the acceptor bead binding to methyl group of the lysine tail of the peptide protein. After excitation by laser light at 680 nm, the singlet oxygen release and transfer from the donor bead to the acceptor bead. It results in the emission of the fluorescent light signal at 615 nm. The light signal emit is proportional to the amount of methylated G9a present in the assay.

4.5 Experiments and Result

4.5.1 Bioinformatical analysis of the G9a SET domain.

First, a bioinformatical analysis of the G9a was performed. The SET domain structure was stripped from all water molecules and small molecule ligands. Both the crystal structures of the histone lysine methyltransferase G9a with an inhibitor (pdb 3RJW) [279], and GLP's SET domain in complex with SEDnmt3aK44me0 (3SW9) [276] were used. Using FTMap, we analyzed the structure to identify the probable pockets. This revealed two major pockets. One pocket is corresponding to the cofactor binding site, where the SAM molecule is bound. The second pocket corresponds to the histone tail (substrate binding site). Next, we analyzed the amino acid conservation among SET domains. The amino acid sequence of the SET domain of 3RJW was chosen as a starting sequence and using BLAST all homologues sequences were collected and aligned using ClustalX [280]. Only G9a and GLP are closely related. Subsequently, the conservation was mapped onto the 3D 3RJW structure. This revealed that G9a and GLP are completely conserved in the substrate binding site. The cofactor site however revealed two amino acids (H1114L and Q1169R) which are different between G9a

and GLP (see Figure 24). This is stimulating to target the SET domain cofactor site to develop a specific inhibitor for either G9a or GLP.

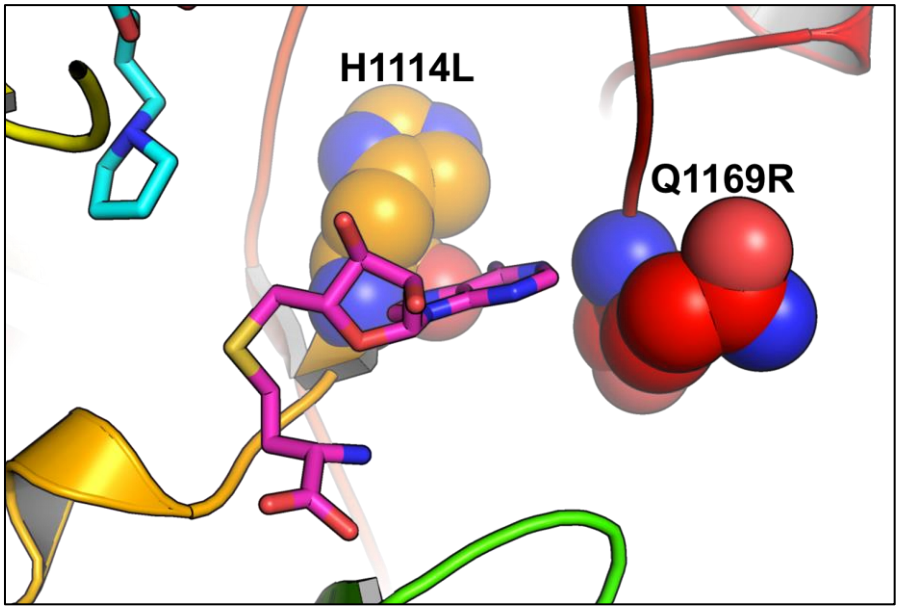


Figure 25 The unconserve region of the G9a and GLP in cofactor site. The figure shows the overview of the SAM molecule cofactor site of both the G9a (3RJW) and GLP (3SW9). The sphere representation of H1114L and Q1169R indicate the important amino acid that are different in the G9a and GLP which give the information for designing the specific inhibitor.

4.5.2 Virtual screening

To target the G9a using virtual screening, we have used a so-called funnel principle VS approach in which multiple complementary VS methods are consecutively used for drug discovery (see Figure 25). The 3RJW was used as a receptor protein in cofactor site, while the 3SW9 was used as a receptor protein for the drug design in protein substrate site since the G9a crystal structures are partially unresolved and the amino acids lining this binding site are 100% identical between G9a and GLP.

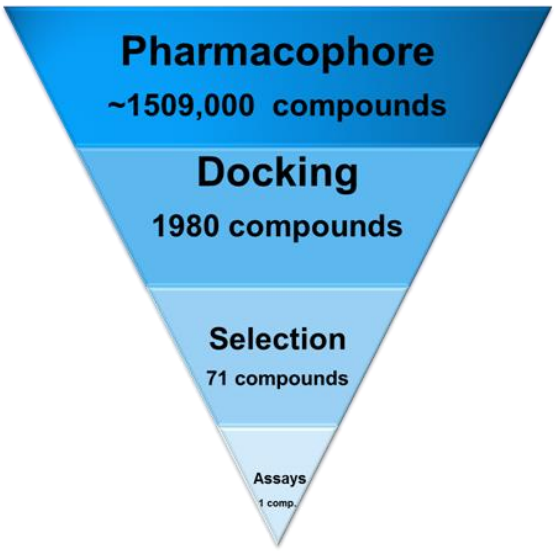


Figure 26 The funnel based virtual screening. Overview of the VS process as applied during the G9a targeted *in silico* screen.

The first step of the funnel based approach was to employ pharmacophore modelling. Receptor based pharmacophore queries were created for both the cofactor and the substrate site by using MOE. In cofactor site, the pharmacophore queries were created based on the key interactions of the SAM molecule. In the substrate site, the pharmacophore query was created by clustering to the common key interactions of the lysine tail of the substrate site with the UNC0638. The key hydrogen bonds, hydrophobic and aromatic interactions were investigated and represented with pharmacophore features in the pharmacophore queries (see Figure 27).

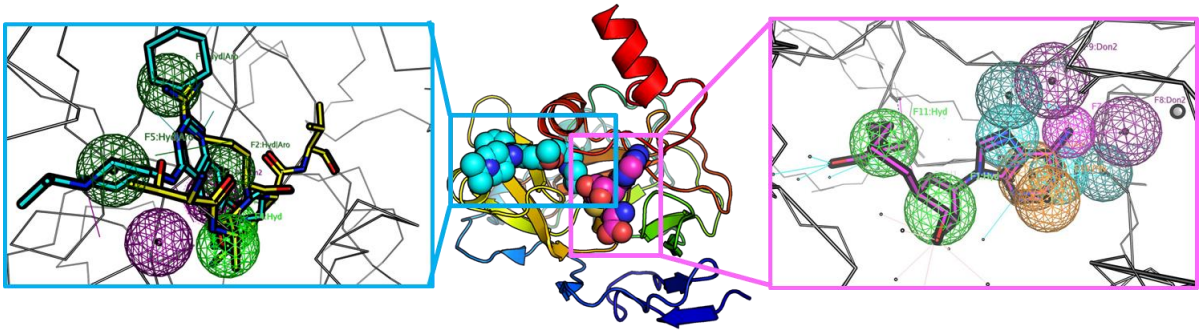


Figure 27 Pharmacophore features of active sites, G9a. The pharmacophore model (in blue) is generated on the protein substrate site of G9a pharmacophore features of substrate site, G9a. This pharmacophore model consists of 7 features: three hydrophobic/aromatic rings (dark green), two hydrophobic groups (green), two hydrogen bond donors (purple) with their respective projected features determining the orientation of the interaction. The pharmacophore model is built on the cofactor site of G9a. The pharmacophore model (in pink) consists of 11 features: two aromatic rings (brown), two hydrophobic groups (light green), three hydrogen bond donors (purple), four hydrogen bond acceptors (cyan) with their respective projected features determining the orientation of the interaction.

With these 2 different pharmacophore queries, the 3D conformational Enamine database was queried in order to identify compounds that harbour the correct chemical functionalities in the correct 3D conformation to potentially interact with respectively the cofactor or the substrate site on G9a. The hits that generated at the cofactor site and the substrate site are 1329 and 1932 respectively.

As pharmacophore modelling does not incorporate the receptor protein explicitly, a molecular docking simulation was chosen as the second step in our funnel approach as this approach does use the receptor structure explicitly. The hits of the pharmacophore queries were docked to their respective sites using GOLD [272] with default virtual screening parameters and PDB entries 3RJW and 3SW9 as G9a receptor structures. In order to identify compounds that do not make the key interactions at their receptor site, the pharmacophore query was used as a post filter with absolute position restraints to remove all docked molecules that do not make the desired interactions.

The remaining docked molecules were post filtered further with EleKit2 to remove undesired binding modes. A final *in cerebro* selection was made using expert knowledge incorporating the docking scores, the electrostatic complementarity and the chemotype composition. At the end, 71 molecules were selected and ordered from for further biological evaluation by using AlphaLISA.

4.5.3 AlphaLISA

The final and the most crucial step in the in silico screen targeting G9a is experimental evaluation of the selected compounds. The selected compounds were acquired via Namiki Shoji and tested at 500 µM. From the 71 molecules, 5 molecules showed significant inhibitory potency. However, as the AlphaLISA assay is sensitive to quenching by chemical compounds, these 5 compounds were assayed using the ALPHA TruHits assay. Finally 3 remaining compounds were identified as true positives. These compounds (structure not shown due to IP reasons) are still weak inhibitors but are a very different chemotype compared to the previously reported inhibitors, and could be of major pharmaceutical interest after optimization. According to our virtual screening approach 2 of the inhibitors likely target the cofactor site, the remaining inhibitor the substrate site.

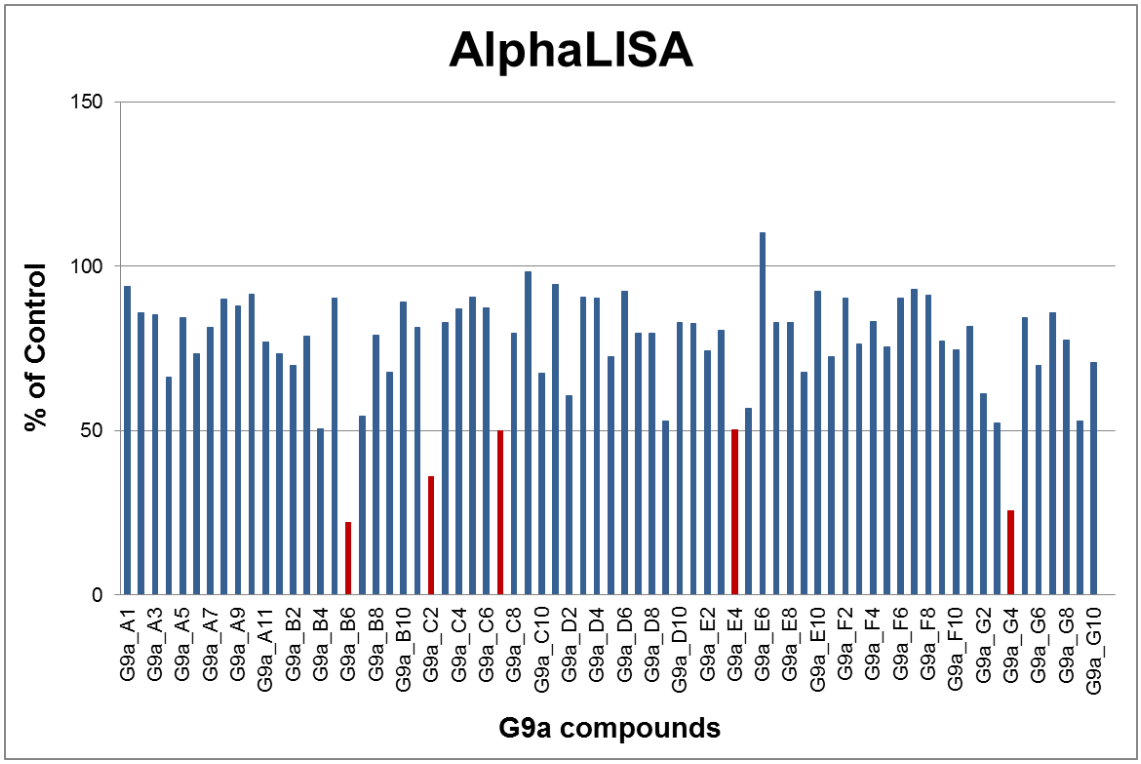


Figure 28 Percentage of control for G9a Compounds. The percentage of control is calculated for each compound. Compounds causing a significant lower signal than the control experiment are considered to be active inhibitors. The red colored bars represent the most promising compounds for which derivatives were sent for testing using bioassays.

4.5.4 SAR by catalog

The Structure-Activity Relationship (SAR) analysis is an essential step for the rational optimization of compounds. Typically a set of similar compounds is required, with a broad range of activity. This set typically originates from organic synthesis of derivative compounds. We however have circumvented the organic chemistry step by following a SAR by Catalog approach where the derivatives are selected from a catalog of commercially available molecules.

Using the MACCS fingerprint [135] and Tanimoto cutoff of 0.7 derivative compounds of the 3 hit compounds were identified from the ZINC database, a collection of all commercially available molecules. Out of the top 100 compounds (per query), a total of 34 diverse compounds were selected and ordered. The 34 derivatives that were selected from these 3 hits were again evaluated using the AlphaLISA assay at a concentration of 100 µM (see Figure 29), as well as a TruHits assay. One molecule was found to be very potent and was subjected to an AlphaLISA based titration experiment to determine the potency. The hit (G9a_der14)

demonstrates a good potency with IC_{50} 13.5 µM. For a graphical depiction of the binding mode see Figure 30.

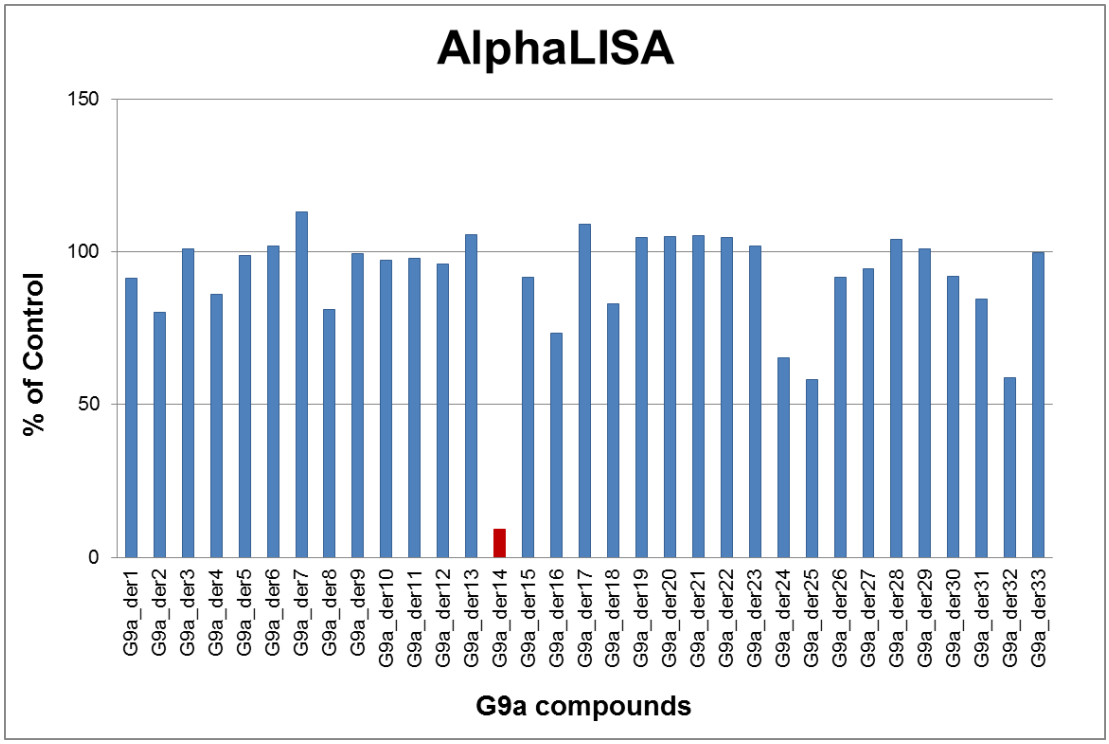


Figure 29 Percentage of control for G9a compounds. The percentage of inhibition is calculated for each compound. Compounds causing a significant higher signal than the control experiment are considered to be active inhibitors. The blue color bars show the AlphaLISA assay while red color bar indicates the stronger hits of the assay. The bars under the 50 % of control represent the most promising compounds for which derivatives are tested in SAR by catalog approach.

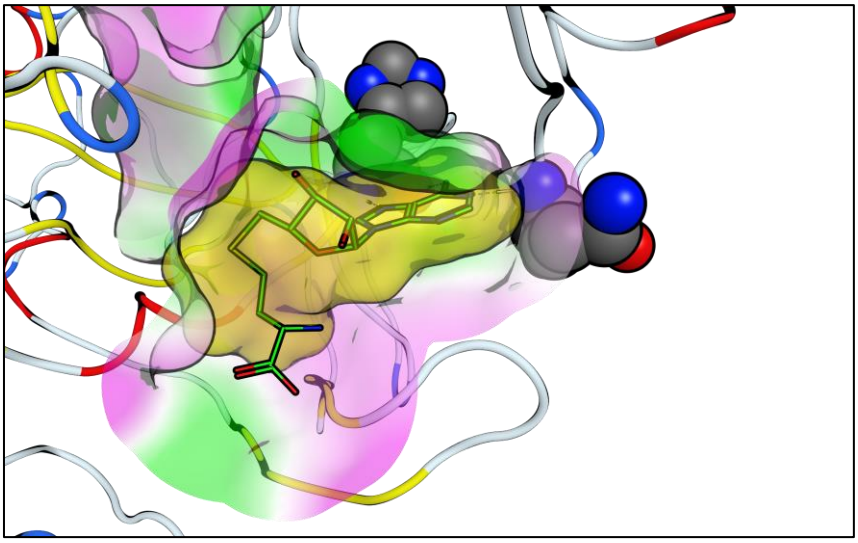


Figure 30 Binding mode of G9a_der14. Zoom in on the SAM cofactor site of G9a. The two amino acids that are different in GLP are depicted as spheres. The SAM cofactor is depicted with green carbon sticks. The G9a_der14 molecule is represented as a yellow surface, the overall structure however remains confidential.

4.6 Discussion

A funnel based virtual screening approach was used for targeting the G9a. The funnel approach incorporates receptor based pharmacophore queries, molecular docking simulations and post filtering based on the pharmacophore queries as well as the complementarity in electrostatic potential fields. We not only targeted the classic G9a SET domain substrate site, but also the G9a SET domain SAM cofactor site.

Compounds binding to the cofactor site would prevent the SAM molecule to enter and preventing the methyl to be transferred to the substrate peptide. Compounds binding to the substrate site would also prevent the substrate peptide to enter the active site of the protein. Compounds for any of these 2 sites would therefore be valuable G9a inhibitors. The protein substrate binding site has already been targeted before [98, 279, 281-283] with many inhibitors reported already. However, these inhibitors all belong to the same chemotype. For the cofactor site, very few inhibitors have been reported so far, and for non of them the mechanism of action has been proven. In this research our aim was to identify novel inhibitors, dissimilar from previously reported inhibitors and preferably having a novel mechanism of action by competing with the SAM molecule as bioinformatical analysis revealed this option to develop a G9A/GLP specific inhibitor.

Using the funnel based virtual screening approach, a total of 71 compounds were identified and evaluated *in vitro*. The AlphaLISA assay was performed to detect and quantify the concentration of the methylated H3K9 histone tail, after processing by G9a.

The initial screen delivered 3 true positive G9a inhibitors at weak potency. It should however be noted that all 3 inhibitors had a molecular weight below 350Da and should be in fact considered fragments rather than druglike molecules. In fragment based drug design, it is often observed that the initial molecules have a very weak potency and can later be optimized into fully active compounds [284]. Next, the derivatives of the three compounds were selected and ordered for further experimental testing.

In the second round, a total of the 34 derivatives were examined by the same procedure with AlphaLISA assay at 100 µM concentration combined with the TruHits assay. The assay was performed at 5 times lower concentration than the first screen to identify the most promising

compound. This revealed 1 potent hit compound. This compound possessed high potency with 13.5 µM. By design, this compound should target the cofactor site. This is a breakthrough as the inhibitor discovered was the first compound with the different chemotype found after the Chaetocin, for which the mechanism of action remains unknown. This novel class of compound could be of pharmaceutical interest, thus a final step of the optimization need to be carried out before clinical testing. As a remark, this compound is still below 350Da and signification improvements can be expected in the next round of optimization.

Currently, this compound has been subject to another round of SAR by catalogue. This time however the SAR information is taken into account and only derivatives which are more likely to be active will be ordered. Furthermore, since this compound is still fragment like rather than druglike, larger and more decorated derivatives will also be identified for subsequent testing.

In parallel, more experiments will be performed to analyze the true mechanism of action. For this substrate peptide and cofactor competition experiments will be performed, as well as attempting co-crystallization.

Chapter 5

ACPC

ACPC, short for Autocorrelation of Partial Charges [284], is a rotation-translation invariant molecular descriptor of partial charges that can be used for ligand-based virtual screening. Similarity searching is a commonly applied method in ligand based virtual screening and multiple methods, either based on molecular fingerprints or molecular descriptors. ACPC creates a molecular descriptor based on the distribution of partial charges 3D. From the similarity search methods currently available, none uses this information. ACPC is an open source tool which uses the effective autocorrelation function and linear binning to code all atoms of a molecule into two rotation-translation invariant vectors. Using a single molecule entry, a database is screened and a novel in house developed scoring function assesses the similarity. Next to the development of this novel approach, ACPC has also been compared with other open sources molecular similarity search tools including MACCS, Pharao and Shape-it. This comparison indicated that ACPC is not only faster than the others; it is also more powerful during retrospective ligand-based virtual screening experiments.

For this research, I have contributed by preparing all the datasets for the experiment. The DUD dataset [285] was used for the configuration of the ACPC. DUD consists of the 40 targets with 2950 ligands. Every single active ligand has 36 decoys (inactive) molecules that are physically similar but topologically different. For this dataset, I have generated the conformers of all the ligands and decoys by using OMEGA v2.4.6 [128]. In this experiment, 2 different dataset were generated, one consisting one conformer (1conf) and another with 25 conformers (25conf) per molecule. The 1conf dataset comprised the lowest energy conformer per molecule. This dataset was first filtered to obtain unique SMILES string (only the first smile string among the similar strings was kept) by Open Babel [286] to prevent multiple entries of the same molecule to be present. Subsequently, the partial charges were assigned to each molecule by using the default force-field (MMFF94x) from MOE [287]. The size of "1 conf" is relatively small and it enables the experiment of the four different software (ACPC

[284], Pharao [288], MACCS [289] and Shape-it [290]) to run smoothly on a cluster environment. The second dataset “25 conf” comprised of 25 low energy conformers per molecule.

Chapter 6

Summary

Electrostatic complementarity plays a key role in molecular recognition. Other than shape complementarity, the long range electrostatic interactions also exhibit complementarity for optimal recognition at the receptor-ligand interface. We believe that electrostatic complementarity can be a crucial filter for virtual screening. Thus, a Poisson Boltzmann toolkit, EleKit2 is developed. EleKit2 is able to reveal the presence of electrostatic complementarity at the receptor-ligand interface. The electrostatic complementarity was employed to prioritize GOLD docking results. The evaluation of EleKit2 shows that the inclusion of electrostatic complementarity can indeed improve the overall virtual screening process. While the enrichment studies have shown that when applying the ECρ cutoff value of -0.1, the chance of identifying the active ligands are increased by a factor 2. Hence, electrostatic complementarity can act as a robust scoring filter in any ligand placement search algorithms to improve the simulation performance.

In addition, EleKit2 was also utilized during our virtual screening experiment to identify novel inhibitors targeting G9a. G9a is a histone lysine methyltransferase and involved in epigenetics. The methylation of H3K9 will result in numerous epigenetic changes and causing the major damage of the cell division and development which related to cancer.

There are few reported inhibitors targeting the G9a protein, and they are mostly belonging to the same chemotype and bind to the G9a SET domain substrate site. In the second part of my thesis, I have performed a virtual screening that relied on pharmacophore modelling, combined with molecular docking simulations and EleKit2-based post-filtering. This resulted in 3 initial hit compounds. Subsequently derivatives of the first hit compounds were computationally identified and evaluated *in vitro*. The most potent derivative, which is designed to bind into the G9a SET cofactor site, possesses a promising potency with an IC50 of 13.5 μM. This novel inhibitor is still a fragment-like molecule and the optimisation and

further evaluation is still on-going. Nevertheless this compound is a breakthrough for G9a research as currently it is a completely novel chemical class, with very promising potency and a novel mechanism of action.

References

1. Waddington, C.H., *The Epigenotype.* International Journal of Epidemiology, 2012. **41**(1): p. 10-13.
2. Mayr, E., *This Is Biology: The Science of the Living World.* MA: Belknap of Harvard UP. Cambridge, 1997.
3. Egger, G., et al., *Epigenetics in human disease and prospects for epigenetic therapy.* Nature, 2004. **429**(6990): p. 457-463.
4. Shen, Q., H. Jin, and X. Wang, *Epidermal Stem Cells and Their Epigenetic Regulation.* International Journal of Molecular Sciences, 2013. **14**(9): p. 17861-17880.
5. Sandoval, J., et al., *Epigenetic biomarkers in laboratory diagnostics: emerging approaches and opportunities.* Expert Review of Molecular Diagnostics, 2013. **13**(5): p. 457-471.
6. Feinberg, A.P. and B. Tycko, *The history of cancer epigenetics.* Nat Rev Cancer, 2004. **4**(2): p. 143-153.
7. Feinberg, A.P., R. Ohlsson, and S. Henikoff, *The epigenetic progenitor origin of human cancer.* Nat Rev Genet, 2006. **7**(1): p. 21-33.
8. Herceg, Z. and P. Hainaut, *Genetic and epigenetic alterations as biomarkers for cancer detection, diagnosis and prognosis.* Molecular Oncology, 2007. **1**(1): p. 26-41.
9. Zaratiegui, M., D.V. Irvine, and R.A. Martienssen, *Noncoding RNAs and Gene Silencing.* Cell, 2007. **128**(4): p. 763-776.
10. Flanagan, J. and L. Wild, *An epigenetic role for noncoding RNAs and intragenic DNA methylation.* Genome Biology, 2007. **8**(6): p. 307.
11. Esteller, M., *Epigenetics in Cancer.* New England Journal of Medicine, 2008. **358**(11): p. 1148-1159.
12. Jones, P.A. and S.B. Baylin, *The Epigenomics of Cancer.* Cell, 2007. **128**(4): p. 683-692.
13. Weichenhan, D. and C. Plass, *The evolving epigenome.* Hum Mol Genet, 2013.
14. Jenuwein, T. and C.D. Allis, *Translating the Histone Code.* Science, 2001. **293**(5532): p. 1074-1080.
15. Kouzarides, T., *Chromatin Modifications and Their Function.* Cell, 2007. **128**(4): p. 693-705.
16. Li, B., M. Carey, and J.L. Workman, *The Role of Chromatin during Transcription.* Cell, 2007. **128**(4): p. 707-719.
17. Allfrey, V.G., R. Faulkner, and A.E. Mirsky, *ACETYLATION AND METHYLATION OF HISTONES AND THEIR POSSIBLE ROLE IN THE REGULATION OF RNA SYNTHESIS.* Proceedings of the National Academy of Sciences, 1964. **51**(5): p. 786-794.
18. Eberharter, A. and P.B. Becker, *Histone acetylation: a switch between repressive and permissive chromatin.* EMBO reports, 2002. **3**(3): p. 224-229.
19. Goll, M.G. and T.H. Bestor, *Histone modification and replacement in chromatin activation.* Genes Dev, 2002. **16**(14): p. 1739-1742.
20. Sterner, D.E. and S.L. Berger, *Acetylation of Histones and Transcription-Related Factors.* Microbiology and Molecular Biology Reviews, 2000. **64**(2): p. 435-459.
21. Zhang, Y. and D. Reinberg, *Transcription regulation by histone methylation: interplay between different covalent modifications of the core histone tails.* Genes Dev, 2001. **15**(18): p. 2343-2360.

22. Shilatifard, A., *Chromatin Modifications by Methylation and Ubiquitination: Implications in the Regulation of Gene Expression.* Annual Review of Biochemistry, 2006. **75**(1): p. 243-269.
23. Nowak, S.J. and V.G. Corces, *Phosphorylation of histone H3: a balancing act between chromosome condensation and transcriptional activation.* Trends in Genetics, 2004. **20**(4): p. 214-220.
24. Nathan, D., et al., *Histone sumoylation is a negative regulator in Saccharomyces cerevisiae and shows dynamic interplay with positive-acting histone modifications.* Genes Dev, 2006. **20**(8): p. 966-976.
25. Hassa, P.O., et al., *Nuclear ADP-Ribosylation Reactions in Mammalian Cells: Where Are We Today and Where Are We Going?* Microbiology and Molecular Biology Reviews, 2006. **70**(3): p. 789-829.
26. Cuthbert, G.L., et al., *Histone Deimination Antagonizes Arginine Methylation.* Cell, 2004. **118**(5): p. 545-553.
27. Nelson, C.J., H. Santos-Rosa, and T. Kouzarides, *Proline Isomerization of Histone H3 Regulates Lysine Methylation and Gene Expression.* Cell, 2006. **126**(5): p. 905-916.
28. Arrowsmith, C.H. and e. al., *Epigenetic protein families: a new frontier for drug discovery.* 2012. **11**(5): p. 384-400.
29. Egger, G., et al., *Epigenetics in human disease and prospects for epigenetic therapy.* Nature, 2004. **429**(6990): p. 437-463.
30. Barski, A., et al., *High-Resolution Profiling of Histone Methylations in the Human Genome.* Cell, 2007. **129**(4): p. 823-837.
31. Liu, F., et al., *Discovery of an in Vivo Chemical Probe of the Lysine Methyltransferases G9a and GLP.* Journal of Medicinal Chemistry, 2013. **56**(21): p. 8931-8942.
32. Milner, C.M. and R.D. Campbell, *The G9a gene in the human major histocompatibility complex encodes a novel protein containing ankyrin-like repeats.* Biochem J, 1993. **290 (Pt 3)**: p. 811-8.
33. Milner, C.M. and R.D. Campbell, *The G9a gene in the human major histocompatibility complex encodes a novel protein containing ankyrin-like repeats.* Biochem. J. , 1993. **290**: p. 811-818.
34. Noma, K.-i., C.D. Allis, and S.I.S. Grewal, *Transitions in Distinct Histone H3 Methylation Patterns at the Heterochromatin Domain Boundaries.* Science, 2001. **293**(5532): p. 1150-1155.
35. Litt, M.D., et al., *Correlation Between Histone Lysine Methylation and Developmental Changes at the Chicken β-Globin Locus.* Science, 2001. **293**(5539): p. 2453-2455.
36. Tachibana, M., et al., *G9a histone methyltransferase plays a dominant role in euchromatic histone H3 lysine 9 methylation and is essential for early embryogenesis.* Genes Dev, 2002. **16**(14): p. 1779-91.
37. Jones, D.O., I.G. Cowell, and P.B. Singh, *Mammalian chromodomain proteins: their role in genome organisation and expression.* BioEssays, 2000. **22**(2): p. 124-137.
38. Wallrath, L.L., *Unfolding the mysteries of heterochromatin.* Current Opinion in Genetics & Development, 1998. **8**(2): p. 147-153.
39. Cavalli, G. and R. Paro, *Chromo-domain proteins: linking chromatin structure to epigenetic regulation.* Current Opinion in Cell Biology, 1998. **10**(3): p. 354-360.
40. Gazzar, M.E., et al., *G9a and HP1 Couple Histone and DNA Methylation to TNFα Transcription Silencing during Endotoxin Tolerance.* Journal of Biological Chemistry, 2008. **283**(47): p. 32198-32208.

41. Dormann, H.L., et al., *Dynamic Regulation of Effector Protein Binding to Histone Modifications: The Biology of HP1 Switching.* Cell Cycle, 2006. **5**(24): p. 2842-2851.
42. Tachibana, M., et al., *Histone methyltransferases G9a and GLP form heteromeric complexes and are both crucial for methylation of euchromatin at H3-K9.* Genes Dev, 2005. **19**(7): p. 815-26.
43. Ogawa, H., et al., *A Complex with Chromatin Modifiers That Occupies E2F- and Myc-Responsive Genes in G0 Cells.* Science, 2002. **296**(5570): p. 1132-1136.
44. Weiss, T., et al., *Histone H1 variant-specific lysine methylation by G9a/KMT1C and Glp1/KMT1D.* Epigenetics & Chromatin, 2010. **3**(1): p. 7.
45. Tachibana, M., et al., *Set domain-containing protein, G9a, is a novel lysine-preferring mammalian histone methyltransferase with hyperactivity and specific selectivity to lysines 9 and 27 of histone H3.* J Biol Chem, 2001. **276**(27): p. 25309-17.
46. Tschiersch, B., *The protein encoded by the Drosophila position-effect variegation suppressor gene Su(var)3-9 combines domains of antagonistic regulators of homeotic gene complexes.* EMBO J., 1994. **13**: p. 3822-3831.
47. Jones, R.S. and W.M. Gelbart, *The DrosophilaPolycomb -group gene Enhancer of zeste contains a region with sequence similarity to trithorax.* Mol. Cell. Biol., 1993. **13**: p. 6357-6366.
48. Stassen, M.J., et al., *The Drosophilatrithorax protein contains a novel variant of the nuclear receptor type DNA binding domain and an ancient conserved motif found in other chromosomal proteins.* Mech. Dev., 1995. **52**: p. 209-223.
49. Jenuwein, T., et al., *SET-domain proteins modulate chromatin domains in eu- and heterochromatin.* Cell. Mol. Life Sci., 1998. **54**: p. 80-93.
50. O'Carroll, D., et al., *Isolation and Characterization ofSuv39h2, a Second Histone H3 Methyltransferase Gene That Displays Testis-Specific Expression.* Mol Cell Biol, 2000. **20**(24): p. 9423-9433.
51. Yoon, K.-A., et al., *Novel polymorphisms in the SUV39H2 histone methyltransferase and the risk of lung cancer.* Carcinogenesis, 2006. **27**(11): p. 2217-2222.
52. Strahl, B.D., et al., *Methylation of histone H3 at lysine 4 is highly conserved and correlates with transcriptionally active nuclei in Tetrahymena.* Proceedings of the National Academy of Sciences, 1999. **96**(26): p. 14967-14972.
53. Cui, X., *Association of SET domain and myotubularin-related proteins modulates growth control.* Nature Genet., 1998. **18**: p. 331-337.
54. Trievel, R.C., et al., *Structure and Catalytic Mechanism of a SET Domain Protein Methyltransferase.* Cell, 2002. **111**(1): p. 91-103.
55. Lachner, M., et al., *Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins.* Nature, 2001. **410**(6824): p. 116-120.
56. Kouzarides, T., *Histone methylation in transcriptional control.* Current Opinion in Genetics & Development, 2002. **12**(2): p. 198-209.
57. Jenuwein, T., et al., *SET domain proteins modulate chromatin domains in eu- and heterochromatin.* Cellular and Molecular Life Sciences CMLS, 1998. **54**(1): p. 80-93.
58. Purcell, D.J., et al., *A Distinct Mechanism for Coactivator versus Corepressor Function by Histone Methyltransferase G9a in Transcriptional Regulation.* Journal of Biological Chemistry, 2011. **286**(49): p. 41963-41971.
59. Feldman, N., et al., *G9a-mediated irreversible epigenetic inactivation of Oct-3/4 during early embryogenesis.* Nat Cell Biol, 2006. **8**(2).
60. Tachibana, M., et al., *G9a histone methyltransferase plays a dominant role in euchromatic histone H3 lysine 9 methylation and is essential for early embryogenesis.* Genes Dev, 2002. **16**: p. 1779 - 1791.

61. Tachibana, M., et al., *Histone methyltransferases G9a and GLP form heteromeric complexes and are both crucial for methylation of euchromatin at H3-K9.* Genes Dev, 2005. **19**: p. 815 - 826.
62. Wagschal, A., et al., *G9a Histone Methyltransferase Contributes to Imprinting in the Mouse Placenta.* Mol Cell Biol, 2008. **28**(3): p. 1104-1113.
63. Xin, Z., et al., *Role of Histone Methyltransferase G9a in CpG Methylation of the Prader-Willi Syndrome Imprinting Center.* Journal of Biological Chemistry, 2003. **278**(17): p. 14996-15000.
64. Ohhata, T., et al., *X-inactivation is stably maintained in mouse embryos deficient for histone methyl transferase G9a.* Genesis, 2004. **40**(3): p. 151-156.
65. Lee, D.Y., et al., *Histone H3 Lysine 9 Methyltransferase G9a Is a Transcriptional Coactivator for Nuclear Receptors.* Journal of Biological Chemistry, 2006. **281**(13): p. 8476-8485.
66. Huang, N., et al., *Two distinct nuclear receptor interaction domains in NSD1, a novel SET protein that exhibits characteristics of both corepressors and coactivators.* EMBO J, 1998. **17**(12): p. 3398-3412.
67. Thomas, L.R., et al., *Functional analysis of histone methyltransferase g9a in B and T lymphocytes.* J Immunol, 2008. **181**(1): p. 485-93.
68. Feldman, N., et al., *G9a-mediated irreversible epigenetic inactivation of Oct-3/4 during early embryogenesis.* Nat Cell Biol, 2006. **8**(2): p. 188-94.
69. Schaefer, A., et al., *Control of Cognition and Adaptive Behavior by the GLP/G9a Epigenetic Suppressor Complex.* Neuron, 2009. **64**(5): p. 678-691.
70. Cormier-Daire, V., et al., *Cryptic terminal deletion of chromosome 9q34: a novel cause of syndromic obesity in childhood?* Journal of Medical Genetics, 2003. **40**(4): p. 300-303.
71. Verhoeven, W.M.A., T. Kleefstra, and J.I.M. Egger, *Behavioral phenotype in the 9q subtelomeric deletion syndrome: A report about two adult patients.* American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 2010. **153B**(2): p. 536-541.
72. Kleefstra, T., et al., *Further clinical and molecular delineation of the 9q subtelomeric deletion syndrome supports a major contribution of EHMT1 haploinsufficiency to the core phenotype.* Journal of Medical Genetics, 2009. **46**(9): p. 598-606.
73. Jamie, M.K., et al., *Epigenetic Regulation of Learning and Memory by Drosophila EHMT/G9a.* PLoS Biology, 2011. **9**(1).
74. Kramer, J.M. and H. van Bokhoven, *Genetic and epigenetic defects in mental retardation.* The International Journal of Biochemistry & Cell Biology, 2009. **41**(1): p. 96-107.
75. Kondo, Y., et al., *Alterations of DNA methylation and histone modifications contribute to gene silencing in hepatocellular carcinomas.* Hepatology Research, 2007. **37**(11): p. 974-983.
76. Huang, J., et al., *G9a and Glp Methylate Lysine 373 in the Tumor Suppressor p53.* Journal of Biological Chemistry, 2010. **285**(13): p. 9636-9641.
77. Watanabe, H., et al., *Deregulation of histone lysine methyltransferases contributes to oncogenic transformation of human bronchoepithelial cells.* Cancer Cell International, 2008. **8**(1): p. 15.
78. Yuan, Y., et al., *Gossypol and an HMT G9a inhibitor act in synergy to induce cell death in pancreatic cancer cells.* Cell Death Dis, 2013. **4**: p. e690.
79. Vandel, L. and D. Trouche, *Physical association between the histone acetyl transferase CBP and a histone methyl transferase.* EMBO reports, 2001. **2**(1): p. 21-26.

80. Steindl, T.M., et al., *Parallel Screening: A Novel Concept in Pharmacophore Modeling and Virtual Screening†.* Journal of Chemical Information and Modeling, 2006. **46**(5): p. 2146-2157.
81. Lengauer, T. and M. Rarey, *Computational methods for biomolecular docking.* Current Opinion in Structural Biology, 1996. **6**(3): p. 402-406.
82. Gardiner, D.M., P. Waring, and B.J. Howlett, *The epipolythiodioxopiperazine (ETP) class of fungal toxins: distribution, mode of action, functions and biosynthesis.* Microbiology, 2005. **151**(4): p. 1021-1032.
83. Strahl, B.D. and C.D. Allis, *The language of covalent histone modifications.* Nature, 2000/01/06. **403**(6765).
84. Kubicek, S., et al., *Reversal of H3K9me2 by a Small-Molecule Inhibitor for the G9a Histone Methyltransferase.* Mol Cell. **25**(3): p. 473-481.
85. Kubicek, S., et al., *Reversal of H3K9me2 by a Small-Molecule Inhibitor for the G9a Histone Methyltransferase.* Mol Cell, 2007. **25**(3): p. 473-481.
86. Chang, Y., et al., *Structural basis for G9a-like protein lysine methyltransferase inhibition by BIX-01294.* Nat Struct Mol Biol, 2009. **16**(3): p. 312-7.
87. Quinn, A.M., et al., *A chemiluminescence-based method for identification of histone lysine methyltransferase inhibitors.* Molecular BioSystems, 2010. **6**(5): p. 782-788.
88. Vedadi, M., et al., *A chemical probe selectively inhibits G9a and GLP methyltransferase activity in cells.* Nat Chem Biol, 2011/08. **7**(8).
89. Vedadi, M., et al., *A chemical probe selectively inhibits G9a and GLP methyltransferase activity in cells.* Nat Chem Biol, 2011. **7**(8): p. 566-74.
90. Liu, F., et al., *Discovery of a 2,4-diamino-7-aminoalkoxyquinazoline as a potent and selective inhibitor of histone lysine methyltransferase G9a.* J Med Chem, 2009. **52**(24): p. 7950-3.
91. Liu, F., et al., *Protein Lysine Methyltransferase G9a Inhibitors: Design, Synthesis, and Structure Activity Relationships of 2,4-Diamino-7-aminoalkoxy-quinazolines.* J Med Chem, 2010. **53**(15): p. 5844-5857.
92. Rea, S., et al., *Regulation of chromatin structure by site-specific histone H3 methyltransferases.* Nature, 2000/08/10(6796).
93. Ruthenburg, A.J., et al., *Multivalent engagement of chromatin modifications by linked binding modules.* Nat Rev Mol Cell Biol, 2007. **8**(12): p. 983-94.
94. Liu, F., et al., *Optimization of Cellular Activity of G9a Inhibitors 7-Aminoalkoxy-quinazolines.* J Med Chem, 2011. **54**(17): p. 6139-6150.
95. Isham, C.R., et al., *Chaetocin: a promising new antimyeloma agent with in vitro and in vivo activity mediated via imposition of oxidative stress.* Blood, 2007. **109**(6): p. 2579-2588.
96. Greiner, D., et al., *Identification of a specific inhibitor of the histone methyltransferase SU(VAR)3-9.* Nat Chem Biol, 2005. **1**(3): p. 143-5.
97. Cherblanc, F.L., et al., *Chaetocin is a nonspecific inhibitor of histone lysine methyltransferases.* Nat Chem Biol, 2013. **9**(3): p. 136-7.
98. Kubicek, S., et al., *Reversal of H3K9me2 by a Small-Molecule Inhibitor for the G9a Histone Methyltransferase.* Molecular Cell, 2007. **25**(3): p. 473-481.
99. Quinn, A.M. and A. Simeonov, *Methods for Activity Analysis of the Proteins that Regulate Histone Methylation.* Curr Chem Genomics, 2011. **5**(Suppl 1): p. 95-105.
100. Yuan, Y., et al., *A Small-Molecule Probe of the Histone Methyltransferase G9a Induces Cellular Senescence in Pancreatic Adenocarcinoma.* ACS Chemical Biology, 2012. **7**(7): p. 1152-1157.
101. Newman, D.J. and G.M. Cragg, *Natural Products as Sources of New Drugs over the Last 25 Years⊥.* Journal of Natural Products, 2007. **70**(3): p. 461-477.

102. M Lourenco, A., L. M Ferreira, and P. S Branco, *Molecules of natural origin, semi-synthesis and synthesis with anti-inflammatory and anticancer utilities.* Current pharmaceutical design, 2012. **18**(26): p. 3979-4046.
103. Wikberg, J.E.S., et al., *Chemoinformatics Taking Biology into Account: Proteochemometrics,* in *Computational Approaches in Cheminformatics and Bioinformatics*2011, John Wiley & Sons, Inc. p. 57-92.
104. Reardon, S., *Project ranks billions of drug interactions.* Nature, 2013. **503**(7477): p. 449-450.
105. Hughes, J.P., et al., *Principles of early drug discovery.* British Journal of Pharmacology, 2011. **162**(6): p. 1239-1249.
106. Krasavin, M., et al., *Discovery and Potency Optimization of 2-Amino-5-arylmethyl-1,3-thiazole Derivatives as Potential Therapeutic Agents for Prostate Cancer.* Archiv der Pharmazie, 2009. **342**(7): p. 420-427.
107. Kaul, P., *Drug discovery: Past, present and future,* in *Progress in Drug Research,* E. Jucker, Editor 1998, Birkhäuser Basel. p. 9-105.
108. Veselovsky, A.V., et al., *Computer-aided design and discovery of protein–protein interaction inhibitors as agents for anti-HIV therapy.* SAR and QSAR in Environmental Research, 2014. **25**(6): p. 457-471.
109. Song, C.M., S.J. Lim, and J.C. Tong, *Recent advances in computer-aided drug design.* Briefings in Bioinformatics, 2009. **10**(5): p. 579-591.
110. Ou-Yang, S.S., et al., *Computational drug discovery.* Acta Pharmacol Sin, 2012. **33**(9): p. 1131-40.
111. Taft, C.A., V.B. da Silva, and C.H.T.d.P. da Silva, *Current topics in computer-aided drug design.* Journal of Pharmaceutical Sciences, 2008. **97**(3): p. 1089-1098.
112. Bajorath, J., *Integration of virtual and high-throughput screening.* Nat Rev Drug Discov, 2002. **1**(11): p. 882-94.
113. Hopkins, A.L., et al., *The role of ligand efficiency metrics in drug discovery.* Nat Rev Drug Discov, 2014. **13**(2): p. 105-21.
114. Ballester, P.J., et al., *Hierarchical virtual screening for the discovery of new molecular scaffolds in antibacterial hit identification.* Journal of The Royal Society Interface, 2012.
115. Boyd, M.R., *The position of intellectual property rights in drug discovery and development from natural products.* Journal of Ethnopharmacology, 1996. **51**(1–3): p. 17-27.
116. Thiel, K.A., *Structure-aided drug design's next generation.* Nat Biotechnol, 2004. **22**(5): p. 513-9.
117. Schuffenhauer, A., *Computational methods for scaffold hopping.* Wiley Interdisciplinary Reviews: Computational Molecular Science, 2012. **2**(6): p. 842-867.
118. Sun, H., G. Tawa, and A. Wallqvist, *Classification of scaffold-hopping approaches.* Drug Discov Today, 2012. **17**(7–8): p. 310-324.
119. Schneider, G., P. Schneider, and S. Renner, *Scaffold-Hopping: How Far Can You Jump?* QSAR & Combinatorial Science, 2006. **25**(12): p. 1162-1171.
120. Langdon, S.R., et al., *Scaffold-Focused Virtual Screening: Prospective Application to the Discovery of TTK Inhibitors.* Journal of Chemical Information and Modeling, 2013. **53**(5): p. 1100-1112.
121. Li, A.P., *Screening for human ADME/Tox drug properties in drug discovery.* Drug Discov Today, 2001. **6**(7): p. 357-366.
122. Yu, H. and A. Adedoyin, *ADME–Tox in drug discovery: integration of experimental and computational technologies.* Drug Discov Today, 2003. **8**(18): p. 852-861.

123. Ekins, S., et al., *Towards a new age of virtual ADME/TOX and multidimensional drug discovery.* Molecular Diversity, 2000. **5**(4): p. 255-275.
124. Agrafiotis, D.K., et al., *Recent Advances in Chemoinformatics.* Journal of Chemical Information and Modeling, 2007. **47**(4): p. 1279-1293.
125. Valerio, L.G. and S. Choudhuri, *Chemoinformatics and chemical genomics: potential utility of in silico methods.* Journal of Applied Toxicology, 2012. **32**(11): p. 880-889.
126. Vogt, M. and J. Bajorath, *Chemoinformatics: A view of the field and current trends in method development.* Bioorganic & Medicinal Chemistry, 2012. **20**(18): p. 5317-5323.
127. Bhattacharya, A., R. Tejero, and G.T. Montelione, *Evaluating protein structures determined by structural genomics consortia.* Proteins: Structure, Function, and Bioinformatics, 2007. **66**(4): p. 778-795.
128. Hawkins, P.C.D. and A. Nicholls, *Conformer Generation with OMEGA: Learning from the Data Set and the Analysis of Failures.* Journal of Chemical Information and Modeling, 2012. **52**(11): p. 2919-2936.
129. Johnson, M.A. and G.M. Maggiora, *Concepts and Applications of Molecular Similarity.*1990, New York.: Wiley.
130. Shemetulskis, N.E., et al., *Stigmata: An Algorithm To Determine Structural Commonalities in Diverse Datasets.* Journal of Chemical Information and Computer Sciences, 1996. **36**(4): p. 862-871.
131. Gardiner, E.J., et al., *Effectiveness of 2D fingerprints for scaffold hopping.* Future Med Chem, 2011. **3**(4): p. 405-414.
132. Riniker, S. and G. Landrum, *Open-source platform to benchmark fingerprints for ligand-based virtual screening.* Journal of Cheminformatics, 2013. **5**(1): p. 26.
133. Daylight Chemical Infromation System, I., 2003: 120 Vantis, Suite 550, Aliso Viejo, CA 92656.
134. Mauri, A., et al., *Dragon software: An easy approach to molecular descriptor calculations.* MATCH Commun Math Comput Chem, 2006. **56**: p. 237 - 248.
135. *MACCS structural keys.* 2011.
136. Kearsley, S., et al., *Chemical similarity using physiochemical property descriptors[dagger]*. J Chem Inf Comput Sci, 1996. **36**(1): p. 118 - 127.
137. Raymond, J. and P. Willett, *Effectiveness of graph-based and fingerprint-based similarity measures for virtual screening of 2D chemical structure databases.* Journal of Computer-Aided Molecular Design, 2002. **16**(1): p. 59-71.
138. Rogers, D.J. and T.T. Tanimoto, *A Computer Program for Classifying Plants.* Science, 1960. **132**(3434): p. 1115-8.
139. Jenkins, J.L., M. Glick, and J.W. Davies, *A 3D Similarity Method for Scaffold Hopping from Known Drugs or Natural Ligands to New Chemotypes.* J Med Chem, 2004. **47**(25): p. 6144-6159.
140. Fontaine, F., et al., *Fast 3D shape screening of large chemical databases through alignment-recycling.* Chemistry Central Journal, 2007. **1**(1): p. 12.
141. Hawkins, P.C.D., A.G. Skillman, and A. Nicholls, *Comparison of Shape-Matching and Docking as Virtual Screening Tools.* J Med Chem, 2006. **50**(1): p. 74-82.
142. Walters, W.P., M.T. Stahl, and M.A. Murcko, *Virtual screening—an overview.* Drug Discovery Today, 1998. **3**(4): p. 160-178.
143. Morris, G. and M. Lim-Wilby, *Molecular Docking*, in *Molecular Modeling of Proteins*, A. Kukol, Editor 2008, Humana Press. p. 365-382.
144. Kurogi, Y. and O.F. Guner, *Pharmacophore Modeling and Three-dimensional Database Searching for Drug Design Using Catalyst.* Curr Med Chem, 2001. **8**(9): p. 1035-1055.

145. Stahura, F.L. and J. Bajorath, *Virtual screening methods that complement HTS.* Comb Chem High Throughput Screen, 2004. **7**(4): p. 259-69.
146. Kuntz, I.D., et al., *A geometric approach to macromolecule-ligand interactions.* Journal of Molecular Biology, 1982. **161**(2): p. 269-288.
147. Morris, G.M., et al., *AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility.* Journal of Computational Chemistry, 2009. **30**(16): p. 2785-2791.
148. Verdonk, M.L., et al., *Improved protein–ligand docking using GOLD.* Proteins: Structure, Function, and Bioinformatics, 2003. **52**(4): p. 609-623.
149. Friesner, R.A., et al., *Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy.* J Med Chem, 2004. **47**(7): p. 1739-1749.
150. McGann, M., *FRED Pose Prediction and Virtual Screening Accuracy.* Journal of Chemical Information and Modeling, 2011. **51**(3): p. 578-596.
151. Cross, S.S.J., *Improved FlexX Docking Using FlexS-Determined Base Fragment Placement.* Journal of Chemical Information and Modeling, 2005. **45**(4): p. 993-1001.
152. Bienstock Rachelle, J., ed. *Library Design, Search Methods, and Applications of Fragment-Based Drug Design*. ACS Symposium Series. Vol. 1076. 2011, American Chemical Society. 0.
153. Jain, A.N., *Surflex: Fully Automatic Flexible Molecular Docking Using a Molecular Similarity-Based Search Engine.* J Med Chem, 2003. **46**(4): p. 499-511.
154. Lang, P.T., et al., *DOCK 6: Combining techniques to model RNA–small molecule complexes.* Rna, 2009. **15**(6): p. 1219-1230.
155. Alisaraie, L., L.A. Haller, and G. Fels, *A QXP-Based Multistep Docking Procedure for Accurate Prediction of Protein−Ligand Complexes.* Journal of Chemical Information and Modeling, 2006. **46**(3): p. 1174-1187.
156. Venkatachalam, C.M., et al., *LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites.* J Mol Graph Model, 2003. **21**(4): p. 289-307.
157. Abagyan, R., M. Totrov, and D. Kuznetsov, *ICM—A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation.* Journal of Computational Chemistry, 1994. **15**(5): p. 488-506.
158. Liu, M. and S. Wang, *MCDOCK: a Monte Carlo simulation approach to the molecular docking problem.* J Comput Aided Mol Des, 1999. **13**(5): p. 435-51.
159. Mitchell, M., *An Introduction to Genetic Algorithms*1998, Cambridge, MA: MIT Press
160. Böhm, H.-J., *Prediction of binding constants of protein ligands: A fast method for the prioritization of hits obtained from de novo design or 3D database search programs.* Journal of Computer-Aided Molecular Design, 1998. **12**(4): p. 309-309.
161. Eldridge, M.D., et al., *Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes.* J Comput Aided Mol Des, 1997. **11**(5): p. 425-45.
162. Wang, R., Y. Lu, and S. Wang, *Comparative Evaluation of 11 Scoring Functions for Molecular Docking.* J Med Chem, 2003. **46**(12): p. 2287-2303.
163. Choi, V., *Yucca: An Efficient Algorithm for Small-Molecule Docking.* Chemistry & Biodiversity, 2005. **2**(11): p. 1517-1524.
164. Gehlhaar, D.K., et al., *Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming.* Chem Biol, 1995. **2**(5): p. 317-324.

165. Yin, S., et al., *MedusaScore: An Accurate Force Field-Based Scoring Function for Virtual Drug Screening.* Journal of Chemical Information and Modeling, 2008. **48**(8): p. 1656-1662.
166. Sotriffer, C.A., et al., *SFCscore: scoring functions for affinity prediction of protein-ligand complexes.* Proteins, 2008. **73**(2): p. 395-419.
167. Bohm, H.J., *LUDI: rule-based automatic design of new substituents for enzyme inhibitor leads.* J Comput Aided Mol Des, 1992. **6**(6): p. 593-606.
168. Thomsen, R. and M.H. Christensen, *MolDock: A New Technique for High-Accuracy Molecular Docking.* J Med Chem, 2006. **49**(11): p. 3315-3321.
169. Gohlke, H., M. Hendlich, and G. Klebe, *Knowledge-based scoring function to predict protein-ligand interactions.* J Mol Biol, 2000. **295**(2): p. 337-56.
170. Huang, S.-Y. and X. Zou, *A knowledge-based scoring function for protein-RNA interactions derived from a statistical mechanics-based iterative method.* Nucleic Acids Res, 2014.
171. Jones, D.T., W.R. Taylor, and J.M. Thornton, *A new approach to protein fold recognition.* Nature, 1992. **358**(6381): p. 86-9.
172. Vajda, S., M. Sippl, and J. Novotny, *Empirical potentials and functions for protein folding and binding.* Curr Opin Struct Biol, 1997. **7**(2): p. 222-8.
173. Torda, A.E., *Perspectives in protein-fold recognition.* Curr Opin Struct Biol, 1997. **7**(2): p. 200-205.
174. Li, X. and J. Liang, *Knowledge-Based Energy Functions for Computational Studies of Proteins*, in *Computational Methods for Protein Structure Prediction and Modeling*, Y. Xu, D. Xu, and J. Liang, Editors. 2007, Springer New York. p. 71-123.
175. Huang, S.-Y., 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, 2010. **12**(40): p. 12899-12908.
176. Jones, G., et al., *Development and validation of a genetic algorithm for flexible docking.* Journal of Molecular Biology, 1997. **267**(3): p. 727-748.
177. Sotriffer, C.A., H. Gohlke, and G. Klebe, *Docking into knowledge-based potential fields: a comparative evaluation of DrugScore.* J Med Chem, 2002. **45**(10): p. 1967-70.
178. Velec, H.F., H. Gohlke, and G. Klebe, *DrugScore(CSD)-knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction.* J Med Chem, 2005. **48**(20): p. 6296-303.
179. Grinter, S.Z., et al., *Automated large-scale file preparation, docking, and scoring: evaluation of ITScore and STScore using the 2012 Community Structure-Activity Resource benchmark.* J Chem Inf Model, 2013. **53**(8): p. 1905-14.
180. Muegge, I. and Y.C. Martin, *A General and Fast Scoring Function for Protein−Ligand Interactions: A Simplified Potential Approach.* J Med Chem, 1999. **42**(5): p. 791-804.
181. Mitchell, J.B.O., et al., *BLEEP—potential of mean force describing protein–ligand interactions: I. Generating potential.* Journal of Computational Chemistry, 1999. **20**(11): p. 1165-1176.
182. Oloff, S. and I. Muegge, *kScore: a novel machine learning approach that is not dependent on the data structure of the training set.* J Comput Aided Mol Des, 2007. **21**(1-3): p. 87-95.
183. Zhang, C., S. Liu, and Y. Zhou, *Accurate and efficient loop selections by the DFIRE-based all-atom statistical potential.* Protein Sci, 2004. **13**(2): p. 391-9.

184. DeWitte, R.S. and E.I. Shakhnovich, *SMoG: de Novo Design Method Based on Simple, Fast, and Accurate Free Energy Estimates. 1. Methodology and Supporting Evidence.* J. Am. Chem. Soc., 1996. **118**(47): p. 11733 - 11744.
185. Garmendia-Doval, A.B., S.D. Morley, and S. Juhos, *Post Docking Filtering Using Cartesian Genetic Programming*, in *Artificial Evolution*, P. Liardet, et al., Editors. 2004, Springer Berlin Heidelberg. p. 189-200.
186. Deng, Z., C. Chuaqui, and J. Singh, *Structural Interaction Fingerprint (SIFt): A Novel Method for Analyzing Three-Dimensional Protein−Ligand Binding Interactions.* J Med Chem, 2003. **47**(2): p. 337-344.
187. Athanasiadis, E., Z. Cournia, and G. Spyrou, *ChemBioServer: a web-based pipeline for filtering, clustering and visualization of chemical compounds used in drug discovery.* Bioinformatics, 2012. **28**(22): p. 3002-3003.
188. Voet, A.D., et al., *Combining in silico and in cerebro approaches for virtual screening and pose prediction in SAMPL4.* Journal of Computer-Aided Molecular Design, 2014. **28**(4): p. 363-373.
189. Gund, P., *Three-Dimensional Pharmacophoric Pattern Searching*, in *Progress in Molecular and Subcellular Biology*, F. Hahn, et al., Editors. 1977, Springer Berlin Heidelberg. p. 117-143.
190. Wermuth, C.G., et al., *Glossary of terms used in medicinal chemistry (IUPAC Recommendations 1998)*, in *Pure and Applied Chemistry*1998. p. 1129.
191. Ngan, C.H., et al., *FTMAP: extended protein mapping with user-selected probe molecules.* Nucleic Acids Res, 2012. **40**(Web Server issue): p. W271-5.
192. Hawkins, P.C.D., et al., *Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database.* Journal of Chemical Information and Modeling, 2010. **50**(4): p. 572-584.
193. Perola, E. and P.S. Charifson, *Conformational Analysis of Drug-Like Molecules Bound to Proteins: An Extensive Study of Ligand Reorganization upon Binding.* J Med Chem, 2004. **47**(10): p. 2499-2510.
194. Voet, A., F. Berenger, and K.Y. Zhang, *Electrostatic similarities between protein and small molecule ligands facilitate the design of protein-protein interaction inhibitors.* PLoS One, 2013. **8**(10): p. e75762.
195. Voet, A.D., et al., *Combining in silico and in cerebro approaches for virtual screening and pose prediction in SAMPL4.* Journal of Computer-Aided Molecular Design, 2014: p. 1-11.
196. Dolinsky, T.J., et al., *PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations.* Nucleic Acids Res, 2004. **32**(Web Server issue): p. W665-7.
197. Baker, N.A., et al., *Electrostatics of nanosystems: application to microtubules and the ribosome.* Proc Natl Acad Sci U S A, 2001. **98**(18): p. 10037-41.
198. Chau, P.L. and P.M. Dean, *Electrostatic complementarity between proteins and ligands. 2. Ligand moieties.* Journal of Computer-Aided Molecular Design, 1994. **8**(5): p. 527-544.
199. Chau, P.L. and P.M. Dean, *Electrostatic complementarity between proteins and ligands. 3. Structural basis.* Journal of Computer-Aided Molecular Design, 1994. **8**(5): p. 545-564.
200. Chau, P.L. and P.M. Dean, *Electrostatic complementarity between proteins and ligands. 1. Charge disposition, dielectric and interface effects.* Journal of Computer-Aided Molecular Design, 1994. **8**(5): p. 513-525.

201. McCoy, A.J., V. Chandana Epa, and P.M. Colman, *Electrostatic complementarity at protein/protein interfaces.* Journal of Molecular Biology, 1997. **268**(2): p. 570-584.
202. Honig, B. and A. Nicholls, *Classical electrostatics in biology and chemistry.* Science, 1995. **268**(5214): p. 1144-1149.
203. McCoy, A.J., V. Chandana Epa, and P.M. Colman, *Electrostatic complementarity at protein/protein interfaces.* Journal of Molecular Biology, 1994. **268**: p. 570-584.
204. Tanford, C. and J.G. Kirkwood, *Theory of Protein Titration Curves. I. General Equations for Impenetrable Spheres.* J Am Chem Soc, 1957. **79**(20): p. 5333-5339.
205. Braden, B.C. and R.J. Poljak, *Structural features of the reactions between antibodies and protein antigens.* The FASEB Journal, 1995. **9**(1): p. 9-16.
206. Demchuk, E., et al., *Receptor binding properties of four-helix-bundle growth factors deduced from electrostatic analysis.* Protein Science, 1994. **3**(6): p. 920-935.
207. Hendsch, Z.S. and B. Tidor, *Do salt bridges stabilize proteins? A continuum electrostatic analysis.* Protein Science, 1994. **3**(2): p. 211-226.
208. Lescar, J., et al., *Crystal Structure of a Cross-reaction Complex between Fab F9.13.7 and Guinea Fowl Lysozyme.* Journal of Biological Chemistry, 1995. **270**(30): p. 18067-18076.
209. Karshikov, A., et al., *Electrostatic interactions in the association of proteins: An analysis of the thrombin–hirudin complex.* Protein Science, 1992. **1**(6): p. 727-735.
210. Novotny, J. and K. Sharp, *Electrostatic fields in antibodies and antibody/antigen complexes.* Prog Biophys Mol Biol, 1992. **58**(3): p. 203-224.
211. Chau, P.L. and P.M. Dean, *Electrostatic complementarity between proteins and ligands. 1. Charge disposition, dielectric and interface effects.* J Comput Aided Mol Des, 1994. **8**(5): p. 513-25.
212. Chau, P.L. and P.M. Dean, *Electrostatic complementarity between proteins and ligands. 2. Ligand moieties.* J Comput Aided Mol Des, 1994. **8**(5): p. 527-44.
213. Chau, P.L. and P.M. Dean, *Electrostatic complementarity between proteins and ligands. 3. Structural basis.* J Comput Aided Mol Des, 1994. **8**(5): p. 545-64.
214. Lee, B. and F.M. Richards, *The interpretation of protein structures: Estimation of static accessibility.* Journal of Molecular Biology, 1971. **55**(3): p. 379-IN4.
215. Dobosh, P.A., *QCPE Program No. 141,Bloomington, IN.* 1968.
216. Scheraga, H.A., *QCPE Program No. 286,Bloomington, IN.* , 1975.
217. Mulliken, R.S., *Electronic Population Analysis on LCAO–MO Molecular Wave Functions. I.* The Journal of Chemical Physics, 1955. **23**(10): p. 1833-1840.
218. Gouy, G., *Constitution of the electric charge at the surface of an electrolyte.* Journal de Physique, 1910. **9**: p. 457-467.
219. Chapman, D.L., *A contribution to the theory of electrocapillarity. . 1913;25:475.* Philos Mag Ser 6 1913. **25**: p. 475.
220. Debye, P. and E. Hückel, *Zur theorie der electrolyte.* Zeitschrift fur Physik A., 1923. **24**: p. 185–206.
221. Grochowski, P. and J. Trylska, *Continuum molecular electrostatics, salt effects, and counterion binding—A review of the Poisson–Boltzmann theory and its modifications.* Biopolymers, 2008. **89**(2): p. 93-113.
222. Gronwall, T.H., V.K. La Mer, and K. Sandved, *Über den Einfluss der Sogenannten Höhere Glieder in der Debye–Hückelschen Theorie der Losung Starker Electrolyte,* Phys.Zeithschr, 1928. **29 (1928)** p. 358–393.
223. Cortis, C.M. and R.A. Friesner, *Numerical solution of the Poisson–Boltzmann equation using tetrahedral finite-element meshes.* Journal of Computational Chemistry, 1997. **18**(13): p. 1591-1608.

224. Baker, N., M. Holst, and F. Wang, *Adaptive multilevel finite element solution of the Poisson–Boltzmann equation II. Refinement at solvent-accessible surfaces in biomolecular systems.* Journal of Computational Chemistry, 2000. **21**(15): p. 1343-1352.
225. Chen, L., M. Holst, and J. Xu, *The Finite Element Approximation of the Nonlinear Poisson–Boltzmann Equation.* SIAM Journal on Numerical Analysis, 2007. **45**(6): p. 2298-2320.
226. Holst, M.J., *Adaptive Numerical Treatment of Elliptic Systems on Manifolds.* Adv. Comput. Math., 2011. **15**: p. 131-191.
227. Holst, M., N. Baker, and F. Wang, *Adaptive multilevel finite element solution of the Poisson–Boltzmann equation I. Algorithms and examples.* Journal of Computational Chemistry, 2000. **21**(15): p. 1319-1342.
228. Holst, M.J. and F. Saied, *Numerical solution of the nonlinear Poisson–Boltzmann equation: Developing more robust and efficient methods.* Journal of Computational Chemistry, 1995. **16**(3): p. 337-364.
229. Gilson, M.K., K.A. Sharp, and B.H. Honig, *Calculating the electrostatic potential of molecules in solution: Method and error assessment.* Journal of Computational Chemistry, 1988. **9**(4): p. 327-335.
230. Warwicker, J. and H.C. Watson, *Calculation of the electric potential in the active site cleft due to α-helix dipoles.* Journal of Molecular Biology, 1982. **157**: p. 671-679.
231. Klapper, I., et al., *Focusing of electric fields in the active site of Cu-Zn superoxide dismutase: Effects of ionic strength and amino-acid modification.* Proteins: Structure, Function, and Bioinformatics, 1986. **1**(1): p. 47-59.
232. Nichols, A., R. Bharadwaj, and B. Honig, *GRASP−graphical representation and analysis of surface properties.* Biophys. J., 1993. **64**(2): p. A166
233. Zhou, Y.C., M. Feig, and G.W. Wei, *Highly accurate biomolecular electrostatics in continuum dielectric environments.* Journal of Computational Chemistry, 2008. **29**(1): p. 87-97.
234. Grant, J.A., B.T. Pickup, and A. Nicholls, *A smooth permittivity function for Poisson–Boltzmann solvation methods.* Journal of Computational Chemistry, 2001. **22**(6): p. 608-640.
235. Bashford, D., *An object-oriented programming suite for electrostatic effects in biological molecules. InScientific Computing in Object-Oriented Parallel Environments.Lecture Notes in Computer Science*, 1997, Springer: Berlin, Germany. .
236. al., M.e., *Electrostatics and diffusion of molecules in solution: simulations with the University of Houston Brownian Dynamics program.* Computer Physics Communications, 1995. **91**(1-3): p. 57-95.
237. Im, W., D. Beglov, and B. Roux, *Continuum solvation model: Computation of electrostatic forces from numerical solutions to the Poisson-Boltzmann equation.* Computer Physics Communications, 1998. **111**(1-3).
238. Bajaj, C., S. Chen, and A. Rand, *An Efficient Higher-Order Fast Multipole Boundary Element Solution for Poisson–Boltzmann-Based Molecular Electrostatics.* SIAM Journal on Scientific Computing, 2011. **33**(2): p. 826-848.
239. Zauhar, R.J. and R.S. Morgan, *The rigorous computation of the molecular electric potential.* Journal of Computational Chemistry, 1988. **9**(2): p. 171-187.
240. Zhou, H.X., *Boundary element solution of macromolecular electrostatics: interaction energy between two proteins.* Biophysical Journal, 1993. **65**(2): p. 955-963.
241. Zauhar, R.J. and R.S. Morgan, *A new method for computing the macromolecular electric potential.* Journal of Molecular Biology, 1985. **186**(4): p. 815-820.

242. Dolinsky, T.J., et al., *PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations.* Nucleic Acids Res, 2007. **35**(suppl 2): p. W522-W525.
243. Dolinsky, T.J., et al., *PDB2PQR: an automated pipeline for the setup of Poisson–Boltzmann electrostatics calculations.* Nucleic Acids Res, 2004. **32**(suppl 2): p. W665-W667.
244. Unni, S., et al., *Web servers and services for electrostatics calculations with APBS and PDB2PQR.* Journal of Computational Chemistry, 2011. **32**(7): p. 1488-1491.
245. Huey, R., et al., *A semiempirical free energy force field with charge-based desolvation.* Journal of Computational Chemistry, 2007. **28**(6): p. 1145-1152.
246. Morris, G.M., et al., *Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function.* Journal of Computational Chemistry, 1998. **19**(14): p. 1639-1662.
247. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-242.
248. Pearlman, D.A., et al., *AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules.* Computer Physics Communications, 1995. **91**(1-3): p. 1-41.
249. Humphrey, W., A. Dalke, and K. Schulten, *VMD: Visual molecular dynamics.* Journal of Molecular Graphics, 1996. **14**(1): p. 33-38.
250. DeLano, W.L. and C.A. Palo Alto, *The PyMOL Molecular Graphics System,* 2002.
251. Sanner, M.F., *Python: A Programming Language for Software Integration and Development.* J. Mol. Graphics Mod, 1999. **17**: p. 57-61.
252. MacKerell, A.D., et al., *All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins†.* The Journal of Physical Chemistry B, 1998. **102**(18): p. 3586-3616.
253. Wang, J., P. Cieplak, and P.A. Kollman, *How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules?* Journal of Computational Chemistry, 2000. **21**(12): p. 1049-1074.
254. Sitkoff, D., K.A. Sharp, and B. Honig, *Accurate Calculation of Hydration Free Energies Using Macroscopic Solvent Models.* The Journal of Physical Chemistry, 1994. **98**(7): p. 1978-1988.
255. Tan, C., L. Yang, and R. Luo, *How Well Does Poisson−Boltzmann Implicit Solvent Agree with Explicit Solvent? A Quantitative Analysis.* The Journal of Physical Chemistry B, 2006. **110**(37): p. 18680-18687.
256. Czodrowski, P., et al., *Development, validation, and application of adapted PEOE charges to estimate pKa values of functional groups in protein–ligand complexes.* Proteins: Structure, Function, and Bioinformatics, 2006. **65**(2): p. 424-437.
257. Baker, N.A., et al., *The adaptive multilevel finite element solution of the Poisson-Boltzmann equation on massively parallel computers.* IBM J. Res. Dev., 2001. **45**(3-4): p. 427-438.
258. Holst, M., *FEtk: Finite Element ToolKit.* 2003.
259. Bank, R. and M. Holst, *A New Paradigm for Parallel Adaptive Meshing Algorithms.* SIAM Review, 2003. **45**(2): p. 291-323.
260. Holst, M., *Adaptive Numerical Treatment of Elliptic Systems on Manifolds.* Advances in Computational Mathematics, 2001. **15**(1-4): p. 139-191.
261. Holst, M. and F. Saied, *Multigrid solution of the Poisson—Boltzmann equation.* Journal of Computational Chemistry, 1993. **14**(1): p. 105-113.

262. Ren, J., et al., *Opal web services for biomedical applications.* Nucleic Acids Res, 2010. **38**(suppl 2): p. W724-W731.
263. Krishnan, S., et al. *Design and Evaluation of Opal2: A Toolkit for Scientific Software as a Service*. in *Services - I, 2009 World Conference on*. 2009.
264. Carbó, R., L. Leyda, and M. Arnau, *How similar is a molecule to another? An electron density measure of similarity between two molecular structures.* International Journal of Quantum Chemistry, 1980. **17**(6): p. 1185-1189.
265. Hodgkin, E.E. and W.G. Richards, *Molecular Similarity Based on Electrostatic Potential and Electric-Field.* International Journal of Quantum Chemistry, 1987. **32**(S14): p. 105-110.
266. Pearson, K., *Proceedings of the Royal Society of London.* Vol. 58. 1895, Royal Society of London
267. Tanimoto, T.T., *IBM Internal Report 17th Nov see also Jaccard, P. (1901) Bulletin del la Société Vaudoisedes Sciences Naturelles*, 1957. p. 241-272.
268. Leroy, X., et al., *The OCaml system release 3.12 Documentation and user's manual*, 2011, INRIA,France.
269. Danelutto, M. and R. Di Cosmo, *A "Minimal Disruption" Skeleton Experiment: Seamless Map & Reduce Embedding in OCaml.* Procedia Computer Science, 2012. **9**(0): p. 1837-1846.
270. Warren, G.L., et al., *Essential considerations for using protein–ligand structures in drug discovery.* Drug Discovery Today, 2012. **17**(23–24): p. 1270-1281.
271. Huang, N., B.K. Shoichet, and J.J. Irwin, *Benchmarking Sets for Molecular Docking.* J Med Chem, 2006. **49**(23): p. 6789-6801.
272. Jones, G., P. Willett, and R.C. Glen, *Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation.* J. Mol. Biol., 1995 **245**(43-53.).
273. Truchon, J.-F. and C.I. Bayly, *Evaluating Virtual Screening Methods: Good and Bad Metrics for the "Early Recognition" Problem.* Journal of Chemical Information and Modeling, 2007. **47**(2): p. 488-508.
274. Baker, N.A., et al., *Electrostatics of nanosystems: Application to microtubules and the ribosome.* Proceedings of the National Academy of Sciences, 2001. **98**(18): p. 10037-10041.
275. Hawkins, P. and A. Nicholls, *Conformer generation with OMEGA: learning from the data set and the analysis of failures.* J Chem Inf Model, 2012. **52**(11): p. 2919 - 2936.
276. Chang, Y., et al., *MPP8 mediates the interactions between DNA methyltransferase Dnmt3a and H3K9 methyltransferase GLP/G9a.* Nat Commun, 2011. **2**: p. 533.
277. Eglen, R.M., et al., *The use of AlphaScreen technology in HTS: current status.* Curr Chem Genomics, 2008. **1**: p. 2-10.
278. Du, Y., F.R. Khuri, and H. Fu, *A homogenous luminescent proximity assay for 14-3-3 interactions with both phosphorylated and nonphosphorylated client peptides.* Curr Chem Genomics, 2008. **2**: p. 40-7.
279. Vedadi, M., et al., *A chemical probe selectively inhibits G9a and GLP methyltransferase activity in cells.* Nat Chem Biol, 2011. **7**(8): p. 566-574.
280. Thompson, J.D., T.J. Gibson, and D.G. Higgins, *Multiple sequence alignment using ClustalW and ClustalX.* Curr Protoc Bioinformatics, 2002. **Chapter 2**: p. Unit 2 3.
281. Chang, Y., et al., *Structural basis for G9a-like protein lysine methyltransferase inhibition by BIX-01294.* Nat Struct Mol Biol, 2009. **16**(3): p. 312-317.
282. Chang, Y., et al., *Adding a Lysine Mimic in the Design of Potent Inhibitors of Histone Lysine Methyltransferases.* Journal of Molecular Biology, 2010. **400**(1): p. 1-7.

283. Liu, F., et al., *Discovery of a 2,4-Diamino-7-aminoalkoxyquinazoline as a Potent and Selective Inhibitor of Histone Lysine Methyltransferase G9a.* Journal of Medicinal Chemistry, 2009. **52**(24): p. 7950-7953.
284. Berenger, F., et al., *A rotation-translation invariant molecular descriptor of partial charges and its use in ligand-based virtual screening.* Journal of Cheminformatics, 2014. **6**(1): p. 23.
285. von Korff, M., J. Freyss, and T. Sander, *Comparison of Ligand- and Structure-Based Virtual Screening on the DUD Data Set.* J Chem Inf Model, 2009. **49**(2): p. 209 - 231.
286. O'Boyle, N., et al., *Open Babel: an open chemical toolbox.* J Cheminformatics, 2011. **3**(1): p. 33.
287. *Molecular Operating Environment (MOE)*, 2013.08, Chemical Computing Group Inc: 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2013.
288. Taminau, J., G. Thijs, and H. De Winter, *Pharao: pharmacophore alignment and optimization.* J Mol Graph Model, 2008. **27**(2): p. 161-9.
289. Christie, B.D., et al., *Database structure and searching in MACCS-3D.* Tetrahedron Computer Methodology, 1990. **3**(6, Part C): p. 653-664.
290. Delaneau, O., C. Coulonges, and J.F. Zagury, *Shape-IT: new rapid and accurate algorithm for haplotype inference.* BMC Bioinformatics, 2008. **9**: p. 540.

Appendix 1 Iridium benchmarking set

The Iridium dataset that are used including both the highly trustworthy and mildly trustworthy. Both dataset are computed under the solvated and desolvated condition respectively.			
	Target	**Desolvated**	**Solvated**
1	1a28	-0.23	-0.31
2	1ai5	-0.83	-0.56
3	1b9v	-0.62	0.17
4	1br6	-0.45	-0.50
5	1c1b	-0.29	-0.22
6	1ctr	-0.67	-0.15
7	1cvu	0.13	-0.07
8	1cx2	-0.13	-0.08
9	1dds	-0.84	0.01
10	1exa	-0.68	0.00
11	1ezq	-0.05	-0.07
12	1f0s	-0.16	-0.18
13	1f0t	-0.25	-0.49
14	1f0u	-0.38	-0.38
15	1fcx	-0.68	-0.04
16	1fcz	-0.73	-0.11
17	1fh8	-0.35	-0.11
18	1fh9	-0.10	-0.24
19	1fhd	0.05	-0.39
20	1fjs	-0.46	-0.42
21	1fl3	-0.53	-0.12
22	1fm6	0.16	-0.04
23	1fm9	0.19	0.11
24	1fq5	-0.16	-0.17
25	1frp	-0.66	-0.15
26	1fvt	-0.10	0.13
27	1g9v	-0.87	0.14
28	1gm8	-0.51	-0.21
29	1gwx	-0.01	-0.41
30	1h1p	-0.20	-0.15
31	1h1s	-0.32	-0.42

32	1hgg	-0.42	-0.39
33	1hgh	-0.44	-0.10
34	1hgi	-0.49	-0.27
35	1hgj	-0.73	-0.36
36	1hnn	-0.42	-0.10
37	1hp0	-0.03	-0.22
38	1hwi	-0.71	-0.21
39	1ivb	-0.35	-0.28
40	1ivd	-0.28	-0.45
41	1ive	-0.50	-0.20
42	1ivf	-0.39	-0.10
43	1iy7	-0.88	0.18
44	1jla	-0.48	-0.07
45	1k1j	-0.39	-0.41
46	1k3u	-0.87	0.18
47	1ke5	-0.39	-0.33
48	1l2s	-0.45	-0.28
49	1l7f	-0.53	-0.39
50	1lpz	-0.11	-0.37
51	1lqd	-0.19	-0.73
52	1m2z	-0.14	0.01
53	1ml1	-0.10	-0.33
54	1mq6	0.04	-0.28
55	1mts	-0.20	-0.54
56	1n1m	-0.12	-0.36
57	1n2j	-0.26	-0.35
58	1n2v	-0.33	-0.47
59	1n46	-0.10	-0.15
60	1of1	-0.29	-0.18
61	1of6	-0.47	0.55
62	1owe	-0.41	-0.14
63	1oyt	-0.33	0.37
64	1p62	-0.42	0.02
65	1pmn	-0.27	-0.01
66	1pso	-0.64	0.02
67	1q1g	0.06	-0.20
68	1q41	-0.05	0.07
69	1qhi	-0.56	-0.08

70	1r9o	-0.50	-0.18
71	1rob	-0.24	-0.12
72	1s19	-0.21	-0.12
73	1sq5	-0.58	-0.34
74	1tow	-0.73	-0.33
75	1tt1	-0.85	-0.08
76	1u4d	-0.53	-0.34
77	1ukz	-0.76	-0.17
78	1ulb	-0.70	0.06
79	1unl	-0.24	-0.41
80	1uou	-0.57	-0.61
81	1v0p	0.06	-0.24
82	1w2g	-0.14	-0.34
83	1x8x	-0.53	-0.17
84	1xm6	-0.52	-0.17
85	1xoq	-0.24	-0.44
86	1ydr	-0.29	-0.08
87	1yds	-0.56	-0.31
88	1ydt	-0.53	-0.20
89	1yv3	-0.03	-0.13
90	1yvf	-0.36	-0.35
91	1ywr	-0.48	-0.51
92	2ack	-0.17	-0.61
93	2br1	-0.26	-0.45
94	2ctc	-0.47	-0.48
95	2mcp	-0.78	-0.52
96	2pcp	-0.09	-0.14
97	3ptb	-0.36	-0.17
98	4cox	-0.57	-0.44
99	4ts1	-0.25	-0.01
100	13gs	-0.55	-0.25
101	1a07	-0.61	-0.57
102	1a1b	-0.58	-0.46
103	1a1e	-0.49	-0.45
104	1a4k	-0.45	-0.38
105	1a8t	-0.16	0.00
106	1afq	-0.39	-0.37
107	1ake	-0.80	-0.56

108	1avd	-0.59	-0.43
109	1b6j	-0.33	-0.30
110	1b6k	-0.48	-0.29
111	1b6l	-0.27	-0.17
112	1b6m	-0.54	-0.39
113	1bkm	-0.60	-0.53
114	1bmq	-0.70	-0.48
115	1c3i	-0.17	-0.08
116	1c5c	-0.32	-0.05
117	1c5p	-0.32	-0.24
118	1c5x	-0.33	-0.23
119	1cbs	-0.32	-0.55
120	1cdg	0.03	0.04
121	1cet	-0.39	-0.37
122	1cqp	0.01	-0.09
123	1ctt	-0.61	-0.53
124	1dy9	-0.58	-0.37
125	1eve	-0.17	-0.02
126	1f0r	-0.16	-0.13
127	1fbl	-0.41	-0.16
128	1gkc	-0.14	-0.07
129	1gpk	-0.55	-0.44
130	1gsp	-0.49	-0.36
131	1hbv	-0.02	-0.14
132	1i7z	-0.21	-0.16
133	1ibg	-0.13	-0.13
134	1ig3	-0.33	-0.20
135	1igj	-0.50	-0.40
136	1ivc	0.04	-0.04
137	1jje	-0.13	-0.11
138	1k7e	-0.72	-0.46
139	1k7f	-0.86	-0.58
140	1kel	-0.26	-0.19
141	1kzk	-0.25	-0.14
142	1lmo	-0.11	-0.23
143	1lpm	-0.06	-0.16
144	1lyl	-0.86	-0.69
145	1m48	-0.03	-0.11

146	1mq5	-0.32	-0.12
147	1mrg	-0.13	-0.31
148	1mrk	-0.37	-0.29
149	1mtw	-0.46	-0.32
150	1nhu	0.10	-0.24
151	1nhv	0.29	-0.19
152	1odw	-0.52	-0.38
153	1opk	-0.58	-0.45
154	1pgp	0.22	0.18
155	1q4g	-0.85	-0.48
156	1qh7	-0.08	-0.02
157	1r1h	-0.05	-0.12
158	1rnt	0.09	-0.03
159	1s3v	-0.16	0.04
160	1sg0	-0.58	-0.28
161	1sj0	0.16	0.05
162	1sln	-0.02	-0.13
163	1snc	-0.14	-0.16
164	1sqn	-0.22	-0.18
165	1t46	-0.14	-0.13
166	1thy	-0.39	-0.34
167	1tlp	-0.23	-0.21
168	1tpp	-0.29	-0.18
169	1tz8	-0.19	-0.17
170	1u1c	-0.12	-0.16
171	1v48	-0.68	-0.41
172	1v4s	-0.54	-0.42
173	1vcj	-0.52	-0.46
174	1xoz	0.15	0.06
175	1y6b	-0.13	-0.19
176	1ygc	-0.33	-0.23
177	1z95	-0.36	-0.24
178	25c8	-0.38	-0.39
179	2bm2	-0.20	-0.20
180	2bsm	-0.41	-0.18
181	2tpi	-0.69	-0.48
182	3tpi	-0.73	-0.35
183	4hmg	-0.46	-0.28

184	5tim	-0.53	-0.10
185	6tmn	0.19	0.05
	Average	**-0.35**	**-0.23**