12-2018

Modelling the Effects of Disease-Associated Single Amino Acid Variants and Rescuing the Effects by Small Molecules

Yunhui Peng
Clemson University, yunhuipengys@gmail.com

MODELLING THE EFFECTS OF DISEASE-ASSOCIATED SINGLE AMINO ACID VARIANTS AND RESCUING THE EFFECTS BY SMALL MOLECULES

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
Physics

---

by
Yunhui Peng
December 2018

Accepted by:
Dr. Emil Alexov, Committee Chair
Dr. Weiguo Cao
Dr. Feng Ding
Dr. Hugo Sanabria
Dr. Joshua Alper

ABSTRACT

Single nucleotide polymorphism (SNP) is a variation of a single nucleotide in the genome. Some of these variations can cause a change of single amino acid in the corresponding protein, resulting in single amino acid variation (SAV). SAVs can lead to profound alterations of the corresponding biological processes and thus can be associated with many human diseases. This dissertation focuses on integration of existing and development of new computational approaches to model the effects of SAVs with the goal to reveal molecular mechanism of human diseases. Since proton transfer and pKa shifts are frequently attributed to disease causality, the proton transfers in the protein-nucleic acid interactions are investigated and along with development of a new computational approach to predict the SAV's effect on the protein-DNA binding affinity. The SAVs in four proteins: Lysine-specific demethylase 5C (KDM5C), Spermine Synthase (SpmSyn), 7-Dehydrocholesterol reductase (DHCR7) and methyl CpG binding protein 2 (MeCP2) are extensively studied using numerous computational approaches to reveal molecular details of disease-associated effects. In case of MeCP2 protein, the effects of the most commonly occurring disease-causing mutation, R133C, was targeted by structure-based virtual screening to identify the small molecules potentially to rescue the malfunctioning R133C mutant.

DEDICATION

For their love, support and encouragement, I dedicate my thesis to my parents.

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Emil Alexov for his guidance, patience, encouragement and financial support during my PhD study. It's really my great pleasure to work with him and his expertise in biophysics, bioinformatics and structural biology gives me lots of valuable guidance and suggestions for my research. The best experience for me in Clemson is to work with such a great advisor and researcher like him.

I really appreciate the help and guidance from other committee members. Dr. Weiguo Cao helps with experiments for many of my works and we have lots of joint publications. Dr. Feng Ding helps me with my courses and provides me many valuable suggestions for my research and future academia career development. I also thank Dr. Hugo Sanabria and Dr. Joshua Alper for their teaching in biophysics courses. Their highly expertise in the experimental biophysics make me gain lots of knowledge and inspiration from discussion.

I am also very grateful to the mentoring and guidance from Dr. Maria Miteva in University of Paris VII during my internship in France. Without her help in my research, I cannot complete my projects in Paris and gain the knowledge in *in silico* Virtual Screening. I also thank to the Chateaubriand fellowship provided by the Embassy of France for their financial support for my internship in France.

I also would like to thank all lab members in my group. You are all super talented young researchers and it's my pleasure to work and learn from all of you.

TABLE OF CONTENTS

Page

LIST OF TABLES

List of Tables (Continued)

Table Page

LIST OF FIGURES

List of Figures (Continued)

Figure Page

List of Figures (Continued)

Figure Page

List of Figures (Continued)

Figure Page

CHAPTER ONE

INTRODUCTION

Importance of modelling the molecular effect of single amino acid variation:

Human genetic variations result in natural differences among the humans or may cause diseases[1]. Genetic variations originate from subtle differences in DNA and it is well know that humans share 99.5% of DNA code and only the rest 0.5% results in the uniqueness of individuals. However, despite of low occurrence, common genetic variations may contribute significantly to human's susceptibility to common diseases[2-4]. Thus, understanding common human genetic variations and associated functional impact is a very important part of any genetic study and has great potential for direct clinical applications[5, 6].

Genetic differences can be manifested at different levels as a Single Nucleotide Polymorphism (SNPs), which is a change of single nucleotide or as non-synonymous SNP (nsSNP), manifested as amino acid change in the corresponding transcribed product. My dissertation focused on substitutions of single amino acid in the corresponding protein. Following the literature, such a change is termed single amino acid variation (SAV) [4, 7-9]. The SAV can affect the corresponding protein's function and thus may be associated with many human diseases[10-13]. Predicting disease associated SAV's effect and discriminating disease-causing and harmless SAV is of crucial importance for the early diagnostics and medicine [5, 14-17]. However, predicting the effect of disease-associated SAV is not a trivial problem[18, 19], prompting many researchers to develop predictive algorithms and tools[6, 18-23].

Disease-causing SAV can cause malfunctioning of the corresponding protein [13, 18, 24-26]. Some disease-causing SAVs affect protein stability, resulting in unfolded dysfunctional protein [11, 25, 27, 28]. Other disease-causing SAVs that occur in protein binding interface may disrupt the protein interaction network by altering the affinity of interacting partners [24, 29, 30]. The effects on protein folding and binding can be accessed via the changes of folding free energy ($\Delta\Delta G$) and binding free energy ($\Delta\Delta\Delta G$). Many computational and experimental efforts were carried out to determine the changes of folding and binding free energies due to SAVs [30-34]. Roughly speaking, current existing methods can be categorized into sequence-based approaches, structure-based approaches and first principle approaches. Sequence-based approaches utilized machine learning models to perform fast predictions but highly depend on the training datasets. Structure-based approaches consider the potential function or knowledge-based scoring function delivered from protein structural information. First principle approaches, such as the free energy perturbation (FEP) and the thermodynamic integrations (TI) are the most rigorous, but require intensive calculations, which limit their applicability for large-scale analysis.

The experimental studies are relatively limited due to the highly cost in time and expense. Thus, integration of existing *in silico* approaches can help us better understand the molecular mechanism of the disease-associated SAVs, which is crucial for the personalized medicine and identification of potential drugs for treatment. Besides, further development of accurate approaches is also highly in demand for understanding the mechanism of diseases and protein design. Especially for large scales-analysis of mutations' effects,

computational approaches can complement the experimental measurement and provide fast predictions (Chapter III).

Rescue the effects of disease associated single amino acid variation:

With the rapid development of computer techniques, computer-aided approaches have been currently widely applied in aiding early-stage drug discovery in both industrial and academic projects [35-38]. By discovering the potential compounds that target and affect the function of the specific proteins, the biological process can be modulated to mitigate or eliminate the disease-causing effects [36, 38]. Advances in human genome projects have provided vast target proteins for drug discovery projects [39, 40]. Meanwhile, breakthroughs in structural biology have offered in-depth structural information of more and more targets and elucidated the disease mechanisms at molecular level [41-44]. Such advances have further stimulated the application of computational approaches to integrate the available structural information, functional mechanism and physical-chemical properties to drug discovery [37, 45]. Discovery of compounds to mitigate or eliminate the disease-causing effects induced by a specific amino acid mutation is the main goal of Personalized Medicine [1].

Importance of elucidating and clustering the mutations' effects in drug design:

In terms of a drug-design process targeting specific disease-causing mutations, elucidation of mutations' effects is of great importance, especially for the approaches requiring information about target protein structure. Alone with aforementioned computational approaches, it can be integrated into the drug design pipeline (Fig. 1.1). For example, free energy calculation methods are used to determine the dominant effects of

mutations, whether affecting protein stability, protein binding or both. With the in-depth analysis of mutations' effects at molecular-level, the disease-causing mutations in the target proteins can be further clustered by their major effects such as destabilizing mutation, dimerization-affecting mutation, conformational-affecting mutation or catalytic mutation [41, 46-48]. Such type of classification can help designing drugs for certain groups of mutations with similar effects and thus being applicable to broader spectrum of patients.

Structure-based approach in drug design:

Structure-based drug design (SBDD) is the computational approaches that rely on knowledge of the 3D structure of the biological targets to identify or design the potential chemical structure suitable for clinical tests[45, 49] (Fig 1.1). With the explosion of genomic, functional and structural information in last decades, vast of biological targets with 3D structure have been identified and stimulated the applications of structure-based approaches in current design pipeline. SSDB is popular for virtual screening to filter the drug-like compounds from a large library of small molecules, including widely applied approaches such as docking and structure-based pharcorphore [38]. While the established high-throughput screening (HTS) allows for automatic testing of vast compounds (up to millions), the low success rate and high cost limits its applications. Alternatively, one can use computational approaches to reduce the numbers of compounds subjected to testing [36, 50].

Ligand-based approaches in drug design:

In the cases of lacking structural information of target protein, the aforementioned structure-based approaches may not be suitable for drug design. Alternatively, ligand-based drug design (LBDD) can be applied for such cases [51]. Ligand-based methods only focus on the analysis of physico-chemical properties of known ligands that interact with the target of interests. Most popular approaches are quantitative structure activity relationship (QSAR) models and ligand-based pharmacophore modeling [51]. In terms of drug design targeting the mutant proteins, LBDD could be efficient for novel discovered mutations which effects have not yet been investigated (Fig. 1.1).

With the advances in understanding of structural and functional characteristics of biological targets, structure-based approaches have gained popularity. However, it should be pointed out that combining both ligand and structure-based approaches is expected to provide significant advantages [52, 53].

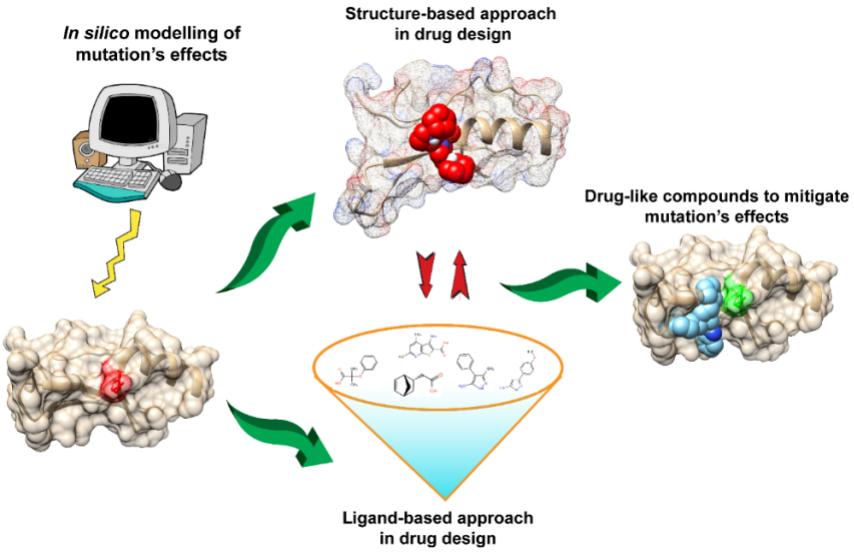


Figure 1.1 Schematic presentation of drug discovery process to mitigate the effects of disease-causing mutations.

CHAPTER TWO

MODELLING THE EFFECTS OF SINGLE AMINO ACID VARIATION

Modelling of SAVs' molecular effects in KDM5C protein:

PART1: Mutations in KDM5C ARID-domain and their association with Syndromic Claes-Jensen-Type Disease:

1. Introduction

Epigenetic processes regulate gene expression and are essential for development and differentiation of cells [1]. Histone proteins are the major components of chromatin, acting as spools around which DNA winds. Particularly, histone lysine methylation is an important epigenetic process which regulates chromatin structure and gene transcription [54, 55]. Due to this, loss of balance of histone lysine methylation has been found to have a profound effect on the diverse biological processes and to be involved in many diseases, including cancer development [56-58].

This work focuses on a particular histone protein, the KDM5C protein of 1560 aa, which is a member of the SMCY homolog family. The KDM5C protein specifically reverses tri- and di-methylation of Lys4 of histone H3 (H3K4), helps maintain the dynamic balance of histone H3K4 methylation states, and also plays a crucial role in functional discrimination between enhancers and core promoters [59-61]. It is a multi-functional protein, which contains highly-conserved domains, including ARID/Bright, JmjN, JmjC, C5HC2 zinc finger, and two PHD zinc finger domains (Figure 2.1). These domains were shown to have specific functions alone or to function in concert with the other KDM5C domains. Thus,

the ARID (A–T rich interaction domain) is a helix–turn–helix motif-based DNA-binding domain, which is highly conserved in all eukaryotic proteins and plays important roles in development, tissue-specific gene expression, and cell growth regulation [62, 63]. The DNA sequence binding preference is still unclear for the ARID domain of KDM5C. The other domain, JmjC, catalyzes demethylation of H3K4me3 to H3K4me1 [59]. The JmjN domain and its interaction with the JmjC catalytic domain are important for the KDM5C function [64]. The N-terminal PHD zinc finger is a histone methyl-lysine binding motif and was shown to have a preferential binding to histone H3K9me3 [59, 65].

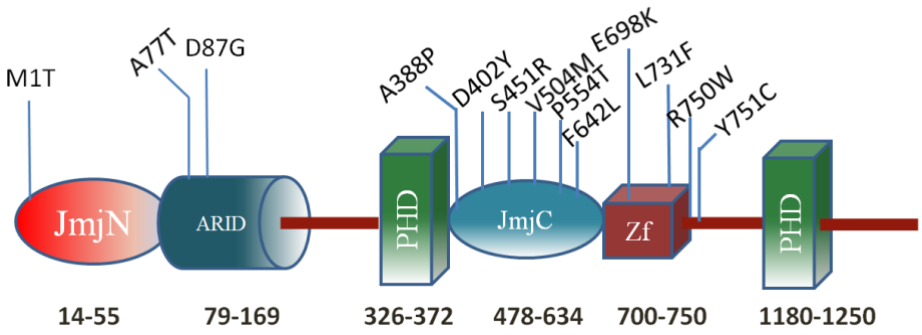


Figure 2.1 KDM5C protein domains. The numbers indicate approximate domain boundaries. The known disease-associated missense mutations are provided as well.

Previous studies have shown that many mutations in the *KDM5C* gene cause X-linked mental retardation (XLMR), the syndromic Claes-Jensen-type disease [59, 66, 67]. Mental retardation (MR) generally causes significant limitations both in intellectual functioning and in adaptive behavior, covering the social and practical skills that originate before the age of 18 years [68]. The estimated prevalence of MR among the general population is around 1%–3% [67, 69]. The frequency of mutations in the *KDM5C* gene approximately

accounts for 2.8% to 3.3% of families with XLMR [70]. Thirteen missense mutations associated with XLMR, the syndromic Claes-Jensen-type disease in *KDM5C* have been reported to date and affected individuals with *KDM5C* mutations show a mild-to-severe range of intellectual disability. Most of mutations are located in JmjC domain, ZF domain (C5HC2 zinc finger domain), and inter-domain regions and affect the demethylation activity [59, 60]. The severity of associated XLMR is roughly related to the cellular demethylase activities of KDM5C mutants [71]. In this study, we focus on the mutations in the ARID domain. Two MR associated mutations (A77T and D87G) are reported in the ARID domain [66, 72]. The D87G mutation causes mild to moderate MR including aggressive behavior, epileptic seizures, and speech impairment, while the A77T results in severe MR including speech impairment, short stature, seizures, microcephaly, hyper reflexia, and aggressive behavior [66, 72]. However, recent work has shown that the D87G has a minimal effect on KDM5C demethylase activity *in vivo* [71] indicating that the disease-associated effect is not demethylation. Combined with the lack of data for the molecular effect of A77T mutation, it can be concluded that the disease-associated effects of both A77T and D87G mutations are unknown. In this work, we extend the list of mutations, which will be investigated, to include three currently non-classified missense mutations in the ARID domain. The non-classified mutations are R108W(rs146232504), N142S(rs377166019), and R178H(rs201805773), taken from the NCBI dbSNP database [73]. They were identified from population cohorts participating in the NHLBI Exome Sequencing Project [74]. This project is designed to identify genetic variants in coding regions of the human genome that are associated with heart, lung, and blood diseases, and

the group included 200,000 individuals. However, there is no data about the linkage of these mutations with a particular disease. This motivates us to investigate the molecular mechanism of all abovementioned mutations, disease-associated and non-classified, and to infer plausible XLMR linkages with some of the non-classified mutations. The allele frequency of the mutations R108W, N142S, R179H are 0.00001151, 0.00001159, and 0.0002497 taken from the ExAC database [75]. The frequency of the other two mutations is not currently available in the database.

Disease-associated mutations are often found to alter protein structure, dynamics and interaction, and cause deficiency of important protein functions [76-80]. Investigating mutations' effects is important for understanding the molecular mechanisms of disease-associated mutations and discriminating disease-causing and harmless mutations. Protein stability and protein interactions can be quantified by folding free energy change ($\Delta\Delta G$) and binding free energy change ($\Delta\Delta\Delta G$). In this study, we analyze the effects of diseasing-associated and currently non-classified mutations on ARID domain stability and ARID-DNA binding affinity utilizing webservers, third-party software, molecular dynamics (MD) and free energy perturbation (FEP) methods. Additionally, our free energy calculations results are further validated by experiments. Urea-induced unfolding monitored by circular dichroism spectroscopy is used to determine the unfolding free energy of the wild-type ARID domain, and the two disease-associated mutants A77T and D87G.

2. Results

2.1. Protein Stability Changes due to Mutations

We applied the free energy perturbation theory (FEP) to analyze two disease associated (A77T, D87G), and three non-classified (R108W, N142S, R179H), mutations. The calculated folding and binding free energy changes caused by mutations are shown in Table 2.1 It can be seen that the energy changes are predicted to be relatively small, being less than 1 kcal/mol in the majority of cases, with the notable exception of FEP calculated folding free energy changes involving Arg residue. A similar effect of over-predicting the magnitude of the change of the folding free energy involving the Arg group was noticed in another study [46]. Further investigations are needed to reveal the source of the over-estimation of the changes caused by Arg mutants, but for completeness, these calculated energies will be used as they are in the present study. The average folding free energy changes predicted by webservers and third-party software are all relatively small, being less than 1 kcal/mol. The FEP calculated binding free energy changes indicate that mutations R108W and R179H cause relatively large changes compared to other mutations.

Mutation	NeEMO (Folding)	PopMusic (Folding)	I-Mutant (Folding)	DUET (Folding)	CUPSAT (Folding)	Foldx (Folding)	FEP (Folding)	Folding (Average)	FEP (Binding)
A77T	−1.03	−0.22	−0.75	−0.76	−0.29	−1.40	0.13	−0.74/−0.62	−0.35
D87G	−0.16	−0.49	−0.47	−0.729	−0.16	−0.60	−0.28	−0.43/−0.41	0.73
R108W	−0.36	−0.86	−1.32	−0.26	0.28	−0.18	−11.29	−0.45/−1.99	−1.44
N142S	−0.21	−0.27	−0.09	−0.02	0.17	−0.3	−0.98	−0.12/−0.24	0.64
R179H	−0.71	0.06	−0.15	−0.72	0.28	0.746	−8.78	−0.08/−1.32	−3.06

Table 2.1. The calculated binding and folding free energy changes due to mutations in kcal/mol. $\Delta\Delta G > 0$ indicates stabilization, while $\Delta\Delta G < 0$ shows destabilization. The "Folding (average)" column shows the average folding free energy changes calculated using the average folding free energy changes predicted by FEP, webservers, and third-party software (left), and folding free energy changes predicted by webservers and third-party software (right). The changes of the binding free energy were obtained only with FEP, since no reliable third-party tool currently exist.

As mentioned above, the mutations were predicted to have a small effect on both the folding and binding free energy (excluding the FEP results for Arg-involving mutations). This suggests that the disease-associated effect may not be related to these energies but may be linked to structural distortion or change of the internal dynamics/flexibility of the ARID domain caused by the mutations. Therefore, we review the structural features of the mutation sites below and elaborate on their possible linkage with the predicted effect of folding and binding free energy.

2.2. Effect of Mutations on Protein Structure

To analyze the mutations' plausible effect on the protein structure, here we investigated the side chains and backbone conformational changes resulting from mutations and discuss them with respect to structural integrity of the ARID domain and its interactions with DNA. The mutant is introduced into the structure using the Mutator Plugin, Version 1.3 in VMD [81] After that, the mutant structures were subjected to 10,000 steps of energy minimization to relax the structure and remove possible conflicting contacts. The structures are then

visualized in UCSF chimera [82]. The side chain conformation of the residues within 5 Å of the WT position or MT position, are shown in Appendix Figures B-1 and B-2, respectively. Appendix Figure B-1 shows side chain conformation of two disease-associated mutations mapped on the KDM5C ARID domain. The A77T mutation involves substitution of a hydrophobic Ala by a polar Thr and is located in a short turn of the ARID N-terminal. The mutation site is far away from the DNA binding interface and it is solvent-exposed. Neither the wild-type A77 nor the mutant T77 were found to be involved in any specific interactions (Appendix Figure B-1a,b) The mutation D87G is located in Helix 1 of the ARID domain and a charged residue, Asp, is substituted by a small residue, Gly. This mutation site is also far away from the DNA binding interface and it is totally solvent-accessible. The wild-type residue, D87, is not involved in any specific interaction and its side chain faces the water (Appendix B Figure B-1 c,d). Based on these structural observations and the results of folding free energy calculations, it can be summarized that these mutations do not solely affect the stability and the structure of the ARID domain. Similarly, since the mutation sites are far away from DNA, the binding interface, and the binding free energy is not predicted to be affected, one can assume that the mutations have minimal effect on ARID-DNA recognition.

Appendix Figure B-2 shows the side chains and backbone conformations of non-classified mutations mapped onto the ARID domain. The R108W is a positively-charged residue, Arg, substituted by an uncharged hydrophobic residue, Trp. This mutation occurs in the loop between Helix 1 and Helix 2 and is located close to the DNA binding interface (Appendix Figure B2-a,b). Since the mutation drastically changes the physico-chemical

property of the wild-type residue, it can be anticipated that this mutation may cause significant conformational changes. To address this possibility, we performed 20 ns MD simulations of the ARID domain and DNA complex and it was found that R108 does not form a direct hydrogen bond with DNA. Thus, the wild-type residue, R108, is probably not involved in specific interactions with DNA but may provide long-range steering towards the negatively-charged DNA. Figure 2.2 shows the electrostatic potential of WT KDM5C ARID domain and the ARID doman with mutation R108W generated by DelPhi software [83-85]. It can be seen that the electrostatic potential at the mutation site is changed from positive to negative upon the mutation. Since the DNA is highly negatively-charged, this electrostatic potential change nearby the DNA binding interface will probably decrease the ARID–DNA binding affinity and specificity, which is consistent with predictions of the protein binding free energy changes. Further, salt bridge analysis indicated that the R108 forms a transient salt bridge with the neighboring amino acid, E74. Figure 2.3b shows the distance between the oxygen atom of E74 and the nitrogen atom of R108 in the MD simulation of the ARID domain and DNA complex. Using a cut-off distance of 4 Å as an indication of formation of a salt bridge, it was found that such a salt bridge is formed in 17.4 ns out of 20 ns (87% of the simulation time). Thus, the mutation R108W will delete the salt bridge and will probably affect the protein's stability, which is consistent with prediction of the protein folding free energy changes. The other mutation, N142S, occurs in a loop between Helix5 and Helix6, and results in a polar uncharged residue, Asn, substituted by another polar uncharged, but smaller, residue, Ser (Appendix Figure B-2c,d). Such a mutation preserves the biophysical characteristics of the mutation site and is

expected not to affect the stability and structural integrity of the ARID domain. The mutation R179H involves a positively-charged residue, Arg, substituted by a polar residue, His. It is located in the loop of the ARID domain C-terminal, which is far from the DNA binding interface and is totally solvent-exposed (Appendix Figure B-2e,f).

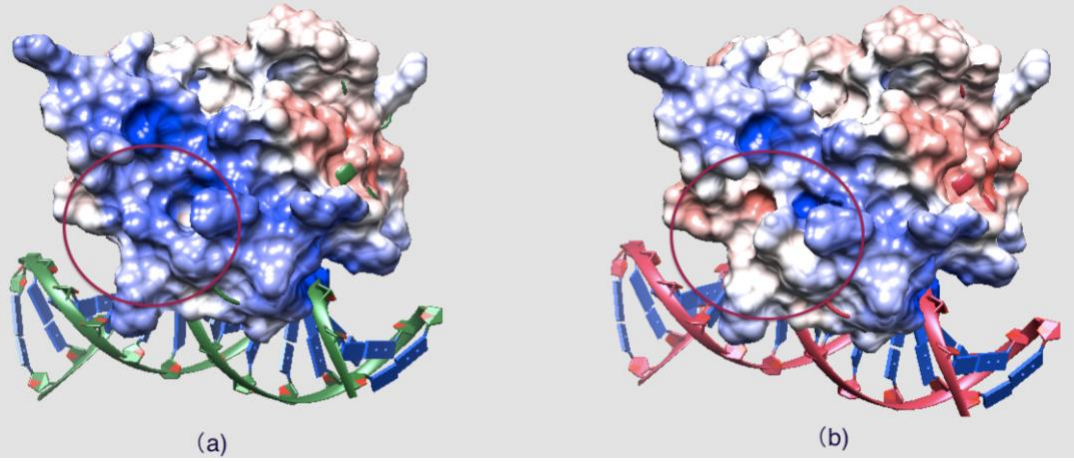


Figure 2.2 (a) Electrostatic potential of the WT KDM5C ARID domain; and (b) the electrostatic potential of the KDM5C ARID domain with mutation R108W. The mutation site is marked with a red circle. The positive potential region is colored with bule and the negative potential region is colored with red.

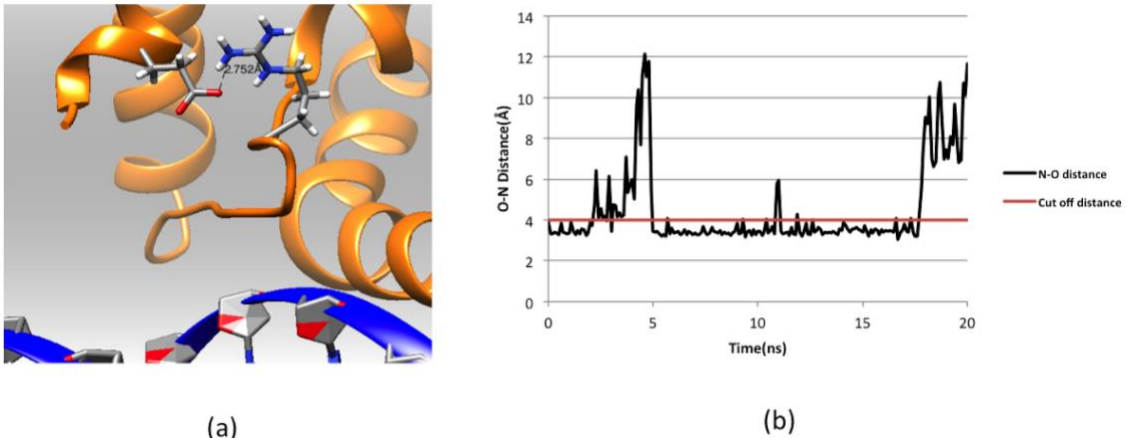


Figure 2.3 (a) Part of the ARID domain zoomed at the salt bridge Glu74-Arg108; and (b) salt bridge analysis for Arg108 and Glu74 in the KDM5C ARID domain: N–O distance shows the distance between oxygen atom of Glu74 and nitrogen atom of Arg108 in the 20 ns simulation. The cutoff distance of forming salt bridge is 4 Å and marked with red line in the graph.

2.3. Residue Conservation via Multiple Sequence Alignment

Further, we investigate the conservation pattern of the KDM5C ARID domain amino acid positions based on the sequence alignment of human ARID domain proteins. The alignment (Appendix Figure B-3) shows that the two disease-associated mutations (A77T and D87G) are conserved in the KDM5 family and D87 is conserved in ARID1, ARID2, and ARID3 families, as well. All non-classified mutations are not conserved in the alignment, including the alignment of only KDM5 family members. However, position 108 is predominantly taken by positively-charged residues, either Arg or Lys. Thus, a substitution to hydrophobic, uncharged Trp may not be tolerable. Combined with the predicted large change of the folding free energy and the change of the electrostatic potential, R108W mutation is predicted to be disease-associated. The other two non-classified mutations, N142S and R179H, occur at sites that are not conserved and there is no pattern to indicate the conservation of physico-chemical property of the wild-type residue. Even more, the substitutions Asn to Ser and Arg to His are found to exist in some family members (ARID3 and ARID4A), which suggest that such substitutions are tolerable.

Overall, the most highly conserved parts of the ARID domain are located on Loop1, Helix2, Helix3, Helix4, Loop2, and Helix5. Recent study showed that the KDM5A ARID domain binds DNA through the motif CCGCCC and the DNA binding interface includes Loop1 and a helix-turn-helix DNA binding motif formed by Helix4, Loop2, and Helix5 [86]. More specifically, six key residues (Pro103, Lys112, Gly123, Gly124, Trp134, and Tyr 157) are conserved in all human ARID-containing proteins, which indicates their importance for protein function.

2.4. Evolutionary Conservation and Protein Interacting Investigation Using the ConSurf Server and IBIS Server

The ConSurf server is a bioinformatics tool for estimating the evolutionary conservation of amino/nucleic acid positions in a protein/DNA/RNA molecule based on the phylogenetic relations between homologous sequences. The ConSurf server result (Figure 2.4) shows that the N-terminal of the ARID domain is one of the most highly-conserved parts in the ARID domain, which is probably essential for protein's function. We also predict the protein interacting partners and binding sites in the KDM5C ARID domain using the NCBI Inferred Biomolecular Interactions Server(IBIS) [87].The results show that Asp87 is a plausible zinc ion binding site. This binding sites is not verified experimentaly, but offer an implication that the N-terminal of ARID may be involved in some currently-unknown function.

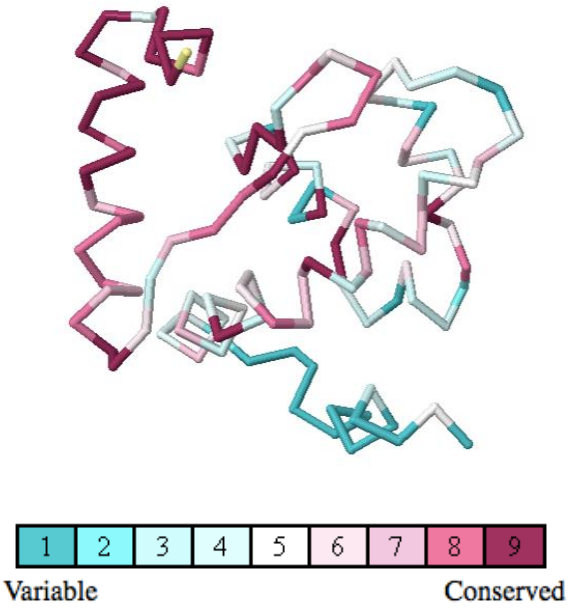


Figure 2.4. Evolutionary conservation analysis of the ARID domain using the ConSurf Server. The conservation grades are color-coded onto each amino acid of the KDM5C ARID domain.

2.5. Experimental Results

The mutations A77T and D87G affect the overall structure of the ARID domain slightly, but the percentage of each secondary structure of the mutants was in the same range as the wild-type (Table 2.2). In general, the effects of both A77T and D87G are the increase of the unordered structure percentage of the protein. While in the A77T mutation, the proportion of the structure shifted from alpha helix and turns to unordered; in D87G the shift came from of alpha helix and beta strand.

	Helix	Strand	Turns	Unordered
WT	14%	31%	20%	35%
A77T	13%	31%	19%	38%
D87G	13%	30%	20%	38%

Table 2.2 Percentage of secondary structures of ARID proteins analyzed by using CONTINNLL [88] with the online tool Dichroweb [89].

The results from the urea denaturation experiments (Table 2.3 and Appendix Figure B-4) indicate that both mutations caused a lower integrity protein structure (lower ΔG, easier to denature) than the wild-type, where the A77T is relatively more stable compared to D87G (but the difference is very small). There is a difference of the free energy of unfolding value of the ARID wild-type and the mutants. The two different methods to calculate the ΔΔG yields different value but the trends are the same, where the two mutants are less stable than the ARID wild-types, and D87G is less stable than A77T. The ΔΔG of the mutants A77T and D87G obtained by urea-induced unfolding monitored by CD are in the same order of magnitude compared to the *in silico* folding free energy predictions (Table 2.1).

Protein	$\Delta G_{app}^{H_2O}$ **(kcal·mol^{-1})**	$\Delta\Delta G_{app}^{H_2O}$ **(kcal·mol^{-1})**	**[urea]$_{½}$ (M)**	$\Delta\Delta G_{app,2}^{H_2O}$ **(kcal·mol^{-1})**
ARID WT	3.51 ± 0.32		3.99 ± 0.02	
A77T	2.41 ± 0.05	1.10	3.07 ± 0.02	0.70

D87G	1.82 ± 0.01	1.70	2.99 ± 0.03	0.76

Table 2.3 Results from an analysis of urea denaturation curves for ARID Wild-Type, A77T, and D87G variants.

3. Methods and Experimental Section

3.1. Structures

The ARID domain contains 90 amino acids and its sequence is mapped onto the KDM5C protein sequence from position 79 to 169. There is an NMR structure of the KDM5C ARID domain (PDB ID: 2JRZ) [90] in the Protein Data Bank (PDB) [91], which was used for modeling the ARID domain stability. The modeling of the effect of mutations on ARID-DNA interactions requires the 3D structure of ARID-DNA complex, which is not available in the PDB and was generated *in silico*. For this purpose, we applied structural alignment between the KDM5C ARID domain (PDB ID: 2JRZ) and all available ARID-DNA complexes in PDB. The lowest RMSD value (2.22 Å) calculated from structural alignment (the alignment between the DNA binding interface of the KDM5C ARID domain and the ARID domain in the available complex structures) was found for the solution structure of the dead ringer ARID-DNA complex (PDB ID: 1KQQ) [92]. The dead ringer and the KDM5C ARID domains' structural similarity (showed the lowest RMSD value (2.22 Å) calculated from structural alignment) was the highest for the residues situated at the protein-DNA interface, which suggested that the binding mode is preserved (Appendix Figure B-5). Thus, the model ARID-DNA complex was built by superimposing the

KDM5C ARID domain onto the dead ringer ARID-DNA complex and replacing the dead ringer ARID domain with the KDM5C ARID domain. Then, we saved the structure of the KDM5C ARID domain and DNA with untransformed coordinates as our model using the UCSF Chimera [82]. The DNA sequence in the model was kept the same as in the dead ringer ARID-DNA complex since the KDM5C ARID was not reported to show a DNA binding preference.

3.2. ARID Folding and Binding Free Energy Changes

We calculated the folding free energy change (ΔΔG) and the binding free energy change (ΔΔΔG) based on free energy perturbation theory (FEP) [93, 94]. The free energy calculations of five mutations (A77T, D87G, R108W, N142S, and R179H) were performed with the NAMD program, version 2.9 [95] using alchemical transformations via the so-called dual topology approach [95, 96], where both the initial and final states were defined concurrently. Periodic boundary conditions and a 12 Å cutoff distance for non-bonded interactions were applied in the system. Each FEP simulation was run using a CHARMM22 force field [97] and each mutation was carried out with one 18 ns run and four 5 ns runs. The initial protein structure used for each run was randomly taken from the trajectory of a 10 ns long equilibration. The results obtained with 18 ns and 5 ns runs were very similar and most of the 5 ns runs showed good convergence comparable with the convergence of 18 ns run. This motivated us to carry the rest of the FEP using 5 ns simulations. Then, the output of FEP simulations was analyzed with the ParseFEP Plugin, Version 1.9 [98] in Visual Molecular Dynamics (VMD) [81]. Also, it has to be pointed out that Gly is a very particular case in FEP calculations since the library of hybrids contains

the dual topologies for amino acids with a true side chain and the alpha carbon of Gly atom has to be modified in the transformation. For that reason, most patches cause problems and mutating glycine caused some angle and dihedral parameters to be duplicated, possibly modifying backbone conformational preferences [99]. Similar problems were also observed in our FEP calculation and, here, the FEP calculations of D87G were carried out for 1 ns with 0.5 fs time steps. For completeness, these calculated energies of D87G are used as they are in the present study.

The calculations of the effects of mutations on the folding free energy were performed utilizing the thermodynamic cycle we have developed in the past [20, 24, 100, 101] (Appendix Figure B-6a). The main assumption in this model is the unfolded state, which is considered to be made of two structural segments: (i) a structural three-residue segment centered at the mutation site; and (ii) the rest of the protein being mutation-independent [20, 24, 100]. This allows for canceling mutation-independent components of the unfolded state. Thus, the folding free energy change due to a mutation was calculated with the following equation:

$$\Delta\Delta G_{folding} = \Delta G_{folding\_WT} - \Delta G_{foldig\_MT} = G_{folded\_WT} - G^{3}_{unfolded\_WT} - G_{folded\_MT} + G^{3}_{unfolded\_MT} \quad (2.1)$$

,where $G^{3}_{\text{unfolded\_}X}$ is the free energy of the unfolded state of the three-residue segments at the center of mutation site, and *x* stands for WT or MT, respectively.

The effect of mutations on the binding free energy was calculated with the following thermodynamic cycle (see refs for more details [20, 102-105]) (Appendix Figure B-6b), and the corresponding equation is provided below:

$$\Delta\Delta G_{binding} = \Delta G_{binding\_WT} - \Delta G_{binding\_MT} = G_{bounded\_WT} - G_{unbounded\_WT} - G_{bounded\_MT} + G_{unbounded\_MT} \quad (2.2)$$

,where the unbounded state means the protein is taken away from its partner and bounded state means the protein forms a complex with its partner protein.

3.3. Utilizing Webservers and Third Party Software

Third-party methods were also used to predict protein folding free energy change, including webservers and stand-alone computer algorithms. The webservers used to predict the folding free energy changes upon single point mutations include NeEMO [106], PopMusic [107], I-Mutant 2.0 [108], DUET [109], and CUPSAT [110]. Additionally, a computer algorithm, FoldX 3.0 Beta3 [111, 112], was used to predict the folding free energy changes upon single-point mutations. Currently, no reliable third-party software or a functioning webserver for predicting binding free energy changes are available.

3.4. Molecular Dynamics Simulation

We carried out MD simulations to investigate mutations' effects on the dynamics on the ARID domain. The simulations were set up within the NAMD program, version 2.9 [95], using the CHARMM22 force field [97]. The PDB structure taken from Protein Data Bank [91] was used as the initial structure. To relax conflicting contacts, energy minimization was performed using the conjugate gradient energy minimization of 10,000 steps. The protein was solvated in a water box with a layer of water extending 10 Å in each direction before the minimization and equilibration with periodic boundary conditions. Temperature and pressure in the simulation were set to 298 K and 1 bar. Each mutation was repeated for

three 100 ns runs using 2 fs time steps. The trajectory files were analyzed by using VMD plugins [81] in order to obtain the RMSD, RMSF, and salt bridges.

3.5. Electrostatic Potential Calculation

The DelPhi program was used to perform the electrostatic potential calculations using the following parameters: scale = 2 grid/Å; percentage of protein filling of the cube = 70%; dielectric constant = 2 for the protein and 80 for the solvent; and water probe radius = 1.4 Å. We outputted the DelPhi-calculated potential map into a file in CUBE format, which was further opened and analyzed in UCSF Chimera.

4. Discussion and Conclusion

The KDM5C ARID domain binds to DNA and the formation of an ARID-DNA complex is important for the KDM5C function in humans [62, 64, 86]. Our analysis shows that A77T and D87G have minimal effect on the ARID domain's DNA binding, which indicates that the disease-associated mechanism is probably not due to the alteration of DNA binding. It is also interesting that both of the disease-associated mutations are located onto the N-terminal of the ARID domain and both of the mutations are far away from the ARID domain's DNA binding interface. We speculate that some not-yet-discovered function of the KDM5C protein is associated with the ARID domain's N-terminal. To test this, we analyzed the KDM5C ARID domain using the ConSurf Server [113-116].The Consurf results support our speculation and show that both of the disease-associated mutations are located in the most highly-conserved part of the ARID domain and possibly cause a change in an important function of the protein. Additionally, D87 is predicted to

be a plausible zinc ion binding site and further supports that some currently-unknown function is linked to N-terminal of the ARID domain.

Previous studies show that KDM5C is a muti-functional protein and inter-domain interactions are identified among the JmjN domain, N-PHD domain, and JmjC domain [59, 64, 71]. The interaction between the JmjC domain is important for the demethelytion activity. The N-PHD domain and JmjC domain can bind to the same histone tail at Lys4 and Lys9. Both of pathogenic mutations happen in the N-terminal of the ARID domain and are close to the linker part between the ARID and JmjN domains. This suggests that A77T and D87G may be involved in some unknown interaction among JmjN, ARID, PHD, and JmjC domains. Currently, only the ARID domain structure is available and the arrangement of the KDM5C domain is unknown.

Our study also evaluates three non-classified mutations' effects on the KDM5C ARID domain. Among them, the R108W causes a loss of a salt bridge, slightly affecting protein's stability and ARID-DNA binding affinity. Therefore, we speculate that R108W is a disease-associated mutation based on altering structural features rather than on the calculated free energy changes. In addition, as demonstrated, R108W changes the electrostatic potential near the DNA binding site which may affect the specificity of ARID-DNA binding.

In our work, protein binding and folding energy changes were calculated with FEP, webservers, and third party software. Limitation about the technical issues in FEP calculations are observed for the Arg- and Gly-involved mutations, possibly causing less reliable predictions. Therefore, other methods, including webservers and third-party

software were also applied in the free energy calculation to compare with the FEP results. Furthermore, the experimental results of the mutants A77T and D87G are obtained by urea-induced unfolding methods, showing the same order of magnitude compared to the folding free energy calculation. Another limitation about this work is that our speculation about the unknown function in the N-terminal of the ARID domain has not been experimentally verified and, currently, the only known function about the ARID domain is the DNA binding interaction. However, our work implicates that the sites 77 and 87 may be involved in some other function or interaction different from cognate ARID-DNA binding. This provides motivation for future studies to further investigating other functions of KDM5C.

PART2: Computational model for quaternary structure of Lysine-specific demethylase 5C (KDM5C) protein

1. Introduction

The epigenetic processes control transcription of genes and result in the widely different gene expression patterns in different tissues and organs [117, 118]. The most frequently occurring histone modifications involve acetylation, phosphorylation, methylation, ubiquitination and crotonylation [119, 120]. Lysine methylation is one of the most important histone modifications among them, and has a crucial role in heterochromatin formation, X-chromosome inactivation and transcriptional regulation. Histone lysine methylation occurs in histones H3 and H4, and this methylation results in lysine residue's three different methylation states (mono-, di-, and tri-), which are associated with different nuclear features and transcriptional states[56].

The *KDM5C* gene (also known as *JARID1C* and *SMCX*) is located on the X chromosome and encodes a ubiquitously expressed 1,560-aa protein, which plays an important role in transcriptional regulation and chromatin remodeling [71]. The KDM5C protein belongs to the JARID subfamily of JmjC-containing proteins. The function of JmjC domain is to specifically demethylate di- and trimethylated lysine 4 on histone 3 [59]. The KDM5C acts as a transcriptional repressor and many mutations in the *KDM5C* gene has been shown to cause X-linked mental retardation (XLMR) and the syndromic Claes-Jensen-type disease [48, 59, 67]. Most of disease-causing mutations are located on JmjC domain, ZF domain (C5HC2 zinc finger domain) and inter-domain regions. These mutations are expected to affect the protein stability and enzymatic activity [59, 121].

Overall KDM5C is a multi-functional protein, consisting structurally of several well-defined domains, including ARID, JmjN, JmjC, ZF, and two PHD zinc finger domains. Little experimental data is available about the 3D structure of the entire protein, individual domains and the interactions with its partners. The first experimental structure of one of the KDM5C domains is the solution structure of ARID domain, released in 2007 [90]. In the past, we used it to generate a 3D model of the ARID-DNA complex [48]. Very recently, the 3D structure of part of the human KDM5C protein (JmjN, JmjC and ZF domains) was released [122]. The same work revealed that ARID and PHD1 domains contribute to the histone substrate recognition despite not being directly required for demethylase activity [122]. However, there is still no experimental or theoretical 3D structure of KDM5C protein which includes JmjN, ARID, JmjC, C5HC2 zinc finger (ZF), and two PHD domains.

Several XLMR-associated mutations were shown to reduce KDM5C demethylase activity and binding to the H3K9me3 peptide [123]. Specifically, ARID domain is a DNA-binding domain in which two missense mutations (A77T and D87G) were reported [67]. The other domain, the JmjC domain, catalyzes demethylation of H3K4me3 to H3K4me1, and three missense mutations (D402Y, S451R and Y642L) are experimentally shown to reduce KDM5C demethylase activity [123]. Several pathogenic mutations (A388P, R731W and Y751C) were reported in the PHD1 domain (a histone methyl-lysine binding motif) and ZF domain and were shown to reduce the demethylase activity as well [123]. However, the molecular mechanism of the above-mentioned disease-causing mutations is mostly unknown. Perhaps this is due to the lack of experimental or theoretical 3D structure of KDM5C protein, which would allow for modeling the effects of mutations on domain stability and inter-domain interactions. The goal of this work is to fill this gap by developing a 3D structural model of KDM5C quaternary structure and using it to model the effects of disease-causing mutations on KDM5C stability, dynamics and inter-domain interactions. It is understood that such multi-functional and multi-domain proteins may adopt various domain arrangements in different functional states. Thus, the quaternary structure that is reported in this work represents one of several domain arrangements of KDM5C that the protein may adopt during its functional cycle.

2. Results and Discussion

The results section has three major components, namely a report of building 3D models of the corresponding KDM5C domains and arranging them into quaternary structure of

KDM5C, and then validating and using the 3D structure of KDM5C protein to predict the effect of XLMD-linked mutations. Below we describe the results in sequential order.

2.1 Modeling quaternary structure of KDM5C protein

2.1.1 Homology model of KDM5C PHD1 domain

As mentioned above, KDM5C is multi-domain protein consisting of well-defined domains as JmjN, ARID, PHD, JmjC and ZF domains. For most of them, JmjN, ARID, JmjC and ZF domains, there is experimental 3D structure available (Protein Data Bank (PDB) ID:2JRZ and 5FWJ)[90]. However, the 3D structure of PHD1 domain is not available and must be modeled. For this purpose, we used homology modeling, since high homology templates do exist. Thus, the amino acid sequence of KDM5C was submitted to the PSI-Blast [124] and the search was performed against the sequences of proteins in PDB database [91]. The best template for KDM5C PHD1 domain is the solution structure of another PHD domain within JARID family (PDB: 2E6R), with 84% sequence identity [125]. Then, we used SWISS-MODEL server [126] to generate the homology model for KDM5C PHD1 domain.

2.1.2 Quaternary structure of KDM5C catalytic core

We began the modeling of quaternary structure of KDM5C protein by taking advantage of recently released experimental structure (PDB ID: 5FWJ), which includes JmjN, JmjC and ZF domains. The experimental structure revealed a previously known fact that JmjN interacts with JmjC domain and that this interaction is crucial for the KDM5C protein stability and catalytic function [64, 122]. Furthermore, it is known that the ZF domain is required for catalytic activity of JARID proteins [127].

While it is known that JmjC domain demethylates di- and trimethylated lysine 4 on histone 3 (H3K4me2 and H3K4me3) along with the cofactors Ferrous ion (Fe2+) and alpha–ketoglutarate (2-oxoglutaric acid), the binding mode of histone peptide to JmjC domain is still unknown for the JARID family. Developing such a model is one of the goals of this work and we intend to generate it by taking the advantage of the experimental structures of other JmjC domain containing proteins. For this purpose, we first collected all existing experimental structures of JmjC domain bound to histone peptide. The experimental structures include the structure of KDM2A bound to H3K36me1 (PDB ID: 4QXH) [128], the structure of JMJD2B complexed with H3K9me3 (PDB ID: 4LXL) [129], the structure of human JMJD2D/KDM4D in complex with an H3K9me3 peptide (PDB ID:4HON) [130], the structure of KDM6B bound with H3K27me3 peptide (PDB ID: 4EZH) [131], the structure of KDM7A from C.elegans complexed with H3K4me3 peptide, H3K9me2 peptide and NOG (PDB ID: 3N9O) [132], and the complex structure of JMJD2A and trimethylated H3K36 peptide (PDB ID:2P5B) [133]. The structural alignment was applied using the Chimera [82] for all collected structures. Figure 2.5A shows the results of structural alignment and it can be observed that the structures are highly conserved among different JmjC containing proteins. The positions of the histone peptides are also similar among different proteins, especially for the regions close to the bound Lys4 residue. Therefore, we took the average coordinates of these histone peptide backbones, which were manually assigned to the histone 3 peptide residues from 1 to 10. The side chain of each residue was generated using the most probable rotamer from the Dunbrack backbond-dependent rotamer library [134] as implemented in Chimera [82].

Further, Lys4 and Lys9 in the peptide were modified to H3K4me3 and H3K9me3 using Avogadro [135], since *in vivo* they are methylated as they bind to JmjC and PHD1 domains. Finally, auto-optimization was performed for the peptide with Avogadro [135] to correct for bad contacts and improper bonds.

Since the template experimental structure contains Mn2+ ion and inhibitor MMK instead of enzymatic cofactors Fe2+ ion and 2-oxoglutaric acid, the enzymatic cofactors must be placed in the model to replace the template cofactors. This was done by superimposing template structure (structure of rice JMJ703 in complex with alpha-KG, PDB ID: 4IGO) onto the structure of KDM5C catalytic core and then replacing the template cofactors with enzymatic cofactors. The final model of KDM5C catalytic core bound with histone peptide and enzymatic cofactors is shown in Appendix Figure 5B.

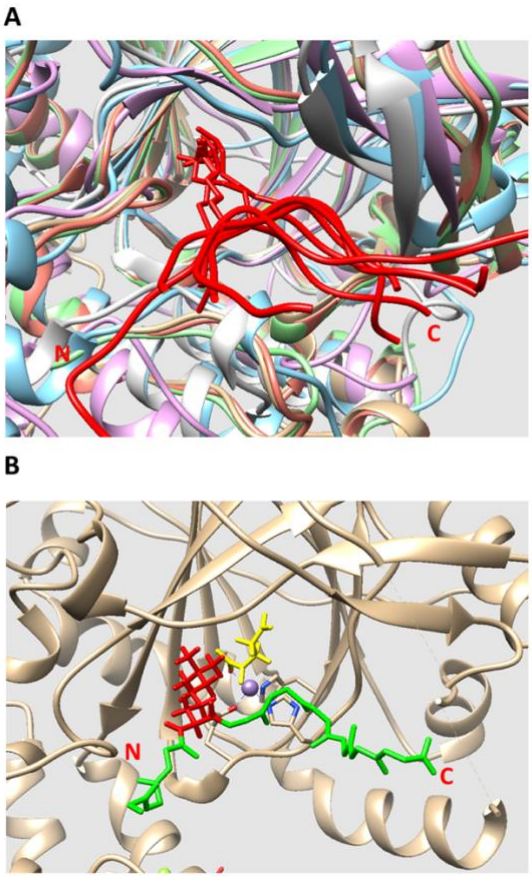


Figure 2.5: (A) Structural alignment of all existing JmjC domains bound to histone peptide. The histone peptides are marked in red. The JmjC domains are marked with other colors. The C and N terminals of the histone peptide are labeled as C and N, respectively. (B) The model of KDM5C catalytic core bound to histone peptide and enzymatic cofactors. The histone peptide backbone is shown in green and H3K4me3 is shown in red. Enzymatic cofactors Fe2+ ion and 2-oxoglutaric acid are marked with purple and yellow, respectively. The C and N terminals of the histone peptide are also labeled.

2.1.3 Modeling ARID domain's interactions in quaternary structure of KDM5C

ARID domain is the DNA binding domain of KDM5C protein and a NMR structure is currently available (PDB ID: 2JRZ). Also, in our previous work [48], a model of ARID-DNA complex has been generated. The presence of DNA provides constrains of how ARID interacts with the rest of KDM5C domains. It should be noted that the ARID domain is the second domain of the KDM5C sequence (after the JmjN domain) and a link consisting of 23 amino acids connects these two domains. This provides further constrains of mutual orientation and positioning of ARID and JmjN domains. Since the quaternary structure of JmjN, JmjC and ZF domains is already available (see above), the next question is to predict ARID domain position and orientation with respect to the JmjN, JmjC and ZF domains.

Thus, we first applied the ZDOCK server [136] to search possible binding modes of the ARID domain to the quaternary structure of JmjN, JmjC and ZF domains. The ten best predictions were collected and analyzed. It should be reiterated that JmjN and ARID domains are connected via a linker of 23 amino acids. Thus, the first consideration that was made in the analysis of the ten best binding modes was to remove binding modes that result in ARID position and orientation that makes it impossible for ARID and JmjN to be connected by 23-amino acid linker. It should be pointed out that the linker region between JmjN and ARID domain is also included in the crystal structure [122] and it is simply wrapped around the JmjC domain. It is tempting to use the linker in the experimental structure of JmjN, JmjC and ZF domains (PDB ID 5FWJ) as a guide in positioning the ARID domain. However, it should be clarified that the linker adopts crystallographic conformation in absence of the ARID domain and thus its conformation in 5FWJ may be misleading. Due to this, we decided to delete the linker region from the experimental

structure and rebuild 3D structure of the linker connecting ARID and JmjC domains using the LOOPY program [137]. The results showed that the 23 amino acid linker is not long enough to connect JmjN and ARID domain in six out of ten binding modes. Thus, these six models were deleted. To compare the rest of the four binding modes, we computed the RMSD among them (Table C-1). The RMSDs among four models is relatively small, ranging from 1.80 to 3.46Å. This indicates that these four binding modes are quite similar and that the binding interface is almost identical (Appendix Figure B-7A).

To further select the best model and test its stability, we run molecular dynamic (MD) simulations for each binding mode. Three 10ns parallel runs were performed for each model with CHARMM22 force field [97] in NAMD [95]. Default charges of titratable residues were assigned, since the pKa analysis predicted no ionization changes in physiological pH. We first calculated RMSD of ARID and JmjC domains to observe the overall structural change along the simulation time (Appendix Figure B-7B). The RMSD ranged from 5 to 7.5 Å and the overall structure became stable after 5ns for all models. To further study the binding mode stability, we also calculated the RMSD of the interfacial residues of each binding mode. We identified the interfacial residues by calculating the solvent accessible surface area (SASA) change of each residue in the complex and unbound domains. A residue is defined as an interfacial residue if the SASA change is not equal to 0. Since the identified interfacial residues were slightly different in each model due to the structure difference, we only selected the interfacial residues common in all models for our analysis and the RMSD results are shown in Appendix Figure B-7C. The RMSD of

interfacial residues ranges from 3.5 to 7 Å and reaches stable value after 5ns simulation time.

To finalize the model of ARID bound to KDM5C catalytic core, we took the twelve MD generated trajectories (Note that for each of the four binding modes we generated three MD trajectories) and selected 50% with lowest RMSD among them. The last 5ns of the trajectories were taken to calculate the averaged structure using VMD tcl script [138]. The averaged structure was subjected to 10,000 steps of energy minimization to relax the structure. The finalized model of ARID bound to KDM5C catalytic core is shown in Appendix Figure B-7D.

2.1.4 Modeling PHD1 domain's interaction in the quaternary structure of KDM5C

The KDM5C PHD1 domain is close in sequence to the JmjC domain (the linker is 13 amino acids long) and it is expected to bind to tri-methylated H3K9 residue (H3K9me3) [59, 139]. Indeed, a recent study indicated that the PHD1 domain is not required for the demethylase activity but does contribute to the recognition of the substrate peptide [59, 122]. Currently, there is no available experimental structure of the KDM5C PHD1 domain bound to H3K9me3 and KDM5C catalytic core.

Above we described the modeling of quaternary structure of KDM5C catalytic core bound to histone peptide and ARID domain. Since a homology model of KDM5C PHD1 domain was already generated (see above), we applied the ZDOCK server [136] to predict the binding modes of PHD1 domain and the model of KDM5C catalytic core bound to histone peptide. The Z-dock server predicted the ten most plausible binding modes. We evaluated them by applying constraint of the linker length between PHD1 and JmjC domain.

Since PHD1 and JmjC domains are connected with a 13-residue-long liner (we built a 3D structure of the linker connecting the PHD1 and JmjC domains using the LOOPY program [137]), we only selected the binding mode in which a stretch of 13 amino acids is able to connect the two domains. This resulted in three possible binding modes (Appendix Figure B-8A). The PHD1 domain in all binding models is close to the substrate peptide but adopts different orientations.

To further test the binding stability in the rest of the three binding modes, we performed MD simulations. Since the substrate peptide is very flexible and tends to move away from the JmjC catalytic core, the position of H3K4me3 residue was fixed to make the substrate peptide stay in the catalytic core during the simulation time. The enzymatic cofactors, Fe2+ ion and 2-oxoglutaric acid, were also fixed in the simulation to prevent them from moving away from the JmjC catalytic core. Five 10ns parallel runs were performed for each model using AMBER ff14SB force field [140] in NAMD [95]. As mentioned above, default charges were used for titratable groups since no protonation changes were predicted by the pKa analysis. The averaged RMSDs calculated for the complex JmjC domain, substrate peptide and the PHD1 domain are shown in the Appendix Figure B-8B. The RMSD calculated for model1 and model3 is around 7 Å and the PHD1 domain stays bound with the JmjC domain during the simulation time. At the same time, model2 shows much larger RMSD values. It was observed that the PHD1 domain moves away from the JmjC domain and substrate peptide. Therefore, model2 was removed from our protocol, while model1 and model3 were subjected to further considerations.

To further select the best binding mode among mode1 and model3, we took advantage of experimental data that indicates that PHD1 domain binds substrate peptide at H3K9me3. Therefore, we identified all residues in PHD1 domain, which have any atom within 6 Å distance from H3K9me3 in the last 2.5ns simulation time (Appendix Figure B-8C). Comparing with model1, the analysis of the trajectories indicated that the PHD1 domain in model3 is farther away from the H3K9me3. Due to this, we selected model1 for further investigations. In particular, we paid attention to plausible stabilizing charge-charge interactions between PHD1 domain and peptides. Since the methylation of Lys does not neutralize the charge, we searched for acidic residue in PHD1 domain within a distance of 6 Å from H3K9me3. Thus, we identified two acidic residues: Glu360 and Glu375. Further N-O distance analysis did not indicate stable salt bridges between H3K9me3 and acidic residues in PHD1 domain. However, two other salt bridges within PHD1, Glu381–Histone ARG2 (H3R2) and PHD1 Glu375–Histone Arg8 (H3R8), were identified (shown in Appendix Figure B-8D). It should be mentioned that the structural segment (the linker) connecting ARID domain and PHD1 domain was not included in the model due to lack of homology template and thus PHD1 domain may appear to be more flexible in the simulations than it actually is and cause us to observe less salt-bridges. Noting the consistence with experimental data, we took model1 as the most plausible binding mode (the structure is shown in Appendix Figure B-8F).

2.1.5 Quaternary structure of KDM5C JmjN, ARID, PHD1, JmjC and ZF domains bound with DNA, substrate histone peptide and enzymatic cofactors

In our previous work, we generated the 3D structure of ARID-DNA complex [48]. Here, we took advantage of our previous work and included the DNA in the KDM5C quaternary structure model discussed above. Combining all the models described above, we finally generated the quaternary model of KDM5C protein - including JmjN, ARID, PHD1, JmjC, ZF domains, substrate histone peptide, enzymatic cofactors and DNA (Figure 2.6).

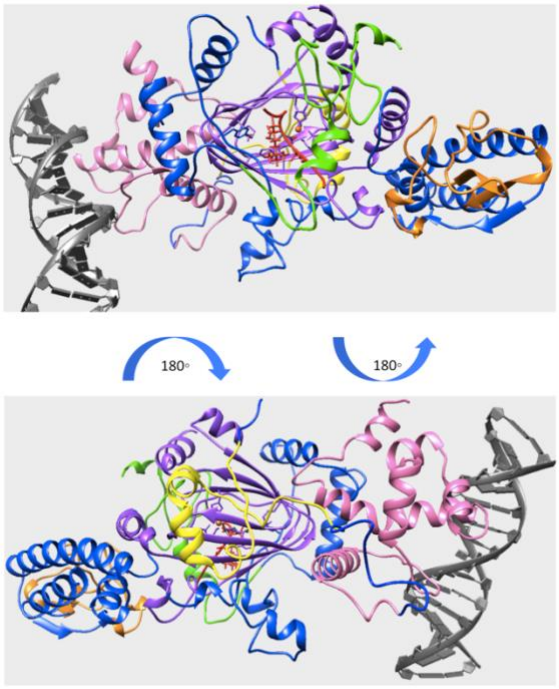


Figure 2.6: Finalized model of quaternary structure for the KDM5C including JmjN, ARID, PHD1, JmjC and ZF domains bound to DNA, substrate histone peptide and enzymatic cofactors. KDM5C JmjN, ARID, PHD1, JmjC, ZF domains, inter domain region, substrate histone peptide, enzymatic cofactors and DNA are marked with yellow, pink, green, purple, orange, blue, red, orange, red, and gray, respectively.

2.2 Validation of KDM5C quaternary structure model

In this paragraph we will investigate domain spatial arrangement within the quaternary structure of KDM5C model using biophysical considerations (charge complementarity, inter-domain salt bridges, binding motif preservation, MD simulations and experimental observations) and predictions of interfacial patches. With our work focusing on predicting spatial positions of ARID and PHD1 domains within KDM5C quaternary structure, we will continue to focus on validating ARID and PHD1 interactions with the rest of the KDM5C domains.

ARID domain: The binding interface is mostly made of ARID domain's N-terminal and C-terminal regions, including Helix1 and Helix6 (Appendix Figure B-9A). Our previous evolutionary conservation study showed that the N-terminal of the KDM5C ARID domain is the most highly conserved region, indicating its essential function for KDM5C protein [48] and thus implicating its involvement in some important interactions. There exists an interface of the ARID domain in which the domain binds the DNA. In the model, it is Loop1 and a helix-turn-helix DNA binding motif formed by Helix4, Loop2, and Helix5, which is typical for DNA binding motifs. In the previous study [48], we demonstrated that DNA binding interface is positively charged, which is expected for interface binding to negatively charged DNA. Similarly, we calculated electrostatic potential of ARID domain and KDM5C JmjC and JmjN domain using Delphi [84]. The binding interface of JmjC domain is highly negatively charged while the binding interface of ARID domain is slightly positively charged (Appendix Figure B-9B). This indicates electrostatic complementarity and further validates the model. Furthermore, we submitted the sequence to cons-PPISP

webserver [141, 142], a consensus neural network method for predicting protein-protein interaction sites. The webserver predicts the residues which form binding sites for another protein. Here, we separately submitted ARID, PHD and JmjC domains to the server to predict the interaction sites on each domain. The results are shown in Appendix Figure B-10D,E. The predicted interaction sites in ARID domain are mostly in the N terminal helix, which is highly consistent with our model (Appendix Figure B-10D). We also perfomed MD simulations on the model of ARID bound with KDM5C catalytic core (see above). The RMSD analysis of the interfacial residues indicated that the binding is stable (Appendix Figure B-7C). Lastly we investigated plausible salt bridges formed at the interface of ARID and PHD1 domains, because salt bridges are frequently observed at transient domain-domain interfaces [143, 144]. Since we had previously performed 5 parallel runs, we only selected the salt-bridges observed in no less than two runs. The list of identified interfacial salt bridges is shown in Table 2.2. It can be seen that ARID interface is rich of interfacial salt bridges, indirectly indicating that the interface is correctly predicted.

PHD1 domain: The predicted binding interface between PHD1 and JmjC domains is shown in Appendix Figure B-10A and marked with green. It includes a short helix and two loops. The KDM5C PHD1 domain binds to H3K9me3 and reduction of its binding decreases enzyme activity [59, 139]. This mechanism is still unknown. However, a recent study indicated that PHD1 domain contributes to the histone substrate recognition, despite being not directly required for demethylase activity [122]. In our model, the PHD1 domain simultaneously binds to peptide and JmjC domain. Here we investigated whether PHD1

domain binds with the substrate histone peptide alone, or whether it also interacts with JmjC domain to help further stabilize its binding to substrate peptide. For this purpose, we carried out MD stimulations and monitored RMSDs (Appendix Figure B-8). It is shown that the PHD1 domain is highly flexible - the salt bridge formed between the substrate peptide and the PHD1 domain were not very stable during the simulations as well (Appendix Figure B-8D,E). To better understand this finding, we carried out the same type of MD simulations of closely related complexes with experimentally available 3D structures. Thus, we collected the experimental structure of the KDM5B PHD1 finger (having 61% sequence identity with KDM5C PHD1 domain) in complex with H3K4me0 [145] and ran 10ns simulation to study the RMSD of PHD1 domain and histone peptide. The results showed that the peptide moves away from PHD1 domain around 5ns (Appendix Figure B-10B). The overall PHD1 domain is very flexible as well (Appendix Figure B-10C). Thus, the simulations indicate that the binding between the substrate histone peptide and PHD domain alone is not stable. This is perhaps due to the high flexibility of both PHD domain and substrate peptide. This advocates that the binding of histone peptide is supported by PHD1 interactions with JmjC catalytic core. Indeed, very recent study of domain arrangement in KDM5B protein by small-angle X-ray scattering (SAXS) and rigid-body modeling [122] provided an evidence of the interaction between PHD1 domain and JmjC catalytic core. The model derived from SAXS indicates that the PHD1 domain is in close contact with the JmjC catalytic core and suggested cooperation between PHD1 and the catalytic core in KDM5 enzymes [122]. Considering both these experimental results and our model [59, 122], we speculate that the PHD1 domain interacts with the JmjC

domain to help stabilize the substrate peptide and itself, and further to position the H3K4 in the JmjC catalytic core. Furthermore, as we did for the ARID domain, we submitted the sequence to cons-PPISP webserver [141, 142]. The cons-PPISP predicted interaction sites in PHD domain are located in the binding interface in our model (Appendix Figure B-10E). Lastly, we identified interfacial salt bridges at the PHD1 interface (Table C-2).

2.3 Effects of disease-associated mutations on KDM5C domain's stability and interactions

Currently there are eleven amino acid mutations in KDM5C protein known to be causing XLMD. They were investigated to predict their effects on the KDM5C domain's stability and interactions. First we predict the effect on folding free energy (domain stability) using various webservers (Table 2.4). The results indicate that most of the disease-associated mutations substantially decrease domain stability, particularly mutations A77T, D87G, F642L, E698K and Y751C. Furthermore, mutations D87G and D402Y are located at the binding interface of the ARID and JmjC domains and their effect on domain-domain interactions was predicted using several webservers (Table 2.5). It can be seen that the mutation D87G significantly decreases the binding affinity while D402Y slightly increases it. Asp87 is in the binding interface (shown in Figure 2.5A) and participates in two salt bridges (Asp87-Lys459, Asp87-Arg460) across the binding interface. Replacing negatively charged D87 with small, uncharged residue will alter the salt bridges and reduce the ARID-JmjC binding. The other mutation, D402, involves negatively charged residue in which the wild type KDM5C is predicted to form a salt bridge with R159 of JmjC domain. Preplacing it with neutral residue is expected to alter the salt bridge and to decrease binding affinity. However, the mutation is predicted to slightly increase binding affinity of

ARID domain to JmjC domain. Perhaps this is due to the fact that R159 is involved in another salt bridge (Table C-2) and the D402-R159 salt bridge does not contribute to inter-domain interactions. However, this bridge may be essential for forming the wild type binding pose. Overall, all investigated mutations are predicted to alter the wild type domain stability and inter-domain interactions.

	mCSM	SDM	DUET	SAAFEC	Average
A77T	0.88	2.48	0.83	1.44	1.41
D87G	0.84	1.9	0.86	5.01	2.16
A388P	0.44	2.12	0.57	-0.13	0.75
D402Y	1.03	-0.8	1.04	-7.93	0.42
S451R	0.52	0.25	0.43	-3.74	0.40
V504M	0.16	0.88	0.16	-0.14	0.27
F642L	1.43	1.05	1.56	1.71	1.44
E698K	0.22	3.27	0.29	4.16	1.99
L731F	1.5	0.31	1.69	0.01	0.88
R750W	-0.56	-1.74	0.76	1.01	0.40
Y751C	1.66	-1.76	1.57	1.44	1.56

Table 2.4: Folding free energy change upon mutations (kcal/mol). Positive value indicates that the mutation decreases domain stability. The most contradictory predictions are underlined and not used. The averaged prediction is shown in the last column.

	BeAtMuSiC	SAAMBE	MutaBind	Average
D87G	1.65	-0.54	0.42	1.04
D402Y	-0.22	-0.66	0.18	-0.23

Table 2.5: Binding free energy change upon mutations (kcal/mol). Positive value indicates that the mutation decreases binding affinity. The most contradictory predictions are underlined and not used. The averaged prediction is shown in the last column.

3. Materials and Methods

3.1 Sequence alignment and homology modeling

The homologues of the KDM5C protein and their domains were retrieved from the Protein Data Bank database using the protein basic local alignment search tool [124], applying Position-Specific Iterated BLAST. The homologues with the highest sequence similarity were used to generate the homology model of KDM5C domains using the Swiss-Model webserver [126].

3.2 Protein docking and inter domain linker building

The KDM5C domains interactions were predicted with ZDOCK 3.0.2 [136], which searches all possible binding modes in the translational and rotational space between two proteins/domains and evaluates each pose using an energy-based scoring function. The linker regions were predicted with LOOPY [146], a computer algorithm to predict loop conformation provided the amino acid sequence.

3.3 Molecular Dynamics simulation

Molecular dynamics (MD) simulations were performed with the NAMD program, version 2.11b [95]. The force field used in the simulation, including the substrate peptide and enzymatic cofactors, was Amber force field in the Amber tools 15 [79]. The inpcrd and prmtop files were generated with Amber tools 15 [147]. Other simulations, with only standard protein residues, were performed with CHARMM22 force field [97]. A 10,000 steps minimization was performed for all simulations to relax plausible overlaps. Generalized Born implicit solvent (GBIS) was applied in the simulations and the time step

was set to 2 fs. The temperature in the simulation was set to 298 K. The trajectory files were investigated using VMD 1.9.2 [138] with related plugins in order to analyze the RMSD, RMSF, and salt bridges.

3.4 Change of folding and binding free energies upon missense mutations

The binding free energy changes upon missense mutations (ΔΔΔG) were predicted with webservers including BeAtMuSiC [148], MutaBind [149] and SAAMBE [30, 150]. The folding free energy changes upon missense mutations (ΔΔG) were evaluated with mCSM [151], SDM [152], DUET [109] and SAAFEC [153] servers.

3.5 Electrostatic Potential Calculation

The DelPhi program was used to perform the electrostatic potential calculations. The following parameters were applied in the calculation: scale = 2 grid/Å, percentage of protein filling of the cube = 70%, dielectric constant = 2 for the protein and 80 for the solvent, and water probe radius = 1.4 Å.

3.6 pKa shifts analysis

To investigate the possibility that some titratable residues may undergo protonation change upon formation of quaternary structure of KDM5C, we performed the pKa calculations with DelPhiPKa [154, 155], which surface-free Poisson-Boltzmann based approach to calculate the pKa values of protein ionizable residues, nucleotides of RNA and DNA. We first calculated the pKa's of titratable residues for unbound ARID, PHD1 domains and KDM5C catalytic core and then repeated the calculations using quaternary KDM5C structure. The pKa shifts are calculated by subtracting the pKas of quaternary KDM5C structure and the pKas of individual domains (details are provided in SI).

4. Conclusion

The KDM5 protein is of significant interest for the biomedical community due to its relevance to X-linked mental retardation [70] and importance in oncological drug development [156]. KDM5C protein is a multi-functional protein of 1560-aa length and 3D structure of the entire protein is currently unavailable. Only the 3D structure of ARID domain, and very recently of catalytic core, is available [122]. Here we fill this gap by reporting a 3D model of KDM5C protein quaternary structure, including JmjN, ARID, PHD1, JmjC and ZF domains bound to DNA, substrate histone peptide and enzymatic cofactors. The model was used to infer the effect of disease-causing mutations of domain stability and domain-domain interactions. From the model, it was demonstrated that the mutations significantly alter wild type domain stability and inter-domain interactions. This suggests that domain stability and domain spatial arrangement with KDM5C protein are essential for its wild type function.

Modelling of SAVs' molecular effects in Spermine Synthase:

1. Introduction

Polyamines are cationic polymers that play multiple important roles in a wide range of cell growth and development processes. [157-159]. This study focuses on human spermine synthase (*SpmSyn*), a protein whose function is to convert spermidine (SPD) into spermine (SPM). The reactant (SPD) and the product (SPM) are both polyamines, which are essential for normal mammalian cell growth and development [160, 161]. A previous study has illustrated that mutations in SpmSyn are associated with Snyder–Robinson syndrome (OMIM #309583, SRS) [13, 162, 163]. Mutations can affect SpmSyn's dimer and monomer stability and alter the wild-type hydrogen bond network, which is important for the enzymatic functionality [13, 101, 164]. All of these alterations cause the disruption of SpmSyn function and thus result in an abnormal SPM/ SPD ratio and SRS. SRS is a rare form of X-linked intellectual disability characterized by mild to moderate mental retardation, asthenic body build (marfanoid habitus), diminished muscle bulk, osteoporosis, kyphoscoliosis, dysmorphisms (facial asymmetry, full lower lip, long great toes) and nasal or dysarthric speech [162, 163]. Significantly decreased SpmSyn activity results in low levels of intracellular SPM and a decreased SPM/ SPD ratio for Snyder–Robinson syndrome patients. SpmSyn is thus an important drug target to restore the protein's function [24, 164].

SpmSyn consists of two structural domains, C-domain and N-domain, connected via a linker-domain. Structural and biochemical analyses have shown that the biological unit of

SpmSyn is a homo-dimer instead of a monomer [101, 165]. The C-terminal domain is the catalytic domain, which carries out the catalysis of SPD to SPM. In contrast, the N-terminal domain is not involved in catalytic function but plays a crucial role for the dimerization. Most of the binding interface is formed with the N-terminal domain and deletion of the N-domain disables dimerization and results in the lack of activity [24, 160]. Missense mutations occurring in SpmSyn can directly affect the wild-type properties of the active site in the C-domain or alter the binding interface in the N-domain to lower dimer affinity [24, 101]. Since the SRS is caused by various molecular mechanisms, combined *in silico* and *in vitro* investigations are necessary to reveal molecular effects of missense mutations in SpmSyn in order to identify drug-like small molecules for disease treatment [24, 164, 166, 167].

Here, we investigate the molecular effect of five SRS causing mutations located within the N-domain of SpmSyn: M35R, G56S, F58L, G67E and P112L. Since these mutations are away from the active center of SpmSyn, they are not expected to directly affect the catalytic function of SpmSyn, but rather to alter SpmSyn activity indirectly by perturbing other biophysical properties. Here we focus on two of them, the stability and the dimerization of SpmSyn.

Some of the abovementioned mutations were previously investigated; others are reported in this work for the first time. Thus, M35R was identified at the Greenwood Genetic Center from a patient diagnosed with SRS. The P112L is the SRS-causing mutation included in this work due to personal communication with Raymond family. The G56S (rs121434610), which occurs at a highly conserved residue within the N-domain

region of SpmSyn, greatly reduces SpmSyn activity and leads to severe epilepsy and cognitive impairment [168]. The F58L (rs397515549) also greatly reduces SpmSyn activity and leads to mental retardation along with severe osteoporosis [169]. The G67E (rs397515553) causes an ectopic kidney and early-onset epilepsy in addition to features characteristic of Snyder-Robinson syndrome and completely destroys SpmSyn activity in the patient's lymphoblastoid cells [162].

2. Results

2.1. Effect of Missense Mutation on Monomer Stability (in Silico Modeling)

Table 2.6 shows the results of monomer stability changes (changes of the folding free energy) due to missense mutations calculated with webservers and stand-alone computer algorithms. For most of the cases, predictions made with different algorithms are in good agreement. The most controversial prediction is made by FoldX, where F58L is predicted to stabilize the monomer while other tools give opposite results. The five disease-causing mutations are all predicted to decrease monomer stability. Specifically, M35R, G67E and G56S are predicted to dramatically decrease monomer stability.

Mutations	**PoPMuSiC**	**DUET**	**FOLDX**	**I-Mutant 2.0**	**SDM**	**SD**	**AV**
M35R	−0.93	−0.47	−0.42	−1.81	−2.93	1.06	−1.31
G56S	−1.99	−0.52	−3.50	−2.16	−3.51	1.24	−2.34
F58L	−1.77	−0.95	2.12	−2.72	−0.18	1.84	−0.7
G67E	−1.99	−1.26	−1.36	−0.18	−1.34	0.65	−1.22
P112L	−0.87	−0.06	−0.43	−0.83	−1.04	0.40	−0.65

Table 2.6 Predictions of monomer stability change due to missense mutations. The calculated folding free energy changes are in kcal/mol. $\Delta\Delta G > 0$ indicates stabilization while $\Delta\Delta G < 0$ indicates destabilization. Average value (AV) of folding free energy changes is given in the last column of the table. Standard deviation (SD) is also calculated to quantify the variation of energy changes.

2.2. Effect of Missense Mutation on Monomer Stability (in Vitro Experiments)

The patient samples showed a reduced level of SpmSyn protein for all the patients either by native or denatured western blot analysis as compared to the control (Figure 2.7). After native gel electrophoresis, the dimer form of SpmSyn was only detectable in the lane for the G67E alteration. Its level was only detectable upon long exposure.

On denatured western blots, the P112L alteration was detected at about 20% of the control; F58L was detected at about 7% of the control, and G67E was detected at about 5% of the control. M35R and G56S were barely detectable (Table C-3). The implied stability order from this data is: WT > P112L > F58L > G67E > G56S > M35R.

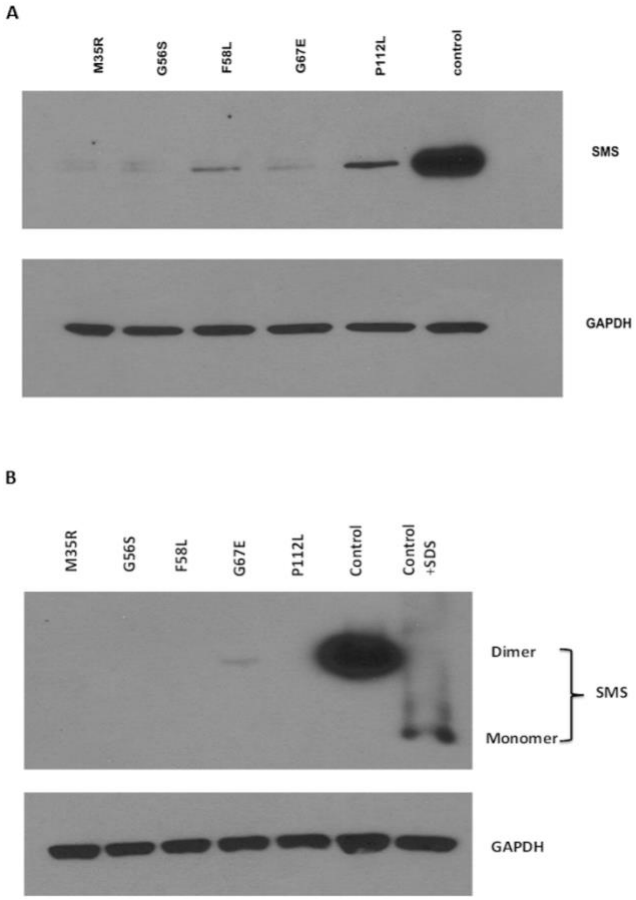


Figure 2.7 Western blot analysis of SMS levels in patient lymphoblast cell lines. (A) Denatured SMS blot. 10 µg of lymphoblast lysate was prepared in Lamelli sample buffer. The buffer was parted on a 4%–20% sodium dodecyl sulfate polyacrylamide gel (SDS–PAGE). Furthermore, the buffer was probed for SMS and glyceraldehyde 3-phosphate dehydrogenase (GAPDH). The control was GAPDH. Densitometry of the blots was analyzed using NIH Image J. SMS expression levels of the mutants were normalized to the control. (B) Native SMS blot. 10 ug of lymphoblast cell lysate was prepared in native sample buffer, separated on a native PAGE gel and probed for SMS and GAPDH. Densitometry of the blots was analyzed by NIH Image J.

2.3. Effect of Missense Mutation on Dimer Affinity

Mutations	BeAtMuSiC (AB)	BeAtMuSiC (CD)	Foldx (AB)	Foldx (CD)	SAAMBE (AB)	SAAMBE (CD)	SD (AB)	MEAN (AB)	SD (CD)	MEAN (CD)
M35R	0.05	0.24	0.17	−0.96	−0.27	−0.19	0.22	0.11	0.60	−0.30
G56S	−1.84	−1.34	−8.64	−11.87	1.58	−4.12	5.20	−5.24	5.45	−5.78
F58L	−2.74	−2.28	−0.1	−1.26	2.20	7.46	2.47	−1.42	5.35	1.31
G67E	−0.78	−0.83	0.372	0.16	−1.86	9.91	1.12	−0.76	5.93	3.08
P112L	−0.11	−0.17	−4.59	−3.32	−0.38	3.39	2.51	−2.35	3.35	−0.03

Table 2.7 Predictions of dimer affinity change due to missense mutations. The calculated binding free energy changes are in kcal/mol. $\Delta\Delta G > 0$ indicates stabilization while $\Delta\Delta G < 0$ indicates destabilization. The calculation is performed for "AB" dimer and "CD" dimer for comparison. The mean value is the average of all calculated results for each mutation. SD is also calculated to quantify variation of values.

Table 2.7 shows the results for dimer affinity change (binding free energy change) due to missense mutations calculated with webservers and stand-alone computer algorithm. The calculations for all investigated mutations are performed using "AB" dimer and "CD" dimer. For most cases, the predictions made by different algorithms are in agreement. Among disease causing mutations, G56S, F58L, G67E and P112L, are predicted to substantially decrease dimer affinity while M35R is calculated to have negligible effect.

2.4. Result of Multiple Sequence Alignment Analysis(MSA)

We investigated the evolutional conservation of the WT residues involved in the mutations based on MSA. The SpmSyn proteins used for MSA are taken from ten different

species, which include seven mammals and three non-mammals. Figure 2.8 shows the result of multiple sequence alignment of SpmSyn among different species. It can be seen that the residues involved in SRS are almost totally conserved across all different species, indicating that these residues are probably important for protein function. The substitution of these highly conserved residues will probably have a large impact on the protein's functionality.

Figure 2.8. Sequence alignment of SpmSyn among different species. The mutation sites considered in this study are represented in bold letters and the position of the residue in human SpmSyn is shown at the top. The mammals considered in this study are represented in bold letters. Multiple sequence alignment (MSA) is performed with Cobalt Constraint-based Multiple Protein Alignment Tool (COBALT).

3. Discussion

Dimerization is essential for the normal function of SpmSyn and the N-terminal domain plays a crucial role in the dimerization. In this work, we investigated the molecular effect of five mutations which causes SRS located within the N-domain of SpmSyn, focusing

mainly on the stability and the dimerization of SpmSyn. To analyze the effects of these mutations on SpmSyn stability and binding affinity, we investigated the structural features of the side chains involved in the mutations and made connections to computational and experimental results. We performed such an analysis for each mutation separately. Appendix Figure B-11 shows the side chain conformations of the wild type and the mutant residue for the five disease-causing mutations studied in this work.

M35R: The M35R is substitution of a hydrophobic residue Met by a positive charged residue, Arg. The M35 site is totally buried in the protein's interior. As it is shown in Appendix Figure B-11a, the structure around M35 site is very well packed and there is no room to accommodate the large Arg side chain. In addition, placing a charged residue, Arg, in the hydrophobic protein interior is energetically very costly and is typically referred as to desolvation penalty. It can be seen (Appendix Figure B-11b) that the mutant R35 does not establish any hydrogen bonds or other type of favorable interaction. In terms of binding affinity, the M35 site is far away from the interface and a direct effect of the mutation on the binding affinity is not expected. This is consistent with the energy calculations and experimental observation, which showed that M35R has large effect on the monomer stability but negligible effect on the dimer affinity.

G56S: The mutation G56S is located ina sharp loop connecting two β strands of the N-domain (Appendix Figure B-11c). The G56 site is almost totally exposed to the water in monomer state. However, it is well known that Gly is typically found in tight turns and any replacement may cause steric clashes. This is the structural argument for predicting that a substitution with a relatively long side chain of Ser is not favorable (Appendix Figure B-

11d). The G56 is at the periphery of the dimer binding interface and direct effect on binding is expected for any residue substitution. This is consistent with the energy calculations reported here and in previous work [24, 101], showing that the G56S mutation decreases both monomer stability and dimer affinity. The effects were also experimentally confirmed [101].

F58L: The mutation F58L is located in a β strand at the dimer binding interface; (Appendix Figure B-11e,f.) The energy calculations predict that the mutation destabilizes the monomer and has a large effect on dimer affinity. The experiments show that the mutation decreases monomer stability as well. As for the dimer affinity, Phe58 is totally buried at the dimer binding interface formed by two β sheets and there is no room for accommodating side chain of different volume—thus the substitution ofLeu is predicted to greatly decrease dimer affinity.

G67E: The mutation G67E is a substitution of a neutral residue Gly by a negative charged residue, Glu, and is located in a sharp loop connecting two β strands of N-domain (Appendix Figure B-11g). Similar to the effect of the above mentioned mutation G56S, the replacement of Gly by Glu, which is a negative charged residue with long side chains and does not form any favorable interactions (Appendix Figure B-11h) , in a tight turn in structures will cause steric clashes and probably destabilize the protein. This predicted effect is consistent with the energy calculation that G67E significantly destabilizes the monomer stability. The experiments also confirm the computational prediction. In terms of dimer affinity, G67E is at the periphery of the dimer binding interface and the prediction is that G67E mutation will also destabilize the dimer.

P112L: The last disease causing mutation, P112L is also located at the binding interface (Appendix Figure B-11i,j). The P112 site is totally exposed to the water in the monomeric state. However the substitution is predicted to decrease stability of the monomer. The same is observed experimentally. Furthermore, the P112L is predicted to have a large effect on the dimer affinity as well. It can be seen that P112 site is located at a turn between two β strands and the specific characteristics of Pro residue cannot be mimicked by any other amino acid. This is consistent with the energy analysis indicating that P112L mutation will significantly destabilize the dimer.

From an evolutionary stand-point, the five disease-causing mutations are in sequence positions that are highly conserved across different species in multiple sequence alignment analyses, indicating that they are important for SpmSyn function.

Overall, the study revealed the molecular mechanism of the five SRS-causing mutations: It was shown that all mutations greatly affect SpmSyn stability and dimerization. Thus, the disease-causing effect alters the structural integrity of SpmSyn and thus, the protein is either unfolded, and therefore subjected to degradation, or present in very small quantities with impaired ability to form a dimer. The functional result of either of these events would be a dysfunctional protein resulting in SRS.

4. Materials and Methods

4.1. Protein Structure

The wild type structure of human SpmSyn (PDB ID: 3C6K)[165] was downloaded from Protein Data Bank (PDB) [91]. The structure contains four chains and the biological unit

is taken as a homo-dimer made of chain "A" and "B" or Chain "C" and "D". Since there is a small structural difference between the "AB" dimer and the "CD" dimer, both structures are used for energy calculation for comparison. The mutant structure was generated *in silico* by side chain replacement with VMD Mutator Plugin, Version 1.3 [81].

4.2. Protein Binding and Folding Free Energy Prediction

Webservers and stand-alone computer programs were applied to assess monomer folding free energy change and dimer affinity change due to mutations. The webservers used for energy calculation included BeAtMuSiC [148], NeEMO [106], PopMusic [107], I-Mutant 2.0 [108], SDM [152], DUET [109] and CUPSAT [110]. Also a computer algorithm, FoldX 3.0 [111, 112] was used to predict the folding free energy changes and dimer affinity change upon single point mutations. Another program developed in our lab, SAAMBE [30], was also used to calculate dimer affinity change.

4.3. Multiple Sequence Alignment

To investigate the evolutional conservation of the above mentioned mutations, multiple sequence alignment (MSA) was performed with Cobalt Constraint-based Multiple Protein Alignment Tool (COBALT) [170] among different species. The sequence of different species was downloaded from UniProtKB/Swiss-Prot Database [171] with FASTA format and include seven mammals: human (Homo sapiens; UniProtKB/Swiss-Prot: P52788), mouse (Mus musculus; UniProtKB/Swiss-Prot: P97355), bovin (Bos taurus; UniProtKB/Swiss-Prot: Q3SZA5), rat (Rattus norvegicus; UniProtKB/Swiss-Prot: Q3MIE9), calja (Callithrix jacchus; UniProtKB/Swiss-Prot: U3FPX7), pantr (Pan

troglodytes; UniProtKB/Swiss-Prot: P97355), and maceu (Macropus eugenii; UniProtKB/Swiss-Prot: B3VFB4) and three non-mammals: danre (Danio rerio; UniProtKB/Swiss-Prot: Q9YGC9), ophha (Ophiophagus hannah; UniProtKB/Swiss-Prot: V8NLB9), and ictpu (Ictalurus punctatus; UniProtKB/Swiss-Prot: W5U5T7).

Modelling of SAVs' molecular effects in DHCR7 protein:

1. Introduction

Smith-Lemli-Opitz syndrome (SLOS) is an inherited disorder of cholesterol synthesis characterized by intellectual disability and multiple malformations, including facial and genital abnormalities and syndactyly and was first described by Smith and coworkers [172]. The reported incidence of SLOS varies widely depending on the heterogeneity of the population studied, the biochemical methods used and the alleles assessed. Current estimates of SLOS carrier frequency in Caucasian populations lie between 1% and 3% [173-175]. SLOS is more prevalent in individuals of northern and eastern European descent and is rarely described in individuals of Asian or African descent [176]. Reports that up to 80% of affected fetuses, likely those heterozygous for null mutations, die before birth and that milder cases of the disease may not be diagnosed, conceivably prevent accurate determination of frequency [177-179]. The majority of "classical" SLOS patients are compound heterozygotes with one severe null mutation and a second missense mutation which retains some enzyme functionality. Milder cases often possess two less severe missense mutations [180].

SLOS is linked to mutations in 7-dehydroxycholesterol reductase (DHCR7), which is the rate-limiting enzyme in the cholesterol synthesis pathway [181]. DHCR7 reduces the C7–C8 double bond of 7-dehydrocholesterol (7DHC), the precursor molecule to cholesterol [182]. Cholesterol, though harmful in high levels, is essential to life since it is involved in membrane structure and permeability, synthesis of steroid hormones and proper fetal development. The loss of functionality of the DHCR7 enzyme in individuals with SLOS results in a significant decrease in cholesterol levels and possibly toxic buildup of 7DHC and other cholesterol precursors [183]. It was shown that accumulation of 7DHC in the brains of rats is associated with intellectual and learning disabilities [13,14].

In addition to its role in cholesterol synthesis, 7DHC is also required for vitamin D3 production. Exposure to sunlight cleaves the C9–C10 bond of 7DHC in the skin, resulting in vitamin D3. Vitamin D3 is essential for calcium absorption and bone health [184]. As DHCR7 activity decreases the amount of 7DHC available for vitamin D3 synthesis, there is a potential heterozygote advantage to carriers of *DHCR7* mutations, which typically decrease enzymatic activity [185, 186]. This may explain the prevalence of mutations originating in areas with decreased sun exposure such as northern Europe and northeast Asia [178, 187].

The *DHCR7* gene maps to chromosome 11q13.2–13.5 [17–19] and consists of nine exons with the initiation codon located in exon three. The gene is expressed in all tissues with peak expression in adrenal glands, liver and brain [188]. *DHCR7* encodes a 475 amino acid polypeptide with a molecular weight of 54.5 kDa, which is a transmembrane protein located in the endoplasmic reticulum (ER) membrane, the location of cholesterol synthesis.

The first *DHCR7* mutations were identified in 1998 by several groups and the early years of the 21st century resulted in more advanced molecular tests to rapidly identify *DHCR7* mutations [188, 189]. Most mutations are identified through sequence analysis of coding exons and flanking intronic sequences [5,17]. To date, more than 160 *DHCR7* mutations have been reported [176]. The most common mutation with a prevalence of ~30% of reported SLOS patients is the IVS8AS G > C − 1 splice acceptor site mutation. This results in the inclusion of 134 base pairs of intronic sequence into the transcript and a non-functional protein. Other common mutations include T93M, W151X, V326L and R404C.

The majority of pathogenic *DHCR7* mutations occur in the highly conserved C-terminus region of the protein. In their molecular model of the DHCR7 protein, Li and coworkers predicted two overlapping binding sites: one for docking of the sterol 7DHC and one for binding of the coactivator NADPH [190]. As both binding sites are critical for proper protein function, it can be speculated that mutations affecting these areas would be most likely to result in disease. In support of this hypothesis, Waterham and Hennekam conducted a systematic review of published SLOS patients and compared genotype with phenotype [176]. They concluded that the most severely affected patients presented with two null alleles or two mutations in the 8–9 cytoplasmic loop while a milder phenotype was associated with mutations in the 1–2 loop or one mutation in the N- or C-terminus [176].

In the present study, we obtained variations in the *DHCR7* gene from online databases and modelled their effects on the corresponding protein to make predictions about SLOS

phenotype. We demonstrate that structural and conservation properties are good discriminators between pathogenic and non-pathogenic mutations, while folding free energy changes (ΔΔGs) are not. This is consistent with previous observations [191] that current methodology for computing ΔΔGs are not accurate enough when applied to membrane proteins. Furthermore, based on detailed analysis of selected mutants, we predict that the currently non-classified mutation, R228Q, is pathogenic.

2. Results and Discussion

2.1. Mapping Missense Mutations onto the 3D Structure of DHCR7 Protein

The dataset of *DHCR7* missense mutations includes three types of mutations: pathogenic, non-pathogenic and mutations of unknown effect. The mutations were visualized by mapping them onto the DHCR7 structure (Figure 2.9A). Pathogenic mutations are predominantly located in transmembrane and ligand-binding regions while non-pathogenic mutations are primarily situated outside the membrane. This observation indicates that pathogenic mutations occur at protein sites that are either buried or directly involved in protein function, which corroborates the findings of previous investigations [41, 48, 78, 192]. To investigate the linkage between structural and evolutionary features of DHCR7 protein, we obtained the evolutionary conservation score (EC score) for each residue from multiple sequence alignment and mapped them onto the 3D structure of DHCR7 (Figure 2.9B). The transmembrane and ligand-binding regions appear to be highly conserved. Thus, most pathogenic mutations are located in highly conserved positions, while non-pathogenic mutations are less conserved. To further quantitatively assess the

mutations' effects, we computed the relative solvent accessible surface area (rSASA), evolutionary conservation score (EC score) and folding free energy change (ΔΔG) for all mutations studied in this work (Table C-4). Pathogenic mutations tend to have lower rSASA values and higher EC scores compared with non-pathogenic mutations. However, ΔΔG results show no obvious tendency to discriminate pathogenic from non-pathogenic mutations. The predictions made with different servers frequently contradict each other resulting in large standard deviation (SD) when averaging these predictions (Table C-4). As DHCR7 is a transmembrane protein and recent work [191] demonstrated that current tools of ΔΔG predictions are not accurate when applied to membrane proteins, this may explain why ΔΔG fails to discriminate pathogenic from non-pathogenic mutations in this case. In addition, we also performed Polyphen predictions on all types of mutations (Table C-4). Almost all the pathogenic mutations are predicted to be probably damaging by Polyphen. However, Polyphen overestimated the deleteriousness of the non-pathogenic mutations. About half of the non-pathogenic mutations were classified as possibly or probably damaging. Thus, Polyphen has limited accuracy in discriminating the pathogenic mutations from the mutations with unknown effects for this particular protein.

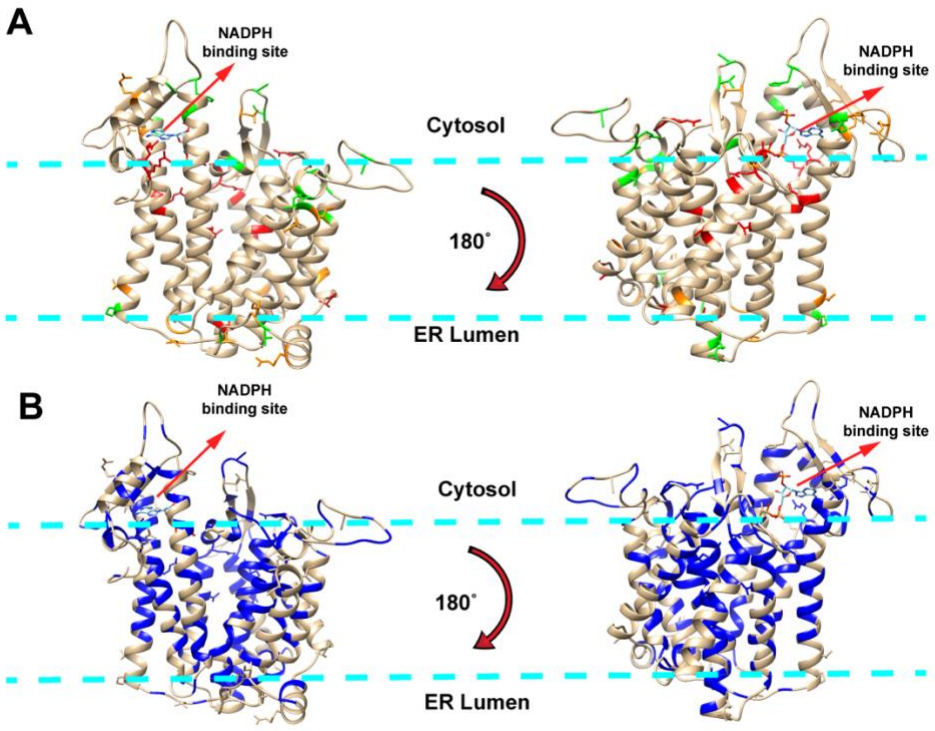


Figure 2.9. (A) Visualization of mutations mapped onto DHCR7 protein. Red, orange and green colored sites represent pathogenic, unknown effects and non-pathogenic mutations, respectively. The membrane boundaries are schematically shown with light blue dashed lines; (B) Most highly evolutionarily conserved residues mapped onto DHCR7 protein. Residues with EC score > 0.9 are marked with blue and all mutation-affected residues are shown with side chain. The membrane boundaries are schematically shown with light blue dashed lines.

2.2. Classification of the Mutations with Unknown Effects Using KNN Model

One of the goals of this study was to identify biophysical features allowing us to distinguish between pathogenic and non-pathogenic mutations, and thus to make predictions about unclassified mutations. Above, we outlined several biophysical features,

namely rSASA, EC score, PD and ΔΔG, which will be used in conjunction with the K-nearest neighbors (KNN) method (see Method section). The dataset includes 16 pathogenic mutations and 23 non-pathogenic mutations. These 39 mutations were randomly partitioned into training dataset (29 mutations) and test dataset (10 mutations) and then subjected to the KNN classifications. As the ΔΔG was shown to be less successful in distinguishing between pathogenic and non-pathogenic mutations, we performed the KNN classification with and without the ΔΔG (Table C-5). The classification shows better performance without using the ΔΔG and the accuracy is 100% when K value is within 5 to 9. Here, we select $K = 7$ (the median of the K value corresponding to highest accuracy). Finally, KNN model with $K = 7$ and using properties: rSASA, EC score and PD applied to classify the mutations with unknown effects (Table 2.8). Thus, we predict that among all currently known unclassified mutations, only R228Q is pathogenic. In Table 2.8 we also compared our KNN classification results with the predictions from Polyphen. Consistent with our results, Polyphen predicted R228Q to be probably damaging. However, Polyphen gives contradictory predictions for eight additional mutations (predicted to be probably damaging), which are classified as non-pathogenic by our KNN classification. Overestimation of mutation deleteriousness was also observed when applying Polyphen to the known non-pathogenic mutations (Table C-4).

Mutation	KNN Classification	Polyphen	Mutation	KNN Classification	Polyphen
A41V	N	Benign	R228Q	P	Probably damaging
I44T	N	Benign	V330M	N	Probably damaging
A67T	N	Possibly damaging	V338M	N	Benign

I75F	N	Benign	F361L	N	Probably damaging
R81W	N	Probably damaging	T364M	N	Probably damaging
A97T	N	Possibly damaging	R367C	N	Probably damaging
V126I	N	Probably damaging	G424S	N	Probably damaging
V134L	N	Benign	G425S	N	Benign
A162V	N	Possibly damaging	R461C	N	Probably damaging

Table 2.8. KNN classifications and Polyphen predictions of the mutations with unknown effects. P and N represent pathogenic and non-pathogenic mutations, respectively.

2.3 Case Study of Selected Mutations Using Molecular Dynamics (MD) Simulations

The above classification and analyses were performed using fast computational approaches and were applied to the entire dataset. We selected a subset of mutations for extensive MD simulations to investigate the possibility that pathogenic and non-pathogenic mutations have different effects on DHCR7 protein conformational dynamics. For this purpose, we selected 10 representative mutations including five pathogenic mutations (T154R, E288K, T289I, G303R and R404C), two non-pathogenic mutations (R260Q and A452T) and three mutations with unknown effects (V134L, R228Q and F361L). These mutations are localized to different regions of protein structure. Five mutations (T154R, R228Q, E288K, T289I and G303R) are located in the transmembrane region and are buried in the membrane, two mutations (F361 and R404C) occur near the ligand-binding site and potentially affect ligand binding, and the remaining three mutations (V134, R260Q and A452T) are in neither the transmembrane region nor the ligand binding site.

Since our focus was on protein conformational dynamics, we calculated the corresponding RMSDs and RMSFs for the wild type and mutant proteins. The average

RMSD data shows no obvious difference between wild type protein and proteins with non-pathogenic or pathogenic mutations. However, the average RMSF indicates some differences between the wild type and mutants. For example, in the mutant A452T, cytosol loops (CL) 2 and 4 and transmembrane domain (TM) 10 regions are more rigid compared to the wild type (Table 2.9). However, no apparent patterns were identified to differentiate pathogenic mutations and non-pathogenic mutations by simply observing the graphs. A previous study of the AGAL protein has indicated a correlation between the protein's flexibility and the severity of a mutant's pathogenicity [193]. Thus, to identify such potential correlation in DHCR7 protein, we mapped the pathogenic and non-pathogenic mutations on the average RMSF of the wild type proteins (shown in Figure B-12). We observed that most pathogenic mutations are located on the low RMSF region while the non-pathogenic mutations show the opposite trend. As the low RMSF residues are mostly transmembrane, such observed correlation is expected when majority of the pathogenic mutations are located on the transmembrane region. In addition, further analysis was performed by grouping the residues into different regions and then summing up the RMSF of residues in that region to get a region-RMSF. Based on DHCR7 protein structure information [194], residues were grouped into regions: TM1 (residues 40–60), TM2 (residues 94–115), TM3 (residues 145–164), TM4 (residues 176–191), TM5 (residues 235–256), TM6 (residues 268–288), TM7 (residues 302–326), TM8 (residues 332–352), TM9-10 (residues 408–442), CL1 (residues 116–144), CL2 (residues 198–234), CL3 (residues 289–301), CL4 (residues 354–407) and CTD (residues 443–475). The topology of the cytosol loops (CL), the C terminal domain (CTD) and transmembrane domains (TM)

mapped with selected mutations are further represented for better visualization of the DHCR7 structure (Figure 2.10).

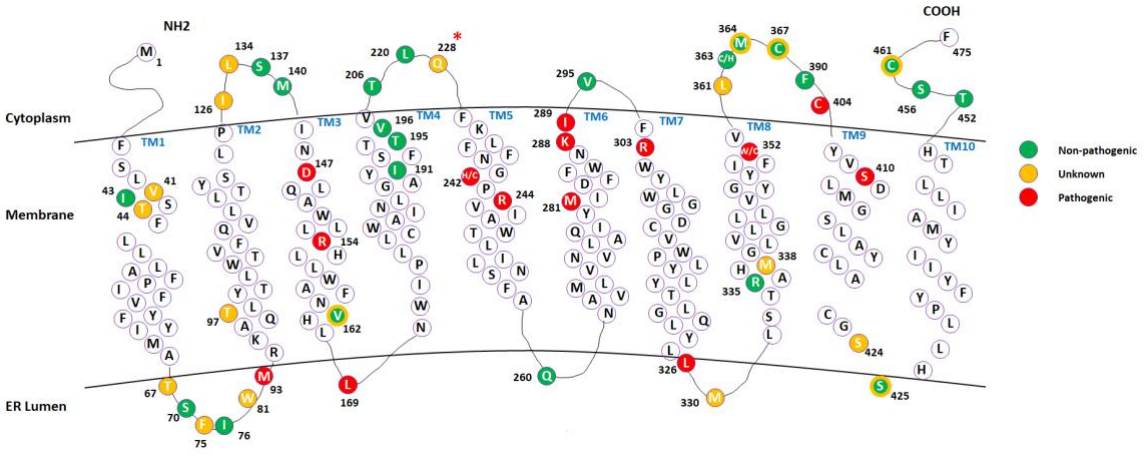


Figure 2.10. The topology of the cytosol loops (CL), the C terminal domain (CTD) and transmembrane domains (TM) in DHCR7 structure. Mutation sites are mapped with different colors according to mutation type (double color is applied for sites with unknown and non-pathologic classification). The unclassified mutation R228Q, which we predict to be pathogenic, is highlighted with a red asterisk.

Table 2.9 shows the region-RMSFs. Pathogenic mutations tend to decrease the flexibility in the TM1, TM2 and CL2 regions and increase the flexibility in the TM7 and TM9-10 regions. Very little is known about DHCR7 function and structural changes occurring during chemical reactions, so we used the above observation to suggest an empirical formula that discriminates between pathogenic and non-pathogenic mutations, which were subjected to MD simulations (ideally, one should perform such an analysis for mutations analyzed in this manuscript, but this is too computationally demanding). For the wild type and each mutant, we sum the RMSFs of TM1, TM2 and CL2 and then subtract

the RMSFs of TM7 and TM9_10 (last column in Table 2.9). We refer to this quantity as cumulative RMSF. The wild type and non-pathogenic mutants have cumulative RMSFs larger than 50 Å while all pathogenic mutants have a cumulative RMSF less than or equal to 46 Å. Among non-classified mutations, V134L is confirmed to be non-pathogenic, while R228Q and F361L show the same cumulative RMSFs as pathogenic mutations. Thus, it is encouraging to observe that R228Q is independently confirmed to be pathogenic mutation (see KNN classification above), while F361L cannot be classified with high confidence and additional investigations are reported in the next section.

Pathogenic Missense Mutations															
	TM1	TM2	TM3	TM4	TM5	TM6	TM7	TM8	TM9_10	CL1	CL2	CL3	CL4	CTD	TM1+TM2-TM7-TM9_10+CL2
T154R	22.9	17.6	15.5	9.1	14.5	17.4	17.2	17.8	31.6	53.1	48.6	8.8	78.8	30.8	40.3
E288K	19.6	16.4	16.0	10.5	13.1	14.3	16.0	18.4	26.1	49.2	38.6	13.6	77.5	32.0	32.7
T289I	25.2	18.0	17.8	12.8	14.1	14.3	19.5	17.5	28.7	50.1	47.9	10.1	71.1	30.4	42.9
G303R	21.1	18.9	16.8	11.0	13.3	16.0	18.2	17.0	30.2	50.5	49.2	10.4	65.0	30.1	40.9
R404C	23.4	16.5	16.0	10.8	16.0	16.8	20.8	23.3	31.6	48.9	57.4	10.0	80.1	32.7	44.9
Missense Mutations with Unknown Effects															
	TM1	TM2	TM3	TM4	TM5	TM6	TM7	TM8	TM9_10	CL1	CL2	CL3	CL4	CTD	TM1+TM2-TM7-

															TM9_10+CL2
V134L	20.1	21.1	18.8	11.0	14.7	13.6	16.6	16.2	27.7	52.9	53.4	11.0	79.7	35.6	50.3
R228Q	17.6	17.0	15.9	8.5	13.6	13.0	15.6	16.9	27.6	53.2	54.2	10.7	75.8	36.7	45.6
F361L	19.4	17.4	14.8	9.9	14.2	14.0	18.3	16.6	28.8	54.7	50.8	11.7	74.8	33.5	40.5
Non-Pathogenic Missense Mutations															
	TM1	TM2	TM3	TM4	TM5	TM6	TM7	TM8	TM9_10	CL1	CL2	CL3	CL4	CTD	TM1+TM2-TM7-TM9_10+CL2
R260Q	19.7	18.6	15.5	9.4	12.9	14.6	15.4	17.2	24.4	58.4	52.1	11.4	79.3	28.1	50.6
A452T	20.9	19.6	17.8	10.5	13.6	16.2	16.4	17.2	26.0	55.2	52.8	8.7	66.6	30.1	51.0
Wild Type															
	TM1	TM2	TM3	TM4	TM5	TM6	TM7	TM8	TM9_10	CL1	CL2	CL3	CL4	CTD	TM1+TM2-TM7-TM9_10+CL2
WT	18.2	18.3	17.9	10.7	16.3	16.0	18.5	18.1	31.1	51.9	65.1	13.0	80.4	37.8	52.0

Table 2.9. RMSF values per structural region (see text for details) for each of the mutants. The RMSFs are given in Å units. The last column reports the RMSF calculated as the sum of RMSFs of TM1, TM2 and CL2 subtracted by RMSF of TM7 and TM9-10. Values larger than 50 Å are underlined.

2.4. Analysis of Mutations' Pathogenic Effects:

2.4.1. Ligand Binding

Here, we investigated the possibility that mutations may change DHCR7 functionality by altering the binding affinity towards its ligand NADPH. For this purpose, we compared the effects of F361L (non-classified) and R404C (pathogenic mutation), both located near the NADPH binding site. It is anticipated that NADPH binding will cause structural rearrangement of the binding site and the conformational flexibility of the binding pocket is essential for proper protein function. We tested the effects of F361L and R404C on binding pocket flexibility by comparing them with the wild type protein. This was done using the MD trajectories obtained above and computing the residue cross-correlation for each trajectory with Bio3D [195]. These types of analyses were successfully used to elucidate the effects of a single mutation on the human β2-microglobulin's protein dynamics [196]. For each mutation and wild type, we calculated the average cross-correlation from three independent MD runs. Finally, the residue cross-correlation changes for mutations F361L and R404C are shown in Figure 2.11A,B, which is the subtraction of the averaged cross-correlation map between mutant and wild type proteins. Significant changes of the cross-correlation coefficient near the NADPH binding site were found for R404C, highlighted with a circle in Figure 2.11, but not for F361L.

We also performed MM/PBSA analysis to investigate the effect of mutations on NADPH binding affinity (Figure 2.11D). Mutation R404C results in a large increase of the binding affinity by about 15 kcal/mol. As shown in the literature [78, 197, 198], any large deviation from wild type characteristics may be deleterious. In this case, R404C mutations

contribute to disease by altering the binding affinity of NADPH. Compared to the effect of F361L, we observe that binding affinity is much less affected. This, combined with correlation analysis, allows us to speculate that F361L is a non-pathogenic mutation.

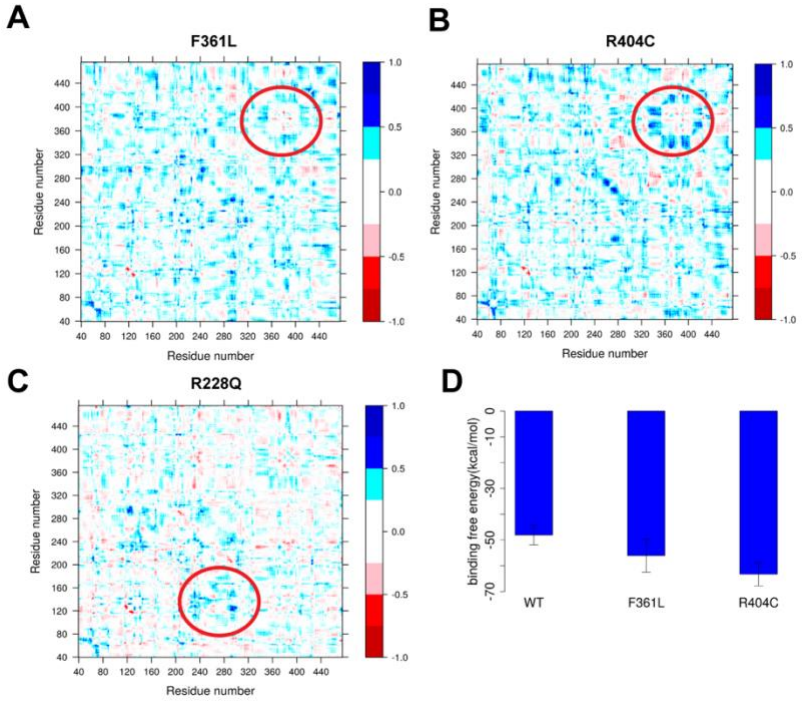


Figure 2.11. (A–C) The changes in residue cross-correlation for mutations F361L, R404C and R228Q; (D) NADHP binding free energy for WT and mutations F361L and R404C.

2.4.2. Protein Dynamics

We further analyzed the selected mutations including our predicted pathogenic mutation R228Q to identify other pathogenic effects on protein functionality. The residue cross-correlation analysis of R228Q (Figure 2.11C) indicates a local conformational change near the mutation site. The R228Q mutation makes the corresponding region more rigid, resulting in local flexibility changes in CL2. Changes in protein dynamics are also

observed in the residue cross-correlation analysis of other pathogenic mutations such as E288K and G303R (shown in Appendix Figure B-13), indicating that alterations in DHCR7 protein dynamics likely contribute to protein dysfunction.

2.5. Allele Frequency Analysis

We compared the frequency distribution of pathogenic mutations and frequently-occurring common mutations among different populations and genders. Figure 2.12A displays the top 40 *DHCR7* mutations of varying types occurring in more than 50 individuals archived in the ExAC database. At the same time, Figure 2.12B shows the distribution of pathogenic missense mutations chosen for this study within the same set of populations. The most frequently-occurring mutations in the general population are found in individuals of non-Finnish European descent followed by South Asian and African and African American descent (Figure 2.12A). Additionally, individuals of non-Finnish European and South Asian descent have the highest frequency of pathogenic mutations as shown in Figure 2.12B. African and African American populations have few cases of SLOS despite high occurrences of *DHCR7* mutations. The low occurrence and frequency of mutations in Europeans of Finnish descent is supported by the extremely low number of SLOS cases in Finland [199].

Interestingly, females in the overall ExAC population possess more *DHCR7* mutations at higher frequencies than males (Figure 2.12C), while this is an opposite for the pathogenic mutations investigated in this manuscript (Figure 2.12D), though no support for this trend has been found in the literature. One can speculate that this is linked to sex hormones and

is embryo lethal, but the observation that females carry more pathogenic mutations than males should be taken with precaution.

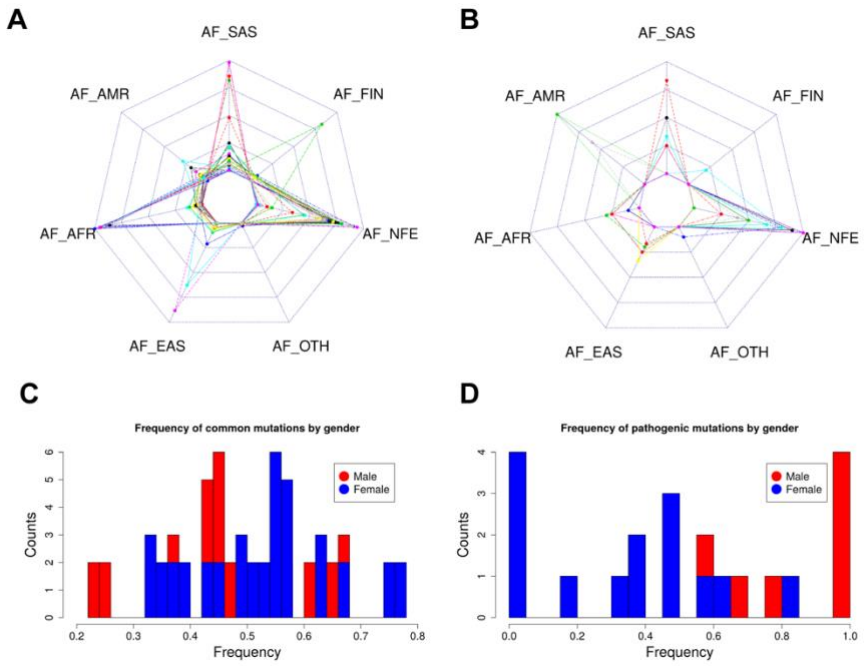


Figure 2.12. The frequency distribution of *DHCR7* mutations. AFR, AMR, EAS, FIN, NFE, SAS and OTH represent African and African American, American, East Asian, Finnish, Non-Finnish European, South Asian and other populations, respectively. (A) The frequency distribution among different populations of the top 40 DHCR7 mutations of varying types occurring in more than 50 individuals archived in the ExAC database; (B) The frequency distribution among different populations of pathogenic missense mutations chosen for this study; (C) The frequency distribution in males and females of the top 40 DHCR7 mutations of varying types occurring in more than 50 individuals archived in the ExAC database; (D) The frequency distribution in males and females of pathogenic missense mutations chosen for this study.

3. Materials and Methods

3.1. Selection of DHCR7 Missense Variants

The missense mutations investigated in this work were selected using ClinVar [200] and ExAC [201] databases. The ClinVar database (https://www.ncbi.nlm.nih.gov/clinvar/) was queried using the search term "DHCR7". The results were further refined by missense mutations consisting of benign (2), likely benign (3), uncertain significance (30), likely pathogenic (15), pathogenic (26) and conflicting reports of pathogenicity (3) (as of 13 November 2017). The ExAC (Exome Aggregation Consortium) Browser (http://exac.broadinstitute.org/) was queried using the search term "DHCR7" and the entries were sorted by allele frequencies in descending order. The missense variants with an allele frequency greater than 0.00001, which were also classified in ClinVar were chosen for further *in silico* analysis. Of the chosen mutations, the variants defined as pathogenic or likely pathogenic in Clinvar database are classified as pathogenic mutations in this study while the others defined as uncertain significance in Clinvar database are classified as mutations with unknown effects. E288K and G303R are previously reported SLOS-causing mutations [202, 203] although they are not classified as pathogenic in the Clinvar database. Thus, E288K and G303R were treated as pathogenic mutations in this study. Overall, 16 pathogenic mutations and 18 mutations with unknown effects are classified for this study.

3.2. Selection of Non-Pathogenic DHCR7 Mutations

We first obtained the missense mutations in *DHCR7* gene from the ExAC database [201], including the whole genome sequencing data from 60,706 unrelated individuals. In total, 280 missense mutations in *DHCR7* were identified. The ExAC database also provides the corresponding allele frequency data from the 1000 Genomes Project and the NHLBI-GO Exome Sequencing Project (ESP) for each mutation. Individuals participating in the 1000 Genomes Project were all healthy while the objective of the ESP is discovery of novel genes and mechanisms contributing to heart, lung and blood disorders. As our goal was to select non-pathogenic mutations from the ExAC database, we applied the following selection criteria: (a) mutations with allele frequency >0 in the 1000 Genomes Project; (b) mutations with allele frequency of 0 in the ESP. Thus, we classified the mutations identified from the healthy population of 1000 Genomes Project but not from the ESP as non-pathogenic mutations in this study. In total, 23 non-pathogenic missense mutations were identified.

3.3. Obtaining Allele Frequency and Gender Occurrence

The allele frequency and gender data of *DHCR7* mutations were obtained from EXAC database [201]. The most recent database version was downloaded from the FTP site (http://exac.broadinstitute.org/downloads) and mutations affecting the DHCR7 protein as well as their corresponding allele frequencies and gender data were obtained. The frequency of mutation by gender is calculated by the number of carrier females or males divided by the total number of carrier individuals.

3.4. Generation of the 3D Model for DHCR7

The 3D structure of the DHCR7 protein was generated by homology modeling due to lack of an existing experimental structure. Structure of the integral membrane sterol reductase from *Methylomicrobium alcaliphilum* (PDB: 4QUV) [190] was used as a template and subjected to MODELLER [204] for homology modeling. The sequence identity between the template and DHCR7 is 37% (sequence alignment is shown in Appendix Figure B-14) and thus high structural similarity was observed between the generated model and template. The model with lowest DOPE score was selected for this study and further subjected to automatic loop refinement with MODELLER [204].

3.5. Property Distance (PD)

To quantify the physical-chemical property differences between the wild type and mutant residues, we used the property distance (PD) as a parameter to quantitatively describe such changes. In this study, we describe physical-chemical properties of a particular residue using a property vector which includes two elements: hydrophobicity and charge. The hydrophobicity of the residues are taken from an experimentally determined hydrophobicity scale [205, 206]. R and K carry +1 charge while E and D have −1 charge. All other residues are considered neutral. PD represents the Euclidean distance of the property vector between the wild type and mutant residues (shown in Equation (2.3)). The *PD* between all types of residues are shown as a matrix in Appendix Figure B-15.

$$PD(x,y) = \sqrt{((H(x) - H(y))^2 + (Q(x) - Q(y))^2} \quad (2.3)$$

,where *x* and *y* represent two types of residues; *H* and *Q* are corresponding hydrophobicity and charge for a particular residue.

3.6. Evolutionary Conservation Score (EC Score) Calculation

The DHCR7 sequence from 35 different species were collected from UnitProt [171] and subjected to multiple sequence alignment with the T-Coffee webserver [207]. The EC score of each residue in the human DHCR7 sequence was calculated using the multiple sequence alignment with the following equation:

$$\mathbf{EC\ score(i)} = \frac{\mathbf{N(i)_{identity}}}{\mathbf{N(i)_{total}}} \quad (2.4)$$

,where $N(i)_{\text{identity}}$ is the number of the species sharing identical residues in position *i* of the human DHCR7 sequence and $N(i)_{\text{total}}$ is the total number of the species in the multiple sequence alignment.

3.7. Folding Free Energy Change (ΔΔG) and Relative Solvent Accessible Surface Area (rSASA) Calculation

Several webservers were used to predict the effect of mutations on protein stability (folding free energy change (ΔΔG)) using the generated homology model of DHCR7 protein. The webservers used in this study include DUET [109], Eris [208], mCSM [151], SDM [209], Foldx [210] and SAAFEC [34]. The SASA were calculated using VMD [138]. As DHCR7 is a transmembrane protein, the membrane was also included when calculating the SASA. Thus, only the amino acids exposed to water were treated as exposed and the transmembrane regions were treated as buried in the calculation. The rSASA for residues were calculated using the following equation:

$$rSASA(i) = \frac{SASA(i)}{SASA(i)_{max}} \quad (2.5)$$

,where *SASA*(*i*) is the SASA measured for particular residue *i* and $SASA(i)_{max}$ is the maximum SASA obtained for a free residue (entire residue taken off the protein).

3.8. Molecular Dynamic Simulations

The membrane-protein-ligand system was built primarily using the CHARMM-GUI [211] tools. The DHCR7 protein with ligand structure was obtained from previous homology modeling. Ten mutant (V134L, T154R, R228Q, R260Q, E288K, T289I, F361L, G303R, R404C and A452T) structures were derived from the wild type DHCR7 protein structure using VMD 1.9.3 [138] mutator package. The protein was embedded in a POPC bilayer using the CHARMM-GUI website. The protein was oriented to align with 4QUV structure in the OPM [212] database. When the oriented protein was placed into the membrane, the z axis of the protein matched the z axis of the membrane. The whole system was solvated with 0.15 M KCl. The final system was 89.13 × 89.13 × 96.64 $Å^3$ with a total of about 70,800 atoms.

Molecular dynamic simulation (MDS) was performed using NAMD2.11 [95]. The system first underwent energy minimization for 10 ps, then equilibrated through 6 cycles where harmonic constraints were applied to keep original positions of: (a) lipid head groups (force constants were gradually reduced from 5 $kcal{\cdot}mol^{-1}{\cdot}Å^{-2}$ to 0 $kcal{\cdot}mol^{-1}{\cdot}Å^{-2}$), (b) protein backbone (force constants were gradually reduced from 10 $kcal{\cdot}mol^{-1}{\cdot}Å^{-2}$ to 0 $kcal{\cdot}mol^{-1}{\cdot}Å^{-2}$), and (c) protein sidechains (force constants were gradually reduced from 5 $kcal{\cdot}mol^{-1}{\cdot}Å^{-2}$ to 0 $kcal{\cdot}mol^{-1}{\cdot}Å^{-2}$). In addition, dihedral restraints were applied to keep

cis double bonds and c2 chirality (force constants were gradually reduced from 500 kcal·mol^{-1}·Å^{-2} to 0 kcal·mol^{-1}·Å^{-2}). A 1 fs timestep was used in the first few cycles and then switched to 2 fs for wild type whereas much smaller timesteps such as 0.01 fs were used for mutants to prevent restraints from failing. In the first two cycles, NVT simulation was performed and then switched to NPT simulation in the later cycles. Temperature was held at 303.15 K using a Langevin thermostat with a damping coefficient of 10 ps^{-1} and velocity rescaling thermostat. The pressure was maintained at 1 atm using a Langevin piston barostat with an oscillation period of 50 fs and a damping time constant of 25 fs. Electrostatic interactions between charged atoms were calculated using the particle mesh Ewald method. Van der Waals interactions were truncated at 12 Å with a switching function applied from 10 Å. RATTLE is used to constrain the length of all bonds involving a hydrogen atom. This stage of equilibration lasts for tens of ps to hundreds of ps. Then three 10 ns equilibration and 10 ns production runs with no constraints were performed for the wild type and each mutant. A 2 fs timestep was used. No velocity rescaling thermostat was used. Other conditions are the same as the previous stage. RMSD and root mean square fluctuation (RMSF) with the structure at the beginning of the 10 ns run as the reference structure were calculated using VMD 1.9.3.

3.9. MM/PBSA Analysis

To estimate the binding affinity of the DHCR7 protein with the ligand NADPH, we calculated the binding free energy using the MM/PBSA approach. For this purpose, we performed three independent 20 ns MD simulations as described above. We took the frames with an interval of 20 ps from the last 10 ns and a total of 500 frames were selected

from each trajectory. All ions, water and lipids were removed before MM/PBSA energy calculations. All the energy terms were averaged over 500 frames for each trajectory and the mean and standard deviation of binding free energy were calculated for wild type and mutant structures. The internal energy and van der Waals interactions were calculated using NAMD2.11b [95] by subjecting the structure to a one step equilibration at 300 K using dielectric constant = 2 for protein and = 80 for solvent. The electrostatic components of the binding free energy (Coulombic and solvation energy) were calculated by solving the Poisson Boltzmann (PB) equation using the Delphi program [84] with dielectric constant = 2 for protein and = 80 for solvent. The solvent accessible surface area (SASA) was calculated by VMD [138] with the solvent and lipid. The non-polar component of the solvation was further calculated with the following widely-used equation:

$$G_{sasa} = \alpha \cdot SASA + \beta \quad (2.6)$$

, where α = 0.0054 and β = 0.92 kcal/mol.

3.10. K-Nearest Neighbors (KNN) classIfication

K-Nearest Neighbors algorithm was used to classify the missense mutations with unknown effects in DHCR7 protein. The dataset includes 16 pathogenic missense mutations and 23 non-pathogenic missense mutations (non-classified/unknown effect mutations were excluded). The dataset was randomly partitioned into a training dataset (29 mutations) and a testing dataset (10 mutations). The KNN classification was performed using R program and various numbers of K values were tested to obtain the best performance.

4. Conclusions

We investigated the effects of mutations causing SLOS on the biophysical characteristics of DHCR7 protein with the goal of identifying methods allowing the discrimination of pathogenic mutations from non-pathogenic mutations. We found that pathogenic mutations are located either within the transmembrane region or are near the ligand-binding site and are highly conserved between species. In contrast, non-pathogenic mutations observed in the general population are located outside the transmembrane region and have different effects on the conformational dynamics of DHCR7. Our analyses confirmed the inability of folding free energy modeling to deliver reliable results and to be used to discriminate pathogenic from non-pathogenic mutations in membrane proteins. Future investigations may include modeling the effects of *DHCR7* mutations on melting temperature (Tm) via MD simulations conducted at different temperatures using the methodology adopted from recent work on NBD1 domain [213]. As mentioned in the work of Estacio et al. [213], the decrease of Tm may cause the protein to adopt partially misfolded states that become targeted for degradation.

In this work, using three characteristics: solvent exposure of the mutation site, residue conservation and physico-chemical descriptors, we were able to distinguish between pathogenic and non-pathogenic mutations. This observation, along with extensive MD simulations and MM/PBSA modeling, was used to classify R228Q as a pathogenic mutation.

Taken together, these observations suggest that the non-classified mutation R228Q is in fact pathogenic. The analyses performed indicate that pathogenic effects may be of

different origin, from affecting protein stability and dynamics to altering binding affinity and flexibility of the binding site.

CHAPTER THREE

PKA SHIFT AND PROTON TRANSFER IN PROTEIN-NUCLEIC ACID INTERACTION AND DEVELOPMENT OF COMPUTATIONAL APPROACH TO PREDICT SAV'S EFFECT IN PROTEIN-DNA BINDING

Computational investigation of proton transfer, pKa shift and pH-optimum in protein-nucleic acid interaction:

1. Introduction:

Protein-nucleic acid interactions are common in various biological reactions and play a crucial role in cell life [214-216]. These interactions are mediated by various forces and effects, such as electrostatic interactions, hydrogen bonding, hydrophobic effect, and base stacking [215, 217, 218]. Particularly, electrostatic interaction plays a crucial role in protein-nucleic acid binding, since nucleic acids are predominantly negatively charged, while the binding protein interface is typically positively charged – this results in charge complementarity[219-223]. It has been demonstrated, in the case of protein-DNA interactions, that the protein recognizes a specific DNA sequence via formation of hydrogen bonds with specific bases (primarily in the major groove) and that the subsequent binding results in sequence-dependent deformations of the DNA helix [224, 225]. Furthermore, it was shown that the narrow minor groove of DNA strongly enhances the negative electrostatic potential of the DNA phosphate groups and thus facilitates the binding of positively charged arginine residues [224, 226]. Similarly, stacking, electrostatics, and hydrogen bonding play important roles in ssRNA recognition, providing affinity and sequence-specificity during the binding process [227].

In the past, existing structures of protein-nucleic acid complexes were utilized to predict protein binding hot spots and elucidate the mechanism of the binding [215, 217, 223, 227, 228]. However, no attempts were made to evaluate the pKa change induced by the binding, even though it is well recognized that any binding results in a change of electrostatic environment [229, 230]. Thus, the pKa values of the titratable groups may shift upon the complex formation and these pKa shifts can be used as an indicator of the electrostatic energy contribution to the binding [230-232].

Most properties of biological macromolecules are pH dependent, and are tuned towards a particular cellular or sub-cellular pH [233]. Stability and binding are among the basic biophysical characteristics of these macromolecules. It was previously indicated that the stability of monomeric proteins is adapted to cellular and sub-cellular characteristic pH [217, 234, 235]. Similarly, our past investigations have demonstrated that the pH-optimum of binding and the pH-optimum of folding are correlated [231, 236, 237]. At the same time, the pH dependence of protein-nucleic acid binding has not attracted much attention.

Our work took advantage of a recent development: a Poisson-Boltzmann based pKa calculation approach, the DelPhiPKa [154, 155]. The DelPhiPKa is capable of performing rapid pKa calculations of protein ionizable residues, and of nucleotides of RNA and DNA. Complex structures from a large protein-nucleic acid interaction database (NPIDB database) [238, 239] were used for the modeling. Our work aims at revealing plausible proton transfers and pKa shifts induced by protein-nucleic acid interactions. Furthermore, we investigate whether or not the pH-optimum protein-nucleic acid binding is correlated with the stability of the corresponding binding protein.

2. Materials and Methods:

2.1 Protein-nucleic acid structures used in the study:

Protein-nucleic complex structures were downloaded from the NPIDB database [238, 239]. The NPIDB is a large protein-nucleic acid interaction database, which contains 5,547 structures of protein-nucleic acid complexes in the PDB format. The database also includes classification of complexes based on the protein domains using Pfam[240] and SCOP[241] families.

Structural analysis of these 5,547 complex structures showed that there are many entries with very similar structures. This is due to either a particular protein-nucleic acid complex being reported at different experimental conditions, structural resolution, or the existence of highly homologous binding domains. These identical or highly similar structures would result in common protonation state changes in our analysis, and would cause overrepresentation of such protein-nucleic acid interaction types. To eliminate structural bias, we took advantage of the existing Pfam and SCOP classification in the NPIDB database. One representative structure from each Pfam/SCOP family was elected based on the best resolution. We then created two datasets resulting in 112 protein-DNA complex structures and 56 protein-RNA complex structures using SCOP classification, along with 99 protein-DNA complex structures and 105 protein-RNA complex structures using Pfam classification. In this investigation, they are referred as “NPIDB Pfam dataset” and “SCOP dataset”.

2.2 pKa calculations:

The calculations of pKa values were performed with DelPhiPKa [154, 155], which is a Poisson-Boltzmann based approach to calculating the pKa values of protein ionizable residues and nucleotides of RNA and DNA. The profix program, a software module within the JACKAL package (http://wiki.c2b2.columbia.edu/honiglab_public/index .php/Software:Jackal_General_Description) was used to generate missing atoms/residues of the original structures. The ligands and ions were removed from the structures. For each protein-nucleic acid complex, one pKa's calculation was performed for the entire complex structure and then another two calculations were run for the protein and nucleic acid component respectively. This provides pKa values of the titratable groups in bound and unbound states. The pH range in the calculations was set from 0 to 14 with an interval of 1.

2.3 Proton uptake and pH dependence of folding and binding energies:

We calculated the pH dependence of the stability of the complexes and their monomers using the following equation [230, 237, 242]:

$$\Delta G\left(pH_f\right) = 2.3RT \int_{pH_i}^{pH_f} \left(Q_f(pH) - Q_u(pH)\right) d(pH) \qquad (3.1)$$

,where $Q_{f(pH)}$ and $Q_{u(pH)}$ are the total net charge of folded and unfolded states. R is the universal gas constant, taken as 8.314J/(mol*K) and T is the temperature (in K). Similarly, in the case of pH dependence of binding energy, $Q_{f(pH)}$ and $Q_{u(pH)}$ represent the net charge of the complex and the sum of the net charges of the unbound protein and nucleic acid

components. Typically, the difference of these net charges is referred as "proton uptake or release"[242].

The net charge of the folded state for complexes and their components in the pH range were calculated with the DelPhiPKa [154, 155]. In this work, the unfolded state was modeled as a chain of non-interacting residues [237, 242]. Thus, the net charge of the unfolded state was calculated with the Henderson-Hasselbalch equation:

$$Q_u(pH) = \sum_{i=1}^{N} \frac{10^{-2.3y(i)(pH-pKa(i))}}{1+10^{-2.3y(i)(pH-pKa(i))}} \quad (3.2)$$

,where the summation is over all titratable groups in the system and y(i) is +1 for basic groups and -1 for acidic groups.

2.4 Determination of the interfacial residues and classification of nucleotides in DNA/RNA:

A residue is defined to be interfacial residue if its solvent accessible surface area (SASA) changes upon complex formation. The SASA of all residues in the complexes and components was calculated using the VMD plugin [138]. The probe radius was taken as 1.4 Å. For statistical analysis of pKa shifts in DNA and RNA, we classified nucleotides into three different types: phosphate group binding type, base group binding type and O-type [231]. These classifications were based on the different binding modes as described below. These different interaction types were identified by calculating the SASA change of phosphate and base groups upon the complex formation. In our work, we are focused on the effects on the protonation state changes for the N1 and N3 atom in the bases of adenine and cytosine [243]. Instead of the entire base group, we only carry out SASA calculations

on N1, N3, and two bound carbon atoms. For the phosphate group, the SASA calculations were restricted to the P, OP1, and OP2 atoms. Thus, the relative SASA change for each group of atoms of interest was calculated as:

$$\Delta rSASA(residue\ i) = \frac{|SASA(residue\ i\ in\ monomer) - SASA(residue\ i\ in\ complex)|}{SASA(residue\ i\ in\ monomer)} \quad (3.3)$$

Finally, the classification of the binding mode was done using the following rules:

$$nucleotides = \begin{cases} base\ binding\ type, if\ \Delta rSASA(base) \geq 25\% \\ phosphate\ binding\ type, if\ only\ \Delta rSASA(phosphate) \geq 25\% \\ O - type, if\ \Delta SASA(base) and\ \Delta rSASA(phosphate) < 25\% \end{cases} \quad (3.4)$$

3. Results and Discussion:

In the results section, we will first report general frequency patterns of ionizable residues in datasets, as well as statistical analysis of pKa shifts induced by the binding. Furthermore, different pKa shift origins are classified for all ionizable groups based on different chemical-physical properties and binding modes. The pKa shifts among different binding modes are then analyzed. Finally, we investigate the pH dependence of the net charge of binding, complexes, and their components – and reveal how the optimum pH values are correlated (pH-optimum is the pH at which the binding or folding free energy is most favorable, see refs [235, 236] for details). Below we describe the results in sequential order.

3.1 The frequency patterns of ionizable residues in the datasets:

It is expected that proteins binding to negatively charged DNA or RNA should be positively charged, so that the electrostatic interactions are able to guide the protein toward

its binding partner. To investigate this expectation, we carried out statistical analysis of amino acid composition of the proteins in both Pfam and SCOP datasets (Appendix Figure B-16). We considered only Arg and Lys residues to be carrying positive charge, as His is typically neutral. The acidic groups were Glu and Asp. It can be seen (Appendix Figure B-16) that the frequency of Arg and Lys residues are almost the same as the acidic residues in both datasets. Thus, the total net charge of these binding proteins is close to zero in the neutral pH range, which is somewhat unexpected.

We now expand the analysis to the interfacial ionizable residues. The frequency patterns of different residue types were shown in Appendix Figure B-17. In contrast to overall amino acid composition (Appendix Figure B-16), the interfacial regions are enriched with basic residues, resulting in highly positive charged interfacial patches. This confirms our expectations, since both RNA and DNA are highly negatively charged in neutral pH. The important role of electrostatics in protein nucleic acid binding is indicated by our observations that the overall net charge is almost zero, but interfaces are positively charged. It provides guidance for correct orientations of binding partners. It should be mentioned that interfaces of DNA binding proteins are typically more positively charged when compared with RNA binding protein (Appendix Figure B-17).

3.2 Statistics of pKa shifts induced by the binding:

Protein-protein binding frequently involves pKa shifts of ionizable groups as previously demonstrated both computationally [230, 231, 242, 244] and experimentally [242, 245-249]. Here we address the same question for protein-DNA and protein-RNA

binding, including the pKa changes of DNA/RNA bases, using computational methods. The calculations were done for all interfacial residues and bases separately for the complex and monomers alone. The pKa shifts were calculated with the following equation:

$$\Delta pKa^z = pKa^z_{complex} - pKa^z_{monomer} \quad (3.5)$$

,where Z stands for the ionizable group in the protein or DNA/RNA.

The pKa shifts of all interfacial residues were calculated in both Pfam and SCOP datasets and the results are shown in Appendix Figures B-18 and B-19. The corresponding pKa shifts are reported for all protein acidic interfacial residues, protein basic interfacial residues, and nucleic acid ionizable groups (bases) separately. The results are grouped by complex type: protein-RNA complex, protein-double stranded DNA (protein-dsDNA) complex, and protein-single stranded (protein-ssDNA) complex. This is done to facilitate the analysis of the effect of different binding modes. It can be seen (Appendix Figures B-18 and B-19) that complex formation is predicted to cause positive pKa shifts for both acidic and basic protein titratable residues. This is in sharp contrast with the statistical observation made for protein-protein complexes[231]. An opposite shift is predicted for nucleic acid bases – they are predicted to lower their pKa values upon complex formation. These tendencies are similar for all types of complexes (such as protein-RNA, protein-dsDNA, and protein-ssDNA complexes) and remain similar across both datasets (Pfam and SCOP). These pKa shifts originate from the different intrinsic properties of these groups, different binding modes, and different structural features (which will be discussed later).

Further analysis of the pKa shift distribution indicates that the overall pKa shifts of protein basic residues are slightly larger when compared with the pKa shifts of acidic

residues. This may indicate that protein basic groups are frequently involved in direct interactions with negatively charged phosphate groups of DNA/RNA. Nucleic bases of RNA and ssDNA are predicted to undergo larger pKa shifts than those of dsDNA. Perhaps this is due to the double helix structure of dsDNA, where the base groups make hydrogen bonds with their partners and are buried before binding. Due to this, the base groups of RNA and ssDNA are involved in more direct interactions with the corresponding binding protein.

3.3 Analysis of the pKa shift origins:

In this section, we will outline common reasons for predicted pKa shifts and categorize them into several distinctive classes. Since protein titratable groups and DNA/RNA bases have different physical-chemical properties, the origins of their pKa shifts will be discussed separately.

Protein pKa shifts:

pKa shifts are caused by various factors, the most prominent being interactions with other charges and de-solvation penalty (upon complex formation). Based on the comparison of these energy components, we will consider two common scenarios: (a) complex formation is not affected by much solvation energy change (small de-solvation penalty) but provides strong favorable interactions supporting the charged state of the protein titratable group (termed C-type); and (b) complex formation does not greatly affect the solvation energy (small de-solvation penalty) while resulting in strong unfavorable interactions suppressing the charged state of the protein titratable group (termed N-type). A representative example

for the first case, C-type residue, is shown in Figure 3.1A ,depicting a fragment of the binding interface of Archaeosine tRNA-Guanine Transglycosylase complexed with lambda-form tRNA(PDB: 1j2b) [250]. Upon the complex formation, Lys430 of chain A forms a new salt-bridge with the RNA Gua927's phosphate group. In the unbound state, the pKa of Lys430 was calculated to be 10.22, a slight deviation from the standard pKa value. In the complex formation, the de-solvation energy slightly increases by 0.04kcal/mol, since the degree of the burial of the residue does not undergo a large change. However, the interaction energy is changed by -1.52kcal/mol – a contribution from the salt-bridge formed by Lys430 and phosphate group of RNA Gua927. As a result of favorable electrostatic interactions between the protein interfacial basic residue and phosphate group in RNA, the pKa of Lys430 shifts from 10.22 to 12.47 at the complex formation. This type of pKa shift was found in many cases, thus explaining the positive pKa shifts predicted for protein interfacial basic residues. The second common scenario is shown in Figure 3.1B for the structure of PVUII Endonuclease complexed with cognate DNA (PDB: 3pvi) [251]. In unbound protein, the Glu68 residue of chain A is exposed to the water and the side chain is stabilized by the interaction with the nearby Lys70. Upon the complex formation, Glu68 side chain points to the phosphate group of DNA Cyt9. As shown in the corresponding Figure, the oxygen-oxygen distance between the Glu68 side chain and DNA Cyt9 phosphate group is only 3.5 Å. This results in strong unfavorable interactions opposing the charged state of Glu68. The existing interactions of the Glu68 and Lys70 are additionally weakened in the complex as Lys 70 forms new interactions with a phosphate group. According to the energy calculation of DelphiPka, the interaction energy is increased by

0.64 kcal/mol and the de-solvation energy is only slightly increased, since the degree of the burial of the residue does not change much. As result, the Glu68 pKa value is shifted from 3.93 to 5.11. Therefore, we refer to these kinds of residues (which are under unfavorable interactions in the complex) as N-type protein residues. As shown in the previous statistical analysis of pKa shifts, the majority of the protein acidic residues are affected by different degrees of positive pKa shifts. Most of these cases can be classified as N-type residues due to the unfavorable electrostatic interactions between the acidic residue and the phosphate group of the DNA/RNA nucleotides.

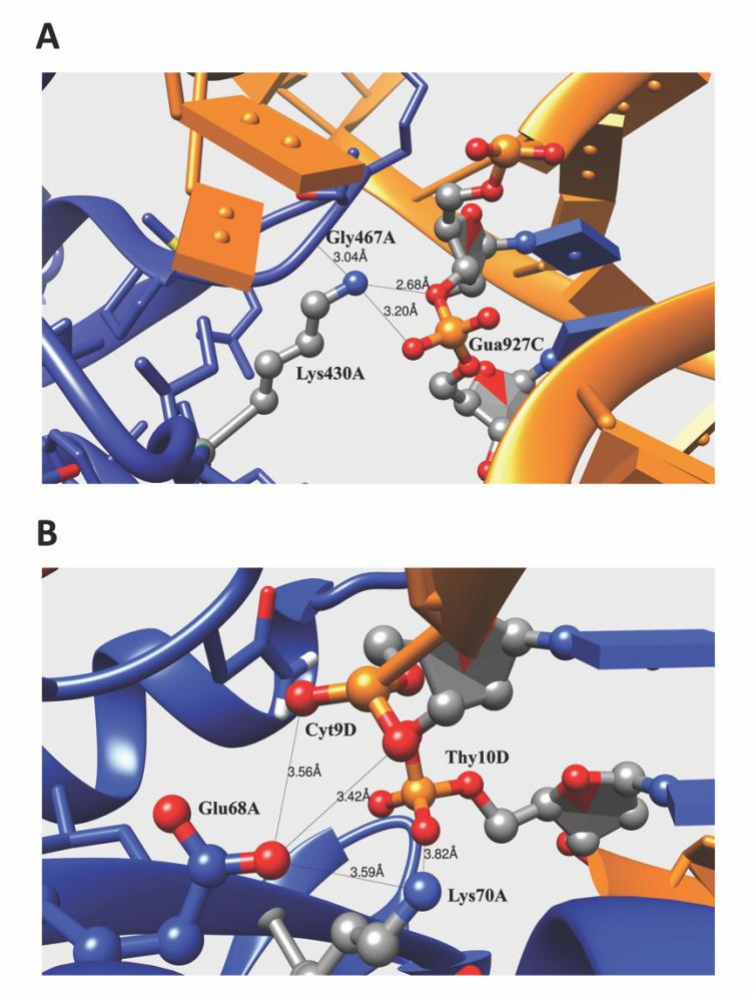


Figure 3.1: (A) Fragment of binding interface of Archaeosine tRNA-Guanine Transglycosylase complexed with lambda-form tRNA (PDB: 1j2b). (B) Fragment of binding interface of PVUII Endonuclease complexed with cognate DNA (PDB: 3pvi). The side chains of the residues directly contributing to the electrostatic interactions or H-bonding are shown with balls and sticks. The protein and DNA/RNA are marked as blue and orange for comparison. The distance between the atom pairs are shown in Å.

DNA/RNA:

In this investigation, the pKa values of Cys and Ade bases are predicted for bound and unbound states. These bases are typically neutral at physiological pH (pH about 7), but can be protonated and positively charged in some cases [154]. As shown from the above statistical analysis, the majority of DNA/RNA bases are predicted to undergo negative pKa shifts due to the binding. Therefore, the base groups are less likely to be protonated at physiological pH values. We will group the common pKa shift scenarios into several categories: (a) bases experiencing large de-solvation penalty and forming H-bonding or electrostatic interactions (termed B-type), (b) bases experiencing electrostatic interactions and negligible de-solvation penalty (termed L-type), and (c) bases experiencing small de-solvation penalty (O-type). Typically, the B-type residue is a base group in which the nucleotide is buried into the binding interface and directly participates in hydrogen bonding or electrostatic interactions upon the complex formation. One representative example is shown in Figure 3.2A: a CCA-adding enzyme complexed with tRNA (PDB: 3ovb) [252]. The atom N3 of the base group of the RNA Cyt33 is predicted to have standard pKa of 4.35 in unbound RNA. In chain C of the complex, the atom N3 of Cyt33 makes a hydrogen bond with the backbone atom of His93 of chain A of the complex. Since the N3 atom plays the role of a proton acceptor, such an interaction increases the energy cost of protonation. Cys 33 is also buried at the binding interface, and thus the charged form of Cys 33 pays the de-solvation penalty. Combining these two effects, the RNA Cys 33 pKa shifts from 4.56 to 1.14. Another example is shown in the Figure 3.2B: TilS complexed with tRNA (PDB: 3a2k)[253]. The base group of RNA Ade36 is buried into the binding interface and

surrounded by a pocket of three Arg residues. Although the base group is not directly forming interactions with nearby residues, it is still affected by electrostatic interactions of the three positively charged Arg residues. This results in a shifting equilibrium towards the de-protonated form. The degree of burial for Ade36 is increased upon complex formation, resulting in a de-solvation penalty. Finally, the pKa value of Ade36 is predicted to decrease from 4.90 to 2.75.

Another common type of pKa shift, listed above as L-type, is shown in Figure 3.2C. This shift occurs in case of restriction endonuclease MspI on its palindromic DNA recognition site (PDB: 1sa3)[254]. The Cyt13 of chain D resides in a double strand structure, and its base group makes a hydrogen bond with its base pair. The base group of Cyt13 is pre-buried in unbound DNA and its degree of burial is almost unchanged in the complex. Therefore, Cys13 does not pay a de-solvation penalty upon complex formation. However, a nearby positively charged Lys261 protein residue does interact with the base of Cyt13. This unfavorable interaction energy is calculated to be about 0.5 kcal/mol, and along with other smaller contributions, results in pKa shift of -0.83.

Finally, the common cases referred to above as “O-type” are represented by many other residues that are not involved in strong interactions with charged residues upon complex formation. Their pKa shifts are relatively small ($|\Delta pKa| < 0.5$) and are mostly due to a de-solvation penalty upon complex formation.

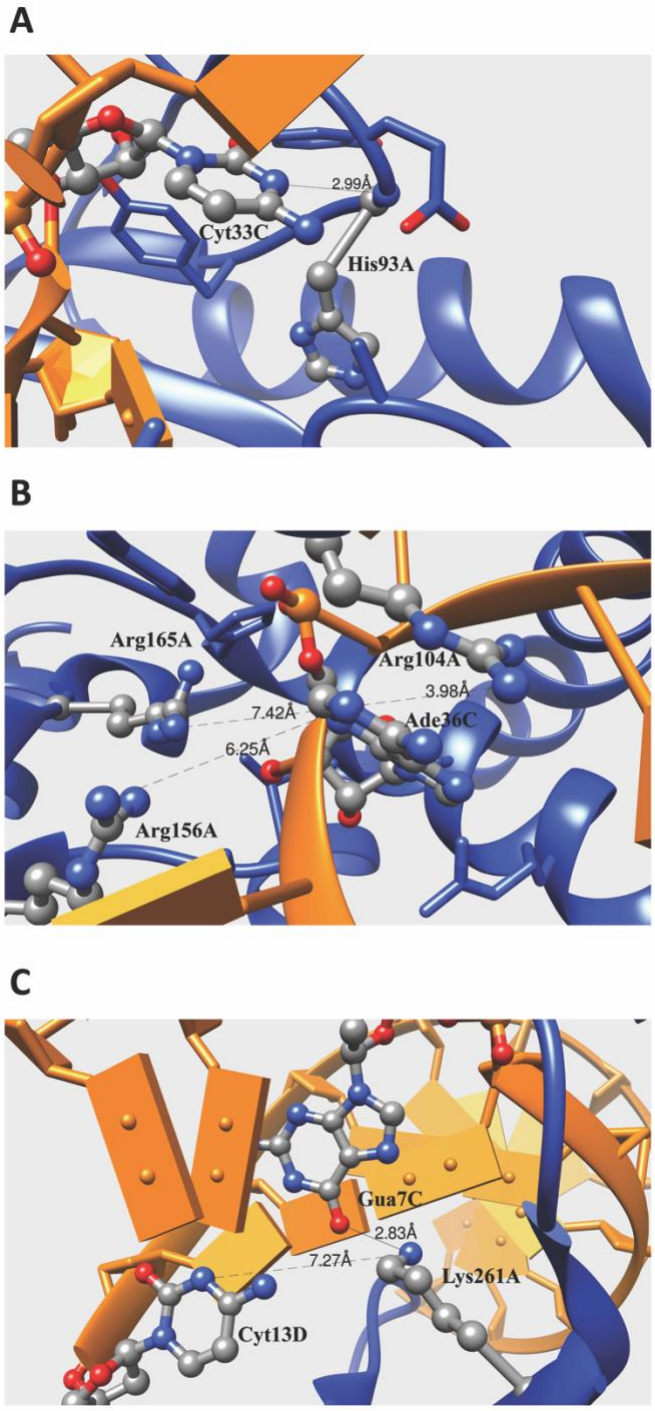


Figure 3.2: (A) Fragment of binding interface of CCA-adding Enzyme complexed with tRNA (PDB: 3ovb). (B) Fragment of binding interface of TilS complexed with tRNA

(PDB: 3a2k). (C) Fragment of binding interface of restriction endonuclease MspI on its palindromic DNA recognition site (PDB: 1sa3). The side chains of the residues directly contributing to the electrostatic interactions or H-bonding are shown with balls and sticks. The protein and DNA/RNA are marked as blue and orange. The distances between atom pairs are shown in Å.

3.4 pKa shifts and binding mode:

In this section we investigate the effect of different binding modes on previously discussed pKa shifts. Here, we classify the binding modes into three categories: (a) phosphate group binding mode (protein interacts mostly with phosphate groups), (b) base group binding mode (protein interacts mostly with base groups), and (c) others (O-type mode: categorization is outlined in Method section).

Appendix Figures B-20 and B-21 show the distributions of pKa shifts in different binding modes for both Pfam and SCOP datasets. The most significant pKa shifts are predicted for the base group binding mode. In base group binding modes, the base groups directly participate in the interactions across the interface and bases are buried at the binding interface, thus paying a large de-solvation penalty. According to the above categorization strategy, these bases are grouped as B-type nucleotides. Both the de-solvation penalty and interaction energy oppose the charged form of the bases and thus result in significant pKa shifts, lowering the pKa of the bases. Phosphate group binding modes result in L-type bases, as the base groups are not at the interface and do not experience burial change upon complex formation (but are affected by long-range

electrostatic interactions). In the phosphate group binding mode, the bases are predicted to have relatively less significant pKa shifts. The rest of these cases are mostly of O-type, and are predicted to have the smallest pKa shifts as they are not involved in strong interactions and do not experience a de-solvation penalty upon binding.

3.5 pH-optimum of binding:

Previous studies investigated protein stability and interactions as a function of pH, and referred to the pH-optimum as the pH of maximal stability and interactions [235-237]. This optimum pH can be obtained by finding the pH value at which the net charge difference of the folded and unfolded states, or bound and unbound states, is zero. We illustrate this with a particular example from our dataset. The pH dependence of the net charge difference and the pH dependence of the binding free energy for a bacteriophage lambda cII protein in complex with dsDNA (PDB: 1zs4) [255] is shown in Figure 2.15. Three distinctively different pH regions can be clearly identified. The first region is in the acidic pH range, where both the net charge difference ($\Delta Q < 0$) and the free energy decrease (the free energy of binding becomes more favorable) with an increase in pH. Proton release occurs in this pH range upon complex formation or protein folding which involves mostly acidic groups. The third region is in the basic pH range, where $\Delta Q > 0$ and the free energy increases with the pH (the binding free energy becoming less favorable). Proton uptake involving predominantly basic groups occurs in this pH range. The second region is in the intermediate pH range, where ΔQ is close to 0 and remains almost unchanged with the increase in pH. The titration of acidic or basic groups with non-standard pKa values occurs

in this pH range. Since most of the proteins perform their function in this intermediate pH range, it is the most interesting pH region for the study. The optimum pH can be determined by finding the pH corresponding to the minimum free energy. This is usually located at the border of first and third pH regions. As shown in Figure 3.3, the ΔQ in the second region is frequently very small, practically close to zero. Thus, the results from this pH region are very sensitive to the imperfections of computational protocol and applied methodology. To reduce the error in finding the optimum pH, we introduce a threshold value Q_t and assume that $\Delta Q = 0$ if $abs(\Delta Q) < Q_t$. We explored different values for Q_t (from 0.1, 0.2, 0.3 to 0.4) and the best results (in terms of obtaining the best correlation coefficient, explained below) were obtained with the 0.1 value. In cases of very flat intermediate pH regions, the pH-optimum was taken to be the center of the intermediate pH range. In Figure 3.3(B) and (D), the intermediate pH regions for binding and protein stability both range from pH 4 to 9. Thus, the optimum pH for the binding and stability of the protein component is taken to be 6.5. Several other approaches were explored as outlined below.

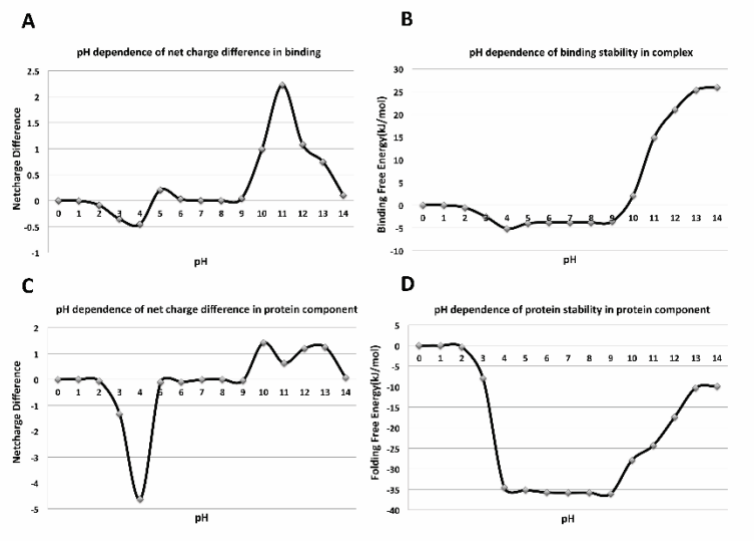


Figure 3.3: pH dependence of the net charge difference and free energy for a bacteriophage lambda cII protein in complex with dsDNA (PDB: 1zs4). (A) and (B) show pH dependence of net charge difference (proton uptake/release) and the corresponding pH dependence of the binding free energy. (C) and (D) show pH dependence of net charge difference and folding free energy of the protein component.

3.6 Correlation of pH-optimum of binding and protein stability:

Protein-DNA/RNA interaction is a pH-dependent process, with the binding affinity reaching a maximum at the pH-optimum. In vivo, the monomers and their complex coexist in the same subcellular environment and thus should be adapted to the corresponding subcellular pH [233]. Indeed, it was demonstrated that the optimum pH of binding and folding are correlated [231, 236]. In this work, we investigate the possibility that the pH-optimum of protein-DNA/RNA binding is correlated with the pH-optimum of the folding of the corresponding binding protein. We do not address the same question for RNA/DNA stability, since our approach considers only basic titration of RNA/DNA titratable groups and therefore the titration is monotonic with pH.

The optimum pH was determined by using the above-discussed strategy for both SCOP and Pfam datasets, and $\overline{Q_t}$ was taken as 0.1 in the calculations. A fraction of cases did not show clear pH dependence and thus no optimal pH value could be determined. Thus, we excluded these cases from the correlation analysis and the rest of the cases (62 out of 105 cases and 33 out of 56 cases for protein-RNA complexes in Pfam and SCOP datasets respectively, as well as 68 out of 99 cases and 91 out of 112 cases for protein-

DNA complexes in Pfam and SCOP datasets respectively) were subjected to two different protocols to assess pH-optimum: (a) pH-optimum is taken to be the middle of the "flat", almost pH independent region and (b) the pH-optimum of binding is taken within the "flat" pH region, with selected pH being the closest to folding of the corresponding binding protein. The results are summarized in Table 3.1 and the corresponding plots are provided in the supplementary material. One can see a weak correlation between the pH-optimum of binding and the pH-optimum folding of the corresponding binding protein.

Scenario (a)			Scenario (b)		
Complexes Type	Correlation coefficient for all complexes	Correlation coefficient for STDEV<2	Complexes Type	Correlation coefficient for all complexes	Correlation coefficient for STDEV<2
Protein-RNA in SCOP	0.71	0.66	Protein-RNA in SCOP	0.78	0.83
Protein-DNA in SCOP	0.3	0.58	Protein-DNA in SCOP	0.42	0.83
Protein-RNA in Pfam	0.48	0.56	Protein-RNA in Pfam	0.5	0.77
Protein-DNA in Pfam	0.24	0.27	Protein-DNA in Pfam	0.41	0.74

Table 3.1 Pearson product-moment correlation coefficient between pH-optimum of binding and folding of the corresponding binding protein. Results are shown for both SCOP and Pfam classifications and the two scenarios outlined above. For each complexes type,

Pearson product-moment correlation coefficient is calculated for all complexes and also for complexes in which outliers are excluded (standard deviation > 2 pH units).

4. Conclusion:

In this work, we investigated the electrostatic properties, pKa shifts, proton uptake/release, and pH-optimum of a large number of protein-DNA/RNA complexes with available 3D structures. The analysis of the pKa shifts induced by the complex formations indicated a completely different trend in comparison with previous studies on protein-protein complexes [230, 231, 256]. Protein titratable residues were found to undergo positive pKa shift, thus increasing the pKa values of both basic and acidic groups. Such an opposite trend (opposite to the trend observed for protein-protein complexes) is due to the difference between the electrostatic properties of the corresponding partners. In the case of protein-protein complexes, the interfaces are frequently made up of patches of opposite polarity and thus the given protein may provide a favorable electrostatic environment for both basic and acidic groups [231, 256]. In contrast, most of the binding modes in our dataset consist of cases in which the protein binds to phosphate groups of DNA/RNA. Since phosphate groups are negatively charged, the electrostatic environment for protein titratable groups make the charged state of acidic groups less favorable while promoting the charged state of basic groups. This is the main reason for the observed tendency of protein-DNA/RNA binding to induce positive pKa shifts of protein titratable groups. In contrast, the binding causes pKa values of nucleic acid bases to lower. Most of this effect

is due to unfavorable electrostatic interactions with the positively charged interface of the corresponding binding protein.

Very little proton uptake/release was predicted to accompany the binding. For many cases in the dataset, the proton uptake/release was almost zero for the pH range of 5 to 8. This is also quite different from observations made of protein-protein complexes [229, 230, 236]. Protein-DNA/RNA binding seems to be less pH dependent than protein-protein binding. This likely reflects the fact that protein-protein interactions occur in more diverse environments than protein-DNA/RNA binding.

A weak correlation was found between the pH-optimum of binding affinity and the folding free energy of unbound protein. The correlation is not as significant as correlations found for protein-protein binding [229-231, 236]. This may be due to the fact that only basic groups of DNA/RNA were treated as titratable residues in our protocol. We anticipate that the inclusion of other groups (phosphate groups, for example) could result in a more significant correlation. Another reason could be computational protocol, which treats the structures of both protein and DNA/RNA as rigid bodies, considers that pKa shift are entirely due to electrostatic energy changes, does not include explicit ions and does not allow for water penetration at the binding interface. Some of the abovementioned deficiencies are intrinsic to continuum approaches [257], others can be handled via continuum approach as explicit ion binding [258-260], but were not implemented in this study in order to reduce computational cost of modeling such large set of complexes. Further insights can be obtained via constant pH molecular dynamics simulations (MD) [261, 262]. Recent constant pH MD study proposed “trap-and-trigger” mechanism was

proposed to accompany protein binding and to involve structural rearrangement and water penetration at the interface [263]. Such structural rearrangements upon molecular recognitions are frequently revealed in studies utilizing constant pH MD [264, 265], indicating that rigid body approach may induce significant error in modeling pKa's [257].

Overall, our study indicates that electrostatics play a significant role in protein-DNA and protein-RNA binding and frequently this binding is accompanied by pKa shifts, resulting in little proton uptake/release and weak pH dependence.

Development of computational approach in prediction of SAV's effect on protein DNA binding:

1. Introduction

Protein-DNA interactions are essential for functions of living cells and are involved in many important cellular processes such as transcription, replication, and recombination. For example, the expression level of genes is regulated by a wide number of proteins named transcription factors, which have DNA-binding domains recognizing a specific sequence of DNA [266, 267]. Protein-DNA binding is mediated by many factors such as DNA sequence, hydrogen bonds, van der Waals contacts, DNA shape, protonation states, flexibility and many others [217, 223, 224, 268-270]. While DNA–backbone interactions are important for the stability of protein-DNA complexes, proteins recognize specific DNA sequence by forming hydrogen bonds between amino-acid side chains and DNA bases [217, 224, 225].

Therefore, mutations occurring in DNA binding proteins that alter the physicochemical properties of the binding interfaces will affect binding specificity and affinity [271, 272]. Such mutations are frequently involved in many diseases like neurological disease, heart disease and cancer. Hence, understanding their molecular effects is crucial for deciphering disease origins and pursuing treatment [48, 273-275].

Significant fractions of diseases are caused by the alteration of native binding affinities, which can be quantitatively described by the binding free energy change [26, 198]. There are many experimental techniques capable of measuring protein-DNA binding free energy such as isothermal titration calorimetry (ITC) [276], fluorescence resonance energy transfer (FRET) [277], nuclear magnetic resonance(NMR) [278], surface plasmon

resonance(SPR) [279] and many others. However, these experimental methods are usually time consuming and non-applicable for large-scale studies. Recently, the available experimental data of protein-DNA binding free energy changes caused by amino acid substitutions was compiled and organized in a database, the ProNIT database [32].

Computational approaches can complement experimental techniques and permit large-scale investigations. Among them, the free energy perturbation (FEP) and the thermodynamic integrations (TI) are the most rigorous, but require intensive calculations, which limit their applicability for large-scale analysis. Alternatively to FEP and TI, different physical models and optimized knowledge-based potentials have been developed to carry out fast predictions of protein–DNA binding affinities achieving a good correlation with experimental measurements [280-284]. A structured based approach, the mCSM method, was developed [151, 285] and was shown that it achieves correlation coefficient of 0.673 in benchmarking test against ProNIT database. Very recently, mCSM-NA, an improved version of mCSM method, achieved correlation coefficient of 0.72 in benchmarking against ProNIT database [285]. Even so, the existing approaches for fast prediction of protein-DNA binding affinity changes upon mutations are still very limited, comparing with approaches developed for protein-protein interactions.

The Molecular Mechanics/Poisson Boltzmann Surface Area (MM/PBSA) approach is a widely applied method to calculate binding free energies of macromolecules by combining molecular mechanics calculations and continuum solvation models [286-288]. The MM/PBSA method computes a linear combination of energy terms for molecular mechanics, polar and non-polar solvation energy and shows high computational efficiency

comparing with the rigorous methods such as FEP and TI methods. In this work, we developed a new approach termed SAMPDI (**S**ingle **A**mino acid **M**utation binding free energy change of **P**rotein-**D**NA **I**nteraction) to perform fast predictions of binding free energy changes of protein-DNA complexes caused by single mutations on the proteins. Our approach combines modified MM/PBSA based energy terms with additional knowledge-based terms. The method is implemented in a webserver (http://compbio.clemson.edu/SAMPDI/), which allows the users to upload the corresponding protein-DNA structural file, to specify the mutations and to obtain the predicted binding free energy change.

2. Methods

2.1 Dataset preparation

We constructed a dataset, containing experimentally measured binding free energy change upon missense mutations and corresponding PDB structures, by combining the ProNIT database [32] and data from recent references. We applied three criteria in constructing the dataset: 1) Mutations affecting protein DNA binding, but not the quaternary structure of the corresponding protein, like dimerization. 2) The binding site of DNA (DNA sequence of the interface) used in the experiment is exactly identical to the DNA sequence of the corresponding PDB structure. 3) The structures with modified DNA, like methylation were removed and not considered in this study. Finally, the constructed dataset for this study included 105 missense mutations from 13 proteins. (The constructed dataset used in this study is shown in the supplementary material and can be downloaded from URL: http://compbio.clemson.edu/downloads)

2.2 NAMD Simulation protocols

The structures of protein-DNA complexes were downloaded from RCSB Protein Data Bank (PDB) [289]. The biological units were retained and ligands, except ions, were removed from the initial structures. The missing heavy atoms were fixed using the default parameters of the profix module in Jackal package (https://honiglab.c2b2.columbia.edu/software/Jackal/Jackalmanual.htm). The mutant (MT) structures were generated by the VMD Mutator plugin [138] using the topology files from CHARMM36 force field [290, 291]. The energy minimization was performed with the NAMD program, version 2.11b [95] using the conjugate gradient algorithm. The default minimization steps were set to 5000 steps but longer minimization was applied if the variation of the total energy was more than 0.5 kcal/mol. In the minimization, the Generalized Born implicit solvent (GBIS) model and CHARMM36 force field[290, 291] were used. The dielectric constant of the implicit solvent was set to 80 and the various values of the protein-DNA dielectric constant were tested (see results section). Finally, the minimized structures were used to calculate the relevant energies.

2.3 Electrostatic energy calculations

Delphi with the Gaussian-based smooth dielectric function [83, 292, 293] was used to calculate the electrostatic component of the binding free energy in the Protein-DNA binding interaction using the following parameters: scale = 2 grid/Å; percentage of filling for the protein-DNA complex structures = 70%; dielectric constant = 80 for the solvent; salt concentration =0.15 mol/L; Gaussian with sigma=0.93, srfcut=20 and non-linear Poisson-Boltzmann equation (PBE) (non-linear PBE was used because of the high charge

of the DNA). Grid box for protein and DNA monomers were set exactly identical as for their complex by specifying the grid box size and center.

2.4 Binding free energy calculations

This study combines a modified MM/PBSA approach and knowledge based energy terms to calculate the protein-DNA binding free energy change upon single amino acid substitution. MM/PBSA is a widely used approach to calculate the receptor-ligand binding free energy and the thermodynamic cycle of computing the binding free energy change upon single amino acid change is shown in Figure 3.4. In our approach, the unbound monomer structures are taken from the corresponding complex, thus assuming no structural changes upon the binding (called rigid body approach). In addition, a set of knowledge based energy terms, which are derived from analysis of physicochemical properties of the corresponding protein-DNA structures, are combined with the MM/PBSA approach (more details are provided in refs [30, 34]). All individual energy terms are combined via weighted linear scoring function and optimal weighted coefficients are determined via multiple linear regression against experimental data. Below, we will describe the protocols of computing each energy terms, including the MM/PBSA and knowledge based ones.

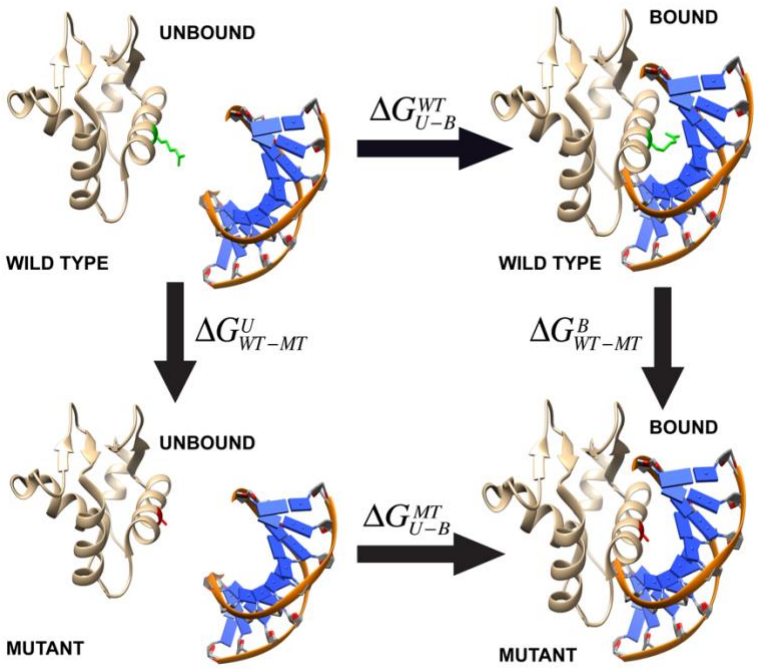


Figure 3.4: Thermodynamic cycle for binding free energy change calculations. The side chain of wild type and mutant residues are show in green and red color, respectively.

2.4.1 The MM/PBSA-based energy terms

The MM/PBSA components of the change of the binding free energy are in a linear combination of the five components shown in the following equation:

$$\Delta\Delta G^{MM/PBSA} = w_0 + w_1 \cdot \Delta\Delta IE + w_2 \cdot \Delta\Delta CE + w_3 \cdot \Delta\Delta PS + w_4 \cdot \Delta\Delta VE + w_5 \cdot \Delta\Delta NS \quad (3.6)$$

,where IE is the internal energy, CE is the Coulombic energy, PS is the polar component of the solvation energy, VE is the van der Waals energy, NS is the non polar component of the solvation energy and *wi* are weight coefficients. The energy difference for each energy term is computed using the following equation:

$$\Delta\Delta E = \left(E_{complex}^{MT} - E_{protein}^{MT} - E_{DNA}^{MT}\right) - \left(E_{complex}^{WT} - E_{protein}^{WT} - E_{DNA}^{WT}\right) \quad (3.7)$$

,where MT and WT represent the mutant and wild-type structures. The structures of unbound protein and DNA are taken from the complex structures. Below we describe each energy component (more details can be found in [30]).

IE and VE energies were calculated using the NAMD program. Since the rigid body approach was applied and no structural changes are considered in the binding, ΔΔIE calculated by equation 3.7 will result in zero. In our methodology development, we have tried to minimize the complex structure and unbound monomer structure separately to take into account the structural changes induced by the binding. However, the results showed weaker correlation between the predicted value and experimental data comparing with applying the rigid body approach, thus *w*1 was set to zero. VE energy was obtained with NAMD by subjecting the corresponding minimized structure to an one step equilibration.

CE and PS were calculated using the Delphi program with Gaussian-based smooth dielectric function, an accurate and fast Poisson-Boltzmann Equation (PBE) solver [83, 292]. In Gaussian Delphi, the solute and water phase are treated as an inhomogeneous dielectric medium by using a smooth Gaussian-based dielectric function, which showed better performance comparing with the traditional two-dielectric model (the traditional two dielectric model treats biomolecule and water as two distinctive media with two different dielectric constants with a sharp dielectric border between the two media). The performance of the traditional two-dielectric model and the smooth Gaussian-based model were tested and the Gaussian-based model showed better results as benchmarked against experimental data.

NS was calculated via the solvent accessible surface area (SASA) using the equation 3.8. The SASA was computed using the NACCESS software with default atom radius parameters [294]. The constants α and β in equation 3.8 were incorporated into to the weight coefficient in equation 3.6.

$$NS = \alpha SASA + \beta \tag{3.8}$$

2.4.2 Knowledge-based energy terms

Many knowledge-based energy terms were tested in this study among which entropy (S) and hydrogen bond (HB) showed highest impact. The impact was evaluated based on the p-test indicating that S and HB are the terms showing highest correlation with experimental measured binding free energy changes (see supplementary material). Finally, the knowledge-based energy terms (ΔΔGKW) are a linear combination of the two components shown in the following equation:

$$\Delta\Delta G^{KW} = w_1 \cdot \Delta\Delta S + w_2 \cdot \Delta\Delta HB \tag{3.9}$$

where S is the entropy, and HB is the number of hydrogen bonds. The energy differences for each term are also computed using equation 3.7.

The entropy of protein's residue is calculated using the following empirical formula originally developed in our previous work [30].

$$S = ln[rSASA(i) \cdot (R(i) - 1) + 1] \tag{3.10}$$

,where rSASA(i) represents the relative solvent accessibility of residue i (calculated by the NACCESS software [294]) and small rSASA(i) values (close to 0) indicate that the residue is buried and only a few side chain rotamers can be sampled, which results in a small entropy contribution; R(i) is the maximum number of the rotamers for residue i (R(i) for

all types of residues are shown in the Table C-6). The entropy change upon mutation is calculated by subtraction of the entropy for the wild-type residue and mutant residue.

The number of hydrogen bonds (HBs) is calculated using the VMD plugin with a cut-off distance 3.0 Å and a cut-off angle of 30 degrees. We tried two protocols to compute the number of HB: 1) compute the total number of the HBs for the entire structures (including intra and inter HBs); 2) only compute the number of HBs near the mutation site and choose to count the HBs within 6 Å of the mutation site (different cut-off values were tested and 6 Å showed the best correlation). The second protocol was applied in our calculation since it showed much better correlation with the experimental $\Delta\Delta G$ in the p-test (see supplementary material).

3. Results

3.1 Finding optimal value of dielectric constant

In our protocol we used an implicit model to minimize protein-DNA structures and to calculate the MM/PBSA energy terms. Different dielectric constant values affect the energy minimization and the energy terms calculated with both Delphi and NAMD programs. Our previous works showed that selecting an optimal dielectric constant value for proteins results in improved correlation coefficient for binding/folding free energy calculation [30, 34]. Here, we tested various dielectric constants for the protein-DNA complex to identify the optimal value corresponding to the highest correlation coefficient against experimental data. Figure 3.5 shows the dependence of correlation coefficient on the value of the dielectric constant of the protein-DNA complex. We varied the dielectric

constant of protein-DNA from 1 to 5 for NAMD program (this was done for testing purposes, while understanding that dielectric constant value of 1 is physically sound) and 1 to 20 for Delphi program with a step of 1. Multiple linear regression was performed for each set of values of dielectric constants using VDW energy, Coulomb energy and the polar component of the solvation energy to obtain the correlation coefficient (Figure 3.5). The results indicate the dielectric constant value used in NAMD modeling highly affects the correlation coefficient (Figure 3.5). Summarizing, the correlation coefficient reaches the highest value with a dielectric constant for NAMD =1 and for Delphi =14 and these values will be used in our protocol.

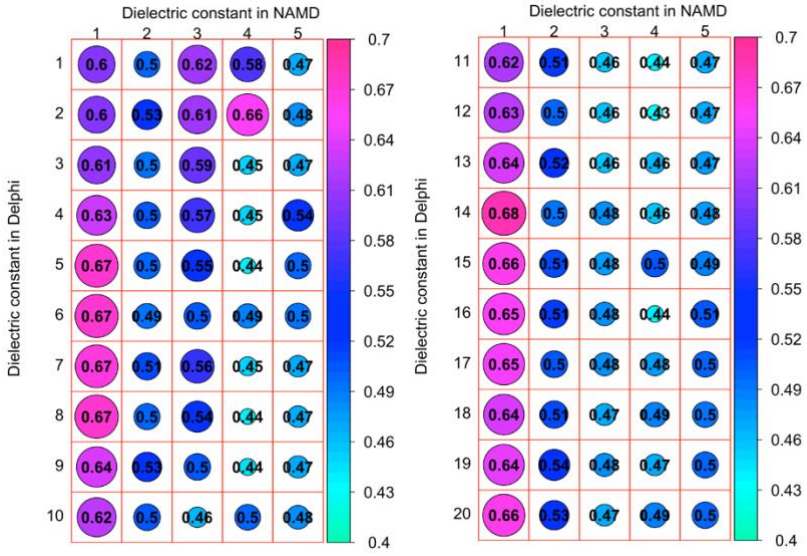


Figure 3.5. The correlation coefficient calculated with various dielectric constants used in Delphi and NAMD. Panel A shows the correlation coefficient dependence of dielectric constant from 1 to 5 in NAMD and 1 to 10 in Delphi while Panel B shows the dependence of dielectric constant from 1 to 5 in NAMD and 11 to 20 in Delphi. The size and color of

circle are representing the correlation coefficients for a particular dielectric constant selection.

3.2 Determination of optimal values of the weight coefficients

As discussed in the Method section, the linear function of binding free energy changes contains 6 terms and 7 weight coefficients:

$$\Delta\Delta G = w_0 + w_1 \cdot \Delta\Delta CE + w_2 \cdot \Delta\Delta PS + w_3 \cdot \Delta\Delta VE + w_4 \cdot \Delta\Delta SASA + w_5 \cdot \Delta\Delta S + w_6 \cdot \Delta\Delta HB \quad (3.10)$$

Then, the weighted coefficients are determined from the multiple linear regression (MLR) between experimentally measured ΔΔG and calculated binding free energy changes. The resulting optimized weight coefficients are shown in Table C-7. The correlation coefficient from MLR is 0.72 over 105 cases. The plot of experimentally measured binding free energy changes and predicted binding free energy changes is shown in Figure 3.6.

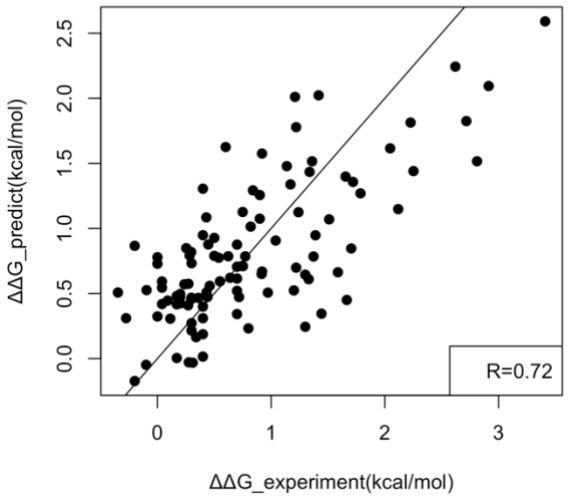


Figure 3.6. A plot of experimentally measured binding free energy changes and predicted binding free energy changes. The corresponding linear fit and correlation coefficient are shown as well.

3.3 Performance and Validation

3.3.1 5-fold cross validation

In our study the datasets used for training and testing are relatively small due to limited available experimental data. To address the problem of overfitting, we further performed 5-fold cross validation by randomly partitioning the dataset into five subgroups of approximately equal sizes. For each round, four subgroups are used for training and the rest one is used for testing. The results are shown in the Table Appendix C-8 and Figure 3.7A. The Root Mean Square of the Error (RMSE) in each fold varies a little and the resulting average is 0.54 kcal/mol. At the same time, Pearson correlation coefficient (CC) varies significantly probably due to the limited number of data points (20 data points for each fold and the corresponding CC shows significant variation even if with roughly same RMSE). We also analyzed the variation of the weighting coefficients for each energy terms in 5-fold cross validation and the results are shown in Appendix Table C-9. The standard deviation of the weighting coefficients is relatively small and indicate the variation is not significant across each fold. We further compared the average weighting coefficients in 5-fold cross validation with the previous determined weighting coefficients from MLR and the results in Appendix Table C-9 shows that the differences for all the energy terms are very small. Overall, the testing indicates that overfitting is not significant.

3.3.2 Receiver operating characteristic (ROC)

To evaluate the performance of SAMPDI, we further performed ROC analysis to distinguish large and small effects on binding free energy changes. Here, we classify the large effects as $|\Delta\Delta G| > 1$kcal/mol and small effects as $|\Delta\Delta G| < 1$kcal/mol. Figure 3.7B shows the ROC curve of SAMPDI for 105 experimentally measured binding free energy changes. The area under the curve is 0.76, indicating the capability of SAMPDI to distinguish different types of mutations.

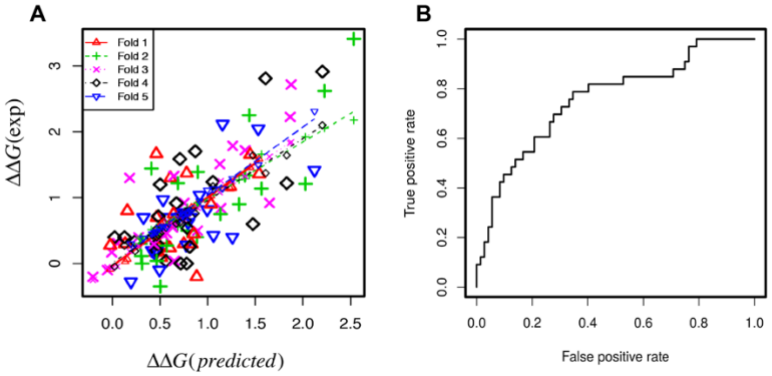


Figure 3.7. (A) Plot of predicted ΔΔG and experimental ΔΔG in 5-fold cross validation. (B) Receiver operating characteristic curve of classification of large effects ($|\Delta\Delta G| > 1$kcal/mol) and small effects ($|\Delta\Delta G| < 1$kcal/mol).

3.3.3 Multicollinearlity analysis

It may be anticipated that some energy terms may reflect similar phenomena. To address such a possibility, we performed multicollinearlity analysis to study the correlation across each term and the variance inflation factors (VIF) from MLR. The results shown in Appendix Table C-10 indicate a strong correlation between CE and PS. This is due to the well-known fact that the PS originates from the CE. In addition, SASA has relative high

correlation with VDW, CE and PS. The rest, the VIFs of SASA, VDW, S and HB are within relative low multicollinearlity (VIF < 4). Removing highly correlated terms from eq. (5) results in decrease of prediction accuracy, but the change is not large. For example, removing the CE in the MLR leads to the decrease of correlation coefficient from 0.72 to 0.65. Thus, these highly conserved terms were kept in our final protocol to achieve optimal accuracy.

3.3.4 Case studies: consistent and inconsistent predictions comparing with experimental data.

To further investigate the factors affecting the predictions, representative examples of consistent and inconsistent predictions will be discussed below. The results of six single mutations shown in Table 3.2 will be discussed.

Protein PDB (Mutation)	ΔΔG (EXP)	ΔΔG (PRED)	ΔΔSASA	ΔΔVE	ΔΔCE	ΔΔPS	ΔΔS	ΔΔHB
1B3T (R469A)	3.4	2.6	260.7	16.0	-61.4	106.9	1.8	-11.0
1B3T (Y518A)	2.6	2.2	72.6	14.3	1.3	-0.9	1.2	-9.0
1MSE (C130I)	0.3	0.2	3.6	-2.0	-2.3	3.8	0.0	0.0

1MSE E141A	-0.1	-0.1	0.8	-1.8	7.6	-19.4	0.0	0.0
1TN9 K54A	1.3	0.6	-18.0	-3.5	-20.5	38.4	0.9	-2.0
2A0I E187A	2.1	1.2	24.3	7.4	7.1	-11.5	0.5	0.0

Table 3.2 Cases of consistent and inconsistent predictions. Mutations in protein 1B3T and 1MSE are the cases of consistent predictions (underlined), while the rest are inconsistent prediction cases. The ΔΔGs are in kcal/mol and positive value indicates destabilization (lowering protein-DNA affinity) while negative indicates stabilization. The ΔΔE for each terms is shown as MT-WT.

- Predictions consistent with experimental data:

Epstein-Barr nuclear antigen 1 (EBNA1) binds to the recognition site of the minimal origin of latent DNA replication of Epstein-Barr virus and results in activation of the latent-phase replication of the viral genome [295]. Here, we outline two single mutations (R469A and Y518A) of a permanganate-sensitive DNA site bound by EBNA1. Both mutations occur on the binding interface (PDB: 1B3T, Figure 3.8A) and dramatically destabilize the protein-DNA binding according to the experimental measurement (3.4 and 2.6 kcal/mol, respectively). The wild type residue R469 interacts with the DNA backbone and forms strong electrostatic interactions upon binding. Our calculations predict that a substitution to ALA will result in dramatic energy change of 61.44 kcal/mol of CE and 16.02 kcal/mol of VE upon binding along with a large effect on the SASA, HB and S (Table 3.2). Taking all together we predicted that R469A would cause decrease of 2.6 kcal/mol of binding free energy, which is very close to experiment. Another mutation, Y518A is also located at the

binding interface, which leads to a large change of VE along with decrease of HB and S. For both mutations, the experimental measured free energy changes are dramatic and destabilize DNA binding, which is reproduced by the SAMPDI. Another representative example are two single mutations (C130I and E141A) in the structure of a specific DNA complex of the Myb DNA-binding domain with cooperative recognition helices (PDB: 1MSE, Figure 3.8B) [296]. Both mutations are not in the binding interface and experimental measurement indicates minimal effects on the binding affinity. As shown in our energy calculations (Table 3.2), no large changes were computed for all energy terms resulting in minimal binding free energy change predictions, which is consistent with experiment.

- Predictions inconsistent with experimental data:

The first case is the mutation K54A in the structure of the Tn916 integrase-DNA complex. (PDB: 1TN9, Figure 3.8C) [297]. Experimental measurement indicated destabilization of binding and our calculation underestimated the binding free energy change by 0.72 kcal/mol (Table 3.2). In the wild-type structure, K54 is located in a flexible loop and does not directly form H-bond with nearby residue. It is feasible that K54 forms H-bonds in unbound protein or other specific interactions, which would not be captured in our rigid-body protocol and this could be the reason for discrepancy between experiment and modeling. Another case is the single mutation E187A in the complex structure of F Factor TraI Relaxase Domain bound to F oriT Single-stranded DNA (PDB: 2A0I, Figure 3.8D). The experimental data indicates that the mutation destabilizes the binding by 2.12

kcal/mol while the effect is underestimated by SAMPDI. The corresponding reference [298] reporting the structure of protein-DNA complex indicates that there is significant uncertainty for the position of the Glu187 side chain. It is indicated that such a large free energy change is unexpected as the Glu187 side chain appears to only contact with Thy1 5-methyl with its carboxylate [298]. The SAMPDI is a structure-based approach and thus strongly depends on the accuracy of the experimental structures.

The reasons that in some cases SAMPDI predictions are good or bad, as compared with experimental data stem from various sources. It should be reiterated that the SAMPDI protocol is a structure-based rigid-body approach and the accuracy is expected to be sensitive to the conformational changes upon binding and the resolution of experimental structures. Thus, mutations that do not induce large conformation changes are expected to be predicted with higher accuracy compared with mutations causing significant conformational changes. Another reason could be that the protocol does not take into account some non-specified experimental conditions, as non-reported specific ion binding, proton release/uptake and many others.

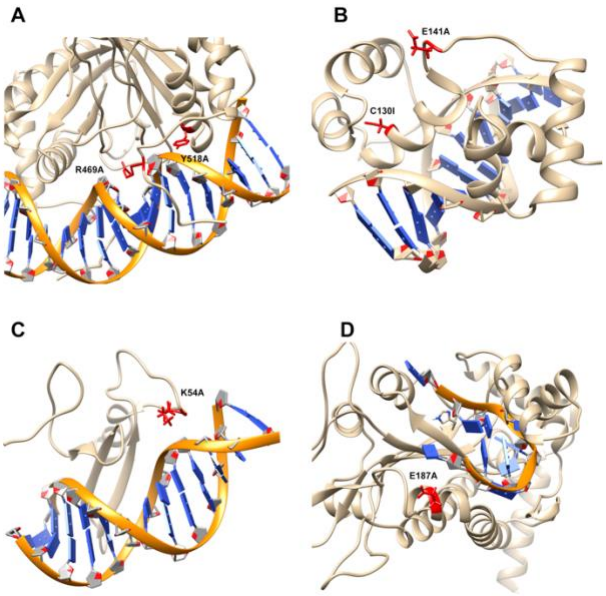


Figure 3.8. Case study of consistent and inconsistent predictions. The backbone of DNA is marked as orange while protein is shown as brown. Mutation site is labeled as red along with the side chain of the wild-type residue. (A) The estrogen receptor DNA-binding domain bound to DNA (PDB: 1HCQ). (B) DNA complex of the Myb DNA-binding domain (PDB: 1MSE). (C) TN916 integrase n-terminal domain/DNA complex (PDB: 1TN9). (D) F Factor TraI Relaxase Domain bound to F oriT Single-stranded DNA (PDB: 2A0I).

4. Implementation

4.1. SAMPDI Webserver architecture

The design of SAMPDI webserver consists of three components: the user interface, the local server and the job backend (The flowchart is shown in Figure 3.9A). The user interface is implemented using the HTML (http://compbio.clemson.edu/SAMPDI/), which provides users with a webpage interface to upload all required input files and fill in

parameters for the free energy calculations. In the webpage, users are firstly asked to upload an input PDB file from a local computer. In addition, the job parameters including chain ID, mutation position, original amino acid and mutated amino acid are provided by the users. Detailed descriptions of all the input parameters are provided as tooltips. Once the job is submitted, users are provided with an URL link to the result page, which will automatically refresh itself every 30s to return the latest results from the backend. The local server part is run on a light-duty computer server, which obtains the PDB files and parameters from the user interface. All the jobs in the backend are executed on the Clemson University Palmetto Cluster. The jobs are executed using multiple nodes with MPI parallel runs to attain the capability for large-scale analysis. Large arrays of independent jobs are permitted to be submitted to the server and are sequentially executed on the Palmetto cluster according to the order of submission.

4.2. Webserver performance

To verify the capability of the SAMPDI server for large-scale analysis, we tested the execution time for different sizes of the proteins ranging from tens of residues up to more than 1000. The execution time linearly increases with the size of proteins (Figure 3.9B). For proteins with less than 200 residues, the results are returned to users within ten minutes. Execution time for middle size proteins is about 20 to 30 minutes and reaches maximum of an hour for large proteins with about and more than 1300 residues.

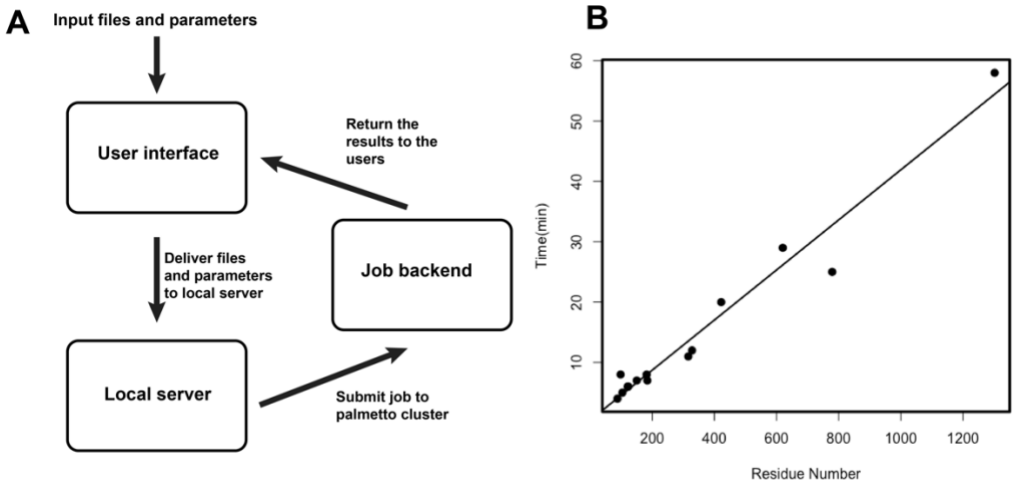


Figure 3.9. (A) Work flowchart of SAMPDI webserver. (B) Performance of SAMPDI webserver showing the execution time for different size proteins.

5. Discussion

Development of computational approaches for large-scale predictions of effect of mutations on macromolecular binding is not a trivial problem [30, 151]. There are multiple available tools and servers for predicting protein-protein binding affinity changes upon single mutations [30, 148, 149, 151, 210, 299]. However, there is still lack of resources for predicting affinity changes of protein-DNA complexes. Currently, the only available method capable of quantitatively predicting binding affinity changes upon single mutation of protein-DNA binding, is the mCSM method [151] and its recent improved version mCSM-NA [285]. The mCSM was benchmarked against the ProNIT database [32] and was reported to result in a correlation coefficient of 0.673. However, the benchmarking was done on the entire ProNIT database without taking into consideration that in ProNIT database (a) some proteins interact with DNA as dimers and mutations could indirectly affect the binding by altering the quaternary structure of the corresponding protein dimer

instead of altering the binding; (b) in some cases, the binding affinity energy change upon mutations was experimentally measured using DNA which does not match the sequence of DNA in ProNIT database. This may indicate that mCSM is not very sensitive to the DNA sequence and may be over fitted, and thus alternative resources are needed. Its recent improved version, mCSM-NA method, enhanced the original method by including a pharmacophore modelling and information of nucleic acid properties into graph-based signatures and benchmarked against the new release of ProNIT database , achieving improved coefficient of 0.70 [285].

In this work, we developed a new approach named SAMPDI, and benchmarked it against purged experimental data from the latest verison of ProNIT database and data from recent references. Comparing with existing mCSM and mCSM-NA approach, SAMPDI shows improvements in the accuracy and result in an improved correlation coefficient 0.72. The SAMPDI method was implemented in a user-friendly webservers, which shows good performance and capability for large-scale analysis.

The SAMPDI applies the so-called rigid body approach, which is based on the assumption that the structures do not undergo conformational changes upon binding. It should be mentioned that in the development of the SAMPDI method we also tested a scenario such that, complex protein-DNA structure and unbound monomeric structures were separately minimized to take into account plausible structural changes induced by the binding, the final results indeed were worse and thus the rigid body approach was applied. In the standard MM/PBSA approach, long time-consuming MD simulations are required to explore the conformational space and intensive sampling of the entire conformation space

is still very challenging. The SAMPDI approach is a trade-off between extensive conformational sampling and execution time since one of the main goals of the SAMPDI method is to allow for large-scale analysis. Future expansion of the method could include fast conformation sampling method for protein and DNA to improve the accuracy of prediction.

CHAPTER FOUR

RESCUING THE R133C RETT SYNDROME CAUSING MUTATION BY SMALL MOLECULE BINDING

1. Introduction

Rett Syndrome (RTT) is another severe neurodevelopmental disease manifested by loss of hand skills, impaired mobility and speech, and development of stereotypical hand movement [300, 301]. RTT exclusively develops in females, affecting one in 10,000 to 15,000 females with 50,000 RTT patients worldwide and no treatment is available now [302]. It was clinically demonstrated that vast majority of RTT cases are caused by mutations in MeCP2 gene [301, 303]. Particularly, the mutations in MeCP2 methyl-CpG-binding (MBD) domain, which specifically binds to a methyl-CpG dinucleotide pair in DNA, were shown by us to affect MBD stability and interactions with DNA [46, 304] (Figure 4.1). Many of the disease-associated mutations were shown to defect the MBD-DNA binding [46] thus modulation of the protein-nucleic interaction could be a promising approach to seek treatment for RTT.

The interactions in biomacromolecule can be modulated via binding of small molecules [164, 305]. Such approaches can be either inhibition [306, 307] or stabilization of the interaction [308, 309]. In this chapter, we focused on rescuing one of the most frequently occurring mutation, R133C, demonstrated both computationally and experimentally that the mutation affects only MBD-DNA interactions[46, 304]. Thus, we are to seek a stabilizer which binding at the periphery of the MBD-DNA interface restores wild-type binding. Structural based virtual screening was applied to screen a large database of compounds to identify potential drug-like compounds. The crystal structure of MeCP2

MBD domain in complex with methylated DNA was used for the structurally-based screening [310]. However, proteins are well known to have highly flexibility and to adopt different conformational states. Such conformational changes should also be taken account into our screening. Thus, molecular dynamics simulations were performed to study the dynamics of the structures and generate the best representative structure for the docking. Three docking programs: Autodock4 [311], Autodock Vina [312] and Dock6 [313] were utilized to dock library of compounds to the structure. We analyzed and compared the docking results and eventually selected the common compound ranked in the top list of the different programs. Lastly, these selected potential compounds are to be subjected to experimental test in the future.

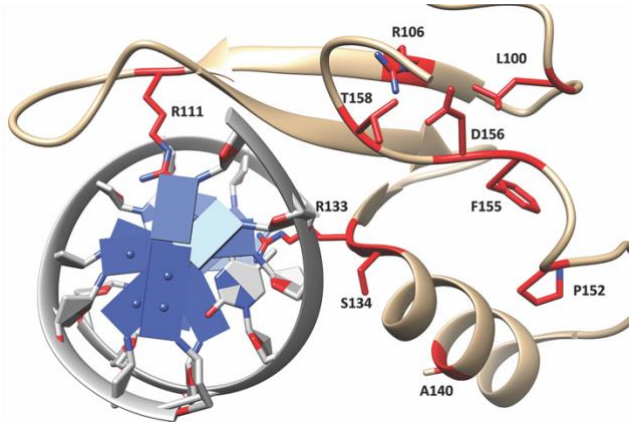


Figure 4.1. Structure of MeCP2 MBD domain bound to DNA. RTT mutations are shown and marked.

2. Method and materials

2.1 Molecular Dynamics simulation

The crystal structure of MeCP2 MBD domain bound to methylated DNA (PDB:3C2i) [310] was used for the sampling of the protein conformations. MD simulations were performed with NAMD 2.11 [95] with Charmm36 force field [290]. The parameter files were prepared with VMD psfgen plugin [138]. Proteins were solvated with 0.15M NaCl in cubic water box with at least 10 Å from the protein to the edge of box. Langevin dynamics with periodic boundary conditions were applied in the simulation. VDW and electrostatic interactions were truncated at 12 Å with a switching function from 10 Å. Particle Mesh Ewald (PME) was applied for long-range electrostatic interaction calculations. First, the system underwent a 5000-step minimization with a fixed backbone, and then a subsequent 5000-step minimization without constraint. Then, all atoms in the protein were fixed for 100 ps equilibration of the water. Harmonic constraint of 1 kcal·mol-1·Å-2 was applied to the protein alpha carbon atoms (CA), and the system was then gradually heated from 0K to 310K with 1000-step/K in the NVT simulation. The system was maintained at 310K for 1ns equilibration with CA constraints and another 2ns equilibration without constraints in NVT system. Finally, the system was switched to an NPT simulation and all constraints were removed for the 10 ns production run.

2.2 Preparation of the compound library for virtual screening

The diversity and size of compound library was critical for the virtual screening. To provide a high-quality library for virtual and vitro screening, we constructed a combined diverse

library using three large commercial libraries: Chembridge library (1159428 compounds), Chemdiv library (1638618 compounds) and LifeChemicals (544924 compunds). The merged compounds library was firstly subjected to the filtering to remove the molecules with undesired physico-chemical properties for a potential drug. We utilized FAF-Drgus4 server [314] for the compounds filtering with the Drug like soft filter, derived from published drug's desired physico-chemical properties [315-319]. The details for the fileting ranges are: 100 < molecular weight < 600, -3 < LogP < 6, HBA (hydrogen bonds acceptors) <= 20, HBD (hydrogen bonds donors) <= 7, tPSA (topological Polar Surface Area) <= 180, Rotatable bonds < 11, Rings <=6. Eventually, 1.34 million compounds remain for further analysis after the physico-chemical properties filtering.

To further enhance the compounds diversity in the merged library, the remaining compounds were further subjected to the clustering using Accelrys Pipeline Pilot [320] with the FCFP-4 fingerprint using similarity cut-off 0.7 and average cluster member size 5. The clustering eventually leads to 0.31 million compounds with highly diversity. Lastly, we generated the 3D structures for the compounds using Corina [320] and further protonated at pH=7 using the ChemAxon.

2.3 Structure-based Virtual Screening

Three popular docking programs were used for the virtual screening of the constructed diverse compound library.

Autodock Vina and Audock4:

Both Autodock Vina and Autodock4 reads PDBQT files as input for docking. We used Autodock tools to set up our systems and prepared the input files for both protein and

compounds. The sample parameter files used for Autodock Vina and Audock4 are shown in appendix Figure B-22 and Figure B-23.

Dock6:

Dock6 reads mol2 files of receptor and ligand for docking. We prepares the receptor mol2 docking files using UCSF Chimera [82] and the input mol2 files for compounds are generated using ChemAxon. The receptor active sites for docking calculation are represented by a subset of spheres within the previous selected druggable pockets. We used grid scores and gbsa hawkins score for primary and secondary ranking of the docking results. The sample parameters file used for docking is shown in Appendix Figure B-24.

3. Results

3.1 Identification of druggable pocket

Since our goal is to identify potential stabilizers to enhance the protein-DNA binding affinity, the small molecules are expected to bind at the periphery of the interface. Thus, we considered two cavities of the interface for the most potential druggable pockets for screening (Figure 4.2). One is located in the major groove of the DNA and in adjacent to the R133C mutation site (Pocket 1). Since the R133C causes loss of two salt bridges with the DNA bases, most promising drug-candidates are expected to form strong interactions to both protein and DNA, like a "clip" to enhance the binding affinity. Another potential druggable pockets are in the DNA minor groove but relatively far away from the mutation site (Pocket 2). Such pocket is a small cavity expected to have small conformational changes induced by the mutation, which was utilized as alternative pocket for our

screening. To fully consider the conformational flexibility in the mutant structure, we perform clustering of the MD trajectory to identify the representative structure. The pockets were clustered via considering both the backbone and the side chain structural differences and the centroid structure from the most protonated cluster was retrieved as the representative structure for further docking analysis.

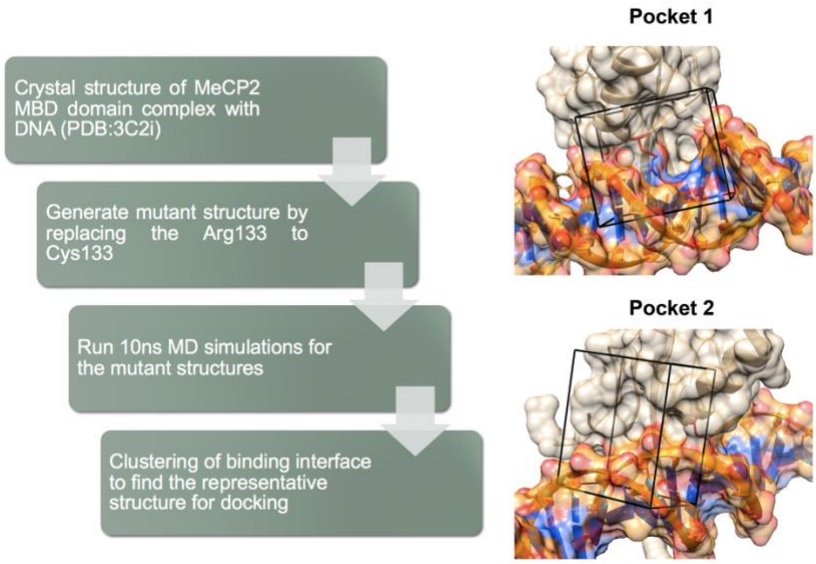


Figure 4.2. Two potential druggable pocket subjected to virtual screening.

3.2 Test of most suitable docking program for screening

The performance of docking programs vary with the targets and it is critical to select the most proper docking program for the screening[321]. Thus, we firstly tested different docking programs' behavior over our systems and choose the best one for the screening over the large compound library. Three docking programs: Autodock4 [311], Autodock Vina [312] and Dock6 [313] were docked into 6000 compounds from the core library of the ChemBridge compound library. Then, the poses are ranked separately with the binding energy from each docking program. To evaluate the results, we took the top ranked 200

compounds from each program and manually compare the poses by considering characters such as charge complement, shape fitness, H-bonding, structural crashes and ligand conformations. Eventually, the overall performance of Autodock4 is best among three docking programs while Autodock Vina ranked least top poses. Thus, we decide to use Autodock4 for the screening of the previous constructed large compound library.

3.3 Virtual Screening

Autodock4 was applied for the docking of the constructed large diverse compound library into the two druggable pockets. For each compound, we retrieve the pose with lowest energy from all clusters. The median binding energy in pocket 1 is -8.08 kcal/mol and relative higher than pocket 2 (-7.44 kcal/mol), as shown in Figure 4.3. Since the R133C directly cause the loss of the H-bonds in the binding interface, compounds which forming H-bonds with both protein and DNA in the interface could have a higher chance to act as a successful stabilizer. Thus, for the first round of selection, we used two constrains: 1) Binding energy <= 8 kcal/mol; 2) Compounds forms at least one H-bond with both protein and DNA. The selection leads to 163528 compounds for pocket 1 and 113654 compounds for pocket 2.

To further reduce the number of candidates for final manual inspection and selection of the poses, we conduct rescoring and ranking using another docking program Dock6 [313]. Since free-energy based approach has been successfully applied in the rescoring of compounds in our previous study [164, 309], we applied Hawkins GB/SA score, an Molecular Mechanics Generalized Born Surface Area (MM/GBSA) approach implemented in Dock6 [313], for the rescoring and reranking. The compounds with

GB/SA score less than -30 kcal/mol are selected, which results in 5543 compounds and 1406 compounds for further analysis.

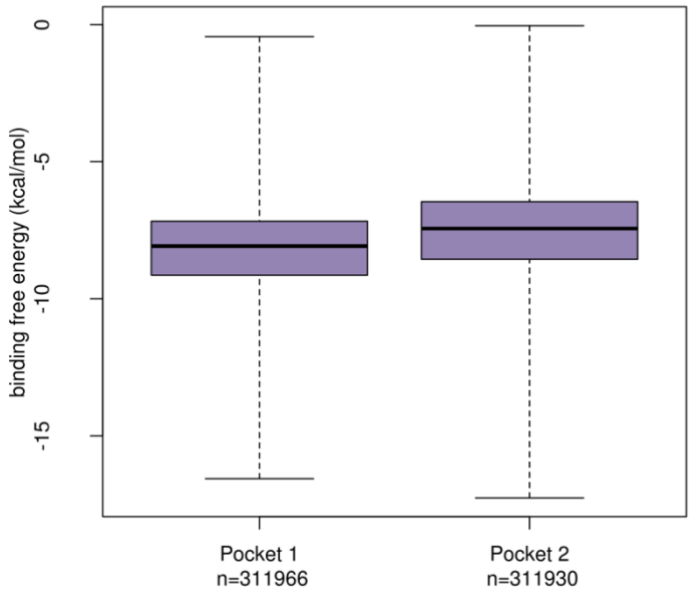


Figure 4.3. The overall binding energy distribution from the virtual screening.

3.4 Manual inspection and selection for the poses

Manual visualization and inspection of poses is very important at the final stage of the virtual screening to identify the best compounds. Thus, we used a visualization tool, PYMOL, to manually select the best poses. The major physico-chemical characteristics which would be critical to the potential stabilizer were fully considered based on our knowledge as shown in Figure 4.4. Besides manual selection, we also considered to eliminate the composes which would act as inhibitor via blocking the protein binding interface. Thus, we docked the selected compounds using a large searching box to include the entire protein and then to remove the compounds which most prefer to bind at the

interface and potentially block the binding (Figure 4.5). Eventually, 80 compounds are selected (shown in Appendix Table C-11) and to be subjected to experimental validation for their effects on the binding affinity.

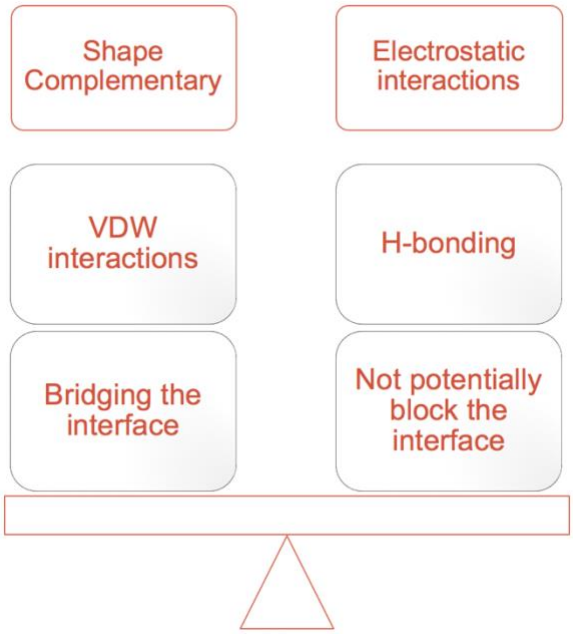


Figure 4.4. The major physico-chemical characteristics considered in manual pose selection.

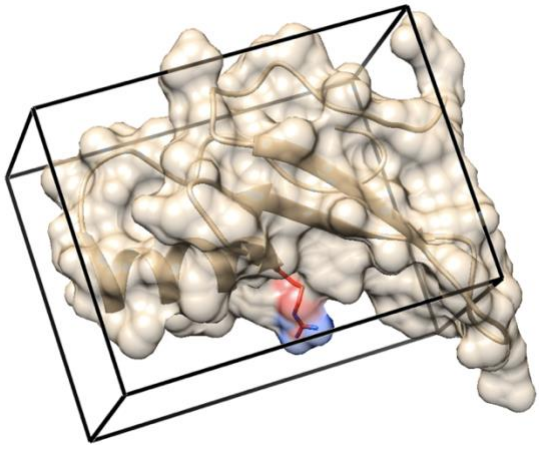


Figure 4.5. The overall binding energy distribution from the virtual screening.

References：

1. Alexov, E., *Advances in Human Biology: Combining Genetics and Molecular Biophysics to Pave the Way for Personalized Diagnostics and Medicine.* Advances in Biology, 2014. **2014**: p. 1-16.
2. Cargill, M., et al., *Characterization of single-nucleotide polymorphisms in coding regions of human genes.* Nat Genet, 1999. **22**(3): p. 231-8.
3. Goldstein, D.B., *Common genetic variation and human traits.* N Engl J Med, 2009. **360**(17): p. 1696-8.
4. Niroula, A. and M. Vihinen, *Classification of Amino Acid Substitutions in Mismatch Repair Proteins Using PON-MMR2.* Hum Mutat, 2015.
5. Suh, Y. and J. Vijg, *SNP discovery in associating genetic variation with human disease phenotypes.* Mutat Res, 2005. **573**(1-2): p. 41-53.
6. Altshuler, D., M.J. Daly, and E.S. Lander, *Genetic mapping in human disease.* Science, 2008. **322**(5903): p. 881-8.
7. Vihinen, M., *Types and effects of protein variations.* Hum Genet, 2015. **134**(4): p. 405-21.
8. Schaafsma, G.C. and M. Vihinen, *VariSNP, a benchmark database for variations from dbSNP.* Hum Mutat, 2015. **36**(2): p. 161-6.
9. Sasidharan Nair, P. and M. Vihinen, *VariBench: a benchmark database for variations.* Hum Mutat, 2013. **34**(1): p. 42-9.
10. Song, C., et al., *Large-scale quantification of single amino-acid variations by a variation-associated database search strategy.* J Proteome Res, 2014. **13**(1): p. 241-8.
11. Kucukkal, T.G. and E. Alexov, *Structural, Dynamical, and Energetical Consequences of Rett Syndrome Mutation R133C in MeCP2.* Comput Math Methods Med, 2015. **2015**: p. 746157.
12. Alexov, E. and M. Sternberg, *Understanding molecular effects of naturally occurring genetic differences.* J Mol Biol, 2013. **425**(21): p. 3911-3.
13. Zhang, Z., et al., *A Y328C missense mutation in spermine synthase causes a mild form of Snyder-Robinson syndrome.* Hum Mol Genet, 2013. **22**(18): p. 3789-97.
14. Casadio, R., et al., *Correlating disease-related mutations to their effect on protein stability: a large-scale analysis of the human proteome.* Hum Mutat, 2011. **32**(10): p. 1161-70.
15. Ramensky, V., *Human non-synonymous SNPs: server and survey.* Nucleic Acids Research, 2002. **30**(17): p. 3894-3900.
16. Niroula, A., S. Urolagin, and M. Vihinen, *PON-P2: prediction method for fast and reliable identification of harmful variants.* PLoS One, 2015. **10**(2): p. e0117380.
17. Vihinen, M., *Proper reporting of predictor performance.* Nat Methods, 2014. **11**(8): p. 781.
18. Ng, P.C. and S. Henikoff, *Predicting the effects of amino acid substitutions on protein function.* Annu Rev Genomics Hum Genet, 2006. **7**: p. 61-80.

19. Kucukkal, T.G., et al., *Computational and experimental approaches to reveal the effects of single nucleotide polymorphisms with respect to disease diagnostics.* Int J Mol Sci, 2014. **15**(6): p. 9670-717.
20. Zhang, Z., et al., *Predicting folding free energy changes upon single point mutations.* Bioinformatics, 2012. **28**(5): p. 664-71.
21. Capriotti, E., R. Calabrese, and R. Casadio, *Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information.* Bioinformatics, 2006. **22**(22): p. 2729-34.
22. Yang, Y., et al., *Structure-based prediction of the effects of a missense variant on protein stability.* Amino Acids, 2013. **44**(3): p. 847-55.
23. Vihinen, M., *How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis.* BMC Genomics, 2012. **13 Suppl 4**: p. S2.
24. Zhang, Z., et al., *Computational analysis of missense mutations causing Snyder-Robinson syndrome.* Hum Mutat, 2010. **31**(9): p. 1043-9.
25. Ferrer-Costa, C., M. Orozco, and X. de la Cruz, *Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties.* J Mol Biol, 2002. **315**(4): p. 771-86.
26. Petukh, M., T.G. Kucukkal, and E. Alexov, *On human disease-causing amino acid variants: statistical study of sequence and structural patterns.* Hum Mutat, 2015. **36**(5): p. 524-34.
27. Guerois, R., J.E. Nielsen, and L. Serrano, *Predicting Changes in the Stability of Proteins and Protein Complexes: A Study of More Than 1000 Mutations.* Journal of Molecular Biology, 2002. **320**(2): p. 369-387.
28. Tokuriki, N. and D.S. Tawfik, *Stability effects of mutations and protein evolvability.* Curr Opin Struct Biol, 2009. **19**(5): p. 596-604.
29. Schreiber, G. and A.R. Fersht, *Energetics of protein-protein interactions: Analysis ofthe Barnase-Barstar interface by single mutations and double mutant cycles.* Journal of Molecular Biology, 1995. **248**(2): p. 478-486.
30. Petukh, M., M. Li, and E. Alexov, *Predicting Binding Free Energy Change Caused by Point Mutations with Knowledge-Modified MM/PBSA Method.* PLoS Comput Biol, 2015. **11**(7): p. e1004276.
31. Moal, I.H. and J. Fernandez-Recio, *SKEMPI: a Structural Kinetic and Energetic database of Mutant Protein Interactions and its use in empirical models.* Bioinformatics, 2012. **28**(20): p. 2600-7.
32. Kumar, M.D., et al., *ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions.* Nucleic Acids Res, 2006. **34**(Database issue): p. D204-6.
33. Peng, Y., et al., *Predicting protein-DNA binding free energy change upon missense mutations using modified MM/PBSA approach: SAMPDI webserver.* Bioinformatics, 2018. **34**(5): p. 779-786.
34. Getov, I., M. Petukh, and E. Alexov, *SAAFEC: Predicting the Effect of Single Point Mutations on Protein Folding Free Energy Using a Knowledge-Modified MM/PBSA Approach.* Int J Mol Sci, 2016. **17**(4): p. 512.

35. Lounnas, V., et al., *Current progress in Structure-Based Rational Drug Design marks a new mindset in drug discovery.* Comput Struct Biotechnol J, 2013. **5**: p. e201302011.
36. Hughes, J.P., et al., *Principles of early drug discovery.* Br J Pharmacol, 2011. **162**(6): p. 1239-49.
37. Michel, J., *Current and emerging opportunities for molecular simulations in structure-based drug design.* Phys Chem Chem Phys, 2014. **16**(10): p. 4465-77.
38. Hung, C.L. and C.C. Chen, *Computational approaches for drug discovery.* Drug Dev Res, 2014. **75**(6): p. 412-8.
39. Sawicki, M.P., et al., *Human Genome Project.* The American Journal of Surgery, 1993. **165**(2): p. 258-264.
40. Genomes Project, C., et al., *A map of human genome variation from population-scale sequencing.* Nature, 2010. **467**(7319): p. 1061-73.
41. Peng, Y., et al., *Revealing the Effects of Missense Mutations Causing Snyder-Robinson Syndrome on the Stability and Dimerization of Spermine Synthase.* Int J Mol Sci, 2016. **17**(1).
42. Li, L., et al., *Forces and Disease: Electrostatic force differences caused by mutations in kinesin motor domains can distinguish between disease-causing and non-disease-causing mutations.* Sci Rep, 2017. **7**(1): p. 8237.
43. Spellicy, C.J., et al., *Key apoptotic genes APAF1 and CASP9 implicated in recurrent folate-resistant neural tube defects.* Eur J Hum Genet, 2018. **26**(3): p. 420-427.
44. Vaidyanathan, K., et al., *Identification and characterization of a missense mutation in the O-linked beta-N-acetylglucosamine (O-GlcNAc) transferase gene that segregates with X-linked intellectual disability.* J Biol Chem, 2017. **292**(21): p. 8948-8963.
45. Chen, W.-T., et al., *Amyloid-beta (Aβ) D7H mutation increases oligomeric Aβ42 and alters properties of Aβ-zinc/copper assemblies.* PloS One, 2012. **7**(4): p. e35807.
46. Yang, Y., et al., *Binding Analysis of Methyl-CpG Binding Domain of MeCP2 and Rett Syndrome Mutations.* ACS Chem Biol, 2016. **11**(10): p. 2706-2715.
47. Peng, Y., et al., *Computational Investigation of the Missense Mutations in DHCR7 Gene Associated with Smith-Lemli-Opitz Syndrome.* International Journal of Molecular Sciences, 2018. **19**(1).
48. Peng, Y., et al., *Mutations in the KDM5C ARID Domain and Their Plausible Association with Syndromic Claes-Jensen-Type Disease.* Int J Mol Sci, 2015. **16**(11): p. 27270-87.
49. Ferreira, L.G., et al., *Molecular docking and structure-based drug design strategies.* Molecules, 2015. **20**(7): p. 13384-421.
50. Bleicher, K.H., et al., *Hit and lead generation: beyond high-throughput screening.* Nat Rev Drug Discov, 2003. **2**(5): p. 369-78.
51. Acharya, C., et al., *Recent Advances in Ligand-Based Drug Design: Relevance and Utility of the Conformationally Sampled Pharmacophore Approach.* Current Computer Aided-Drug Design, 2011. **7**(1): p. 10-22.

52. Wilson, G.L. and M.A. Lill, *Integrating structure-based and ligand-based approaches for computational drug design.* Future Med Chem, 2011. **3**(6): p. 735-50.
53. Drwal, M.N. and R. Griffith, *Combination of ligand- and structure-based methods in virtual screening.* Drug Discov Today Technol, 2013. **10**(3): p. e395-401.
54. Strahl, B.D. and C.D. Allis, *The language of covalent histone modifications.* Nature, 2000. **403**(6765): p. 41-5.
55. Jenuwein, T. and C.D. Allis, *Translating the histone code.* Science, 2001. **293**(5532): p. 1074-80.
56. Martin, C. and Y. Zhang, *The diverse functions of histone lysine methylation.* Nat Rev Mol Cell Biol, 2005. **6**(11): p. 838-49.
57. Blair, L.P., et al., *Epigenetic Regulation by Lysine Demethylase 5 (KDM5) Enzymes in Cancer.* Cancers (Basel), 2011. **3**(1): p. 1383-404.
58. Benevolenskaya, E.V., *Histone H3K4 demethylases are essential in development and differentiation.* Biochem Cell Biol, 2007. **85**(4): p. 435-43.
59. Iwase, S., et al., *The X-linked mental retardation gene SMCX/JARID1C defines a family of histone H3 lysine 4 demethylases.* Cell, 2007. **128**(6): p. 1077-88.
60. Outchkourov, N.S., et al., *Balancing of histone H3K4 methylation states by the Kdm5c/SMCX histone demethylase modulates promoter and enhancer function.* Cell Rep, 2013. **3**(4): p. 1071-9.
61. Grafodatskaya, D., et al., *Multilocus loss of DNA methylation in individuals with mutations in the histone H3 lysine 4 demethylase KDM5C.* BMC Med Genomics, 2013. **6**: p. 1.
62. Patsialou, A., D. Wilsker, and E. Moran, *DNA-binding properties of ARID family proteins.* Nucleic Acids Res, 2005. **33**(1): p. 66-80.
63. Wilsker, D., et al., *Nomenclature of the ARID family of DNA-binding proteins.* Genomics, 2005. **86**(2): p. 242-51.
64. Huang, F., et al., *The JmjN domain of Jhd2 is important for its protein stability, and the plant homeodomain (PHD) finger mediates its chromatin association independent of H3K4 methylation.* J Biol Chem, 2010. **285**(32): p. 24548-61.
65. Mellor, J., *It takes a PHD to read the histone code.* Cell, 2006. **126**(1): p. 22-4.
66. Tzschach, A., et al., *Novel JARID1C/SMCX mutations in patients with X-linked mental retardation.* Hum Mutat, 2006. **27**(4): p. 389.
67. Goncalves, T.F., et al., *KDM5C mutational screening among males with intellectual disability suggestive of X-Linked inheritance and review of the literature.* Eur J Med Genet, 2014. **57**(4): p. 138-44.
68. By Robert L. Schalock, S.A.B.-D., Valerie J. Bradley, Wil H.E. Buntinx, David L. Coulter, Ellis M. (Pat) Craig, ; Sharon C. Gomez, Yves Lachapelle, Ruth Luckasson, Alya Reeve, Karrie A. Shogren, Martha E. Snell, Scott Spreat, Marc J. Tassé, James R. Thompson, Miguel A. Verdugo-Alonso, Michael L. Wehmeyer, and Mark H. Yeager *Intellectual Disability: Definition, Classification, and Systems of Supports* 2010.
69. Kaufman, L., M. Ayub, and J.B. Vincent, *The genetic basis of non-syndromic intellectual disability: a review.* J Neurodev Disord, 2010. **2**(4): p. 182-209.

70. Jensen, L.R., et al., *Mutations in the JARID1C gene, which is involved in transcriptional regulation and chromatin remodeling, cause X-linked mental retardation.* Am J Hum Genet, 2005. **76**(2): p. 227-36.
71. Tahiliani, M., et al., *The histone H3K4 demethylase SMCX links REST target genes to X-linked mental retardation.* Nature, 2007. **447**(7144): p. 601-5.
72. Abidi, F.E., et al., *Mutations in JARID1C are associated with X-linked mental retardation, short stature and hyperreflexia.* J Med Genet, 2008. **45**(12): p. 787-93.
73. Sherry, S.T., et al., *dbSNP: the NCBI database of genetic variation.* Nucleic Acids Res, 2001. **29**(1): p. 308-11.
74. Bullock, A.N., et al., *Thermodynamic stability of wild-type and mutant p53 core domain.* Proceedings of the National Academy of Sciences of the United States of America, 1997. **94**(26): p. 14338-14342.
75. Tan, K.P., et al., *TSpred: a web server for the rational design of temperature-sensitive mutants.* Nucleic Acids Research, 2014. **42**(Web Server issue): p. W277-W284.
76. Schaefer, C., et al., *Disease-related mutations predicted to impact protein function.* BMC Genomics, 2012. **13 Suppl 4**: p. S11.
77. Petukh, M., T.G. Kucukkal, and E. Alexov, *On human disease-causing amino acid variants: statistical study of sequence and structural patterns.* Hum Mutat, 2015.
78. Kucukkal, T.G., et al., *Structural and physico-chemical effects of disease and non-disease nsSNPs on proteins.* Curr Opin Struct Biol, 2015. **32**: p. 18-24.
79. Schuster-Bockler, B. and A. Bateman, *Protein interactions in human genetic diseases.* Genome Biol, 2008. **9**(1): p. R9.
80. Torkamani, A. and N.J. Schork, *Distribution analysis of nonsynonymous polymorphisms within the human kinase gene family.* Genomics, 2007. **90**(1): p. 49-58.
81. Humphrey, W., A. Dalke, and K. Schulten, *VMD: visual molecular dynamics.* J Mol Graph, 1996. **14**(1): p. 33-8, 27-8.
82. Pettersen, E.F., et al., *UCSF Chimera--a visualization system for exploratory research and analysis.* J Comput Chem, 2004. **25**(13): p. 1605-12.
83. Li, C., et al., *Continuous development of schemes for parallel computing of the electrostatics in biological systems: implementation in DelPhi.* J Comput Chem, 2013. **34**(22): p. 1949-60.
84. Li, L., et al., *DelPhi: a comprehensive suite for DelPhi software and associated resources.* BMC Biophys, 2012. **5**: p. 9.
85. Subhra Sarkar, S.W., Jie Zhang, Maxim Zhenirovskyy, Walter Rocchia, and Emil Alexov, *DelPhi Web Server: A comprehensive online suite for electrostatic calculations of biological macromolecules and their complexes.* Commun Comput Phys, 2013 January: p. 13(1): 269–284.
86. Tu, S., et al., *The ARID domain of the H3K4 demethylase RBP2 binds to a DNA CCGCCC motif.* Nat Struct Mol Biol, 2008. **15**(4): p. 419-21.
87. Shoemaker, B.A., et al., *Inferred Biomolecular Interaction Server--a web server to analyze and predict protein interacting partners and binding sites.* Nucleic Acids Res, 2010. **38**(Database issue): p. D518-24.

88. van Stokkum, I.H., et al., *Estimation of protein secondary structure and error analysis from circular dichroism spectra.* Anal Biochem, 1990. **191**(1): p. 110-8.
89. Whitmore, L. and B.A. Wallace, *DICHROWEB, an online server for protein secondary structure analyses from circular dichroism spectroscopic data.* Nucleic Acids Res, 2004. **32**(Web Server issue): p. W668-73.
90. Koehler, C., et al., *Backbone and sidechain 1H, 13C and 15N resonance assignments of the Bright/ARID domain from the human JARID1C (SMCX) protein.* Biomol NMR Assign, 2008. **2**(1): p. 9-11.
91. Berman, H.M., *The Protein Data Bank.* Nucleic Acids Research, 2000. **28**(1): p. 235-242.
92. Iwahara, J., et al., *The structure of the Dead ringer-DNA complex reveals how AT-rich interaction domains (ARIDs) recognize DNA.* EMBO J, 2002. **21**(5): p. 1197-209.
93. Lu, N. and D.A. Kofke, *Accuracy of free-energy perturbation calculations in molecular simulation. I. Modeling.* The Journal of Chemical Physics, 2001. **114**(17): p. 7303.
94. Jorgensen, W.L. and L.L. Thomas, *Perspective on Free-Energy Perturbation Calculations for Chemical Equilibria.* J Chem Theory Comput, 2008. **4**(6): p. 869-876.
95. Phillips, J.C., et al., *Scalable molecular dynamics with NAMD.* J Comput Chem, 2005. **26**(16): p. 1781-802.
96. Pearlman, D.A., *A Comparison of Alternative Approaches to Free Energy Calculations.* The Journal of Physical Chemistry, 1994. **98**(5): p. 1487-1493.
97. MacKerell, A.D., et al., *All-atom empirical potential for molecular modeling and dynamics studies of proteins.* J Phys Chem B, 1998. **102**(18): p. 3586-616.
98. Liu, P., et al., *A Toolkit for the Analysis of Free-Energy Perturbation Calculations.* Journal of Chemical Theory and Computation, 2012. **8**(8): p. 2606-2616.
99. Lockwood, S., B. Krishnamoorthy, and P. Ye, *Neighborhood Properties Are Important Determinants of Temperature Sensitive Mutations.* Plos One, 2011. **6**(12).
100. Ofiteru, A., et al., *Structural and functional consequences of single amino acid substitutions in the pyrimidine base binding pocket of Escherichia coli CMP kinase.* FEBS J, 2007. **274**(13): p. 3363-73.
101. Zhang, Z., et al., *In silico and in vitro investigations of the mutability of disease-causing missense mutation sites in spermine synthase.* PLoS One, 2011. **6**(5): p. e20373.
102. Merz, K.M. and P.A. Kollman, *Free energy perturbation simulations of the inhibition of thermolysin: prediction of the free energy of binding of a new inhibitor.* Journal of the American Chemical Society, 1989. **111**(15): p. 5649-5658.
103. Bjørn O. Brandsdal, F.O.s., Martin Almlöf, Isabella Feierberg, Victor B. Luzhkov and Johan Åqvist*, *Free Energy Calculations and Ligand Binding*, in *Advances in Protein Chemistry*. 2003. p. 123-158.
104. Li, M., et al., *Predicting the Impact of Missense Mutations on Protein-Protein Binding Affinity.* J Chem Theory Comput, 2014. **10**(4): p. 1770-1780.

105. Nishi, H., et al., *Cancer missense mutations alter binding properties of proteins and their interaction networks.* PLoS One, 2013. **8**(6): p. e66273.
106. Giollo, M., et al., *NeEMO: a method using residue interaction networks to improve prediction of protein stability upon mutation.* BMC Genomics, 2014. **15 Suppl 4**: p. S7.
107. Dehouck, Y., et al., *Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0.* Bioinformatics, 2009. **25**(19): p. 2537-43.
108. Capriotti E, F.P., Casadio R *I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure* Nucl Acids Res 2005. **33**: p. W303-W305
109. Pires, D.E., D.B. Ascher, and T.L. Blundell, *DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach.* Nucleic Acids Res, 2014. **42**(Web Server issue): p. W314-9.
110. Parthiban, V., M.M. Gromiha, and D. Schomburg, *CUPSAT: prediction of protein stability upon point mutations.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W239-42.
111. Schymkowitz J1, B.J., Stricher F, Nys R, Rousseau F, Serrano L, *The FoldX web server: an online force field.* Nucleic Acids Res, 2005.
112. Francois Stricher, T.L., Joost Schymkowitz, Frederic Rousseau and Luis Serrano, *FoldX 3.0.* 2007.
113. Celniker, G., et al., *ConSurf: Using Evolutionary Data to Raise Testable Hypotheses about Protein Function.* Israel Journal of Chemistry, 2013. **53**(3-4): p. 199-206.
114. Ashkenazy, H., et al., *ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids.* Nucleic Acids Res, 2010. **38**(Web Server issue): p. W529-33.
115. Landau, M., et al., *ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W299-302.
116. Glaser, F., et al., *ConSurf: Identification of Functional Regions in Proteins by Surface-Mapping of Phylogenetic Information.* Bioinformatics, 2003. **19**(1): p. 163-164.
117. Bonasio, R., S. Tu, and D. Reinberg, *Molecular signals of epigenetic states.* Science, 2010. **330**(6004): p. 612-6.
118. Egger, G., et al., *Epigenetics in human disease and prospects for epigenetic therapy.* Nature, 2004. **429**(6990): p. 457-63.
119. Huang, C., M. Xu, and B. Zhu, *Epigenetic inheritance mediated by histone lysine methylation: maintaining transcriptional states without the precise restoration of marks?* Philos Trans R Soc Lond B Biol Sci, 2013. **368**(1609): p. 20110332.
120. Bannister, A.J. and T. Kouzarides, *Regulation of chromatin by histone modifications.* Cell Res, 2011. **21**(3): p. 381-95.

121. Brookes, E., et al., *Mutations in the intellectual disability gene KDM5C reduce protein stability and demethylase activity.* Hum Mol Genet, 2015. **24**(10): p. 2861-72.
122. Johansson, C., et al., *Structural analysis of human KDM5B guides histone demethylase inhibitor development.* Nat Chem Biol, 2016.
123. Iwase, S., et al., *The X-Linked Mental Retardation GeneSMCX/JARID1C Defines a Family of Histone H3 Lysine 4 Demethylases.* Cell, 2007. **128**(6): p. 1077-88.
124. Altschul, S.F., et al., *Basic local alignment search tool.* Journal of Molecular Biology, 1990. **215**(3): p. 403-410.
125. Kadirvel, S., He, F., Muto, Y., Inoue, M., Kigawa, T., Shirouzu, M., Terada, T., Yokoyama, S., RIKEN Structural Genomics/Proteomics Initiative, *Solution structure of the PHD domain in SmcY protein.*
126. Biasini, M., et al., *SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information.* Nucleic Acids Res, 2014. **42**(Web Server issue): p. W252-8.
127. Yamane, K., et al., *PLU-1 is an H3K4 demethylase involved in transcriptional repression and breast cancer cell proliferation.* Mol Cell, 2007. **25**(6): p. 801-12.
128. Cheng, Z., et al., *A molecular threading mechanism underlies Jumonji lysine demethylase KDM2A regulation of methylated H3K36.* Genes Dev, 2014. **28**(16): p. 1758-71.
129. Wang, W.-C., Chu, C.-H., Chen, C.-C., *Crystal structure of JMJD2B complexed with pyridine-2,4-dicarboxylic acid and H3K9me3.*
130. Krishnan, S. and R.C. Trievel, *Structural and functional analysis of JMJD2D reveals molecular basis for site-specific demethylation among JMJD2 demethylases.* Structure, 2013. **21**(1): p. 98-108.
131. Kruidenier, L., et al., *A selective jumonji H3K27 demethylase inhibitor modulates the proinflammatory macrophage response.* Nature, 2012. **488**(7411): p. 404-8.
132. Yang, Y., et al., *Structural insights into a dual-specificity histone demethylase ceKDM7A from Caenorhabditis elegans.* Cell Res, 2010. **20**(8): p. 886-98.
133. Chen, Z., et al., *Structural basis of the recognition of a methylated histone tail by JMJD2A.* Proc Natl Acad Sci U S A, 2007. **104**(26): p. 10818-23.
134. Dunbrack, R.L., *Rotamer Libraries in the 21st Century.* Current Opinion in Structural Biology, 2002. **12**(4): p. 431-440.
135. Hanwell, M.D., et al., *Avogadro: an advanced semantic chemical editor, visualization, and analysis platform.* J Cheminform, 2012. **4**(1): p. 17.
136. Pierce, B.G., et al., *ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers.* Bioinformatics, 2014. **30**(12): p. 1771-3.
137. Xiang, Z., C.S. Soto, and B. Honig, *Evaluating conformational free energies: the colony energy and its application to the problem of loop prediction.* Proc Natl Acad Sci U S A, 2002. **99**(11): p. 7432-7.
138. Humphrey, W., A. Dalke, and K. Schulten, *VMD: Visual molecular dynamics.* Journal of Molecular Graphics, 1996. **14**(1): p. 33-38.

139. Shi, X., et al., *ING2 PHD domain links histone H3 lysine 4 methylation to active gene repression.* Nature, 2006. **442**(7098): p. 96-9.
140. Maier, J.A., et al., *ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB.* J Chem Theory Comput, 2015. **11**(8): p. 3696-713.
141. Chen, H. and H.X. Zhou, *Prediction of interface residues in protein-protein complexes by a consensus neural network method: test against NMR data.* Proteins, 2005. **61**(1): p. 21-35.
142. Deng, L., et al., *Prediction of protein-protein interaction sites using an ensemble method.* BMC Bioinformatics, 2009. **10**: p. 426.
143. Bairagya, H.R., B.P. Mukhopadhyay, and A.K. Bera, *Role of salt bridge dynamics in inter domain recognition of human IMPDH isoforms: an insight to inhibitor topology for isoform-II.* J Biomol Struct Dyn, 2011. **29**(3): p. 441-62.
144. Gc, J.B., et al., *Interdomain salt-bridges in the Ebola virus protein VP40 and their role in domain association and plasma membrane localization.* Protein Sci, 2016.
145. Zhang, Y., et al., *The PHD1 finger of KDM5B recognizes unmodified H3K4 during the demethylation of histone H3K4me2/3 by KDM5B.* Protein Cell, 2014. **5**(11): p. 837-50.
146. Xu, D. and Y. Zhang, *Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field.* Proteins, 2012. **80**(7): p. 1715-35.
147. Theillet, F.-X., et al., *The alphabet of intrinsic disorder.* Intrinsically Disordered Proteins, 2013. **1**(1): p. e24360.
148. Dehouck, Y., et al., *BeAtMuSiC: Prediction of changes in protein-protein binding affinity on mutations.* Nucleic Acids Res, 2013. **41**(Web Server issue): p. W333-9.
149. Li, M., et al., *MutaBind estimates and interprets the effects of sequence variants on protein-protein interactions.* Nucleic Acids Res, 2016. **44**(W1): p. W494-501.
150. Petukh, M., L. Dai, and E. Alexov, *SAAMBE: Webserver to Predict the Charge of Binding Free Energy Caused by Amino Acids Mutations.* Int J Mol Sci, 2016. **17**(4): p. 547.
151. Pires, D.E., D.B. Ascher, and T.L. Blundell, *mCSM: predicting the effects of mutations in proteins using graph-based signatures.* Bioinformatics, 2014. **30**(3): p. 335-42.
152. Worth, C.L., R. Preissner, and T.L. Blundell, *SDM--a server for predicting effects of mutations on protein stability and malfunction.* Nucleic Acids Res, 2011. **39**(Web Server issue): p. W215-22.
153. Getov, I., M. Petukh, and E. Alexov, *SAAFEC: Predicting the Effect of Single Point Mutations on Protein Folding Free Energy Using a Knowledge-Modified MM/PBSA Approach.* Int J Mol Sci, 2016. **17**(4).
154. Wang, L., L. Li, and E. Alexov, *pKa predictions for proteins, RNAs, and DNAs with the Gaussian dielectric function using DelPhi pKa.* Proteins, 2015. **83**(12): p. 2186-97.
155. Wang, L., M. Zhang, and E. Alexov, *DelPhiPKa web server: predicting pKa of proteins, RNAs and DNAs.* Bioinformatics, 2016. **32**(4): p. 614-5.

156. Rasmussen, P.B. and P. Staller, *The KDM5 family of histone demethylases as targets in oncology drug discovery.* Epigenomics, 2014. **6**(3): p. 277-86.
157. Pegg, A.E., *Mammalian polyamine metabolism and function.* IUBMB Life, 2009. **61**(9): p. 880-94.
158. Thomas, T. and T.J. Thomas, *Polyamines in cell growth and cell death: molecular mechanisms and therapeutic applications.* Cell Mol Life Sci, 2001. **58**(2): p. 244-58.
159. Kusano, T., et al., *Polyamines: essential factors for growth and survival.* Planta, 2008. **228**(3): p. 367-81.
160. Pegg, A.E. and A.J. Michael, *Spermine synthase.* Cell Mol Life Sci, 2010. **67**(1): p. 113-21.
161. Imai, A., et al., *Spermidine synthase genes are essential for survival of Arabidopsis.* Plant Physiol, 2004. **135**(3): p. 1565-73.
162. Peron, A., et al., *Snyder-Robinson syndrome: a novel nonsense mutation in spermine synthase and expansion of the phenotype.* Am J Med Genet A, 2013. **161A**(9): p. 2316-20.
163. Schwartz, C.E., et al., *Spermine synthase deficiency resulting in X-linked intellectual disability (Snyder-Robinson syndrome).* Methods Mol Biol, 2011. **720**: p. 437-45.
164. Zhang, Z., et al., *Rational design of small-molecule stabilizers of spermine synthase dimer by virtual screening and free energy-based approach.* PLoS One, 2014. **9**(10): p. e110884.
165. Wu, H., et al., *Crystal structure of human spermine synthase: implications of substrate binding and catalytic mechanism.* J Biol Chem, 2008. **283**(23): p. 16135-46.
166. Teng, S., et al., *Modeling effects of human single nucleotide polymorphisms on protein-protein interactions.* Biophys J, 2009. **96**(6): p. 2178-88.
167. Barenboim, M., et al., *Statistical geometry based prediction of nonsynonymous SNP functional effects using random forest and neuro-fuzzy classifiers.* Proteins, 2008. **71**(4): p. 1930-9.
168. de Alencastro, G., et al., *New SMS mutation leads to a striking reduction in spermine synthase protein function and a severe form of Snyder-Robinson X-linked recessive mental retardation syndrome.* J Med Genet, 2008. **45**(8): p. 539-43.
169. Lemke, J.R., et al., *Targeted next generation sequencing as a diagnostic tool in epileptic disorders.* Epilepsia, 2012. **53**(8): p. 1387-98.
170. Papadopoulos, J.S. and R. Agarwala, *COBALT: constraint-based alignment tool for multiple protein sequences.* Bioinformatics, 2007. **23**(9): p. 1073-9.
171. UniProt, C., *UniProt: a hub for protein information.* Nucleic Acids Res, 2015. **43**(Database issue): p. D204-12.
172. Smith, D.W., L. Lemli, and J.M. Opitz, *A newly recognized syndromeof multiple congenital anomalies.* The Journal of Pediatrics, 1964. **64**(2): p. 210-217.
173. Battaile, K.P., et al., *Carrier frequency of the common mutation IVS8-1G>C in DHCR7 and estimate of the expected incidence of Smith-Lemli-Opitz syndrome.* Mol Genet Metab, 2001. **72**(1): p. 67-71.

174. Nowaczyk, M.J., J.S. Waye, and J.D. Douketis, *DHCR7 mutation carrier rates and prevalence of the RSH/Smith-Lemli-Opitz syndrome: where are the patients?* Am J Med Genet A, 2006. **140**(19): p. 2057-62.
175. Yu, H., et al., *Detection of a common mutation in the RSH or Smith-Lemli-Opitz syndrome by a PCR-RFLP assay: IVS8-1G ? C is found in over sixty percent of US propositi.* American Journal of Medical Genetics, 2000. **90**(4): p. 347-350.
176. Waterham, H.R. and R.C. Hennekam, *Mutational spectrum of Smith-Lemli-Opitz syndrome.* Am J Med Genet C Semin Med Genet, 2012. **160C**(4): p. 263-84.
177. Kelley, R.I. and G.E. Herman, *Inborn errors of sterol biosynthesis.* Annu Rev Genomics Hum Genet, 2001. **2**: p. 299-341.
178. Kelley, R.I., *The Smith-Lemli-Opitz syndrome.* Journal of Medical Genetics, 2000. **37**(5): p. 321-335.
179. Kelly, M.N., et al., *Brothers with Smith-Lemli-Opitz syndrome.* J Pediatr Health Care, 2015. **29**(1): p. 97-103.
180. Witsch-Baumgartner, M., et al., *Mutational spectrum in the Delta7-sterol reductase gene and genotype-phenotype correlation in 84 patients with Smith-Lemli-Opitz syndrome.* Am J Hum Genet, 2000. **66**(2): p. 402-12.
181. Tint, G.S., et al., *Defective cholesterol biosynthesis associated with the Smith-Lemli-Opitz syndrome.* N Engl J Med, 1994. **330**(2): p. 107-13.
182. Shefer, S., et al., *Markedly inhibited 7-dehydrocholesterol-delta 7-reductase activity in liver microsomes from Smith-Lemli-Opitz homozygotes.* J Clin Invest, 1995. **96**(4): p. 1779-85.
183. Correa-Cerro, L.S. and F.D. Porter, *3beta-hydroxysterol Delta7-reductase and the Smith-Lemli-Opitz syndrome.* Mol Genet Metab, 2005. **84**(2): p. 112-26.
184. Hossein-nezhad, A. and M.F. Holick, *Vitamin D for health: a global perspective.* Mayo Clin Proc, 2013. **88**(7): p. 720-55.
185. Porter, F.D. and G.E. Herman, *Malformation syndromes caused by disorders of cholesterol synthesis.* J Lipid Res, 2011. **52**(1): p. 6-34.
186. Prabhu, A.V., et al., *Cholesterol-mediated Degradation of 7-Dehydrocholesterol Reductase Switches the Balance from Cholesterol to Vitamin D Synthesis.* J Biol Chem, 2016. **291**(16): p. 8363-73.
187. Kuan, V., et al., *DHCR7 mutations linked to higher vitamin D status allowed early human migration to northern latitudes.* BMC Evol Biol, 2013. **13**: p. 144.
188. Moebius, F.F., et al., *Molecular cloning and expression of the human delta7-sterol reductase.* Proc Natl Acad Sci U S A, 1998. **95**(4): p. 1899-902.
189. Fitzky, B.U., et al., *Mutations in the Delta7-sterol reductase gene in patients with the Smith-Lemli-Opitz syndrome.* Proc Natl Acad Sci U S A, 1998. **95**(14): p. 8181-6.
190. Li, X., R. Roberti, and G. Blobel, *Structure of an integral membrane sterol reductase from Methylomicrobium alcaliphilum.* Nature, 2015. **517**(7532): p. 104-7.
191. Kroncke, B.M., et al., *Documentation of an Imperative To Improve Methods for Predicting Membrane Protein Stability.* Biochemistry, 2016. **55**(36): p. 5002-9.

192. Petukh, M., T.G. Kucukkal, and E. Alexov, *On human disease-causing amino acid variants: statistical study of sequence and structural patterns.* Hum Mutat, 2015. **36**(5): p. 524-534.
193. Cubellis, M.V., M. Baaden, and G. Andreotti, *Taming molecular flexibility to tackle rare diseases.* Biochimie, 2015. **113**: p. 54-8.
194. Prabhu, A.V., et al., *DHCR7: A vital enzyme switch between cholesterol and vitamin D production.* Prog Lipid Res, 2016. **64**: p. 138-151.
195. Grant, B.J., et al., *Bio3d: an R package for the comparative analysis of protein structures.* Bioinformatics, 2006. **22**(21): p. 2695-6.
196. Estacio, S.G., E.I. Shakhnovich, and P.F. Faisca, *Assessing the effect of loop mutations in the folding space of beta2-microglobulin with molecular dynamics simulations.* Int J Mol Sci, 2013. **14**(9): p. 17256-78.
197. Witham, S., et al., *A missense mutation in CLIC2 associated with intellectual disability is predicted by in silico modeling to affect protein stability and dynamics.* Proteins, 2011. **79**(8): p. 2444-54.
198. Peng, Y. and E. Alexov, *Investigating the linkage between disease-causing amino acid variants and their effect on protein stability and binding.* Proteins, 2016. **84**(2): p. 232-9.
199. Witsch-Baumgartner, M., et al., *Age and origin of major Smith-Lemli-Opitz syndrome (SLOS) mutations in European populations.* J Med Genet, 2008. **45**(4): p. 200-9.
200. Landrum, M.J., et al., *ClinVar: public archive of interpretations of clinically relevant variants.* Nucleic Acids Res, 2016. **44**(D1): p. D862-8.
201. Lek, M., et al., *Analysis of protein-coding genetic variation in 60,706 humans.* Nature, 2016. **536**(7616): p. 285-91.
202. Romano, F., et al., *A Novel Mutation of the DHCR7 Gene in a Sicilian Compound Heterozygote with Smith-Lemli-Opitz Syndrome.* Molecular Diagnosis, 2012. **9**(4): p. 201-204.
203. Tamura, M., et al., *Novel DHCR7 mutation in a case of Smith-Lemli-Opitz syndrome showing 46,XY disorder of sex development.* Hum Genome Var, 2017. **4**: p. 17015.
204. Webb, B. and A. Sali, *Comparative Protein Structure Modeling Using MODELLER.* Curr Protoc Bioinformatics, 2014. **47**: p. 5 6 1-32.
205. Wimley, W.C. and S.H. White, *Experimentally determined hydrophobicity scale for proteins at membrane interfaces.* Nature Structural Biology, 1996. **3**(10): p. 842-848.
206. Peng, Y., et al., *Predicting protein-DNA binding free energy change upon missense mutations using modified MM/PBSA approach: SAMPDI webserver.* Bioinformatics, 2017.
207. Notredame, C., D.G. Higgins, and J. Heringa, *T-Coffee: A novel method for fast and accurate multiple sequence alignment.* J Mol Biol, 2000. **302**(1): p. 205-17.
208. Yin, S., F. Ding, and N.V. Dokholyan, *Eris: an automated estimator of protein stability.* Nat Methods, 2007. **4**(6): p. 466-7.

209. Topham, C.M., N. Srinivasan, and T.L. Blundell, *Prediction of the stability of protein mutants based on structural environment-dependent amino acid substitution and propensity tables.* Protein Engineering Design and Selection, 1997. **10**(1): p. 7-21.
210. Schymkowitz, J., et al., *The FoldX web server: an online force field.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W382-8.
211. Jo, S., T. Kim, and W. Im, *Automated builder and database of protein/membrane complexes for molecular dynamics simulations.* PLoS One, 2007. **2**(9): p. e880.
212. Lomize, M.A., et al., *OPM: orientations of proteins in membranes database.* Bioinformatics, 2006. **22**(5): p. 623-5.
213. Estacio, S.G., H.F. Martiniano, and P.F. Faisca, *Thermal unfolding simulations of NBD1 domain variants reveal structural motifs associated with the impaired folding of F508del-CFTR.* Mol Biosyst, 2016. **12**(9): p. 2834-48.
214. Varani, G. and K. Nagai, *RNA recognition by RNP proteins during RNA processing.* Annu Rev Biophys Biomol Struct, 1998. **27**: p. 407-45.
215. Lejeune, D., et al., *Protein-nucleic acid recognition: statistical analysis of atomic interactions and influence of DNA structure.* Proteins, 2005. **61**(2): p. 258-71.
216. ML, B., et al., *Quantifying DNA-protein interactions by double-stranded DNA arrays.* Nat Biotechnol, 1999. **17**: p. 573-577.
217. Luscombe, N.M., *Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level.* Nucleic Acids Research, 2001. **29**(13): p. 2860-2874.
218. Rumora, A.E., et al., *Electrostatic and Hydrophobic Interactions Mediate Single-Stranded DNA Recognition and Acta2 Repression by Purine-Rich Element-Binding Protein B.* Biochemistry, 2016. **55**(19): p. 2794-805.
219. K, N., W. S.J, and J. J, *Structural Features of Protein-Nucleic Acid Recognition Sites.* Biochemistry, 1999. **38**: p. 1999-2017.
220. Y, T., R. P.D, and M. C.P, *Thermodynamics of Cro Protein-DNA Interactions.* Proc Natl Acad Sci U S A, 1992. **89(17)**: p. 8180-8184.
221. Bowater, R.P., et al., *Biophysical and electrochemical studies of protein–nucleic acid interactions.* Monatshefte für Chemie - Chemical Monthly, 2015. **146**(5): p. 723-739.
222. Iwakiri, J., et al., *Dissecting the protein-RNA interface: the role of protein surface shapes and RNA secondary structures in protein-RNA recognition.* Nucleic Acids Res, 2012. **40**(8): p. 3299-306.
223. Jones, S., et al., *Protein-DNA interactions: A structural analysis.* J Mol Biol, 1999. **287**(5): p. 877-96.
224. Rohs, R., et al., *The role of DNA shape in protein-DNA recognition.* Nature, 2009. **461**(7268): p. 1248-53.
225. Rohs, R., et al., *Origins of specificity in protein-DNA recognition.* Annu Rev Biochem, 2010. **79**: p. 233-69.
226. West, S.M., et al., *Electrostatic interactions between arginines and the minor groove in the nucleosome.* J Biomol Struct Dyn, 2010. **27**(6): p. 861-6.

227. Auweter, S.D., F.C. Oberstrass, and F.H. Allain, *Sequence-specific binding of single-stranded RNA: is there a code for recognition?* Nucleic Acids Res, 2006. **34**(17): p. 4943-59.
228. Jones, S., *Protein-RNA interactions: a structural analysis.* Nucleic Acids Research, 2001. **29**(4): p. 943-954.
229. Zhang, Z., S. Witham, and E. Alexov, *On the role of electrostatics in protein-protein interactions.* Phys Biol, 2011. **8**(3): p. 035001.
230. Onufriev, A.V. and E. Alexov, *Protonation and pK changes in protein-ligand binding.* Q Rev Biophys, 2013. **46**(2): p. 181-209.
231. Kundrotas, P.J. and E. Alexov, *Electrostatic properties of protein-protein complexes.* Biophys J, 2006. **91**(5): p. 1724-36.
232. Petukh, M., S. Stefl, and E. Alexov, *The role of protonation states in ligand-receptor recognition and binding.* Curr Pharm Des, 2013. **19**(23): p. 4182-90.
233. Garcia-Moreno, B., *Adaptations of proteins to cellular and subcellular pH.* J Biol, 2009. **8**(11): p. 98.
234. Chan, P., J. Lovric, and J. Warwicker, *Subcellular pH and predicted pH-dependent features of proteins.* Proteomics, 2006. **6**(12): p. 3494-501.
235. Talley, K. and E. Alexov, *On the pH-optimum of activity and stability of proteins.* Proteins, 2010. **78**(12): p. 2699-706.
236. Mitra, R.C., Z. Zhang, and E. Alexov, *In silico modeling of pH-optimum of protein-protein binding.* Proteins, 2011. **79**(3): p. 925-36.
237. Alexov, E., *Numerical calculations of the pH of maximal protein stability.* European Journal of Biochemistry, 2003. **271**(1): p. 173-185.
238. Kirsanov, D.D., et al., *NPIDB: Nucleic acid-Protein Interaction DataBase.* Nucleic Acids Res, 2013. **41**(Database issue): p. D517-23.
239. Zanegina, O., et al., *An updated version of NPIDB includes new classifications of DNA-protein complexes and their families.* Nucleic Acids Res, 2016. **44**(D1): p. D144-53.
240. Finn, R.D., et al., *Pfam: the protein families database.* Nucleic Acids Res, 2014. **42**(Database issue): p. D222-30.
241. Andreeva, A., et al., *SCOP database in 2004: refinements integrate structure and sequence family data.* Nucleic Acids Res, 2004. **32**(Database issue): p. D226-9.
242. Alexov, E., *Calculating proton uptake/release and binding free energy taking into account ionization and conformation changes induced by protein-inhibitor association: application to plasmepsin, cathepsin D and endothiapepsin-pepstatin complexes.* Proteins, 2004. **56**(3): p. 572-84.
243. Tang, C.L., et al., *Calculation of pKas in RNA: on the structural origins and functional roles of protonated nucleotides.* J Mol Biol, 2007. **366**(5): p. 1475-96.
244. Nielsen, J.E., M.R. Gunner, and B.E. Garcia-Moreno, *The pKa Cooperative: a collaborative effort to advance structure-based calculations of pKa values and electrostatic effects in proteins.* Proteins, 2011. **79**(12): p. 3249-59.
245. Perez-Canadillas, J.M., et al., *Characterization of pKa values and titration shifts in the cytotoxic ribonuclease alpha-sarcin by NMR. Relationship between*

electrostatic interactions, structure, and catalytic function. Biochemistry, 1998. **37**(45): p. 15865-76.
246. Sharp, K.A., *Electrostatic interactions in hirudin-thrombin binding.* Biophysical Chemistry, 1996. **61**(1): p. 37-49.
247. Richter, H.T., et al., *A linkage of the pKa's of asp-85 and glu-204 forms part of the reprotonation switch of bacteriorhodopsin.* Biochemistry, 1996. **35**(13): p. 4054-62.
248. Schubert, M., et al., *Probing electrostatic interactions along the reaction pathway of a glycoside hydrolase: histidine characterization by NMR spectroscopy.* Biochemistry, 2007. **46**(25): p. 7383-95.
249. Grey, M.J., et al., *Characterizing a partially folded intermediate of the villin headpiece domain under non-denaturing conditions: contribution of His41 to the pH-dependent stability of the N-terminal subdomain.* J Mol Biol, 2006. **355**(5): p. 1078-94.
250. Ishitani, R., et al., *Alternative Tertiary Structure of tRNA for Recognition by a Posttranscriptional Modification Enzyme.* Cell, 2003. **113**(3): p. 383-394.
251. Horton, J.R., et al., *Asp34 of PvuII endonuclease is directly involved in DNA minor groove recognition and indirectly involved in catalysis.* J Mol Biol, 1998. **284**(5): p. 1491-504.
252. Pan, B., Y. Xiong, and T.A. Steitz, *How the CCA-adding enzyme selects adenine over cytosine at position 76 of tRNA.* Science, 2010. **330**(6006): p. 937-40.
253. Nakanishi, K., et al., *Structural basis for translational fidelity ensured by transfer RNA lysidine synthetase.* Nature, 2009. **461**(7267): p. 1144-8.
254. Xu, Q.S., et al., *An asymmetric complex of restriction endonuclease MspI on its palindromic DNA recognition site.* Structure, 2004. **12**(9): p. 1741-7.
255. Jain, D., et al., *Crystal structure of bacteriophage lambda cII and its DNA complex.* Mol Cell, 2005. **19**(2): p. 259-69.
256. Aguilar, B., et al., *Statistics and physical origins of pK and ionization state changes upon protein-ligand binding.* Biophys J, 2010. **98**(5): p. 872-80.
257. Alexov, E., et al., *Progress in the prediction of pKa values in proteins.* Proteins, 2011. **79**(12): p. 3260-75.
258. Petukh, M., M. Zhang, and E. Alexov, *Statistical investigation of surface bound ions and further development of BION server to include pH and salt dependence.* J Comput Chem, 2015. **36**(32): p. 2381-93.
259. Petukh, M., T. Kimmet, and E. Alexov, *BION web server: predicting non-specifically bound surface ions.* Bioinformatics, 2013. **29**(6): p. 805-6.
260. Petukh, M., et al., *Predicting nonspecific ion binding using DelPhi.* Biophys J, 2012. **102**(12): p. 2885-93.
261. Huang, Y., et al., *All-atom Continuous Constant pH Molecular Dynamics With Particle Mesh Ewald and Titratable Water.* J Chem Theory Comput, 2016.
262. Wallace, J.A. and J.K. Shen, *Predicting pKa values with continuous constant pH molecular dynamics.* Methods Enzymol, 2009. **466**: p. 455-75.

263. Wallace, J.A. and J.K. Shen, *Unraveling A Trap-and-Trigger Mechanism in the pH-Sensitive Self-Assembly of Spider Silk Proteins.* J Phys Chem Lett, 2012. **3**(5): p. 658-662.
264. Ellis, C.R., et al., *Constant pH Molecular Dynamics Reveals pH-Modulated Binding of Two Small-Molecule BACE1 Inhibitors.* J Phys Chem Lett, 2016. **7**(6): p. 944-9.
265. Chen, W., Y. Huang, and J. Shen, *Conformational Activation of a Transmembrane Proton Channel from Constant pH Molecular Dynamics.* J Phys Chem Lett, 2016. **7**(19): p. 3961-3966.
266. Orphanides, G. and D. Reinberg, *A Unified Theory of Gene Expression.* Cell, 2002. **108**(4): p. 439-451.
267. Roeder, R.G., *Role of General and Gene-specific Cofactors in the Regulation of Eukaryotic Transcription.* Cold Spring Harbor Symposia on Quantitative Biology, 1998. **63**(0): p. 201-218.
268. Slutsky, M. and L.A. Mirny, *Kinetics of protein-DNA interaction: facilitated target location in sequence-dependent potential.* Biophys J, 2004. **87**(6): p. 4021-35.
269. Peng, Y. and E. Alexov, *Computational investigation of proton transfer, pKa shifts and pH-optimum of protein-DNA and protein-RNA complexes.* Proteins, 2017. **85**(2): p. 282-295.
270. Hogan, M.E. and R.H. Austin, *Importance of DNA stiffness in protein-DNA binding specificity.* Nature, 1987. **329**(6136): p. 263-6.
271. Luscombe, N.M. and J.M. Thornton, *Protein–DNA Interactions: Amino Acid Conservation and the Effects of Mutations on Binding Specificity.* Journal of Molecular Biology, 2002. **320**(5): p. 991-1009.
272. Trelsman, J., et al., *A single amino acid can determine the DNA binding specificity of homeodomain proteins.* Cell, 1989. **59**(3): p. 553-562.
273. Vousden, K.H. and D.P. Lane, *p53 in health and disease.* Nat Rev Mol Cell Biol, 2007. **8**(4): p. 275-83.
274. Garg, V., et al., *Mutations in NOTCH1 cause aortic valve disease.* Nature, 2005. **437**(7056): p. 270-4.
275. Chahrour, M., et al., *MeCP2, a key contributor to neurological disease, activates and represses transcription.* Science, 2008. **320**(5880): p. 1224-9.
276. Velazquez-Campoy, A., et al., *Isothermal titration calorimetry.* Curr Protoc Cell Biol, 2004. **Chapter 17**: p. Unit 17 8.
277. Hillisch, A., M. Lorenz, and S. Diekmann, *Recent advances in FRET: distance determination in protein–DNA complexes.* Current Opinion in Structural Biology, 2001. **11**(2): p. 201-207.
278. Campagne, S., V. Gervais, and A. Milon, *Nuclear magnetic resonance analysis of protein-DNA interactions.* J R Soc Interface, 2011. **8**(61): p. 1065-78.
279. Teh, H.F., et al., *Characterization of protein--DNA interactions using surface plasmon resonance spectroscopy with various assay schemes.* Biochemistry, 2007. **46**(8): p. 2127-35.
280. Jones, S., *Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins.* Nucleic Acids Research, 2003. **31**(24): p. 7189-7198.

281. Liu, Z., et al., *Quantitative evaluation of protein-DNA interactions using an optimized knowledge-based potential.* Nucleic Acids Res, 2005. **33**(2): p. 546-58.
282. Morozov, A.V., et al., *Protein-DNA binding specificity predictions with structural models.* Nucleic Acids Res, 2005. **33**(18): p. 5781-98.
283. Donald, J.E., W.W. Chen, and E.I. Shakhnovich, *Energetics of protein-DNA interactions.* Nucleic Acids Res, 2007. **35**(4): p. 1039-47.
284. Zhang, C., et al., *A knowledge-based energy function for protein-ligand, protein-protein, and protein-DNA complexes.* J Med Chem, 2005. **48**(7): p. 2325-35.
285. Pires, D.E. and D.B. Ascher, *mCSM-NA: predicting the effects of mutations on protein-nucleic acids interactions.* Nucleic Acids Res, 2017.
286. Hou, T., et al., *Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations.* J Chem Inf Model, 2011. **51**(1): p. 69-82.
287. Hou, T., et al., *Assessing the performance of the molecular mechanics/Poisson Boltzmann surface area and molecular mechanics/generalized Born surface area methods. II. The accuracy of ranking poses generated from docking.* J Comput Chem, 2011. **32**(5): p. 866-77.
288. Lee, M.R., Y. Duan, and P.A. Kollman, *Use of MM-PB/SA in estimating the free energies of proteins: Application to native, intermediates, and unfolded villin headpiece.* Proteins-Structure Function and Genetics, 2000. **39**(4): p. 309-316.
289. Rose, P.W., et al., *The RCSB Protein Data Bank: views of structural biology for basic and applied research and education.* Nucleic Acids Res, 2015. **43**(Database issue): p. D345-56.
290. Best, R.B., et al., *Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone phi, psi and side-chain chi(1) and chi(2) dihedral angles.* J Chem Theory Comput, 2012. **8**(9): p. 3257-3273.
291. Denning, E.J., et al., *Impact of 2'-hydroxyl sampling on the conformational properties of RNA: update of the CHARMM all-atom additive force field for RNA.* J Comput Chem, 2011. **32**(9): p. 1929-43.
292. Li, L., C. Li, and E. Alexov, *On the Modeling of Polar Component of Solvation Energy using Smooth Gaussian-Based Dielectric Function.* J Theor Comput Chem, 2014. **13**(3).
293. Jia, Z., et al., *Treating ion distribution with Gaussian-based smooth dielectric function in DelPhi.* J Comput Chem, 2017. **38**(22): p. 1974-1979.
294. S. Hubbard, J.M.T., *'NACCESS', Computer Program.* Department of Biochemistry and Molecular Biology, University College London., 1993.
295. Bochkarev, A., et al., *The 2.2 A structure of a permanganate-sensitive DNA site bound by the Epstein-Barr virus origin binding protein, EBNA1.* J Mol Biol, 1998. **284**(5): p. 1273-8.
296. Ogata, K., et al., *Solution structure of a specific DNA complex of the Myb DNA-binding domain with cooperative recognition helices.* Cell, 1994. **79**(4): p. 639-648.
297. Wojciak, J.M., K.M. Connolly, and R.T. Clubb, *NMR structure of the Tn916 integrase-DNA complex.* Nat Struct Biol, 1999. **6**(4): p. 366-73.

298. Larkin, C., et al., *Inter- and intramolecular determinants of the specificity of single-stranded DNA binding and cleavage by the F factor relaxase.* Structure, 2005. **13**(10): p. 1533-44.
299. Brender, J.R. and Y. Zhang, *Predicting the Effect of Mutations on Protein-Protein Binding Interactions through Structure-Based Interface Profiles.* PLoS Comput Biol, 2015. **11**(10): p. e1004494.
300. Han, Z.A., et al., *Clinical characteristics of children with rett syndrome.* Ann Rehabil Med, 2012. **36**(3): p. 334-9.
301. Percy, A.K., *Rett syndrome: exploring the autism link.* Arch Neurol, 2011. **68**(8): p. 985-9.
302. Laurvick, C.L., et al., *Rett syndrome in Australia: a review of the epidemiology.* J Pediatr, 2006. **148**(3): p. 347-52.
303. Amir, R.E., et al., *Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2.* Nat Genet, 1999. **23**(2): p. 185-8.
304. Kucukkal, T.G., et al., *Impact of Rett Syndrome Mutations on MeCP2 MBD Stability.* Biochemistry, 2015. **54**(41): p. 6357-68.
305. Hollenberg, P.F., *Characteristics and common properties of inhibitors, inducers, and activators of CYP enzymes.* Drug Metab Rev, 2002. **34**(1-2): p. 17-35.
306. Rendic, S. and F.J.D. Carlo, *Human Cytochrome P450 Enzymes: A Status Report Summarizing Their Reactions, Substrates, Inducers, and Inhibitors.* Drug Metabolism Reviews, 2010. **29**(1-2): p. 413-580.
307. Tsoutsikos, P., et al., *Evidence that unsaturated fatty acids are potent inhibitors of renal UDP-glucuronosyltransferases (UGT): kinetic studies using human kidney cortical microsomes and recombinant UGT1A9 and UGT2B7.* Biochemical Pharmacology, 2004. **67**(1): p. 191-199.
308. Liu, X., et al., *Small molecule induced reactivation of mutant p53 in cancer cells.* Nucleic Acids Res, 2013. **41**(12): p. 6034-44.
309. Zhang, Z., et al., *A rational free energy-based approach to understanding and targeting disease-causing missense mutations.* J Am Med Inform Assoc, 2013. **20**(4): p. 643-51.
310. Ho, K.L., et al., *MeCP2 binding to DNA depends upon hydration at methyl-CpG.* Mol Cell, 2008. **29**(4): p. 525-31.
311. Morris, G.M., et al., *AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility.* J Comput Chem, 2009. **30**(16): p. 2785-91.
312. Trott, O. and A.J. Olson, *AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading.* J Comput Chem, 2010. **31**(2): p. 455-61.
313. Lang, P.T., et al., *DOCK 6: combining techniques to model RNA-small molecule complexes.* RNA, 2009. **15**(6): p. 1219-30.
314. Lagorce, D., et al., *FAF-Drugs4: free ADME-tox filtering computations for chemical biology and early stages drug discovery.* Bioinformatics, 2017. **33**(22): p. 3658-3660.
315. Lipinski, C.A., et al., *Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings 1PII of*

original article: S0169-409X(96)00423-1. The article was originally published in Advanced Drug Delivery Reviews 23 (1997) 3–25. 1. Advanced Drug Delivery Reviews, 2001. **46**(1-3): p. 3-26.
316. Oprea, T.I., *Property distribution of drug-related chemical databases.* Journal of Computer-Aided Molecular Design, 2000. **14**(3): p. 251-264.
317. Irwin, J.J. and B.K. Shoichet, *ZINC--a free database of commercially available compounds for virtual screening.* J Chem Inf Model, 2005. **45**(1): p. 177-82.
318. Oprea, T.I., et al., *Is There a Difference between Leads and Drugs? A Historical Perspective.* Journal of Chemical Information and Computer Sciences, 2001. **41**(5): p. 1308-1315.
319. Pihan, E., et al., *e-Drug3D: 3D structure collections dedicated to drug repurposing and fragment-based drug design.* Bioinformatics, 2012. **28**(11): p. 1540-1.
320. Warr, W.A., *Scientific workflow systems: Pipeline Pilot and KNIME.* J Comput Aided Mol Des, 2012. **26**(7): p. 801-4.
321. Warren, G.L., et al., *A critical assessment of docking programs and scoring functions.* J Med Chem, 2006. **49**(20): p. 5912-31.
322. Shapovalov, M.V. and R.L. Dunbrack, Jr., *A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions.* Structure, 2011. **19**(6): p. 844-58.

APPENDICES

Appendix A

Publications resulting from the dissertation

CHAPTER 1:

Peng Yunhui, Sankar Basu, Emil Alexov, “Structural perspective on revealing and altering molecular mechanisms of genetic variants linked with diseases”. Future Medicinal Chemistry, under review

Peng, Yunhui, and Emil Alexov. "Investigating the linkage between disease-causing amino acid variants and their effect on protein stability and binding." *Proteins: Structure, Function, and Bioinformatics* 84.2 (2016): 232-239.

Peng, Yunhui, and Emil Alexov. "Protein Conformational Disease: Visit the Facts at a Glance." *eLS* (2001): 1-7.

CHAPTER 2:

Peng, Yunhui, Jimmy Suryadi, Ye Yang, Tugba Kucukkal, Weiguo Cao, and Emil Alexov. "Mutations in the KDM5C ARID domain and their plausible association with syndromic Claes-Jensen-Type disease." *International journal of molecular sciences* 16, no. 11 (2015): 27270-27287.

Peng, Yunhui, and Emil Alexov. "Cofactors-loaded quaternary structure of lysine-specific demethylase 5C (KDM5C) protein: Computational model." *Proteins: Structure, Function, and Bioinformatics* 84.12 (2016): 1797-1809.

Peng, Yunhui, Joy Norris, Charles Schwartz, and Emil Alexov. "Revealing the effects of missense mutations causing Snyder-Robinson syndrome on the stability and dimerization of spermine synthase." *International journal of molecular sciences* 17, no. 1 (2016): 77.

Peng, Yunhui, Rebecca Myers, Wenxing Zhang, and Emil Alexov. "Computational investigation of the missense mutations in dhcr7 gene associated with smith-lemli-opitz syndrome." *International journal of molecular sciences* 19, no. 1 (2018): 141.

CHAPTER 3:

Peng, Yunhui, and Emil Alexov. "Computational investigation of proton transfer, p K a shifts and p H-optimum of protein–DNA and protein–RNA complexes." *Proteins: Structure, Function, and Bioinformatics* 85.2 (2017): 282-295.

Peng, Yunhui, Lexuan Sun, Zhe Jia, Lin Li, and Emil Alexov. "Predicting protein–DNA binding free energy change upon missense mutations using modified MM/PBSA approach: SAMPDI webserver." *Bioinformatics* 34, no. 5 (2017): 779-786.

CHAPTER 4:

Peng, Yunhui, Maria A. Miteva and Emil Alexov, "Rescuing of R133C Rett Syndrome causing mutation by Small Molecule Binding" In preparation

Appendix B

Supplementary materials: Figures

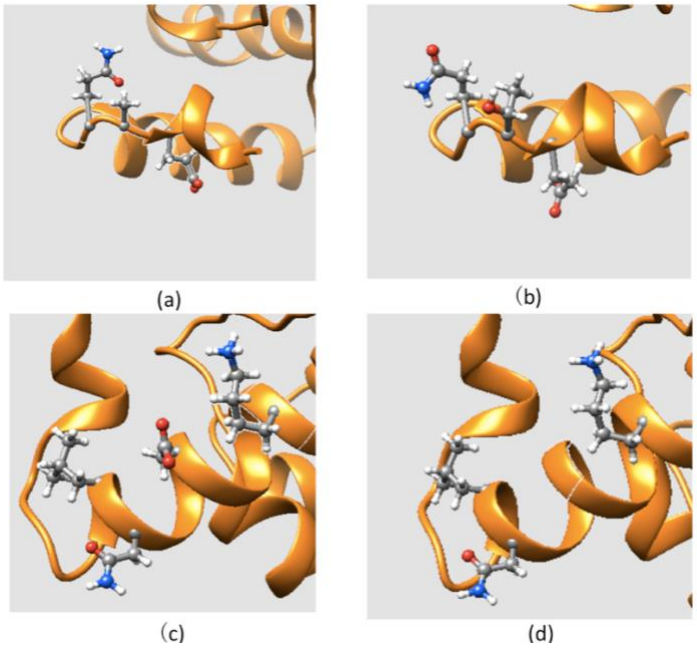


Figure B-1. The side chain conformations of two disease-associated mutations mapped onto the KDM5C ARID domain: (**a**) part of the ARID domain zoomed at the WT position of A77; (**b**) part of thhe ARID domain zoomed at the MT position of T77; (**c**) part of the ARID domain zoomed at the WT position of D87; and (**d**) part of the ARID domain zoomed at the MT position of G87.

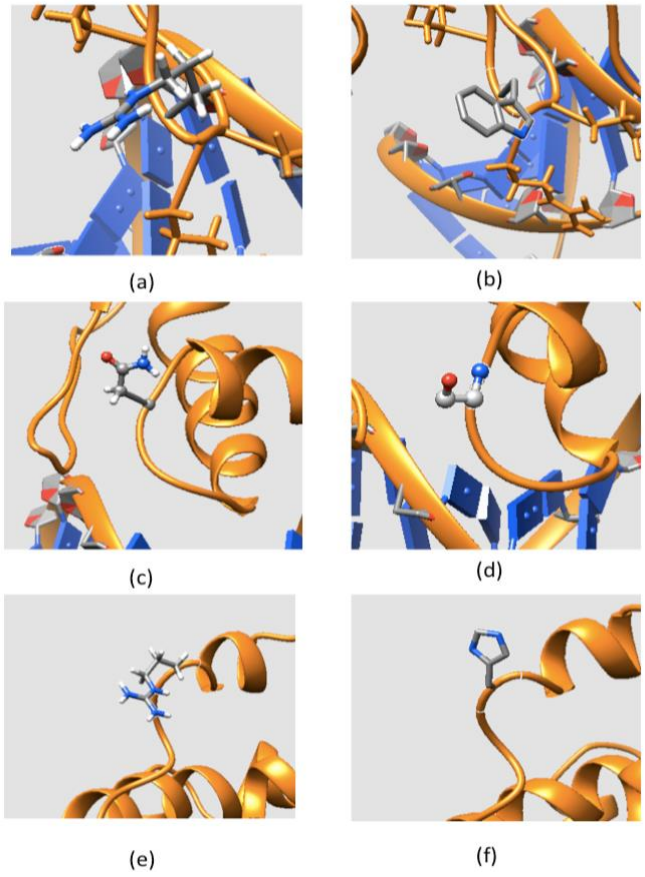


Figure B-2. The side chain conformation of non-classified mutations mapped onto the KDM5C ARID domain: (**a**) part of the ARID domain zoomed at the WT position of Arg108; (**b**) part of the ARID domain zoomed at the MT position of Try108; (**c**) part of the ARID domain zoomed at the WT position of Asn142; (**d**) part of the ARID domain zoomed at the MT position of Ser142; (**e**) part of the ARID domain zoomed at the WT position of Arg179; and (**f**) part of the ARID domain zoomed at the MT position of His179.

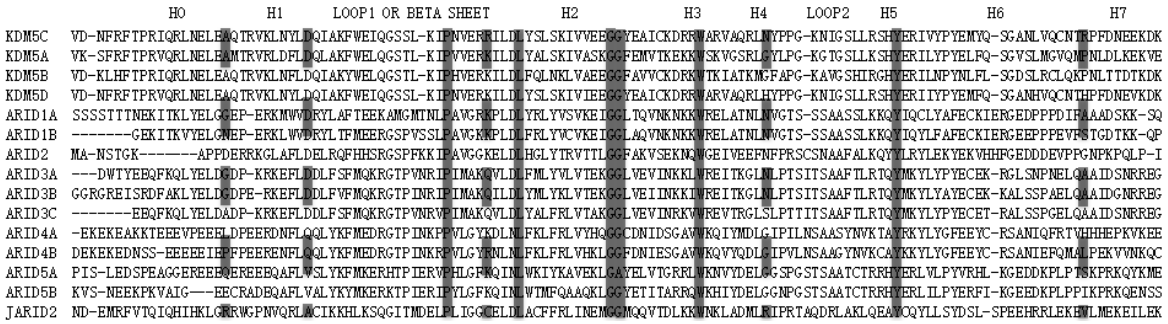


Figure B-3. Sequence alignment of human ARID-containing proteins. The mutation sites considered in this study are marked with grey bash line. The six most highly conserved residues are marked with a grey solid line. The helices from H0 to H7, and loops, are labeled at the top of the figure. The sequences are aligned with T-Coffee [207]. Similar results were obtained using the Clustal Omega webserver.

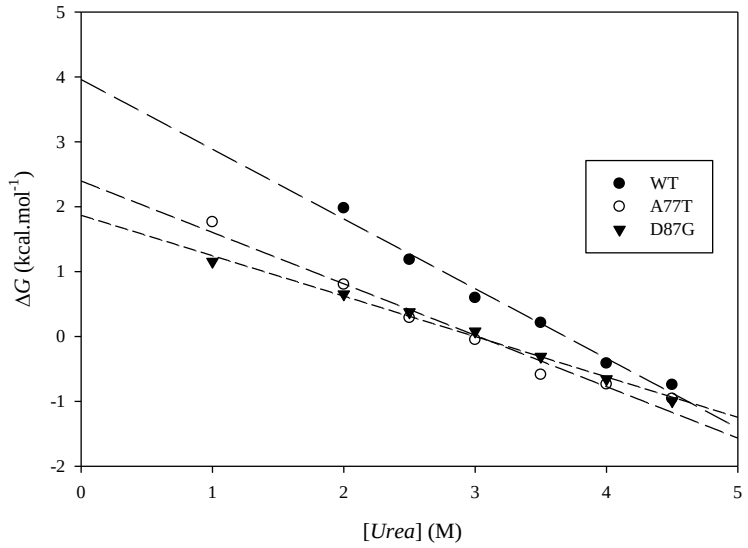
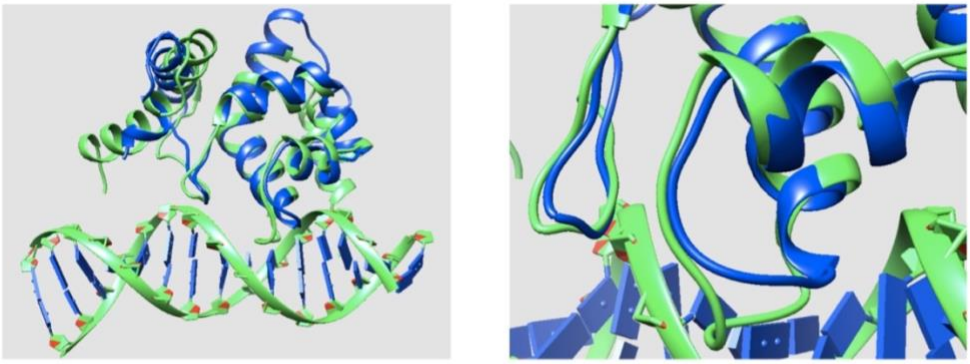
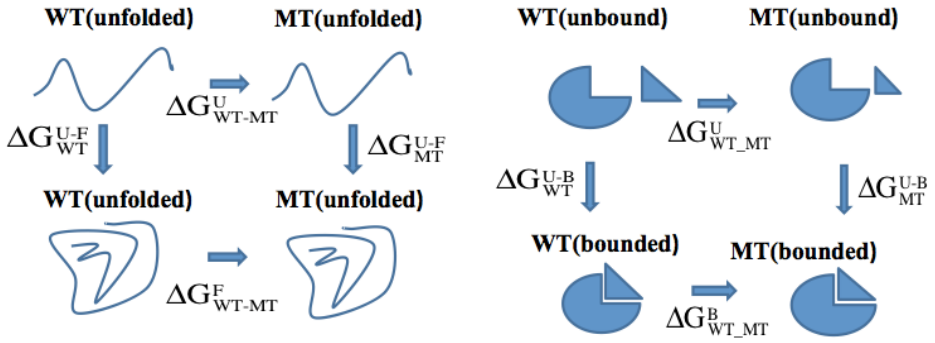


Figure B-4. A representative plot of Δ*G* for ARID WT, A77T, and D87G unfolding as a function of urea concentration.



(a) (b)

Figure B-5. (a) Structural alignment between the KDM5C ARID domain and dead ringer ARID-DNA complex; and (b) part of structural alignment zoomed at DNA binding interface. Dead ringer ARID-DNA complex is marked with green and the KDM5C ARID domain is marked with blue.



(a) (b)

Figure B-6. (a) Thermodynamic cycle for folding free energy changes calculations; and (b) thermodynamic cycle for binding free energy changes calculations.

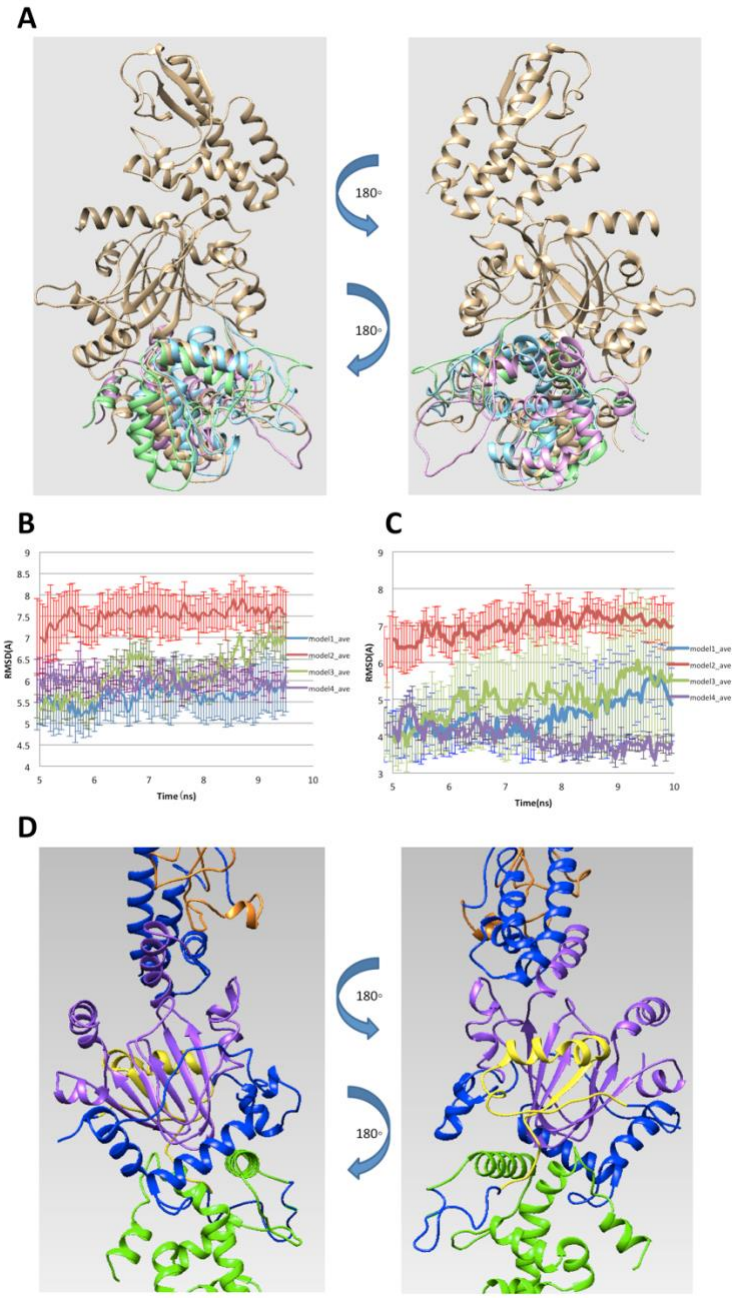


Figure B-7: (A) Four possible binding modes of ARID domain onto 5FWJ structure after applying constraint of linker length. (B) RMSD of ARID domain and JmjC domain. (C) RMSD of the interfacial residues. (D) Finalized model of ARID bound to KDM5C catalytic

core including JmjN, ARID, JmjC, ZF domains and the rest of inter domain regions marked with yellow, green, purple, orange and blue, respectively.

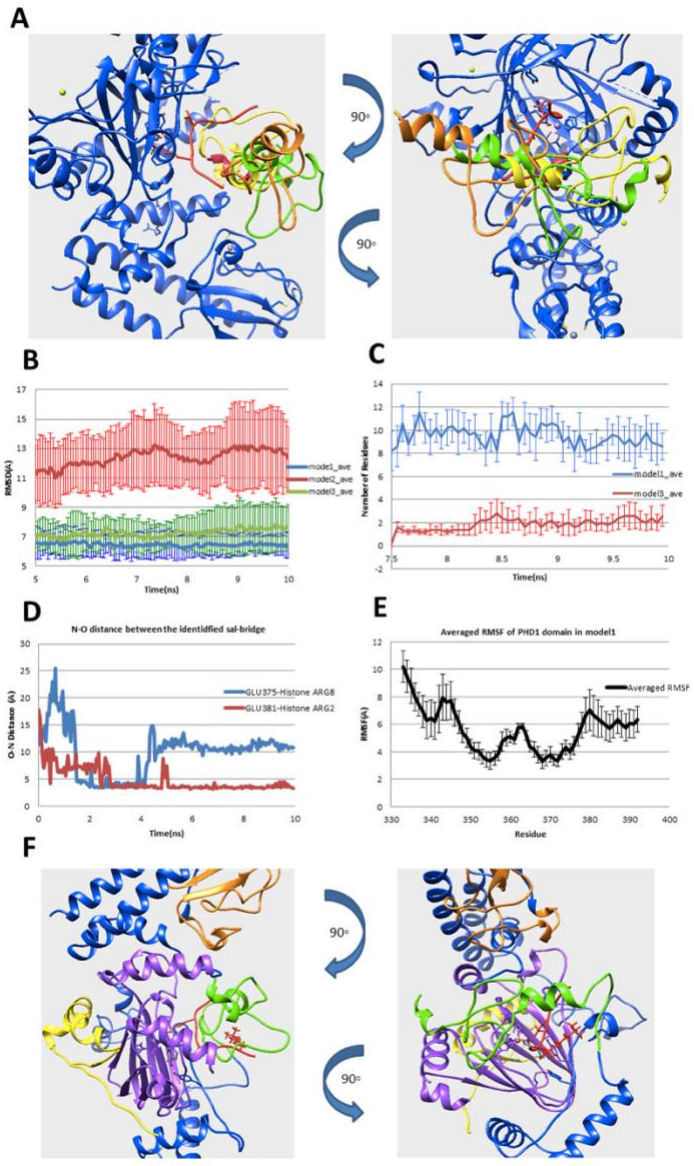


Figure B-8: (A) Three possible binding modes after applying the constraint of linker length. (B) RMSDs of the complex of PHD1and JmjC domains. (C) Number of residues in PHD1 domain, which have any atom within 6Å of H3K9me3 in the last 2.5ns simulation time. (D) The N-O distance of identified salt bridges (Glu375-H3R8 and Glu381-H3R2) between

PHD1 and substrate peptide. (E) Averaged RMSF of PHD1 domain residues calculated from model1 MD simulations. (F) Finalized model of PHD1 domain bound to KDM5C catalytic core including JmjN, PHD1, JmjC, ZF domains, inter domain region and histone peptide marked with yellow, green, purple, orange, blue and red, respectively

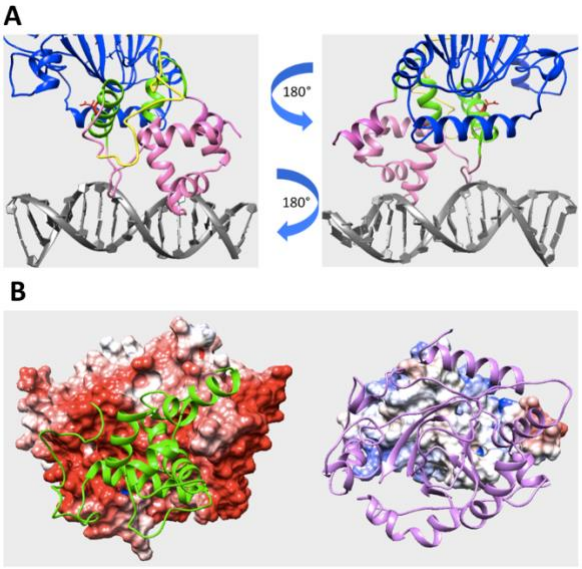


Figure B-9: (A) Predicted interfacial region of ARID domain involved in domain interactions in the KDM5C quaternary structure. The interfacial region of ARID domain is marked with green while the rest of ARID domain is marked with pink. JmjN domain and DNA are marked with yellow and gray. Other regions, including JmjC and ZF domains, are marked with blue. The reside Asp87 side chain is shown with red. (B) The electrostatic potential map of ARID domain and KDM5C catalytic core. The electrostatic potential map of KDM5C catalytic core (including JmjN, JmjC and ZF domains) is shown on the left (the structure of ARID domain is marked with green). The electrostatic potential map of ARID domain is shown on the right (the structure of JmjN and JmjC domains are marked with purple).

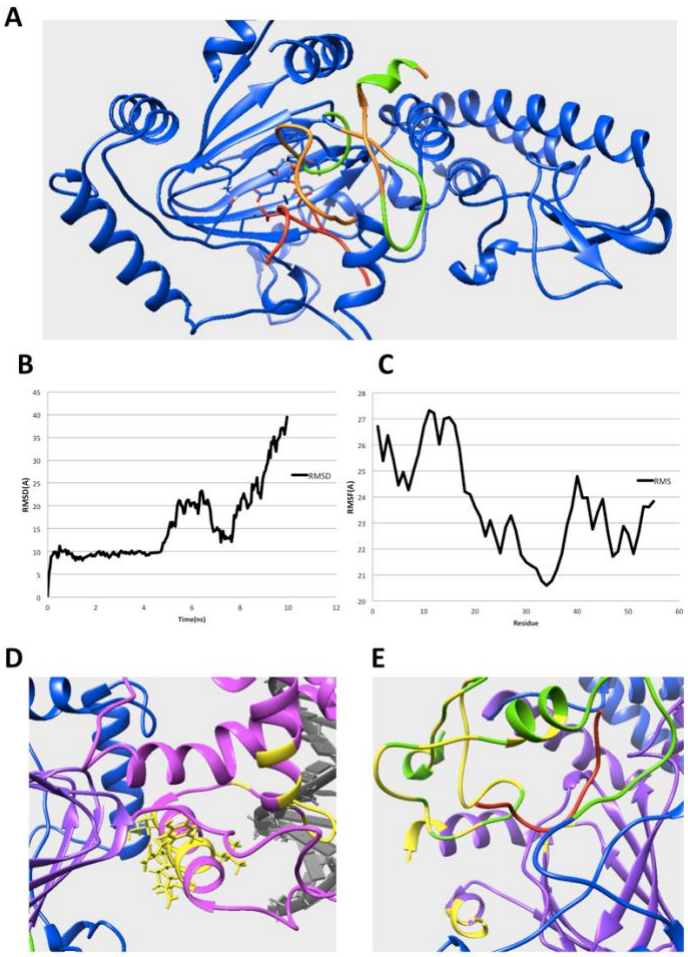
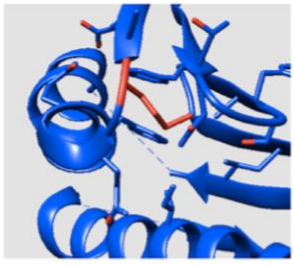
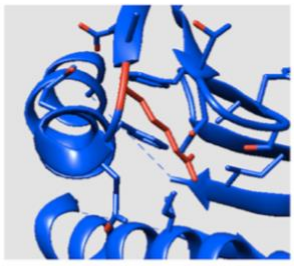


Figure B-10: (A) The interfacial residues of PHD1 domain involved in the interaction in the KDM5C quaternary structure. The interfacial residues of PHD1 domain are marked with green while the rest of the PHD1 domain are marked with orange. The substrate histone peptide is marked with red while the rest of the region including JmjC and ZF domain are marked with blue. (B) The RMSD results for the KDM5B PHD1 domain bound to substrate peptide complex. (C) The RMSF results for the KDM5B PHD1 domain. (D) The predicted interfacial residues in ARID domain. The ARID domain, JmjC domain,

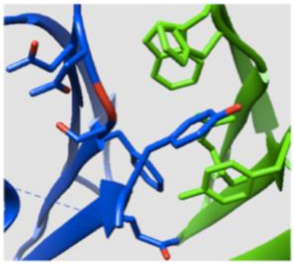
DNA and inter-domain region are marked with pink, purple, gray, and blue, respectively. The predicted interfacial residues in ARID domain are colored with yellow. (E) The predicted interfacial residues in PHD and JmjC domains. The PHD domain, JmjC domain, histone substrate and inter-domain regions are marked with green, purple, red, and blue, respectively. Predicted interfacial residues in PHD and JmjC domains are colored with yellow.



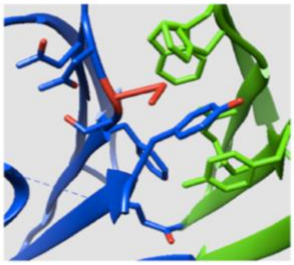
(a) Met 35



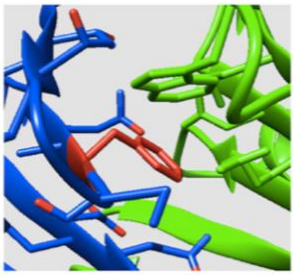
(b) Arg 35



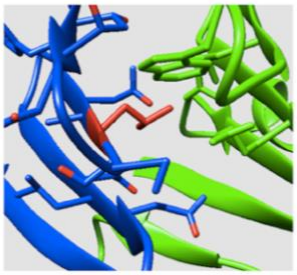
(c) Gly 56



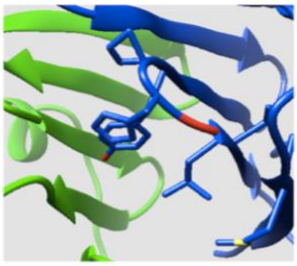
(d) Ser 56



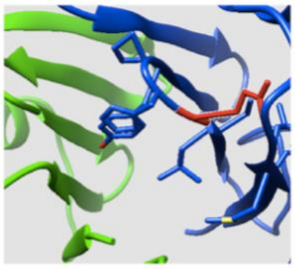
(e) Phe 58



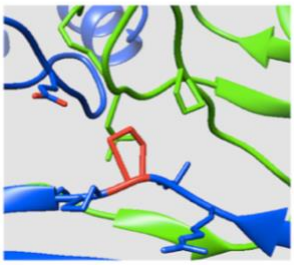
(f) Leu 58



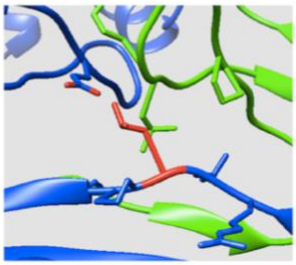
(g) Gly 67



(h) Glu 67



(i) Pro 112



(j) Leu 112

Figure B-11. The side chain conformation of five disease-causing mutations mapped onto SpmSyn: (a) Part of SpmSyn zoomed at WT position of Met35; (b) Part of SpmSyn zoomed at MT position of Arg35; (c) Part of SpmSyn zoomed at WT position of Gly56; (d) Part of SpmSyn zoomed at MT position of Ser56; (e) Part of SpmSyn zoomed at WT position of Phe58; (f) Part of SpmSyn zoomed at MT position of Leu58; (g) Part of SpmSyn zoomed at WT position of Gly67; (h) Part of SpmSyn zoomed at MT position of Glu67; (i) Part of SpmSyn zoomed at WT position of Pro112; (j) Part of SpmSyn zoomed at MT position of Leu112; The side chain of WT and MT position is shown in red. Two different chains of the dimer are shown in blue and green.

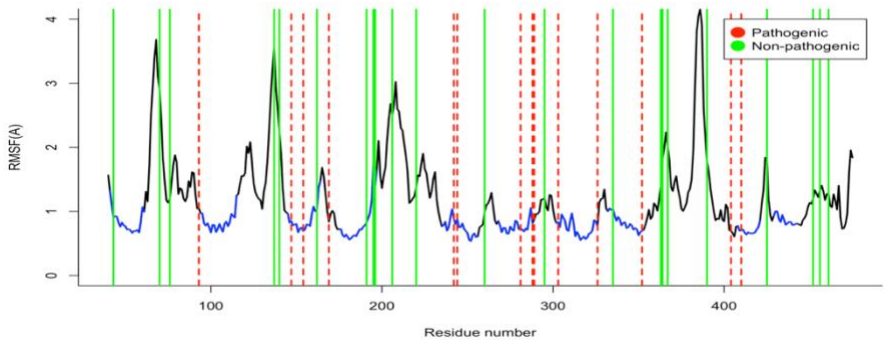


Figure B-12：The pathogenic and non-pathogenic mutation occurring sites mapping on the average RMSF of the WT DHCR7 protein. Pathogenic and non-pathogenic mutation sites are marked with red and green lines. The RMSF of transmembrane region are shown as blue.

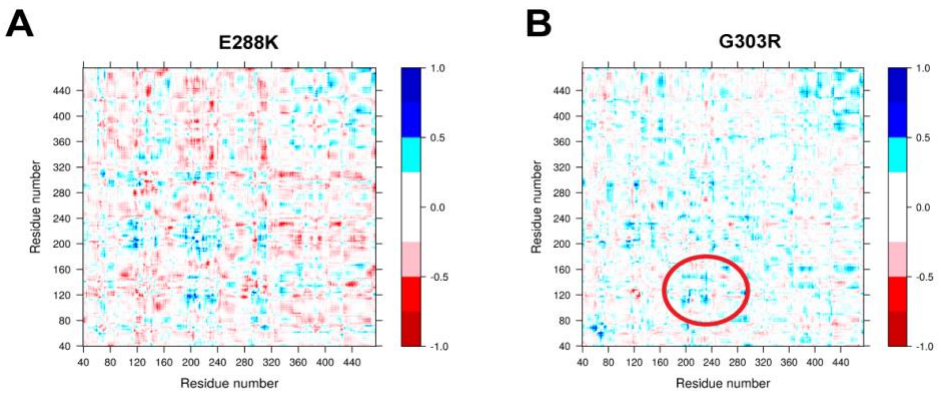


Figure B-13：The changes in residue cross-correlation for mutation E288K and G303R.

```
DHCR7          MAAKSQPNIPKAKSLDGVTNDRTASQGQWGRAWEVDWFSLASVIFLLLFAPFIVYYFIMACDQYSCALTGPVVDIVTGHA
4QUV           MSEQE--------SRDNAAVDAVRQKYGFGFSWLV----------LMIALPPLVYYLWICVTYYQGELV-----FTSDAA
               *: :.         * *..: * . .:  :* :* *          *::  * :***: :.   *.  *.     :.:. *

DHCR7          RLSDIWAKTPPITRKAAQLYTLWVTFQVLLYTSLPDFCHKFLPGYVGGIQEGAVTPAGVVNKYQINGLQAWLLTHLLWFA
4QUV           AWRRFWSHVAPPTWHAAGLYAAWFLGQAALQVWAP-----------GPTVQGMKLPDGSRLDYRMNGIFSFLFTLAVVFG
                   :*::..* * :** **: *.  *. * .  *            *   :*   * *   .*::**: ::*:*  : *.

DHCR7          NAHLLSWFSPTIIFDNWIPLLWCANILGYAVSTFAMVKGYFFPTSARDC-KFTGNFFYNYMMGIEFNPRIGKWFDFKLFF
4QUV           LV-TMGWLDATVLYDQLGPLLTVVNIFTFVFAGFL----YFWGLNGKQWERPTGRPFYDYFMGTALNPRIGS-LDLKLFC
                .  :.*:..*:::*:  ***  .**: :..: *     **:  ..::  : **. **:*:**  :*****. :*:***

DHCR7          NGRPGIVAWTLINLSFAAKQRELHSHVTNAMVLVNVLQAIYVIDFFWNETWYLKTIDICHDHFGWYLGWGDCVWLPYLYT
4QUV           EARPGMIFWLLMNLSMAAKQYELHGTVTVPMLLVVGFQSFYLIDYFIHEEAVLTTWDIKHEKFGWMLCWGDLVWLPFTYT
               :.***:: * *:***:**** ***. ** .*:**  :*::*:**:* :*   *.* ** *::*** * *** ****: **

DHCR7          LQGLYLVYHPVQLSTPHAVGVLLLGLVGYYIFRVANHQKDLFRRTDGRCLIWGRKPKVIECSYTSADGQRHHSKLLVSGF
4QUV           LQAQYLVHHTHDLPVWGIIAIVALNLAGYAIFRGANIQKHHFRRDP-NRIVWGKPAKYIKT--------KQGSLLLTSGW
               **. ***:*. :*..   :.:: *.*.** *** ** **. ***   . ::**: .* *:         :: * **.**:

DHCR7          WGVARHFNYVGDLMGSLAYCLACGGGHLLPYFYIIYMAILLTHRCLRDEHRCASKYGRDWERYTAAVPYRLLPGI-
4QUV           WGIARHMNYFGDLMIALSWCLPAAFGSPIPYFHIVYFTILLLHREKRDDAMCLAKYGEDWLQYRKKVPWRIVPKIY
               **:***:**.**** :*::**... *  :***:*:*::*** **  **:  * :***.** :*   **:*::* *
```

Figure B-14：Sequence alignment between DHCR7 and template 4QUV.

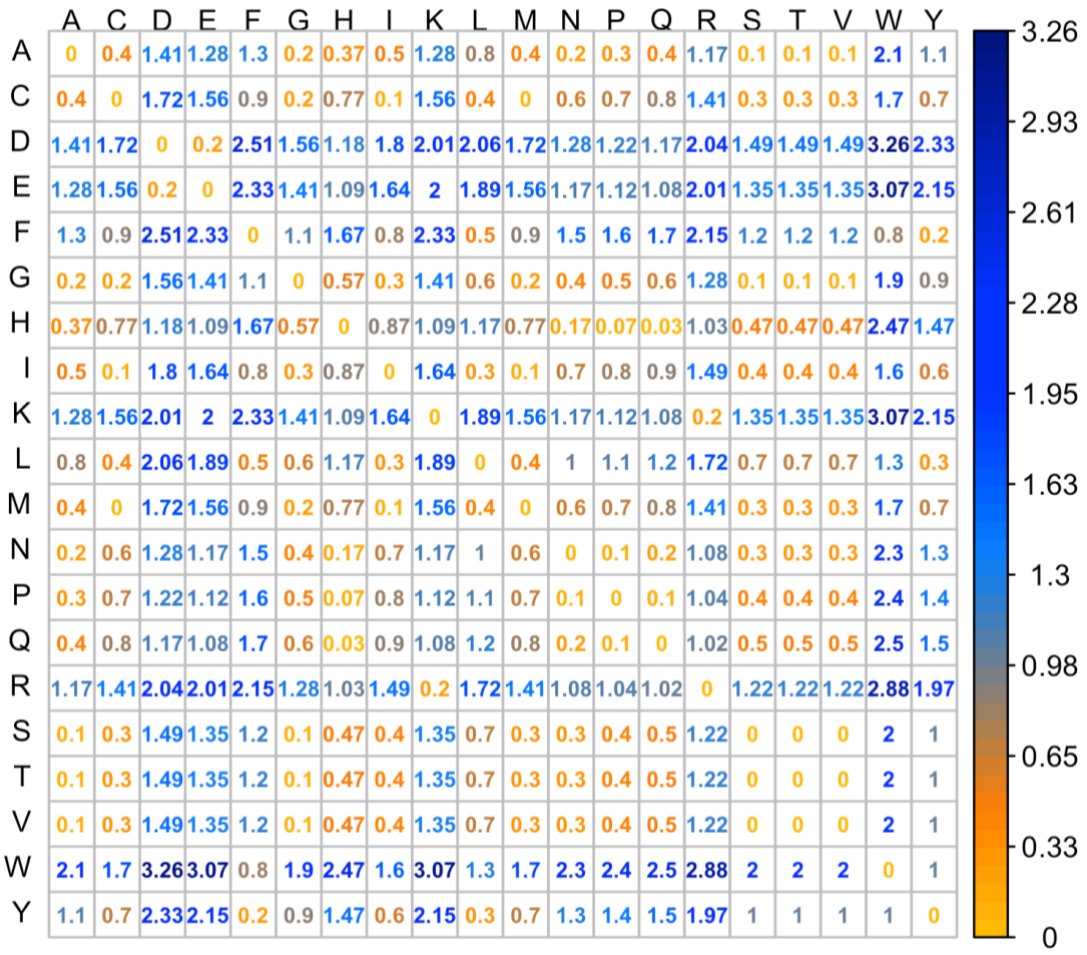


Figure B-15: Property distance for all types of amino acid pairs.

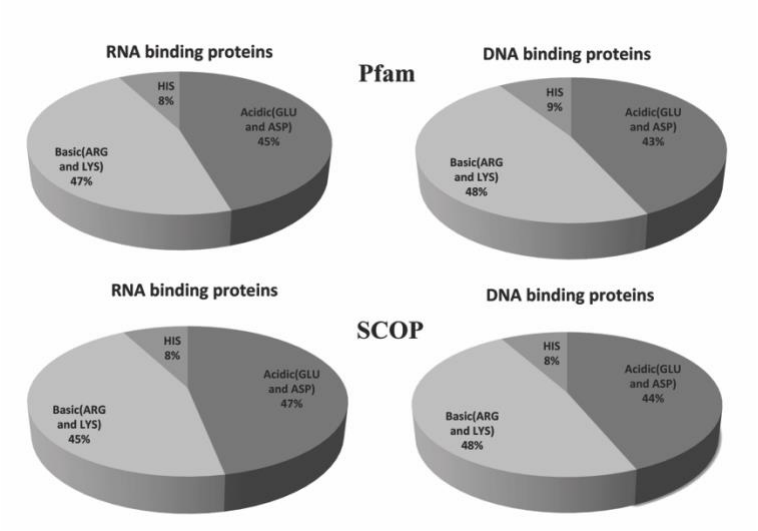


Figure B-16: Frequency patterns of ionizable residues in both Pfam and SCOP datasets.

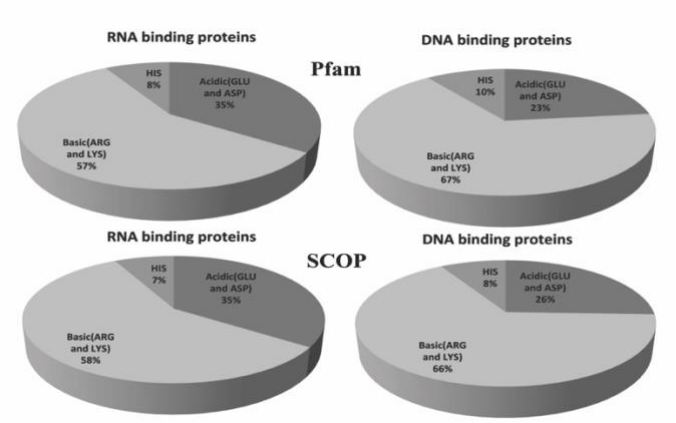


Figure B-17: Frequency patterns of ionizable interfacial residues in both Pfam and SCOP datasets.

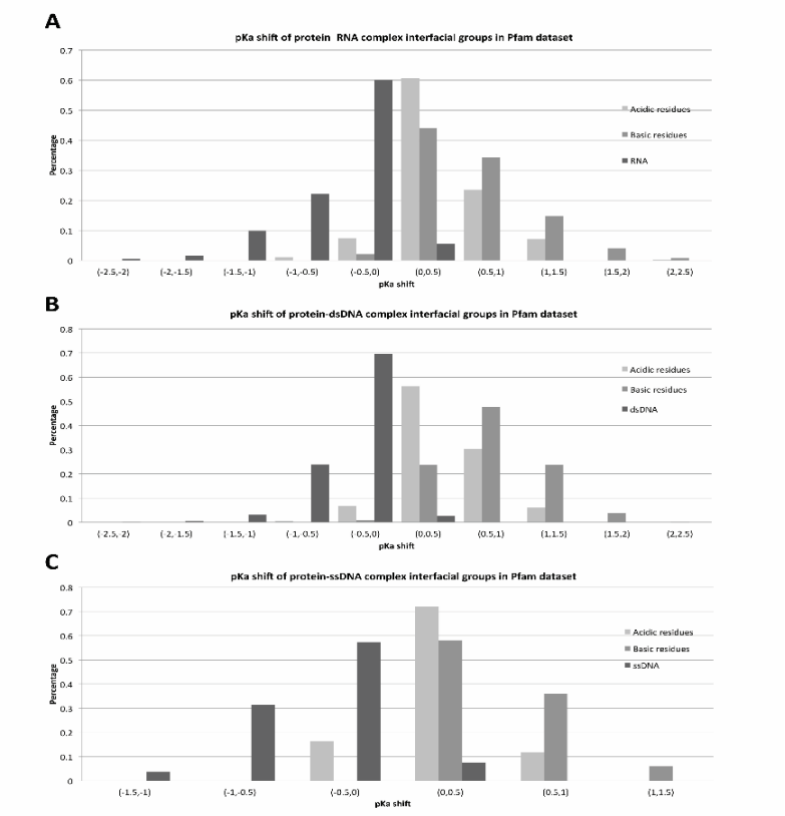


Figure B-18: Distribution of pKa shifts for different types of ionizable groups and different types of complexes in Pfam dataset.

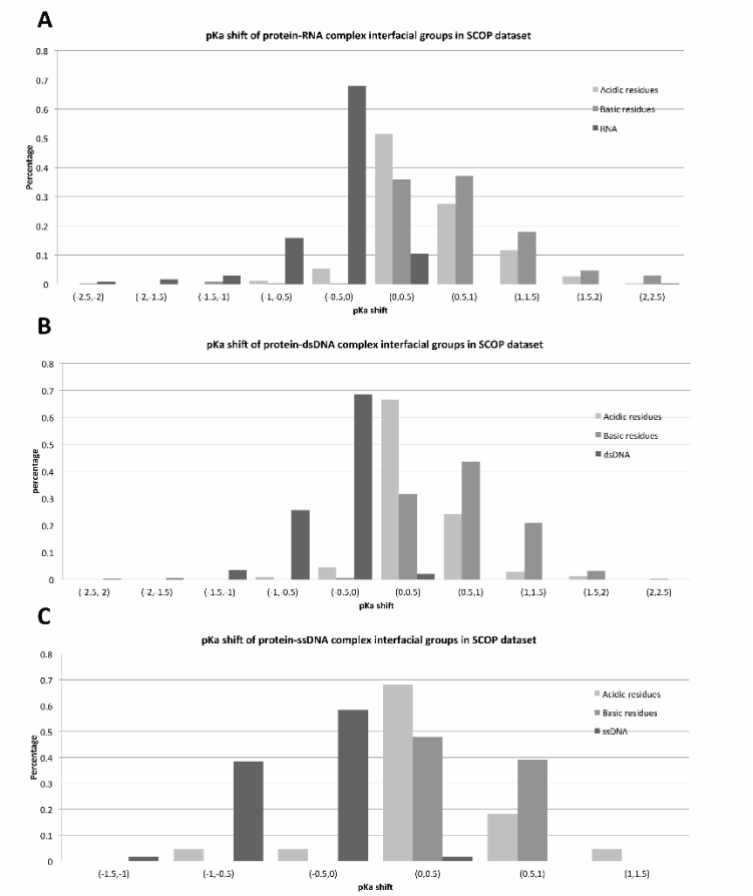


Figure B-19: Distribution of pKa shifts for different types of ionizable groups and different types of complexes in SCOP dataset.

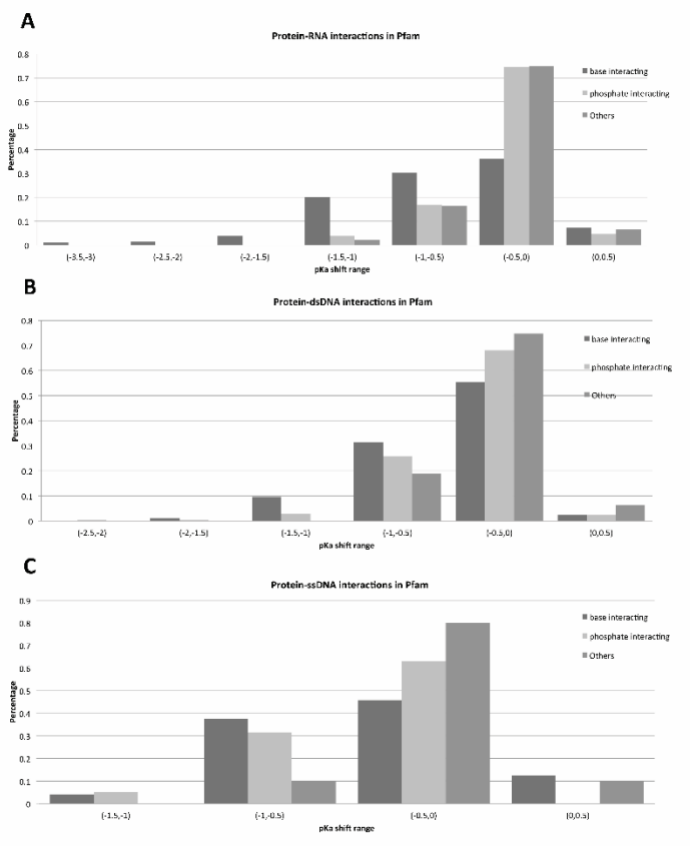


Figure B-20: Distributions of pKa shifts across the different binding modes in the Pfam dataset.

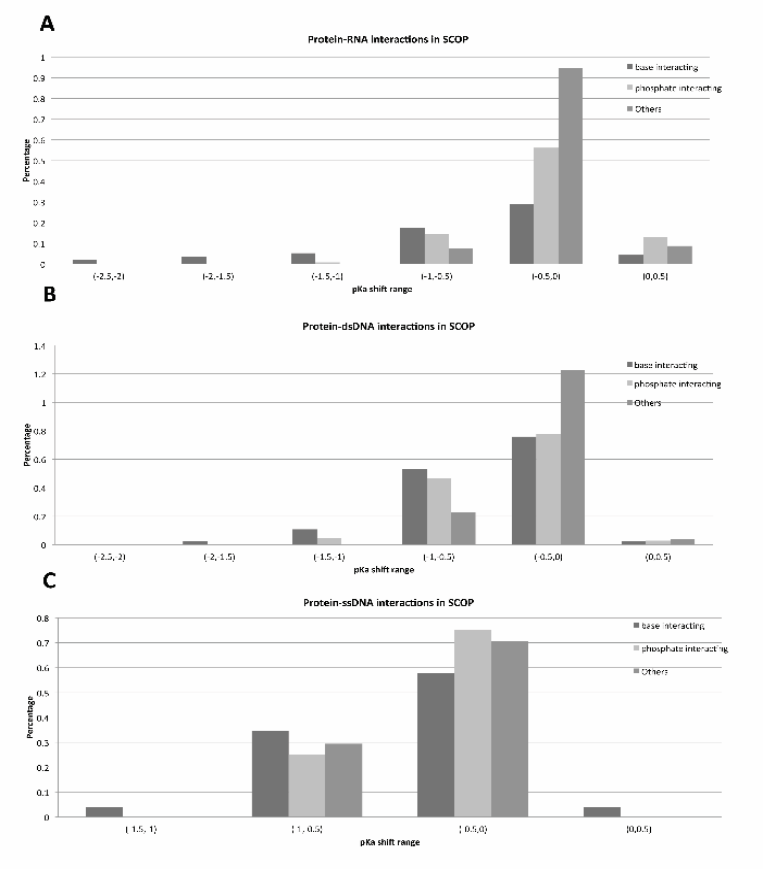


Figure B-21: Distributions of pKa shifts across the different binding modes in the SCOP dataset.

```
receptor = receptor.pdbqt
ligand = ligand.pdbqt
center_x =  1.761
center_y =  6.723
center_z = -35.845

size_x = 46
size_y = 42
size_z = 46

energy_range = 4
num_modes = 9

exhaustiveness = 20
```

Figure B-22: Docking parameter file used for Autodock Vina.

```
autodock_parameter_version 4.2       # used by autodock to validate parameter set
outlev 1                             # diagnostic output level
intelec                              # calculate internal electrostatics
seed pid time                        # seeds for random generator
ligand_types A C F HD N NA OA        # atoms types in ligand
fld cluster.maps.fld                 # grid_data_file
map cluster.A.map                    # atom-specific affinity map
map cluster.C.map                    # atom-specific affinity map
map cluster.F.map                    # atom-specific affinity map
map cluster.HD.map                   # atom-specific affinity map
map cluster.N.map                    # atom-specific affinity map
map cluster.NA.map                   # atom-specific affinity map
map cluster.OA.map                   # atom-specific affinity map
elecmap cluster.e.map                # electrostatics map
desolvmap cluster.d.map              # desolvation map
move test.pdbqt                      # small molecule
about 0.9114 0.1681 0.0628           # small molecule center
tran0 random                         # initial coordinates/A or random
quaternion0 random                   # initial orientation
dihe0 random                         # initial dihedrals (relative) or random
torsdof 2                            # torsional degrees of freedom
rmstol 2.0                           # cluster_tolerance/A
extnrg 1000.0                        # external grid energy
e0max 0.0 10000                      # max initial energy; max number of retries
ga_pop_size 150                      # number of individuals in population
ga_num_evals 25000000                # maximum number of energy evaluations
ga_num_generations 27000             # maximum number of generations
ga_elitism 1                         # number of top individuals to survive to next generat
ga_mutation_rate 0.02                # rate of gene mutation
ga_crossover_rate 0.8                # rate of crossover
ga_window_size 10                    #
ga_cauchy_alpha 0.0                  # Alpha parameter of Cauchy distribution
ga_cauchy_beta 1.0                   # Beta parameter Cauchy distribution
set_ga                               # set the above parameters for GA or LGA
sw_max_its 300                       # iterations of Solis & Wets local search
sw_max_succ 4                        # consecutive successes before changing rho
sw_max_fail 4                        # consecutive failures before changing rho
sw_rho 1.0                           # size of local search space to sample
sw_lb_rho 0.01                       # lower bound on rho
ls_search_freq 0.06                  # probability of performing local search on individual
set_psw1                             # set the above pseudo-Solis & Wets parameters
unbound_model extended               # state of unbound ligand
ga_run 10                            # do this many hybrid GA-LS runs
analysis                             # perform a ranked cluster analysis
```

Figure B-23: Docking parameter file used for Autodock4.

```
conformer_search_type                                flex
min_anchor_size                                      10
pruning_use_clustering                               yes
pruning_max_orients                                  1000
pruning_clustering_cutoff                            100
pruning_conformer_score_cutoff                       100
pruning_conformer_score_scaling_factor               1.0
use_internal_energy                                  yes
internal_energy_rep_exp                              9
internal_energy_cutoff                               100.0
ligand_atom_file                                     ligand.mol2
automated_matching                                   yes
receptor_site_file                                   sphgen_cluster.sph
max_orientations                                     1000
score_molecules                                      yes
contact_score_primary                                no
contact_score_secondary                              no
grid_score_primary                                   yes
grid_score_secondary                                 no
grid_score_rep_rad_scale                             1
grid_score_vdw_scale                                 1
grid_score_es_scale                                  1
grid_score_grid_prefix                               grid
multigrid_score_secondary                            no
gbsa_hawkins_score_secondary                         yes
gbsa_hawkins_score_rec_filename                      cluster.mol2
gbsa_hawkins_score_solvent_dielectric                78.5
gbsa_hawkins_use_salt_screen                         yes
gbsa_hawkins_score_salt_conc(M)                      0.1
gbsa_hawkins_score_gb_offset                         0.09
gbsa_hawkins_score_cont_vdw_and_es                   yes
gbsa_hawkins_score_vdw_att_exp                       6
gbsa_hawkins_score_vdw_rep_exp                       9
grid_score_rep_rad_scale                             1
minimize_ligand                                      yes
minimize_anchor                                      yes
minimize_flexible_growth                             yes
use_advanced_simplex_parameters                      no
simplex_max_cycles                                   1
simplex_score_converge                               0.1
simplex_cycle_converge                               1.0
simplex_trans_step                                   1.0
simplex_rot_step                                     0.1
simplex_tors_step                                    10.0
simplex_anchor_max_iterations                        500
simplex_grow_max_iterations                          500
simplex_grow_tors_premin_iterations                  0
simplex_secondary_minimize_pose                      no
simplex_random_seed                                  0
simplex_restraint_min                                yes
simplex_coefficient_restraint                        10.0
atom_model                                           all
vdw_defn_file                                        vdw_AMBER_parm99.defn
flex_defn_file                                       flex.defn
flex_drive_file                                      flex_drive.tbl
ligand_outfile_prefix                                anchor_and_grow
write_orientations                                   no
num_primary_scored_conformers_rescored               3
write_primary_conformations                          yes
cluster_primary_conformations                        yes
cluster_rmsd_threshold                               1.5
num_clusterheads_for_rescore                         5
num_secondary_scored_conformers                      3
write_secondary_conformations                        yes
rank_primary_ligands                                 yes
max_primary_ranked                                   10000
rank_secondary_ligands                               yes
max_secondary_ranked                                 10000
```

Figure B-24: Docking parameter file used for Dock6.

Appendix C

Supplementary materials: Tables

	Model1	Model2	Model3	Model4
Model 1	0	2.20	2.64	1.80
Model 2	2.20	0	3.46	3.06
Model 3	2.64	3.46	0	2.93
Model 4	1.80	3.06	2.93	0
average	2.21	2.91	3.01	2.60

Table C-1: The RMSD of various ARID binding modes (in Å). The last row shows the average RMSD calculated with respect with other three models.

	Salt bridge involved in the domain interactions
ARID	Arg80-Glu465, Arg80-Glu467, Arg80-Gu468, Asp87-Lys459, Asp87-Arg460 Lys91-Glu465, Lys91-Glu466, Glu94-Arg390, Glu94-Lys459, Arg159-Glu399, Arg159-Asp402, Arg159-Glu419, Arg159-Glu422
PHD1	Asp334-Lys551, Glu334-Lys550, Glu335-Lys551, Glu335-Arg635, Asp336-Arg637, Asp337-Lys550, Asp347-Arg637, Asp347-Lys711

Table C-2: Lists of identified salt bridges involved in interfacial ARID and PHD1 domains interactions.

Mutation	SMS/GAPDH Ratio	% Ctrl
M35R	0.04	1.5
G56S	0.07	2.6
F58L	0.18	6.6
G67E	0.13	4.8
P112L	0.55	20.3
Ctrl	2.71	100

Table C-3. Densitometric analysis of bands present in denatured gel.

Disease-causing missense mutations									
Mutation	**rSASA_mem**	**CV score**	**SAFFEC**	**mCSM**	**SDM**	**DUET**	**FOLDX**	**ΔΔG_ave**	**Polyphen**
T93M	0.21	0.62	0.53	0.10	0.30	-0.03	-0.52	0.08	Probably damaging
G147D	0.01	1.00	3.30	-1.53	-0.35	-1.39	7.40	1.49	Probably damaging
T154R	0.01	0.97	-3.55	-0.59	-1.45	-0.60	1.62	-0.91	Probably damaging
S169L	0.31	0.85	1.39	-0.25	0.69	-0.25	0.62	0.44	Probably damaging
R242H	0.02	1.00	0.17	-2.38	-0.05	-2.59	-0.20	-1.01	Probably damaging
R242C	0.02	1.00	0.40	-1.96	0.39	-2.05	-3.95	-1.43	Probably damaging
G244R	0.01	1.00	0.96	-1.39	-0.10	-1.20	4.72	0.60	Probably damaging
V281M	0.02	0.91	-0.19	-0.28	-0.08	-0.30	0.40	-0.09	Probably damaging

E288K	0.10	1.00	-15.77	-0.27	-1.17	-0.24	0.78	-3.33	Probably damaging
T289I	0.43	0.29	-0.60	-0.09	1.26	0.18	-0.70	0.01	Possibly damaging
G303R	0.04	1.00	-1.48	-1.15	-2.73	-1.22	21.73	3.03	Probably damaging
V326L	0.00	0.94	0.24	-1.08	0.20	-1.00	1.01	-0.13	Benign
R352W	0.06	0.94	0.00	-0.37	2.57	-0.55	0.12	0.35	Probably damaging
R352Q	0.06	0.94	-0.91	-0.81	-0.63	-0.81	0.60	-0.51	Probably damaging
R404C	0.04	0.97	-0.05	-1.97	0.11	-2.18	3.00	-0.22	Probably damaging
G410S	0.01	0.91	-2.71	-1.81	-0.55	-1.69	11.05	0.86	Probably damaging

Missense mutations with unknown effects

Mutation	rSASA_mem	CV score	SAFFEC	mCSM	SDM	DUET	FOLDX	Average	SD
A41V	0.05	0.44	0.03	-0.28	-0.29	0.01	-0.30	-0.17	Benign
I44T	0.18	0.76	0.69	-1.29	-2.79	-1.19	1.40	-0.64	Benign
A67T	0.00	0.18	2.48	-1.59	-0.79	-1.66	5.65	0.82	Possibly damaging
I75F	0.53	0.24	-0.47	-0.64	0.12	-0.61	-0.56	-0.43	Benign
R81W	0.51	0.21	-0.55	-0.49	1.18	-0.58	-0.64	-0.22	Probably damaging
A97T	0.24	0.80	0.18	-1.28	-2.29	-1.19	1.64	-0.59	Possibly damaging
V126I	0.12	0.76	0.60	-0.78	0.74	-0.72	0.04	-0.02	Probably damaging
V134L	0.28	0.32	1.36	-0.63	1.13	-0.43	-1.24	0.04	Benign
A162V	0.13	0.76	1.18	0.00	0.68	0.19	-1.50	0.11	Possibly damaging
R228Q	0.18	0.97	1.27	-1.01	-0.84	-1.13	-0.23	-0.39	Probably damaging
V330M	0.58	0.85	-2.47	-0.52	-0.51	-0.45	-0.57	-0.90	Probably damaging

V338M	0.25	0.41	-0.92	-0.38	0.80	-0.19	-0.70	-0.28	Benign
F361L	0.01	0.91	-0.16	-2.72	-0.17	-2.80	-0.90	-1.35	Probably damaging
T364M	0.25	0.85	-0.63	-0.03	0.30	-0.32	-0.60	-0.26	Probably damaging
R367C	0.56	0.32	4.01	-0.37	0.83	-0.30	0.30	0.89	Probably damaging
G424S	0.08	0.35	-0.48	-0.48	2.60	-0.02	-0.03	0.32	Probably damaging
G425S	0.12	0.56	-0.95	-0.68	2.54	-0.28	-0.10	0.11	Probably damaging
R461C	0.54	0.71	3.02	-0.39	0.16	-0.31	1.25	0.75	Probably damaging

Non-disease-causing missense mutations

Mutation	rSASA_mem	Conser	SAFFEC	mCSM	SDM	DUET	FOLDX	Average	SD
V43I	0.16	0.74	0.02	-0.32	0.59	-0.10	0.61	0.16	Benign
G70S	0.00	0.12	0.67	-1.68	-2.79	-1.95	2.51	-0.65	Benign
V76I	0.43	0.06	-1.00	-0.27	0.01	-0.19	0.75	-0.14	Benign
A137S	0.68	0.97	0.39	-0.81	-1.42	-0.66	-0.19	-0.54	Possibly damaging
V140M	0.37	0.53	-0.75	-0.49	-0.74	-0.59	-0.81	-0.67	Benign
A162V	0.13	0.76	0.94	0.00	0.68	0.19	-1.47	0.07	Possibly damaging
V191I	0.31	0.82	0.43	-0.45	0.59	-0.20	-1.11	-0.15	Possibly damaging
A195T	0.03	0.85	0.37	-0.96	-0.79	-0.96	4.06	0.35	Probably damaging
M196V	0.03	0.41	-0.37	-0.97	-1.11	-1.08	4.19	0.13	Benign
A206T	0.29	0.59	2.01	-1.05	-1.72	-0.99	-0.16	-0.38	Probably damaging
M220L	0.08	0.85	0.27	0.16	0.40	0.47	0.14	0.29	Benign
R260Q	0.25	0.15	-1.37	-0.49	-0.26	-0.28	0.23	-0.43	Benign
I295V	0.09	0.88	0.55	-1.04	-0.59	-0.97	0.50	-0.31	Possibly damaging

P335R	0.32	0.38	-3.49	0.15	1.18	0.40	1.85	0.02	Possibly damaging
R363C	0.54	0.88	2.53	-0.17	0.16	-0.10	0.29	0.54	Benign
R363H	0.54	0.88	2.00	-0.98	-0.41	-0.99	0.29	-0.02	Probably damaging
T364M	0.25	0.85	0.59	-0.03	0.30	-0.32	-0.62	-0.02	Probably damaging
R367C	0.56	0.32	3.65	-0.37	0.83	-0.30	0.28	0.82	Probably damaging
H390T	0.36	0.85	0.49	1.55	1.47	1.48	-0.34	0.93	Benign
G425S	0.12	0.56	1.13	-0.68	2.54	-0.28	-0.10	0.52	Benign
A452T	0.32	0.32	1.32	-1.28	-2.72	-1.18	0.41	-0.69	Possibly damaging
G456S	0.33	0.94	-0.43	-1.30	-2.55	-1.21	4.05	-0.29	Probably damaging
R461C	0.54	0.71	3.71	-0.39	0.16	-0.31	1.25	0.88	Probably damaging

Table C-4: Folding free energy, rSASA and Polyphen predictions for the mutations in DHCR7 protein. ΔΔGs are shown in kcal/mol and average ΔΔG are also calculated using the results from multiple webservers. Mutations A206T and H390T are located on the loop, not present in the template, thus the rSASA is highlighted as red to indicate low confidence in our calculation for these two mutations.

Using rSASA, EC score, PD and ΔΔG				**Using only rSASA, EC score, PD**			
K	TP	TN	Accuracy	K	TP	TN	Accuracy
1	4	2	0.6	1	5	4	0.9
2	4	2	0.6	2	5	4	0.9
3	5	2	0.7	3	6	4	1
4	4	2	0.6	4	6	3	0.9
5	5	2	0.7	5	6	4	1

6	5	2	0.7	6	6	4	1
7	5	2	0.7	7	6	4	1
8	5	3	0.8	8	6	4	1
9	5	2	0.7	9	6	4	1
10	5	0	0.5	10	5	4	0.9

Table C-5. KNN classifications using different properties and K values. True positive (TP), true negative (TN) and accuracy are calculated for each K value.

Residue	***A***	***C***	***D***	***E***	***F***	***G***	***H***	***I***	***K***	***L***
Rotamer	1	3	18	54	18	1	36	9	81	9
Residue	***M***	***N***	***P***	***Q***	***R***	***S***	***T***	***V***	***W***	***Y***
Rotamer	27	36	2	108	81	3	3	3	36	18

Table C-6. Max number of the rotamers for all types of amino acids taken from [322].

	CE	**PS**	**VE**	**NS**	**S**	**HB**	**Y-intercept**
Coefficient	0.078	0.048	0.088	-0.0012	0.14	-0.043	0.445
P-value	2E-05	1E-05	7.13E-08	0.4	0.041	0.10	8.92E-08
Correlation coefficient		0.72		**Number of cases**		105	

Table C-7. The weight coefficients of the linear function for binding free energy changes determined from MLR. The corresponding p-values are shown as well.

	Root Mean Square of the Error (kcal/mol)	*Pearson correlation coefficient*
Fold 1	0.53	0.33
Fold 2	0.57	0.7
Fold 3	0.48	0.76
Fold 4	0.64	0.6
Fold 5	0.49	0.52
Average	0.54	0.58

Table C-8. 5-fold cross validation for the dataset used for the SAMPDI approach.

	Averaged weighting coefficients in 5-fold cross validation	**Standard Deviation of weighting coefficients in 5-fold cross validation**	**Weighting coefficients from MLR in the training**
Y-Intercept	0.46	0.075	0.445
NS	-0.0025	0.0015	-0.0012
VE	0.098	0.016	0.088
CE	0.074	0.017	0.078
PS	0.046	0.01	0.048
S	0.14	0.068	0.14
HB	-0.045	0.025	-0.043

Table C-9: Average weighting coefficients and corresponding standard deviation in 5-fold cross validation for all the energy terms. The determined weighting coefficients from MLR was also shown for the comparison.

Correlation matrixes calculated with Pearson correlation						
	SASA	***VDW***	***CE***	***PS***	***S***	***HB***
SASA	1					
VDW	0.7	1				
CE	0.64	0.29	1			
PS	0.61	0.26	0.99	1		
S	0.32	0.41	0.2	0.2	1	
HB	0.31	0.44	0.17	0.15	0.44	1
Variance inflation factors (VIF)						
	SASA	***VDW***	***CE***	***PS***	***S***	***HB***
VIF	3.31	2.43	47.72	45.03	1.35	1.38

Table C-10. Correlation matrixes and variance inflation factors (VIF) for the energy terms in SAMPDI. Terms with high correlation and VIF values (CC > 0.5 and VIF >4) are underlined.

Compound ID	Chemical NAME
Chembridge:78356510	(2R*,3R*)-3-amino-1'-[(1-isopropyl-4-piperidinyl)methyl]-2,3-dihydrospiro[indene-1,4'-piperidin]-2-ol
Chembridge:63954502	N-[3-(4-methylpiperazin-1-yl)butyl]-1-piperidin-1-ylcyclohexanecarboxamide
Chembridge:82584961	3-{[4-(3-phenylpropyl)-1,4-diazepan-1-yl]methyl}pyrrolidin-3-ol
Chembridge:14214075	1-methyl-4-{[1-(2-piperidin-2-ylethyl)-1H-1,2,3-triazol-4-yl]carbonyl}-1,4,9-triazaspiro[5.5]undecane
Chembridge:51996641	N-[(3-methyl-5,6,7,8-tetrahydro-2,7-naphthyridin-4-yl)methyl]-2-(3-pyrrolidinyl)benzamide
Chembridge:16017655	1-ethyl-4-{[3-(3-methylphenyl)-1-(4-methylphenyl)-1H-pyrazol-4-yl]methyl}piperazine
Chembridge:60958399	N-(2,5-dimethylphenyl)-3-{4-[1-(4-methylpiperazin-1-yl)ethyl]piperidin-1-yl}-3-oxopropanamide
Chembridge:74091868	2-(dimethylamino)-N-[2-methyl-2-(4-methyl-1-piperazinyl)propyl]-2-(3-methylphenyl)acetamide
Chembridge:58607167	(2R*,3R*)-1'-(N,N-dimethylglycyl)-3-(4-morpholinyl)-2,3-dihydrospiro[indene-1,4'-piperidin]-2-ol
Chembridge:71823596	2-methyl-N-{2-[3-(3-pyridinyl)-1H-1,2,4-triazol-5-yl]ethyl}-2,8-diazaspiro[4.5]decane-3-carboxamide
Chemdiv:L295-0542	N-(4-ethylbenzyl)-2-[2-(4-methylphenyl)-4-oxo-2,4-dihydro-5H-pyrazolo[3,4-d]pyrimidin-5-yl]acetamide
Chembridge:54664387	N-(2-methoxyethyl)-2-methyl-3-{[(4-methyl-1,4-diazepan-1-yl)acetyl]amino}benzamide
Chembridge:17062855	2-oxo-N-(4,5,6,7-tetrahydropyrazolo[1,5-a]pyrazin-2-ylmethyl)-1,2,3,4-tetrahydroquinoline-6-sulfonamide
Chembridge:41385842	8-(2-amino-6-methyl-4-pyrimidinyl)-2-[(3,5-dimethyl-4-isoxazolyl)methyl]-2,8-diazaspiro[4.5]decan-3-one
Chembridge:80203307	N-(1,2-diphenylethyl)-3-(4-methylpiperazin-1-yl)propanamide
Chembridge:44531433	2-{1-benzyl-5-[(4-ethylpiperazin-1-yl)methyl]-1H-1,2,4-triazol-3-yl}acetamide
Chembridge:48959062	(4R)-N-(1-ethylpiperidin-4-yl)-4-hydroxy-N-(2-pyridin-2-ylethyl)-D-prolinamide
Chembridge:89707730	(1R,9aR)-1-({[2-(1H-indol-3-yl)ethyl]amino}methyl)octahydro-2H-quinolizin-1-ol
Chembridge:93839784	6-[(6-methyl-2-pyridinyl)methyl]-N-[(4-phenyltetrahydro-2H-pyran-4-yl)methyl]-6-azaspiro[2.5]octane-1-carboxamide
Chembridge:46845072	3-[(dimethylamino)methyl]-1-{[3-(2,5-dimethylphenyl)-1H-pyrazol-4-yl]methyl}-3-piperidinol
Chembridge:19371006	2-(1-phenylcyclohexyl)-6-piperidin-4-ylpyrimidin-4(3H)-one
Chemdiv:G008-5368	5-(3,4-dimethylisoxazol-5-yl)-N-[(1-ethylpyrrolidin-2-yl)methyl]-2-methoxybenzenesulfonamide
Chembridge:35311030	4-{2-[(4-thiomorpholin-4-ylpiperidin-1-yl)carbonyl]phenyl}morpholine
Chembridge:81745001	4-{1-[2-(4-methylbenzyl)benzoyl]piperidin-4-yl}morpholine

Chembridge:15330541	N-[(1-ethyl-2-pyrrolidinyl)methyl]-N-{[1-(2-methoxyethyl)-4-piperidinyl]methyl}-2-(3-methyl-1H-pyrazol-1-yl)acetamide
Chembridge:9294680	N-[2-methoxy-5-({[2-(1-methyl-4-piperidinyl)ethyl]amino}sulfonyl)phenyl]acetamide
Chembridge:30440650	4-{[3-(3-methoxybenzoyl)-1-piperidinyl]methyl}-1,5-dimethyl-2-phenyl-1,2-dihydro-3H-pyrazol-3-one
Chembridge:61839572	N-(5-methyl-2,1,3-benzothiadiazol-4-yl)-3-(4,5,6,7-tetrahydropyrazolo[1,5-a]pyrazin-2-yl)propanamide
Chemdiv:D153-0063	1'-[(2,3,5,6-tetramethylphenyl)sulfonyl]-1,4'-bipiperidine
Chembridge:40078167	2-(3,9-diazabicyclo[4.2.1]non-9-yl)-N-(2-methylbenzyl)acetamide
Chembridge:10864439	N-[3-[(benzylamino)methyl]-5-(1-piperidinylcarbonyl)phenyl]methanesulfonamide
Chembridge:70812278	N-[(4-{[1-(cyclopropylmethyl)piperidin-4-yl]methyl}-5-oxomorpholin-2-yl)methyl]-1H-pyrazole-4-sulfonamide
Chembridge:52194258	5-{[{[3-(1-benzofuran-2-yl)-1-benzyl-1H-pyrazol-4-yl]methyl}(isopropyl)amino]methyl}-2-pyrrolidinone
Chembridge:38093279	9-{3-[(dimethylamino)methyl]benzoyl}-1-oxa-9-azaspiro[5.5]undecan-5-ol
Chembridge:40889750	N-[(1R)-1-(3-methoxyphenyl)ethyl]-2-methyl-6-piperidin-4-ylpyrimidin-4-amine
Chembridge:64562297	N-[(3-methyl-5,6,7,8-tetrahydro-2,7-naphthyridin-4-yl)methyl]-2-phenyl-2-(1H-tetrazol-1-yl)acetamide
Chembridge:23511553	5-({[2-(1-benzylpiperidin-2-yl)ethyl]amino}sulfonyl)thiophene-3-carboxamide
Chembridge:26076097	N-{2-[3-(2-methylphenyl)-1-pyrrolidinyl]ethyl}-2-oxo-1,2-dihydro-4-quinolinecarboxamide
Chembridge:80360739	5-{[3-(diphenylmethyl)-6,7-dihydroisoxazolo[4,5-c]pyridin-5(4H)-yl]carbonyl}-1H-1,2,4-triazol-3-amine
Chemdiv:C255-0943	2-methyl-N-(3-methylbutyl)-1-[2-(4-methylpiperazin-1-yl)ethyl]-5-oxoprolinamide
Chembridge:84959457	N-[(1-benzyl-1H-imidazol-2-yl)methyl]-N-methyl-5-(pyrrolidin-1-ylmethyl)-2-furamide
Chembridge:69031290	1'-{2-[(4-methylphenyl)thio]propanoyl}-1,4'-bipiperidine-4'-carboxamide
Chemdiv:S591-1521	N-[1-(2,1,3-benzothiadiazol-4-ylsulfonyl)-3-azetanyl]-N-ethyl-N-tetrahydro-2H-pyran-4-ylamine
Lifechem:F5017-0127	5-ethyl-N-((1-isopropylpiperidin-4-yl)methyl)thiophene-2-sulfonamide
Chemdiv:1956-0061	1-[(4-methylphenyl)sulfonyl]-3-(pyrrolidin-1-ylacetyl)imidazolidine
Chembridge:56480740	2-(2-pyridin-2-ylethyl)-8-(pyrimidin-5-ylmethyl)-2,8-diazaspiro[5.5]undecan-3-one
Chembridge:77982386	N-(cis-4-aminocyclohexyl)-3-(2-furyl)-4-phenylbutanamide
Chembridge:66763532	N-(1-tert-butylpyrrolidin-3-yl)-2-methyl-5-(1H-pyrazol-1-yl)benzenesulfonamide
Chembridge:20528061	N-(1-methyl-4-piperidinyl)-N-(2-phenylethyl)-1-azepanesulfonamide
Chembridge:78005408	N-{2,4-dimethyl-5-[(4-methylpiperazin-1-yl)methyl]benzyl}-3-(2-methyl-1H-imidazol-1-yl)propan-1-amine
Chembridge:39970175	2-(1-{[5-(pyrrolidin-1-ylmethyl)-2-thienyl]methyl}piperidin-4-yl)propan-2-ol
Chembridge:91750483	N-({1-[2-(2-methylphenyl)ethyl]-4-piperidinyl}methyl)-N-(3-pyridinylmethyl)ethanamine
Chembridge:91245078	2-[(4-benzylpiperazin-1-yl)methyl]-N-(1H-imidazol-2-ylmethyl)-N-methyl-1,3-oxazole-4-carboxamide
Chembridge:43919158	5-[(2-{4-[(dimethylamino)methyl]phenyl}-1-piperidinyl)methyl]-N,N-dimethyl-2-furamide

Chembridge:63028404	7-{2-[3-(dimethylamino)-2-hydroxypropoxy]-5-methoxybenzyl}-3,5,6,7,8,9-hexahydro-4H-pyrimido[4,5-d]azepin-4-one
Chemdiv:S720-1526	(2S,4S)-N-[3-(3,5-dimethyl-4-isoxazolyl)propyl]-4-phenoxy-1-tetrahydro-2H-pyran-4-yltetrahydro-1H-pyrrole-2-carboxamide
Chembridge:81027564	N-{[8-(3-methylbut-2-en-1-yl)-1-oxa-8-azaspiro[4.5]dec-2-yl]methyl}-3-(2-oxoazepan-1-yl)propanamide
Chembridge:45596927	N-{[1-(4-morpholinyl)cyclohexyl]methyl}-4,5,6,7-tetrahydro-1H-imidazo[4,5-c]pyridine-4-carboxamide
Chembridge:11156722	1-methyl-4-[3-(5-methyl-1H-tetrazol-1-yl)propanoyl]-1,4,9-triazaspiro[5.6]dodecan-10-one
Chembridge:65552572	3-{[({[1-(2-methoxyethyl)pyrrolidin-3-yl]methyl}amino)carbonyl]amino}-4-methylbenzenesulfonamide
Chembridge:61821585	N-[2-(dimethylamino)ethyl]-2-(2,4-dimethyl-6-oxo-1,6-dihydropyrimidin-5-yl)-N-(2-methylbenzyl)acetamide
Chembridge:97069383	7-(1H-imidazol-4-ylmethyl)-N-(2-phenylethyl)-6,7,8,9-tetrahydro-5H-pyrimido[4,5-d]azepin-4-amine
Chembridge:62154375	2-(3-{[3-(2-pyridinyl)-1-azetidinyl]methyl}phenoxy)acetamide
Chembridge:31484474	8-[2-(dimethylamino)ethyl]-2-(2-phenylethyl)-2,8-diazaspiro[5.5]undecan-3-one
Chembridge:40851669	N-[3-(4-methylpiperazin-1-yl)butyl]-4-(4H-1,2,4-triazol-4-yl)benzamide
Chembridge:76608419	2-(2-isopropyl-1H-benzimidazol-1-yl)-N-[(7S,8aS)-2-methyloctahydropyrrolo[1,2-a]pyrazin-7-yl]acetamide
Chembridge:89145681	N-ethyl-5-[(4-ethylpiperazin-1-yl)methyl]-N-(2-methylbenzyl)isoxazole-3-carboxamide
Chembridge:21627035	1-methyl-5-[N-methyl-N-(1-methylpiperidin-4-yl)glycyl]-4,5,6,7-tetrahydro-1H-pyrazolo[4,3-c]pyridine-3-carboxamide
Chembridge:75582364	N-[(3-methyl-5,6,7,8-tetrahydro-2,7-naphthyridin-4-yl)methyl]-2-(1H-tetrazol-5-yl)benzamide
Chembridge:98517584	5,7-dimethyl-6-[3-oxo-3-(4-pyridin-3-yl-1,4,6,7-tetrahydro-5H-imidazo[4,5-c]pyridin-5-yl)propyl][1,2,4]triazolo[1,5-a]pyrimidine
Chembridge:25787349	(1R,9aR)-1-{[bis(2-furylmethyl)amino]methyl}octahydro-2H-quinolizin-1-ol
Chembridge:13033623	(2-{4-[2-(2,5-dimethylphenoxy)propanoyl]piperazin-1-yl}ethyl)dimethylamine
Chembridge:88031006	N-[(4-{[1-(cyclopropylmethyl)piperidin-4-yl]methyl}-5-oxomorpholin-2-yl)methyl]-5-oxo-4,5-dihydro-1H-1,2,4-triazole-3-carboxamide
Chembridge:60618842	2-(1-isopropylpiperidin-4-yl)-N-(2-phenylethyl)-N-(pyridin-2-ylmethyl)acetamide
Chembridge:73314389	[1-(2,1,3-benzoxadiazol-4-ylmethyl)-3-(2-phenylethyl)-3-piperidinyl]methanol
Lifechem:F6178-7296	N1-(2-cyanophenyl)-N2-((1-(tetrahydro-2H-pyran-4-yl)piperidin-4-yl)methyl)oxalamide
Chemdiv:SA46-2193	{7-[benzyl(methyl)amino]-5-oxa-2-azaspiro[3.4]oct-2-yl}(1H-indazol-3-yl)methanone
Chembridge:13675749	N-[(4-{[1-(cyclopropylmethyl)piperidin-4-yl]methyl}-5-oxomorpholin-2-yl)methyl]-2-(3,5-dimethyl-1H-pyrazol-1-yl)acetamide
Chembridge:19197720	(1R,9aR)-1-({[2-(2-methyl-1H-imidazol-1-yl)benzyl]amino}methyl)octahydro-2H-quinolizin-1-ol
Chembridge:76559661	2-morpholin-2-yl-N-[2-(2-phenoxyphenyl)ethyl]acetamide

Table C-11. Final selected compounds for experimental verification.