

T.R.
GEBZE TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

**ACCURATE FREE ENERGY CALCULATIONS FOR
BINDING OF SMALL MOLECULES TO SARS-COV-2 MAIN
PROTEASES, HIV-1 PROTEASES AND JNK1 PROTEIN
KINASES WITH MOLECULAR DYNAMICS SIMULATIONS
USING MOLECULAR MECHANICS, QUANTUM
MECHANICS AND MACHINE LEARNING METHODS**

EBRU AKKUŞ
A THESIS SUBMITTED FOR THE DEGREE OF
MASTER OF SCIENCE
DEPARTMENT OF BIOENGINEERING

GEBZE
2023

T.R.
GEBZE TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

**ACCURATE FREE ENERGY CALCULATIONS FOR
BINDING OF SMALL MOLECULES TO SARS-COV-
2 MAIN PROTEASES, HIV-1 PROTEASES AND
JNK1 PROTEIN KINASES WITH MOLECULAR
DYNAMICS SIMULATIONS USING MOLECULAR
MECHANICS, QUANTUM MECHANICS AND
MACHINE LEARNING METHODS**

EBRU AKKUŞ
**A THESIS SUBMITTED FOR THE DEGREE OF
MASTER OF SCIENCE
DEPARTMENT OF BIOENGINEERING**

THESIS SUPERVISOR
ASSIST. PROF. PINAR PIR
II. THESIS SUPERVISOR
ASSOC. PROF. ABDULKADIR KOÇAK

GEBZE

2023

T.C.
GEBZE TEKNİK ÜNİVERSİTESİ
FENBİLİMLERİ ENSTİTÜSÜ

**MM, QM VE ML METOTLARINI
KULLANARAK MOLEKÜLER DİNAMİK
SİMÜLASYONLARIYLA KÜÇÜK
MOLEKÜLLERİN SARS-COV-2 ANA
PROTEAZLAR VE HIV-1 PROTEAZLARA
BAĞLANMASININ HASSAS SERBEST
ENERJİ HESAPLAMALARI**

EBRU AKKUŞ
YÜKSEK LİSANS TEZİ
BİYOMÜHENDİSLİK ANABİLİM DALI

DANIŞMANI
DR. ÖĞR.ÜYESİ. PINAR PİR
II. DANIŞMANI
DOÇ. DR. ABDULKADİR KOÇAK

GEBZE
2023



YÜKSEK LİSANS JÜRİ ONAY FORMU

GTÜ Fen Bilimleri Enstitüsü Yönetim Kurulu'nun 13/04/2023 tarih ve 2023/24 sayılı kararıyla oluşturulan jüri tarafından 20/04/2023 tarihinde tez savunma sınavı yapılan Ebru AKKUŞ'un tez çalışması Biyomühendislik Anabilim Dalında YÜKSEK LİSANS tezi olarak kabul edilmiştir.

JÜRİ

ÜYE
(TEZ DANIŞMANI) : Dr.Öğr.Üyesi PINAR PİR

ÜYE : Prof. Dr. CANAN ATILGAN

ÜYE :Doç. Dr. MEHMET ÖZBİL

ONAY

Gebze Teknik Üniversitesi Fen Bilimleri Enstitüsü Yönetim Kurulu'nun
.../.../.../tarih ve/.....sayılı kararı.

İMZA/MÜHÜR

SUMMARY

Machine learning (ML) tools have emerged in all aspects of science including the potential energy surfaces at quantum mechanical (QM) accuracy. In particular, ML tools could be utilized for estimation of free energy changes upon small organic molecules' binding to bio-macromolecules such as proteins with higher accuracy and lower costs than the conventional methods. A very recent neural network potential (NNP) called ANI (Accurate NeurAl network engINe for Molecular Energies) employs active learning to determine the overall potential energy surfaces of organic molecules using C, H, O, N, F, Cl, and S atoms at the accuracy of the wb97x/6-31G* level.

In this thesis, the success of ANI potentials has been explored in the prediction of binding free energies (BFE) of small organic compounds (inhibitors) to enzymes. BFEs were estimated by the adoption of ANI-2x produced energies in the formalisms of two different strategies: linear interaction energy (LIE) and Molecular Mechanics Poisson Boltzmann Surface Area (MM/PBSA). Although developed for small organic covalent compounds, we show that ANI-2x potentials could indeed be used to produce non-covalently interacting systems by means of single-point interaction energies between ligand–protein and ligand–solvent pairs sampled by molecular dynamics (MD) simulations. Our results on 54 protein–ligand complexes composed of Covid-19 main proteases, HIV-1 proteases, and JNK1 protein kinases show that the LIE and MM/PBSA modifications by ANI (ANI_LIE and ANI_PBSAe) can be much more accurate and fast with a correlation of $R=0.87-0.92$ to the experimental binding free energies respectively, outperforming current end-state methods.

Keywords: Molecular Dynamics Simulation, Binding Free Energy, Quantum Mechanics, Machine Learning, SARS-CoV-2 main protease, HIV-1 protease, JNK1 protein kinase.

ÖZET

Makine öğrenimi (ML), potansiyel enerji yüzeylerini kuantum mekaniği (QM) doğruluğunda hesaplayabilmek de dahil olmak üzere bilimin her alanında yer almaya başlamıştır. Özellikle küçük organik moleküllerin proteinler gibi biyo-makromoleküllere bağlanması ile ilişkili serbest enerji değişimlerinin geleneksel yöntemlerden daha yüksek doğruluk ve daha düşük maliyetlerle birlikte tahmin edilmesi için kullanılabilir. Son zamanlarda ANI (Moleküler Enerjiler için Doğru Nöral ağ motoru) olarak adlandırılan çok yeni bir nöral ağ potansiyeli (NNP) C, H, O, N, F, Cl ve S atomlarını kullanan organik moleküllerin potansiyel enerji yüzeylerini wb97x/6-31G* seviyesi doğruluğunda belirlemek için aktif öğrenmeyi kullanır.

Bu tezde, küçük organik bileşiklerin (inhibitörler) enzimlere bağlanma serbest enerjilerinin (BFE) tahmininde ANI potansiyellerinin başarısı araştırılmıştır. BFE'ler, ANI-2x tarafından üretilen enerjilerin Lineer etkileşim enerjisi (LIE) ve Moleküler Mekanik Poisson Boltzmann Yüzey Alanı (MM/PBSA) formüllerine yapılan adaptasyonlar ile tahmin edilmiştir. Burada ANI-2x potansiyellerinin küçük organik kovalent bileşikler için geliştirilmiş olmasına rağmen, moleküler dinamik (MD) simülasyonları tarafından örneklenen ligand-protein ve ligand-çözücü çiftleri arasındaki tek noktalı etkileşim enerjileri aracılığıyla kovalent olarak etkileşime girmeyen sistemler için de kullanılabileceğini gösteriyoruz. Covid-19 ana proteazlar, HIV-1 proteazlar ve JNK1 protein kinazlardan oluşan 54 protein-ligand sistemlerine ilişkin sonuçlarımız, LIE ve MM/PBSA yaklaşımları yerine ANI ile yapılan modifikasyonların (ANI_LIE ve ANI_PBSAe) çok daha doğru ve hızlı sonuçlar üretebileceğini göstermektedir. Tahminlerin, deneysel bağlanma serbest enerjileri ile korelasyonlarına bakıldığında da bunlar sırasıyla $R=0.87-92$ 'dir ve mevcut son-durum yöntemlerinden çok daha iyi performans gösterir.

Anahtar Kelimeler: Moleküler Dinamik Simülasyonları, Bağlanma Serbest Enerjisi, Quantum Mekanik, Makine Öğrenmesi, Covid-19 ana proteaz, HIV-1 proteaz, JNK1 protein kinaz.



ACKNOWLEDGEMENTS

I would like to extend my heartfelt gratitude to everyone who contributed to this study.

Especially, I would like to express my sincere gratitude to my supervisor, Abdulkadir Kocak, for his constant support and direction throughout many projects. His educated advice and constructive criticism were beneficial in supporting me in achieving my goals.

And I would like to express my gratefulness to my supervisor Pınar Pir for giving me the chance to work with feelings freely. I could not do anything if Pınar Pir not directed me about many beginnings.

I also thank to the Tubitak ULAKBIM because all calculations that are reported in this thesis were performed at TUBITAK ULAKBIM, High Performance and Grid Computing Center (TRUBA resources). This work was supported by Scientific and Technological Research Council of Turkey-TÜBİTAK (project number: 120Z732).

TABLE OF CONTENTS

	<u>Page</u>
SUMMARY	v
ÖZET	vi
ACKNOWLEDGEMENTS	viii
TABLE OF CONTENTS	ix
LIST OF ABBREVIATIONS AND ACRONYMS	xi
LIST OF FIGURES	xiv
LIST OF TABLES	xvii
 1. INTRODUCTION	 1
1.1. Protein-ligand Interactions and Mechanism for Drug Discovery	1
1.2. Computational methods in drug discovery	4
1.2.1. Molecular Docking	5
1.2.2. Classical Molecular Dynamics (MD) Simulations	6
1.2.3. Quantum Mechanical Calculations	8
1.2.4. QM/MM Methods	9
1.2.5. ML methods approaching to DFT accuracy	10
1.3. Conventional Binding Free Energy Calculations	11
1.3.1. End-state Methods	12
1.3.2. Alchemical Perturbation Methods	13
1.3.3. Physical Pathway Methods	18
2. MODIFICATION OF LIE BY ANI FOR BINDING FREE ENERGY CALCULATIONS FROM MD SIMULATIONS	19
2.1. Introduction	19
2.2. Theory	22
2.2.1. Linear Interaction Energy (LIE)	22
2.2.2. ANI/D3 Linear Interaction Energy (ANI_LIE)	24
2.3. Computational Methods	26
2.3.1. Dataset Selection: SARS-CoV-2 main proteases, HIV-1 proteases, JNK1 Proteins	26
2.3.2. System Preparation	28
2.3.3. MD Simulations	28
2.3.4. Free Energy Calculations	29
2.4. Results and Discussion	30
2.4.1. Workflow	30

2.4.2. ANI Results	32
2.4.3. Benchmarking with Other Methods	41
2.5. Summary and Conclusion	46
3. MODIFICATION OF MM/PBSA BY ANI : PERFORMANCE	47
IMPROVEMENTS OF MM-PBSA BY ANI-ML POTENTIALS, MC BASED SURFACE AREA/VOLUME, AND TWO VALUED INTERIOR DIELECTRIC CONSTANT	
3.1. Introduction	47
3.2. Theory	49
3.2.1. MM-P(G)BSA	49
3.2.2. ANI Replacement with MM terms	54
3.3. Computational Methods	55
3.3.1. System Preparation	55
3.3.2. MD Simulations	55
3.3.3. PBSA Calculations	56
3.4. Results and Discussion	56
3.4.1. MMPBSA by SASA and/or SAV	56
3.4.2. ASA/PCAV(POAV) in replacement of SASA/SAV	60
3.4.3. ANI in replacement of MM terms (ANI_PBSA)	62
3.4.4. Bivalent ϵ_{int} in MMPBSA/ANIPBSA calculations	65
3.5. Summary and Conclusion	67
4. CONCLUSION	69
REFERENCES	71
BIOGRAPHY	87
APPENDICES	88

LIST OF ABBREVIATIONS AND ACRONYMS

<u>Abbreviations</u>	<u>Explanation</u>
<u>and Acronyms</u>	
λ	: Number of state
Δ	: Delta
ΔG	: Gibbs free Energy change
$\Delta\Delta G$	: Relative Binding Free Energy
ABFE	: Absolute Binding Free Energy
ASA	: Accessible Surface Area
ANI	: Accurate Neural Network Engine for Molecular Energies
BAR	: Bennett Acceptance Ratio
BFE	: Binding Free Energy
DFT	: Density Functional Theory
FE	: Free Energy
FEP	: Free Energy Perturbation
FM	: Funnel-Metadynamics
H	: Hamiltonian operator
HIV-1	: Human immunodeficiency virus type 1
IC50	: Half maximal inhibitory concentration
JNK1	: c-Jun N-terminal kinases 1
K_i	: Inhibition constant
L	: Ligand

LIE	:	Linear Interaction Energy
LPB	:	Ligand Protein Binding
MBAR	:	Multistate Bennett Acceptance Ratio
MC	:	Monte Carlo
MD	:	Molecular Dynamics
MM	:	Molecular Mechanics
MMGBSA	:	Molecular Mechanics Generalized Born Surface Area
MMPBSA	:	Molecular Mechanics Poisson Boltzmann Surface Area
M ^{pro} ,3Cl ^{pro}	:	Main protease
NMR	:	Nuclear magnetic resonance
NNP	:	Neural Network Potential
NPT	:	Isobaric-isothermal ensemble
NVT	:	Canonical ensemble
P	:	Protein
PCAV	:	Probe Centered Accessible Volume
PDB	:	Protein Data Bank
PL	:	Protein-ligand complex
PL ^{pro}	:	Papain-like protease
PME	:	Particle Mesh Ewald Method
POAV	:	Probe Occupiable Accessible Volume
PP	:	Physical Pathway
QM	:	Quantum Mechanics

QM/MM	:	Quantum Mechanics/Molecular Mechanics
RBFE	:	Relative Binding Free Energy
SARS-CoV	:	Severe acute respiratory syndrome coronavirus
SASA	:	Solvent Accessible Surface Area
SAV	:	Solvent Accessible Volume
SEQM	:	Semi-empirical Quantum Mechanics
SMD	:	Steered Molecular Dynamics
TI	:	Thermodynamic Integration
TMPRSS2	:	Cellular transmembrane serine protease 2
US	:	Umbrella Sampling
WHAMM	:	Weighted Histogram Analysis Method

LIST OF FIGURES

<u>Figure No:</u>	<u>Page</u>
1.1: Representation of binding of a ligand to a protein.	2
1.2: The representation of interaction types exist in force fields to be considered.	7
1.3: Representation of binding strength at different lambda steps.	14
1.4: Representation of the alchemical free energy calculations using a model of the thermodynamic cycle.	15
1.5: Defining several MD techniques mathematically.	17
2.1: deepQM General Workflow for Performing EndState Binding Free Energy Calculations (The first step consists of extracting individual groups and converting to xyz file format. The second step is the calculator. Single or multicalculators can be selected. Finally, the results were printed and combined to predict binding free energy).	31
2.2: Truncated structure in the DFT calculation as a comparison to ANI, in which all the PL system can also be calculated. All the residues were saturated with H atoms to make the structure neutral. All H atoms were ommitted from structuress for better view. The table shows which residues included in the truncated structure.	32
2.3: Absolute deviations of the average P–L interaction energies for selected complexes calculated by ANI, considering 1000 frames and 10 ns simulations as the converged value a) as a function of the number of frames to average and b) as a function of simulation time using only 100 frames in each time interval.	34
2.4: ANI-predicted average interaction energies and classical electrostatic interaction energies between the protein and ligand from PLS simulations are correlated while the energies from Grimme’s D3 functions are correlated to van der Waals interactions.	35
2.5: ANI interaction energies calculated from eq 2.10 with $\beta =$	36

	0.10639 $\pm$ 0.0164 and $\gamma = -4.9875 \pm 0.881$ using average of 100 snapshots of 10 ns MD simulations for 54 protein–ligand complexes, showing 0.87 correlation to the experimental values. The bar lines show uncertainties. The bottom plots show the fits of individual protein families as CoV, HIV-1, and JNK-1.	
2.6:	ANID3 interaction energies calculated from the equation 2.5 with $\alpha = -0.0254 \pm 0.0282$, $\beta = 0.1240 \pm 0.0212$ and $\gamma = -5.5692 \pm 0.7821$ using average of 1000 snapshots of 10 ns MD simulation for 54 protein ligand complexes, showing 0.88 correlation to the experimental values. The bar lines show the mean absolute error. The bottom plots show the fits of individual protein families as CoV, HIV-1 and JNK-1.	37
2.7:	Illustration for the representative structures in protein + ligand + complete/reduced water (PLS) systems. The LIE calculations were performed using reduced water to ease memory cost of the computations. The results are almost identical in both cases.	38
2.8:	LIE calculated by ANI and ANID3 using a) eq 2.12 and b) eq 2.13, in which the solvent effects around the ligand were included in the calculations.	40
2.9:	The LIE calculated by MM electrostatic and van der Waals interactions using equation 2.9. The preassigned parameters available in Gromacs software were used. The bottom plots show the correlation in each individual protein family.	42
2.10:	LIE calculated by MM electrostatic and van der Waals interactions using eq 2.9. The parameters were determined from the fit. The bottom plots show the correlation in each individual protein family.	43
2.11:	MMPBSA results show low correlation due to failure to predict BFEs in HIV-1 protease. The bottom plots show the correlation in each individual protein family.	44
2.12:	MMGBSA results showing higher correlation than LIE, but still lower than ANI_LIE. The bottom plots show the correlation in each individual protein family.	45

2.13:	BAR results for CoV and HIV compared with experimental results when they are treated together or individually.	46
3.1:	The experimental and calculated binding free energies by MMPBSA using SASA only/SPT regimes for non-polar contribution where $\epsilon_{\text{int}}=2.1$ for electrostatic and polar solvation terms.	58
3.2:	The relationship between SASA (or SAV enclosed by SASA) and E_{disp} (calculated by integration of SASA). The values are of averages of PL-P-L differences over 100 MD snapshots.	59
3.3:	Illustration of pore volume produced by Zeo++ showing a) covered regions in PCAV and POAV definition b) how we calculated the molecular surface from pore volume.	61
3.4:	The comparison of MMPBSA_E and ANI_PBSAe by using SASA only and SPT regimes.	63
3.5:	The comparison of the single valued vs two valued ϵ_{int} for coulombic (P-L) and polar solvation term, respectively.	66

LIST OF TABLES

<u>Table No:</u>	<u>Page</u>
2.1: Absolute Energies of Truncated Protein + Ligand Systems (HIV Protease, PDB ID = 5IVS) from MD Simulations (in eV).	31
2.2: MM energy terms and LIE for complete vs reduced systems, showing reducing system hardly affects the results.	37
2.3: Coefficient values derived from equations 2.4-2.7.	40
3.1: The experimental and calculated binding free energies by MMPBSA using SASA only/SPT regimes for non-polar contribution where $\epsilon_{int}=2.1$ for electrostatic and polar solvation terms.	56
3.2 MMPBSA_E by SASA only and MMPBSA_Ex (extended) by SPT models for non polar contribution to solvation free energy.	57
3.3: MMPBSA calculated by Zeo++ produced SA terms for non polar contribution to solvation free energy.	59
3.4: ANI_PBSA calculated by AMBER produced SA terms for non polar contribution to solvation free energy.	61
3.5 ANI_PBSA calculated by Zeo++ produced SA terms for non polar contribution to solvation free energy.	61
3.6 ANI_PBSAe calculated by SPT model for non polar contribution to solvation free energy.	62
3.7 Bivalent epsilon calculated by MMPBSA produced SA terms for non polar contribution to solvation free energy.	64
3.8 Bivalent epsilon calculated by Zeo++ produced SA terms for non polar contribution to solvation free energy.	64

1. INTRODUCTION

In this thesis, linear interaction energy (LIE) and Molecular Mechanics Poisson Boltzmann Surface Area (MMPBSA) methods will be the main focus. It is aimed to correct the Molecular Mechanics (MM) energy terms in these methods by implementation of Quantum Mechanics (QM) accurate energies produced by ANI-2x in the corresponding equations to predict BFEs more accurately yet without too much of computational cost.

This thesis is organized in the following way: In the first chapter (the introduction chapter), a general background information about kinetics and thermodynamics of protein-ligand binding is given and then several computational methods are also briefly mentioned in. In the second chapter, the adoption of ANI-ML potentials in LIE method for BFEs is introduced. The third chapter describes ANI-ML potentials replaced in the MM term of MMPBSA calculations. The conclusion and future directions are discussed in the last chapter.

1.1. Protein-ligand Interactions and Mechanism for Drug Discovery

To carry out their biological tasks, proteins interact with a variety of molecules, including other proteins, peptides, substrates, and small molecules (referred to as ligands in this study) that they can bind to. Studies on interactions between proteins and ligands have opened a number of new perspectives for investigating structural dynamics at the molecular level. These interactions can be covalent or non-covalent. The effectiveness of biological systems depends in part on how strongly ligands attach to their target proteins [1].

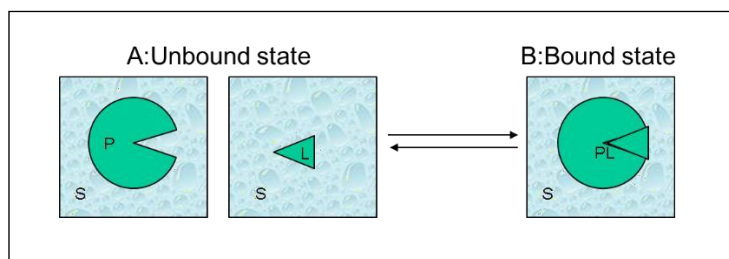
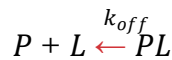


Figure 1.1: Representation of binding of a ligand to a protein.

The kinetic mechanism can be described by association and dissociation events that are dependent on molar concentrations of unbound states, protein (P) and ligand (L) free can also be called as reactants and bound state protein-ligand (PL) together can also be called as products. The equation [1] will be:



The k_{on} and k_{off} are kinetic rate constants that are related with the rate of bound and unbound states and their unit's $\frac{1}{M.s}$, $\frac{1}{s}$ respectively. An equilibrium is the condition in which the P, L, and PL are at a point where there is no obvious change in the system anymore. After reaching equilibrium, the same number will be produced and dissociated at any given time (i.e, the rate of the forward and reverse reactions will be the same). When the system reaches equilibrium, the equation [1] will be:

$$k_{on}[P][L] = k_{off}[PL] \quad (1.2)$$

The brackets represent concentrations of each molecule. And the association (binding) constant K_b is driven by ratio of k_{on} and k_{off} , the dissociation (unbinding) constant K_d is driven by the reverse of this:

$$K_b = \frac{k_{on}}{k_{off}} = \frac{[PL]}{[P][L]} = \frac{1}{K_d} \quad (1.3)$$

Unit of these constants is M^{-1} . If the PL binding is strong, K_b will be high and K_d will be low.

Spontaneous events require the system's Gibbs free energy change, ΔG , to be negative. However, when this energy becomes zero. The equation [1] is:

$$\Delta G = \Delta H - T\Delta S = \Delta G^0 + RT\ln Q_b \quad (1.4)$$

$$\Delta G_{bind}^0 = -RT\ln(K_b \cdot C^0) = RT\ln(K_d/C^0) \quad (1.5)$$

where C^0 is the standard molar concentration of water (for aqueous systems), R represents gas constant ($8.3144598 \frac{\text{J}}{\text{mol.K}}$) and T is absolute temperature (K). As K_b increases ΔG_{bind}^0 becomes more negative. Therefore, magnitude of the negative ΔG_{bind}^0 determines how strongly ligand binds to a protein.

In drug discovery, one of the main strategies is to find a small molecule that target a specific protein and activate or inhibit its function. In this thesis this inhibition property through competitive inhibition mechanism will be focus. In this mechanism, the inhibitor at the tolerable concentrations must have more binding affinity than the protein's natural substrate. In the case of inhibitor, the binding constant is expressed by either K_i and half maximal inhibitory concentration (IC50) values. When the inhibition is competitive, K_i shares the same mean with K_d . Thus, the two constants generally describe the same reaction and K_d can be replaced with the inhibition constant (K_i). In addition to this, for competitive inhibition, K_i and IC50 also have the relation as:

$$IC50 = K_i \left(1 + \frac{[S]}{K_m} \right) \quad (1.6)$$

K_i is the inhibitory concentration necessary to dissociate 50% of the PL complex in an equilibrium state, $[S]$ is concentration of enzyme's substrate, and K_m is the Michaelis Menten constant of the enzyme [2]. The IC50 value, on the other hand, assesses an inhibitor's efficacy by calculating the inhibitor concentration required to suppress the rate of the PL binding process by 50% under a given test setting. To put it another way, K_i is a measurement of how tightly the inhibitor binds to the PL complex, whereas IC50 is a measure of how well an inhibitor suppress the PL binding reaction [3].

$$\Delta G_{bind}^0 \propto RT \ln(IC50) \quad (1.7)$$

In principle, ΔG_{bind}^0 can be calculated for any ligand to estimate the binding strength and thus to understand whether the process of binding is thermodynamically favored or not. It's crucial to remember that, regardless of the pathway separating these states (bound and unbound states), the beginning and final states determine ΔG_{bind}^0 and the BFE is a state function of a system. Therefore, it does not care about which path is preferred. However, as shown in Figure 1.1, binding of ligand in solution is not a simple one step process but rather complex process involving the solvation of protein-ligand complex, isolated protein, and isolated ligand. The existence of explicit solvent molecules complicates the calculations, since MD simulations are stochastic and sufficient sampling is a limitation. The protein also has dynamic properties which are affected when it is isolated (apo) or complexed (holo). A careful consideration of conformational entropy of the protein must be considered. Moreover, the classical MM force fields have somewhat lower accuracy when describing system when compared to QM energies. The classical MD works only on ensemble averages rather than single frame (of P, L and PL complex). Thus, calculation of only a single snapshot at the end-states (bound and unbound states) is not suitable to accurately predict the binding free energy. To calculate the BFEs, the ensemble averages must be utilized in a way that takes the system from the initial state to a final state during the MD. The events which evolve the system from initial to final state occur at equilibrium, these processes are called reversible, and the work produced by the system is maximized. Thus, a good complete thermodynamic cycle, which mimics the reversible process, must be created to calculate the free energy change. There have been several methods with devoting to the different thermodynamic cycles have been developed to evolve the system from initial to final state by infinite changes.

1.2. Computational Methods in Drug Discovery

Crystallographic studies show that protein flexibility is crucial in L binding, but the cost and effort involved make them impractical for many researchers. Thus, computational techniques that can estimate protein structures, properties, and

motions have become popular. These can be divided into five categories as Molecular Docking, Molecular Dynamics (MD) Simulations, Quantum Mechanical (QM) Calculations, Combination of Methods (i.e., QM/MM) and ML methods. While the docking approach can be beneficial to find binding mode, others are more related with binding strength. Numerous theoretical methods are devoted to calculating this strength to create more effective candidates that can take the place of the protein's original substrate. When there is little or no information about the binding site, molecular docking studies can help to place the ligand into the most probable site and correct orientation (pose). However, these studies can be inaccurate when compared to experimental results in terms of binding affinities. Binding free energy (BFE, ΔG_{bind}) calculations at various levels ranging from molecular mechanics (MM) to quantum mechanics (QM) can be used to estimate the binding strength if the crystal structure of the PL complexes is known. In MM based BFE calculations, a specific set of force fields are used in MD simulations to create the conformational ensembles. Following an MD simulation, the PL system can be subjected to BFE computations throughout the equilibrium or non-equilibrium evolution of the system from bound to unbound state.

Although MD can accurately simulate the movement of atoms in many systems, it is not capable of modeling chemical reactions that include changes in electronic structures, such as charge transferring (polarization), bond formation and breaking. QM is the ideal method for describing electronic structure changes, but it brings high computational demands so it can be used for small systems with only a few hundred atoms. Despite this limitation, QM methods are still widely used due to their accuracy and ability to provide detailed insight into chemical reactions and processes. In 1976, Warshel and Levitt introduced the hybrid quantum QM/MM approach [4] for studying large biomolecules. This method was further developed and evaluated by Karplus, who coupled semi-empirical QM methods with MM force fields [5]. The QM/MM approach has become a widely used tool for studying complex biological systems and chemical reactions [6].

1.2.1. Molecular Docking

Molecular docking is well-known in computational-aided drug design (CADD) for prediction of binding affinity/mode between P and L. The important aim of

docking is to find the most stable ligand orientation in target protein as well as binding trend among ligands (although the success for BFE is rather low due to coarse nature of the method [7]). Diversity of docking software/algorithms rely on two key concepts: a search algorithm and a scoring function. The algorithm will search through the site of possible poses of ligand to find ones that are more stable and have a low energy by using sampling techniques like Monte Carlo (MC) and genetic algorithms [8], [9]. Once docking calculation is completed the ranks of all possible conformations about ligand binding by using scoring function will be provided. The most stable poses will be given top of the result with the highest scores.

The strongest feature of docking in general comes from the calculation speed. It is genuinely fast and can perform the millions of compounds at low computational time and it will be useful for virtual screening [10]. The details of docking are beyond the scope of this thesis and will not be discussed further.

1.2.2. Classical Molecular Dynamics (MD) Simulations

Molecular dynamics (MD) simulations are performed to model a system at the most realistic definition. Building a computer model of the chemical system using information from sources like nuclear magnetic resonance (NMR), crystallography, or homology modeling is the first step in the general MD approximation process. The forces operating on each atom are then estimated using this model by applying Newton's equation of motion and considering both bonded and non-bonded interactions. Chemical bonds and atomic angles are approximated by virtual springs, while dihedral angles are represented by a sinusoidal function that quantifies the energy differences between all conceivable conformations throughout the simulation. The model also accounts for non-bonded forces, covers van der Waals interactions and electrostatic interactions that are modeled using the Lennard-Jones potential and Coulomb's law, respectively [11].

All these different interaction terms between a pair of atoms (i.e., pairwise interactions) are described by using a force field. The force field is a series of equations that expresses a system's potential energy as a function of the atomic locations. The potential energy is required to solve the classical equations of motion (Velocity Verlet algorithm) [12]:

$$q(t + \Delta t) = q(t) + \Delta t v(t) - \frac{\Delta t^2}{2m} \frac{dV(t)}{dq} \quad (1.8)$$

$$v(t + \Delta t) = v(t) - \frac{\Delta t}{2m} \left(\frac{dV(t + \Delta t)}{dq} + \frac{dV(t)}{dq} \right)$$

The q is coordinate and v is velocity that are changeable through simulation time. The positions of the atoms are related with bond stretching, angle bending, dihedral (torsion) and non-bonded interaction energies are given at Figure 1.2 as example in one molecule. The green, red, blue lines represent bond, angle, dihedral as the bonded interactions and the pink line represents non-bonded interactions which are van der Waals and coulombic interactions. There are several different types of force fields available, and choosing the appropriate one is an important step in setting up an MD simulation. Figure 1.2 represents example of interactions that are defined in force fields.

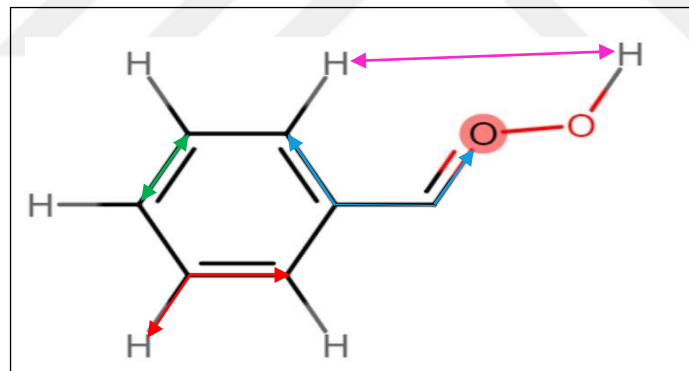


Figure 1.2: The representation of interaction types that exist in the force fields to be considered.

And the overall equation will be as in equation 1.9 [13]:

$$\begin{aligned}
E_{bonded} = & \sum_{bond} K_r (r - r_{eq})^2 + \sum_{angles} K_\theta (\theta - \theta_{eq})^2 + \\
& \sum_{dihedrals} K_\phi [1 + \cos(n\phi - \gamma)] + \sum_{impropers} K_\omega (\omega - \omega_{eq})^2
\end{aligned}
\tag{1.9}$$

$$E_{nonbonded} = \sum_{i < j}^{atoms} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^6} + \frac{q_i q_j}{4\pi\epsilon R_{ij}} \right]$$

$$E_{tot} = E_{bonded} + E_{nonbonded}$$

Bond and angle definitions can be explained by harmonic potentials. These potentials cannot describe bond formation or breaking cases. r is the bond length, θ is the atomic angle, ϕ is improper dihedral angle and ω is the improper dihedral angle. K_ϕ , n , and γ are parameters that represents the shape of potential of the dihedral angle. Between atoms i and j , R_{ij} is the distance. K_r , K_θ , K_ϕ and K_ω are force constants. r_{eq} , θ_{eq} and ω_{eq} are equilibrium positions. n is multiplicity, and γ is phase shifting term and A_{ij} and B_{ij} are the distance at which the potentials reach their minimum and q_i , q_j are the charges on the respective atoms, finally ϵ is dielectric constant [6]. All these parameters are combined to build up the potential energy of the system, that mimics experimental 3D structure of the system.

1.2.3. Quantum Mechanical Calculations

The behavior of matter and energy at the subatomic level is explained by the theory of quantum mechanics. The properties of subatomic particles, atoms, molecules, and molecular assemblies are only a few of the entities that are described by the fundamental theory of physics. The correlations between the molecules' physical properties that come from their wave-like nature are known as quantum coherence [14] so the laws of quantum coherence control interactions between molecules which cause them to behave like waves.

Quantum mechanical calculations are a type of computational techniques to predict the behavior of biomolecular systems at subatomic level by approximation of wavefunction and solution of the Schrödinger equation [15], [16]. A bound electron cannot have any energy and cannot be in any particular place in space, according to quantum physics [17]. These characteristics of a bound electron can be found using the time-independent Schrödinger equation:

$$\hat{H}\psi(x) = \hat{T}\psi(x) + \hat{V}\psi(x) \quad (1.10)$$

$\hat{H}$ stands for the Hamiltonian operator, which is sum of kinetic energy operator, $\hat{T}$ and potential energy operator, $\hat{V}$. With Schrödinger equation, understanding of electron movements and making predictions about molecular structures or energies become to be possible. However, as the equation can only be precisely solved for a hydrogen atom with a single electron, approximations must be made for the others.

QM calculations can be performed using a variety of techniques, such as semi-empirical approaches and ab-initio density functional theory (DFT) [18], [19]. To take into consideration electron correlation, DFT methods require approximations [18]–[20]. Thermodynamics of a reaction, vibrational frequencies, equilibrium molecule structure, and dipole moments are a few examples of features of a system that may be predicted using QM calculations. DFT methods provide reliable results, while wavefunction methods such as electron correlation techniques like coupled cluster single double (Triple) (CCSD (T)) and Møller-Plesset perturbation theory (MP2 is a post-Hartree-Fock approach) could produce even more precise outcomes [18]. As a result, the use of the latter methods is only limited to a few tens of atoms since they are the most time- and resource-intensive methodologies.

1.2.4. QM/MM Methods

The quantum mechanics/molecular mechanics methods are a widely used computational techniques for accurately predicting the electronic and dynamic properties of biomolecules. The capacity of QM approaches to describe systems larger than a few hundred atoms, despite being needed for this kind of study, is constrained. The accuracy of QM approaches and the effectiveness of MM methods about behavior of structural and dynamics of systems can be combined to create

hybrid QM/MM methods, which can help us get beyond this restriction. In this method, the surrounding areas (known as MM region) are treated using the MM method while the chemically active part of the system (known as QM region) is described using the QM method. As a result, we are able to precisely model complicated systems with acceptable computational effort. In order to understand and anticipate atoms and molecules behavior at the quantum level while still being able to model larger systems, hybrid QM/MM approaches are a crucial tool in the study of chemistry [21]. The details of QM/MM are beyond the scope of this thesis and will not be discussed further.

1.2.5. ML Methods Approaching to DFT Accuracy

Through the use of regression algorithms, machine learning (ML) is becoming a widely used technique for developing various types of transferable [22]–[24] and non-transferable [25], [26] atomistic potentials. In chemistry, ML has been used to predict reaction pathways [27], calculate excited state energies [28] using QM, determine formation energies [29], calculate atomic forces, and predict nuclear magnetic resonance chemical shifts [30] and some other things. Furthermore, ML potentials have been used to accurately predict molecular energies, achieving a significant speedup of up to five orders of magnitude over traditional quantum mechanical methods. When developing transferable ML potentials, the choice of molecular representation is critical. This representation is a set of mathematical functions that describe the energy of a molecule based on atomic coordinates. These representations must meet certain criteria outlined by Behler [31] to be helpful for ML models: the representation must give a distinct description of the conformation, they must be rotationally and translationally consistent, and identical atoms must provide the same outcome. Many molecular representations have been proposed [32]–[35]; however, achieving a good accuracy in a wide range of chemical environments remains a challenge, particularly for arbitrary organic molecules. Symmetry functions (SF) were suggested by Behler and Parrinello as a molecular representation approximation in 2007. Through the use of chemical localization, these functions increase the transferability of neural network potentials (NNP) [33].

A lack of functional form for defining distinguishable features, limited atomic number differentiation, and difficulty transferring knowledge from one molecule to

another are some of the limitations of SFs in molecular representation for NN-based predictions. As a result, SFs are only appropriate for studies of chemically symmetric systems with small number of atom types or molecule data sets. In order to get over these limitations, in 2017 the (Accurate NeurAl network engINe for Molecular Energies) short for ANAKIN-ME or ANI [36] was presented as a transferrable deep learning potential. A potential that demonstrated chemical correctness when compared to reference DFT ω B97X/6-31G* calculations level in test situations was trained using the ANI-1 data set [37]. The data set was produced by randomly perturbing 22 million different conformations into 57,000 small molecules containing CHNO. Then, random normal-mode sampling, despite being based on harmonic approximation, provides inadequate coverage of chemical space, particularly in torsion space. Consequently, in order to diversity the data set a query by committee (QBC)-based active learning algorithm was utilized to automatically select new data [38], [39]. For poorly characterized structures, the QBC approach was utilized to produce fresh QM data by using an ensemble of models to estimate the energies of molecules. The ANI-1x potential (was developed after an extensive conformational search using the QBC active learning algorithm, but because it was an early proof-of-concept, it only sampled molecules with a small number of the elements H, C, N, and O due to the size of the chemical space. Finally, because to its broad usefulness in a variety of domains, including protein simulation and drug development, ANI-1x was enlarged as ANI-2x (include 8.9 million molecular conformations) [40] to incorporate elements S, F, and Cl.

1.3. Conventional Binding Free Energy Calculations

Free energy (FE) methods are computational techniques which predict the change in free energy of systems about their state or structure. Over the last few decades, the use of free-energy calculations that are based on MD and MC simulations has gathered more and more attention for their capacity to give precise prediction of ligand protein binding (LPB) affinity and explanation of the binding mechanisms of ligand in drug discovery. FE calculations can be classified into two different categories: equilibrium and non-equilibrium methods. In this thesis, only equilibrium methods are handled. These methods are separated into three categories as end-state methods, alchemical perturbation methods and physical pathway

methods. While the end-state and alchemical perturbation methods are used to predict the LPB affinity given the binding mode of the ligand, the physical pathway methods can be used to uncover the binding mode and mechanism of the ligand binding. Some of these methods supply to predict of absolute PL BFE that is related to free energy difference between the bound position (PL) and unbound position (P and L are free), while others compute the relative binding free-energy (RBFE) that is the difference of BFE between two ligands or more [41].

1.3.1. End-state Methods

1.3.1.1. Linear Interaction Energy (LIE)

In LIE method [42], [43], two kind of simulations should be run. One for ligand in solvent (LS system) the other one for PL in solvent (PLS system). After simulations, the frames will be provided and these can be used to calculate van der Waals interaction energy and the Boltzmann-averaged electrostatic between ligand and residues at LBD (at PLS), solvent (at LS) separately. The linear relation of binding free energy can be provided using these average predictions if two empirical parameters are specified that represent the internal energy of the P and the S but solvation energy of P at PS system and entropic changes of L conformation are not considered in this method.

In terms of accuracy, LIE methods outperform scoring functions when predicting binding free energies, with a typical RMS error of less than ~1 kcal/mol compared to an average error of 2-2.5 kcal/mol for scoring functions [44]. The ANI_LIE section of THEORY will address details on LIE and LRA.

1.3.1.2. MMP(G)BSA, MMP(G)BSAe

When it comes to the contribution of conformational entropy and enthalpy from water molecules in the binding site, MMP(G)BSA have approximation constraints like LIE. The results obtained from both methods can differ depending on the system being studied due to the disregarded enthalpy and entropy contributions, even though MMGBSA has been demonstrated to be faster and more accurate than MMPBSA. Studies have demonstrated that both LIE and MMPBSA outcomes are system-dependent in terms of dependability. In conclusion, LIE is typically more

rapid and precise than MMPBSA, especially when entropic term computations are considered. It should be highlighted, nevertheless, that approaches like MM-PBSA, MM-GBSA, and others of a similar nature can be useful for assessing large-scale energy differences between various binding modes from docking. These techniques might not be accurate enough to be used in studies on drug design based on free energy calculations, though.

A computational technique for estimating the binding affinity of PL complexes relying on the MM/PBSA descriptors was introduced in 2016 and is known as the PBSA_E method [45]. This method has been developed and improved to increase the precision of scoring systems used in the structure-based drug design process to predict PL binding affinity. Following energy minimization processes, a single optimum configuration is obtained in this procedure. The MM/PBSA method is then applied to this improved setup to predict PL binding affinity. The approach has been demonstrated to be more reliable than other standard scoring functions on basically two test sets having 121 and 130 PL complexes. It is viewed as a potential credible technique for a more accurate evaluation of PL binding affinity in the search for new medicines. The ANI_PBSAe section of THEORY will address details on MM/GPBSA and PBSA_E methods.

1.3.2. Alchemical Perturbation Methods

By analyzing simulation trajectories that define a thermodynamic path connecting the bound (PL together) and unbound (P and L free) states using statistical mechanics, alchemical free-energy perturbation methods compute the BFE difference between these two end states. These alchemical states are unphysical that can be connection between bound-unbound states. Free Energy Perturbation (FEP), Thermodynamic Integration (TI) and Bennett Acceptance Ratio (BAR) are the most known and applied alchemical methods, and their theoretical basis can be found in the contributions of Kirkwood, Zwanzig, Bennett, and other experts in the field. These methods all utilize the alchemical transformation of one state into the other using a coupling parameter λ that ranges from 0 to 1 (0 stands for bound state PL at beginning of the simulation and 1 stand for unbound state P, L free at end of the simulation). By summing the alchemical transformations between several different λ states, the final BFE difference between the two end states can be calculated.

TI, BAR, and FEP have some conceptual distinctions in their approach to computing the ligand BFE, but they can still be considered similar due to their shared application field and comparable similar important points. When crystal structure is or ligand binding mode is known, these methods can be used for a precise estimation of the ABFE for TI and BAR, and of the RBFE for FEP [41].

1.3.2.1. Free-Energy Perturbation Method

FEP is a method derived from statistical mechanics to calculate BFE differences with a carefully controlled shifting in potential energy of transformation of one ligand to another (presented in Figure 1) through MD or MC simulations.

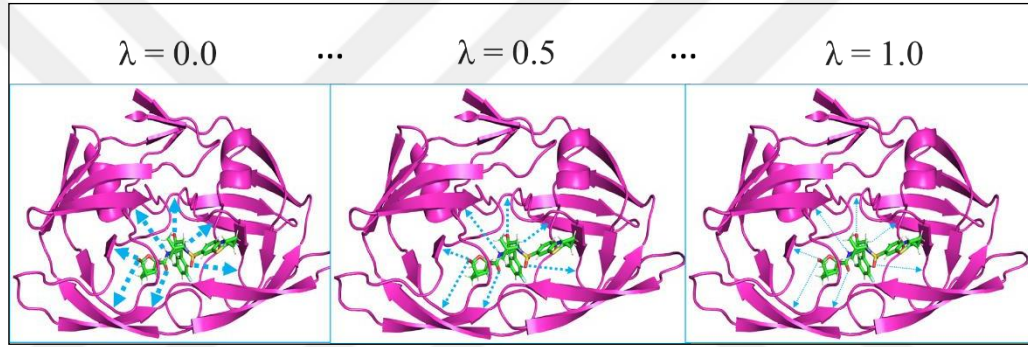


Figure 1.3: Representation of binding strength at different lambda steps.

The transformation process is divided into several windows, which are determined by the coupling parameter λ . A sufficient overlap of the states visited in the adjacent λ windows in phase space is crucial for the calculation to converge. It will work well if the number of the windows is sufficient for sampling and changes between two adjacent windows is very low that depend on λ size. FEP can calculate BFE difference by using Zwanzig equation [46]:

$$\Delta F = F_1 - F_0 = -kT \langle \exp \left[\frac{-(E_1 - E_0)}{kT} \right] \rangle_0 \quad (1.11)$$

ΔF stands for the FE difference between two states of PL that in first and second window, $-k$ is Boltzmann's constant, T is temperature, brackets define average of the equation in it, E_0 and E_1 are energies of first and second window, respectively.

FEP's strong points include precise BFE prediction (1 kcal/mol) and low computational costs associated with operating distinct windows that allow for parallel simulation. However, the method is constrained by the requirement to know the binding mode [47].

1.3.2.2. Thermodynamic Integration

TI and related techniques such as double annihilation and double decoupling approaches [48], [49] allow for the artificial separation of the intra and inter-molecular energy contributions of a single ligand at the PBD, which can then be recombined in the unbound state of the L free in solvent [48]–[50]. TI calculate BFE by decoupling of the L from the L free in solvent and PL complex in solvent by using following equation:

$$\Delta G = \int_0^1 d\lambda \left\langle \frac{\partial U}{\partial \lambda} \right\rangle_{\lambda} \quad (1.12)$$

U stands for total Potential Energy of the given system using non-bonded interactions, λ is window size about non-bonded interactions. According to the traditional alchemical thermodynamic cycle, the BFE (ΔG_{bind}) can be calculated by considered the difference between BFE of the complex ($\Delta G_{complex}$) from the BFE of L free (ΔG_L) [51].

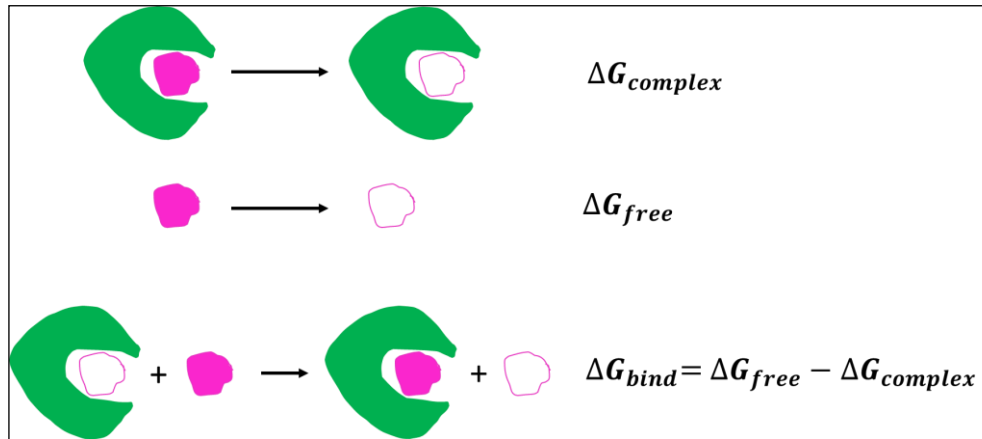


Figure 1.4: Representation of alchemical free energy calculations using a model of the thermodynamic cycle.

This allows for the determination of the energy required for the ligand to bind to the protein.

1.3.2.3. Bennet Acceptance Ratio (BAR) and Multistate BAR (mBAR)

Similar to how TI works, BAR and multistate BAR (MBAR) [52]–[54] can model the shift of a ligand from in the PL complex to free state, calculating the ΔG between adjacent alchemical states using the Bennett approach and the Hamiltonians and Fermi-Dirac distribution of the states. The calculation is considered to have converged when the energy values obtained using the forward and backward approaches are in good agreement for adjacent states [41].

By adding a λ parameter, which gradually alters the contact between the ligand and receptor, from complete interaction ($\lambda = 1$) to no interaction ($\lambda = 0$), the BAR approach is applied in practice. This is accomplished by multiplying the lambda parameter by the interaction energy of the P and L. The chance of detecting each state is determined after the system has been in equilibrium for a specific amount of time at each λ value. The Bennett relation is then used to integrate the free energy changes over the λ values, yielding the ΔG between the two states. The BAR computation solves for the FE (C) that fulfills the connections where i and j are consecutive states and U is the potential energy from a defined state [55]:

$$\Delta G = -kT \ln \left[\frac{\sum_j f(U_i - U_j + C)}{\sum_j f(U_j - U_i + C)} \right] + C \quad (1.13)$$

However, this approach might run into the same problems as if there were no overlapping between nearby states. By utilizing the energy differential from all windows, this has been expanded to the MBAR approach, which addresses the crucial BAR concerns and offers the lowest variation of all FE estimators. The fact that BAR does not necessitate computing the partition function which can be challenging to compute for large systems is one of its key advantages. The procedure, however, can be computationally expensive and may need a lot of simulations to produce reliable results. The method also relies on the presumption that the simulation adequately models the system and that it runs long enough to

reach equilibrium. BAR gives more accurate and trustable results when is compared with TI especially if there is low number of windows for calculations [56].

The visual representation of all end-state and alchemical perturbation methods is given at Figure 1.5:

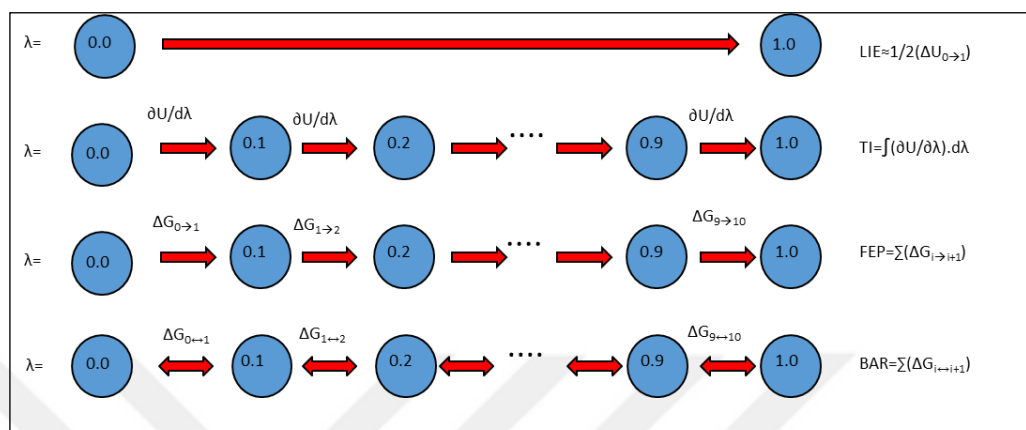


Figure 1.5: Defining several MD techniques mathematically.

Non-covalent interactions between molecules are typically described by two types of energies: electrostatic and van der Waals. The parameter λ_i is adjusted by reducing it in small steps from 0 to 1, separately for each of these interactions. Many windows is ideal for constructing a reversible system. However, the number of windows must be chosen with the trade-off between the number of windows and the computational cost in mind. For an accurate estimate of a ligand's solvation free energy (SFE), a minimum of 10 windows for electrostatic separation ($\lambda_{coul} = 0.0, 0.1, 0.2, \dots, 0.9, 1.0$) and a minimum of 10 windows for van der Waals separation ($\lambda_{vdw} = 0.0, 0.1, 0.2, \dots, 0.9, 1.0$) should be utilized, for a total of 20 windows. More MD simulations are required to calculate the BFE of the P-L at a P-L-Solvent (PLS) systems. These simulations use a constant force field and are run at the MM level. There are various methods proposed in the literature for developing computational techniques. Most of these methodologies aim to enhance accuracy and decrease the computational cost. However, currently there is no established optimal method that has reached a consensus among the scientific community.

1.3.3. Physical Pathway Methods

The primary objective of Physical Pathway (PP) methods is to regenerate the physical pathway process by which L binds to P. PP methods differ from end-point methods and alchemical transformation techniques in that they estimate the free energy of the ligand binding process to the protein, allowing for identification of the energetically bound state that is most stable one. This allows for a more detailed understanding of the ligand binding mode. Thus, they are particularly useful in cases where the possible location of the binding site on the protein is known, but there is little or no information available on the ligand binding mode. These techniques can provide especially valuable insights into the binding process in such cases. The PP methods include Umbrella Sampling (US), Steered Molecular Dynamics (SMD) [57], and Funnel-Metadynamics (FM). FM is the newest one among these techniques [41]. The details of PP methods are beyond the scope of this thesis and will not be discussed.

2. MODIFICATION OF LIE BY ANI FOR BINDING FREE ENERGY CALCULATIONS FROM MD SIMULATIONS

Adopted with permission from {Akkus, E., Tayfuroglu, O., Yildiz, M., & Kocak, A. (2022). Accurate Binding Free Energy Method from End-State MD Simulations. *Journal of Chemical Information and Modeling*, 62(17), 4095-4106”}. Copyright {2022} American Chemical Society (AC) & Creative Common License (CC-BY).

2.1. Introduction

Developing small compounds that target medically relevant protein is one of the main strategies in modern drug discovery studies. There are two core questions that must be addressed in these efforts: (a) what is the binding mode? (i.e., where does it bind from the receptor?) and (b) what is the binding strength? (binding free energy) [41]. Inhibition of an enzyme can be involved in various mechanisms. For instance, an inhibitor can bind to the enzyme from the active site via competitive inhibition or it can bind from an exo/allosteric site by uncompetitive/noncompetitive inhibition [3].

Kinetic studies was conducted to understand the activity loss of the enzyme on a substrate by the inhibitor report values such as inhibition equilibrium constant K_i or half maximal inhibitory concentration IC_{50} . These two experimental parameters are proportional to each other for competitive inhibition (they are equal in special cases) [3], [58]. Thermodynamic equilibrium constant K_i and thus IC_{50} can be approximated to the experimental binding free energy of the system for the competitive inhibitors by:

$$\Delta G_{exp}^0 = -RT \ln(IC_{50}) \quad (2.1)$$

A more correct form is

$$\Delta \Delta G = -RT \ln(IC_{50 \text{ lig1}}/IC_{50 \text{ lig2}}) \quad (2.2)$$

where $\Delta\Delta G$ is the relative binding free energy between two competitive inhibitors. ΔG can also be calculated from the thermodynamic potential. In principle, it is possible to compare experimental binding free energies determined from IC50 and theoretically determined from IC50 predicted by thermodynamic potentials using computational methods for a series of competitive inhibitors. Thus, numerous efforts have been made to develop new strategies that can accurately predict binding free energies. However, a trade-off between computational accuracy and speed must be made.

Methods to calculate the potential binding free energy range from simple end-state methods such as linear response approximation (LRA) to linear interaction energy (LIE) and molecular mechanics Poisson–Boltzmann/generalized Born surface area (MM-(P/G)BSA) [59]–[62] to more sophisticated alchemical perturbation methods either at equilibrium such as free-energy perturbation (FEP), thermodynamic integration (TI) [51], [56], [63]–[70], Bennett acceptance ratio (BAR) [56], [65], [68], [71]–[75] and multistate BAR (MBAR) [76]–[78] or non-equilibrium like Jarzynski’s equality [79], [80] and Crooks fluctuation theorem [81].

Perturbation methods require several intermediate lambda states (λ –), in which the ligand can be decoupled, annihilated, or pulled. They use the explicit water definition, and its prediction for relative binding free energies is quite successful. However, they are still at the level of MM energy terms and visit nonphysical intermediate states. All of these methods require a higher computational cost.

Binding free energy (BFE) calculations from single-step end-states are quite attractive since they require only one MD simulation for each protein–ligand (PL) complex system with less computational cost. However, the accuracy is very low due to over-simplifications such as implicit solvent definition in the case of MM-P(G)BSA and molecular mechanics (MM) definition of the Hamiltonian of the system.

Classical force fields, which do not account for electron polarization, charge transfer, and many-body effects, are very limited in accurately defining the protein–ligand interaction. Thus, the system needs to be described at the quantum mechanical (QM) level to include these essential physical effects [82], [83]. There have been several attempts to replace the MM energies with more accurate QM energies. Recently, Fox et al. [82]–[84] have introduced a “QM-PBSA” method for eight ligands and improved the MMPBSA energies. One caveat to overcome the

prohibitive DFT calculations on thousands of atoms of protein–ligand complexes is to use semiempirical quantum mechanics (SEQM) such as AM1 and SCC–DFTB or QM/MM calculations [85]–[91]. There are also ongoing efforts to use fragmentation-based QM calculation methods. Most of these studies suffer from no/low sampling from MD simulations [84], [90], [99]–[104], [91]–[98].

Machine learning (ML) techniques have attracted great interest for the past decade due to their successful algorithms to scientific questions including but not limited to chemical reactions [105], [106] potential energy surfaces [24], [105], [107], [108], forces [109]–[111], atomization energies [22], [112], and protein–ligand complex scorings [113]. One of the most promising aspects of the ML techniques is that the trained models can be applied to new systems (transferrable) [105]. In particular, ML potentials with their great capability of learning a multidimensional potential energy surface (PES) at the QM level are promising due to their efficiency and scalability to large systems [114]–[116].

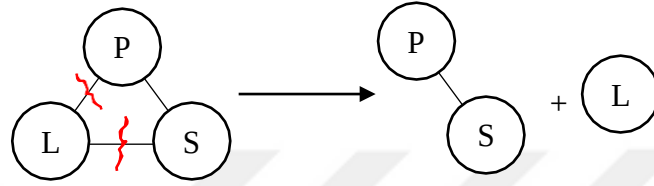
Recently, several ML-based neural network potentials (NNPs) such as ANI [37], [117], [118] have been developed using QM energies and forces to predict the potential energy surfaces of small organic compounds. ANI uses modified Behler and Parrinello symmetry functions to construct single-atom atomic environment vectors (AEVs) and could be considered as a quantum accurate force field without requirement of atomic charges. Thus, it only uses atomic symbols and 3D coordinates as an input for the calculations to estimate pairwise atomic interactions. The initial model (ANI-1x) used the PES of nonequilibrium geometries belonging to a total of ~ 17.2 million conformations generated from $\sim 58k$ small organic compounds composed of C, H, O, and N atoms [36]. The extended version, ANI-2x, has been shown to predict DFT energies of equilibrium or non-equilibrium conformations of molecules containing C, H, O, N, S, F, and Cl atoms. It has been shown to reproduce the energies at the accuracy of the wb97x/6-31G* level million times faster than the actual QM calculations [105].

Herein, we introduce a new strategy to estimate the binding free energies of small organic compounds to proteins using end-state MD simulations. The method utilizes ANI-2x neural network potentials and predicts the binding free energies of the protein–ligand (PL) in the linear interaction energy (LIE) formalism. Our results show that the “ANI_LIE” method outperforms current methods for binding free energies.

2.2. Theory

2.2.1. Linear Interaction Energy (LIE)

In the LIE method, only the protein–ligand aqueous complex (PLS) and free ligand in water (LS) are simulated. Assuming that the interactions are additive from the simulations of the protein (P) + ligand (L) + solvent (S) complex system (PLS), it can be derived as:



$$\begin{aligned}
 E^{PLS} &= E^P + E^L + E^S + \Delta E^{P-S} + \Delta E^{L-P} + \Delta E^{L-S} \\
 E^{PLS} &= E^P + E^L + E^S + (E^{PS} - E^P - E^S) + \Delta E^{L-P} + \Delta E^{L-S} \\
 E^{PLS} - E^{PS} - E^L &= \Delta E^{L-P} + \Delta E^{L-S} \\
 \Delta E^{L-PS} &= \Delta E^{L-P} + \Delta E^{L-S}
 \end{aligned} \tag{2.3}$$

Thus, the interaction of ligand with its surroundings can be calculated by using left-hand side of the equation 2.3, which requires defining L as energy group.

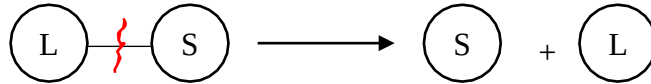
$$\Delta E^{L-surr} = \Delta E^{L-PS} + \Delta E^{L-ions} \cong \Delta E^{L-PS} \tag{2.4}$$

ANI has been trained only for atoms C, O, N, H, Cl, F, and S. Sodium ions cannot be calculated in the current version of ANI-2x. The term ΔE^{L-ions} in this equation is the interaction energy of the ligand with ions (which were added to the simulation box to neutralize the protein), and this term can be neglected assuming that the ions will not be nearby the ligands since there are only a few ions in the box and the ligand is mostly buried in the binding pocket. The ΔE^{L-surr} term can also be calculated from the right-hand side of eq 2.3 by defining two energy groups, P and L.

In that case, the pairwise contributions from each of the separate energy groups of P and L can be extracted and the average interaction between each group is calculated (i.e., ΔE^{L-P} , ΔE^{L-S})

$$\Delta E^{L-surr} = \Delta E^{L-P} + \Delta E^{L-S} + \Delta E^{L-rest} \cong \Delta E^{L-P} + \Delta E^{L-S} \quad (2.5)$$

Similarly, from the free ligand in water simulations



$$\Delta E^{L-surr} = \Delta E^{L-S} \quad (2.6)$$

Originating from the linear response approximation (LRA), the binding free-energy change from the bound (PLS) to the unbound state (LS) is given by

$$\Delta G_{bind} = \beta(\langle \Delta E_{el}^{L-surr} \rangle_{PLS} - \langle \Delta E_{el}^{L-S} \rangle_{LS}) + \gamma \quad (2.7)$$

Here, the brackets and their subscripts, $\langle \rangle$, denote the ensemble average and the components of the simulation, respectively. ΔE_{el}^{L-S} corresponds to the electrostatic interaction energy term between the ligand (L) and surroundings (solvent, S) from the ligand in solvent simulations. Thus, the binding free energy can be calculated from:

$$\Delta G_{bind} = \beta(\langle \Delta E_{el}^{L-P} \rangle_{PLS} + \langle \Delta E_{el}^{L-S} \rangle_{PLS} - \langle \Delta E_{el}^{L-S} \rangle_{LS}) + \gamma \quad (2.8)$$

Including van der Waals terms, the LIE equation becomes:

$$\Delta G_{bind} = \alpha(\langle \Delta E_{vdW}^{L-P} \rangle_{PLS} + \langle \Delta E_{vdW}^{L-S} \rangle_{PLS} - \langle \Delta E_{vdW}^{L-S} \rangle_{LS}) + \beta(\langle \Delta E_{el}^{L-P} \rangle_{PLS} + \langle \Delta E_{el}^{L-S} \rangle_{PLS} - \langle \Delta E_{el}^{L-S} \rangle_{LS}) + \gamma \quad (2.9)$$

Here α , β , and γ are empirical parameters determined from linear fitting to the experimental binding energies.

2.2.2. ANI/D3 Linear Interaction Energy (ANI_LIE)

We explored the success of ANI's performance on estimation of binding energy using these LIE equations after modifications. Here, we replaced the electrostatic terms with ANI energies and van der Waals terms in the MM energies with Grimme's dispersion functions (D3) at the wb97x/6-31G* level. In addition, since all of these equations require two simulations (PLS and LS), we have explored either by ignoring solvent terms or borrowing $\langle E_{ANI}^{L-S} \rangle_{LS}$ term from calculations.

In linear response approximation (LRA), "preorganization energy" (POE) is included in ligand-surrounding electrostatic interaction energy. This term is computed from an ensemble in which the partial charges on the ligand atoms are set to zero. On the other hand, the POE term is ignored in a linear interaction energy (LIE) method. With this aspect, ANI follows LIE rather than LRA.

When the solvent is totally ignored and ANI interaction energies are replaced with MM electrostatic energies (similarly, D3 terms replaced with MM van der Waals terms), eqs 2.8 and 2.9 become:

$$\Delta G_{bind} = \beta \langle \Delta E_{ANI}^{L-P} \rangle_{PLS} + \gamma \quad (2.10)$$

$$\Delta G_{bind} = \beta \langle \Delta E_{ANI}^{L-P} \rangle_{PLS} + \alpha \langle \Delta E_{D3}^{L-P} \rangle_{PLS} + \gamma \quad (2.11)$$

Although these equations ignore the solvent effects, the source of large noise, which stems from the high mobility of solvent molecules, could be avoided, and thus, more precise (not necessarily more accurate) trends in binding free energies could still be observed.

When the solvent is not ignored, these two equations are updated to:

$$\Delta G_{bind} = \beta (\langle \Delta E_{ANI}^{L-P} \rangle_{PLS} + \langle \Delta E_{ANI}^{L-S} \rangle_{PLS} - \langle \Delta E_{ANI}^{L-S} \rangle_{LS}) + \gamma \quad (2.12)$$

$$\begin{aligned}
\Delta G_{bind} = & \beta(\langle \Delta E_{ANI}^{L-P} \rangle_{PLS} + \langle \Delta E_{ANI}^{L-S} \rangle_{PLS} - \langle \Delta E_{ANI}^{L-S} \rangle_{LS}) \\
& + \alpha(\langle \Delta E_{D3}^{L-P} \rangle_{PLS} + \langle \Delta E_{D3}^{L-S} \rangle_{PLS} - \langle \Delta E_{D3}^{L-S} \rangle_{LS}) \\
& + \gamma
\end{aligned} \tag{2.13}$$

Here, we generated conformational sampling of protein-ligand (PL), protein (P) and ligand (L) in the presence of explicit water using a single MD simulation (PLS) in the case of solvent ignored calculations. Then, the single point energies (SPE) of each MD frame have been calculated by ANI-2x and D3 by stripping out water and ions. For each MD frame, three different energies were calculated: the ligand is complexed with protein, E_{ANI}^{PL} ; bare protein that is extracted from the MD frame, E_{ANI}^P and bare ligand that is extracted from the MD frame, E_{ANI}^L . Similarly, dispersion energy of each component (E_{D3}^{PL} , E_{D3}^P , and E_{D3}^L) has been calculated. When

the solvent included, L-S interaction terms in PLS MD simulation along with additional simulation of bare ligand in water (LS) has been considered.

In this study, we used eqs 2.10-2.13 to investigate the ANI's performance of LIE approach (ANI_LIE) in the presence of solvents. Strikingly, we found that the calculations using eq 2.10 outperform conventional approaches by producing BFEs that strongly correlate with experimental results. The additional terms appearing in eqs (2.11)-(2.13) bring subtle improvements to the predictions. Here, our aim is to test the success of the ANI on the interaction energy term; thus, we ignored conformational entropy changes of the ligand and protein upon binding. It has been shown that a single trajectory approach is more precise than the three-trajectory approach due to error cancellation [105]. Although continuum model definitions such as PBSA could be used in eqs 2.12 and 2.13 rather than using explicit solvents, we preferred not to include these terms because the primary aim of this study is to test the success of ANI in the ligand-protein interaction in the LIE approach with explicit waters. It should also be noted that the conformations are already sampled in the presence of explicit waters and ANI predicted energies even in eqs 2.10 and 2.11 belong to equilibrium geometries of aqueous systems (not gas-phase calculations). Therefore, we believe that these conformations inherently reflect the solvent effects at some level. Here, we note that the ANI potentials were used as a postclassical MD

simulation to recalculate the MD frames rather than performing the ANI potentials during the MD simulation. Thus, the method is still limited to the classical force field to generate conformational sampling. In addition, the entropic contributions of the free energies are reflected by the width of the energy distributions as with the nature of the LIE approach.

Due to the power of QM-predicted potential energy surfaces, the ANI method can also be applied to ligands that do not have similar structures as opposed to methods such as BAR, FEP, and MMP(G)BSA, in which calculations mostly require ligands with similar scaffolds. We have embedded the end-to-end script to GitHub for broader access.

2.3. Computational Methods

2.3.1. Dataset Selection: SARS-CoV-2 Main Proteases, HIV-1 Proteases, JNK1 Proteins

The selected data that will be used to test ANI accuracy is generated by collecting severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) main proteases M^{pro} (also known as 3CL^{pro}), human immunodeficiency virus type 1 (HIV-1) proteases, and c-Jun N-terminal kinases 1 (JNK1).

The digestive and respiratory systems of a number of vertebrate hosts, including people, can be harmed by coronaviruses (CoVs), single-stranded ribonucleic acid (RNA) viruses. They are split up into four groups: α , β , γ , and δ . There are seven known human coronaviruses, including severe acute respiratory syndrome coronavirus (SARS-CoV), Middle-East respiratory syndrome coronavirus (MERS-CoV), and severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which have all been linked to lethal respiratory infections. Especially, the virus that causes COVID-19, SARS-CoV-2, has had a serious impact on public health and has killed millions of people worldwide [119]. By attaching to ACE2 with the viral spike (S) protein via ligand binding domain (LBD) and activating the virus S protein with the aid of the cellular transmembrane serine protease 2 (TMPRSS2), SARS-CoV-2 enters human cells [120]. Then it generates two overlapping polyproteins, main protease (M^{pro}) or 3CL^{pro}) and papain-like protease (PL^{pro}) that cleaves itself. During viral replication, the M^{pro} cleaves polyproteins at various sites to create several functional proteins to generate copy of new viruses. The M^{pro} is a class of highly

conserved cysteine hydrolases [121] thus it might be a rational target to generate new drugs to stop viral cycle.

An important turning point in medical history was the identification of human immunodeficiency virus (HIV) as the causative agent of AIDS, and since that time, considerable strides have been achieved in our knowledge of and ability to treat the virus. Since the early 1980s, when the HIV/AIDS epidemic first started, millions of people have been infected or died as a result of the disease, which has had a terrible effect on world health. Despite the ongoing evolution of antiretroviral therapy (ART) for HIV/AIDS treatment, complete clearance of the virus from an infected patient remains elusive. Reverse transcriptase (RT), protease (PR), and integrase (IN) have all been identified as possible targets for antiviral medication as a result of the identification of HIV as the causal agent of AIDS and the finding of crucial molecular steps in its reproduction. HIV type 1 (HIV-1) protease is an important enzyme that is crucial for the development of mature and infectious virions by cleaving viral polyproteins into functional proteins that are required for virus replication.

In detail, by fusing with T-cell membranes, HIV infects them. The RT allows the viral RNA to enter the cell where it is transformed into DNA. When this viral DNA reaches the cell nucleus, IN incorporates it into the DNA of the host cell. The machinery of the host cell is subsequently utilized by HIV to create viral proteins, which PR then transforms into mature forms. When viral RNA and proteins have accumulated at the cell surface, they are subsequently discharged as virions, which are new viruses. HIV-1 protease is an enzyme that breaks down the two specific polyproteins at precise locations to produce the viral structural proteins RT, IN, and PR as well as the viral envelope glycoproteins. The assembly of new virions requires these proteins. One of the most important steps in stopping HIV replication is inhibiting PRs [122].

A subfamily of the mitogen-activated protein kinase (MAPK) superfamily, c-Jun N-terminal protein kinase (JNK), also known as stress-activated protein kinase (SAPK), is essential for the control of a variety of cellular processes, including gene expression, cell proliferation, differentiation, and programmed cell death (apoptosis). JNK1 and JNK2 are the two isoforms of the JNK subfamily, and they are both widely expressed in the body and shared distinct functions. Tumor necrosis factor alpha (TNF- α) and UV light are two such examples of extracellular stimuli that

enable JNKs to become active through a process of successive protein phosphorylation via a MAPK cascade. JSAP1, JIP, and β -arrestin, as well as protein phosphatases, are also a few of the scaffold proteins that interact with one another to regulate the activation of JNK. Inhibition brought on by NF κ -B activity also controls JNK activation, in addition. Long-term JNK activation can cause apoptosis by the induction of TNF- α if NF κ -B activation is not present. The kind of cell, the stimulus that causes apoptosis, the time frame during which JNK is activated, and the activation of some of the other signaling pathways are only a few of the variables that can affect JNK's function in apoptosis [123]. Therefore, JNKs must be inhibited to affect a variety of biological processes and disease conditions. In diseases like cancer, inflammation, and neurodegeneration, JNKs' activity can be inhibited to produce positive results. Additionally, JNK inhibition can aid in the prevention of other disorders like cancer. The inhibition of JNKs is a hot topic for research in the medical industry because of their diverse functions in the body.

2.3.2. System Preparation

For the binding studies, the crystal structures of noncovalent ligands complexed with 3CL^{pro} of SARS-CoV-1, SARS-CoV-2, and HIV-1 proteases, a total of 54 protein–ligand complexes with known IC₅₀ values (Supporting Information), were retrieved from the Protein Data Bank (PDB) except for c-Jun N-terminal kinases, which were retrieved from the work by Khalak et al. [124] using Gaussian 16 software, the model ligands were first optimized at the B3LYP/6-31G* level and ESP charges were generated at the HF/6-31G* level. Using an Antechamber module in AmberTools 2021, RESP charges and GAFF force field atom types were generated. The Amber2gmx module [82] from AmberTools 2021 was used to convert amber-type input files to Gromacs.

2.3.3. MD Simulations

The molecular dynamics simulations were carried out using a Gromacs 2018+ software package [125] with an all-atom model of Amber ff99SB-ILDN75–77 force field implemented in Gromacs. The protein–ligand complex (~6000 atoms) was placed in the center of a dodecahedron box. Each system was solvated in the TIP3P model type water [126] with a cell margin distance of 10 Å for each dimension. The

protein–ligand systems with $\sim 70,000$ atoms were neutralized in 0.15 M NaCl. Classical harmonic motions of all bonds were constrained to their equilibrium values with the LINCS algorithm.

Energy minimization was carried out to a maximum $100 \text{ kJ} \cdot \text{mol}^{-1} \cdot \text{nm}^{-1}$ force using a Verlet cutoff scheme. For both long-range electrostatic and van der Waals interactions, a cutoff length of $12 \text{ \AA}$ was used with the particle mesh Ewald method (PME) (fourth order interpolation) [127]. The neighbor list update frequency was set to 20 ps^{-1} . As with our earlier studies [128]–[130], two-step energy minimization and equilibration schemes were used. Each minimization step consisted of up to 50 000 cycles of steepest descent and a subsequent 50 000 cycles of lbfgs integrators.

After minimization, each system was equilibrated within three steps using Langevin dynamics. The first step consisted of a 1 ns of NVT ensemble. The protein–ligand and the rest of the system were defined as two temperature groups at 310 K. The next step consisted of NPT ensembles, in which the systems were equilibrated to 1 atm pressure by Berendsen for 200 ps and followed by Parrinello-Rahman isotropic pressure coupling for 1 ns to a reference pressure of 1 atm. When systems reached to equilibrium, an MD simulation of 10 ns was carried out at the NPT ensemble.

2.3.4. Free Energy Calculations

Simulations for alchemical free energy calculations were carried out using a Gromacs 2018+ software package [125] with the all-atom model of Amber ff99SB-ILDN [131]–[133] force field implemented in Gromacs. For the alchemical free energy calculations, 10 equal decoupling steps (i.e., decoupling the ligand from the environment) of each Coulombic and van der Waals interaction ($\Delta\lambda = 0.1$, total 21 λ -windows) were used. For the calculations of the protein–ligand complex, a separate decoupling of the ligand from the ligand+water system with the same parameters was performed and subtracted from the decoupling of the ligand from the protein + ligand + water complex system [128]. An Alchemicalanalysis module from the pymbar library⁸³ was used to print alchemical free energies.

MM-PB(GB)SA calculations were performed using gmx_MMPBSA script by AmberTools 2021 v4. For the calculations of protein–ligand complexes, we used

the $\lambda = 0$ trajectories of protein+ligand+water complexes from alchemical free energy calculations. Only the protein + ligand complex part from the trajectories was extracted and default parameters of implicit water with $\epsilon = 80$ and solute with $\epsilon = 2$ were used. Nonpolar solvation parts were separated into dispersion and cavity terms, which is defined by solvent accessible surface area from molecular volume.

ANI and D3 calculations were performed using our custombuilt python scripts deepQM as with our previous study for LS solvation free energies (data unpublished). Since it corresponds to a fully interacting state, we used a trajectory of the first window ($\lambda_{\text{coul}}=0$; $\lambda_{\text{vdw}}=0$) from the alchemical MD simulations for QM accuracy calculations using ANI and Gaussian 16 software (G16). For G16 calculations at the wb97x/6-31G* level, it is computationally too expensive to carry out the entire protein–ligand system. Therefore, we had to truncate the residues around the ligand by 4.0 Å (nearly 900 atoms). Gromacs software was utilized to index energy groups, remove PBC, order the nearest waters around the PL complex, and extract structures from trajectories for ANI calculations while PyMol was used to truncate the nearest residues around the ligand for the QM calculations at Gaussian 16 software.

2.4. Results and Discussion

2.4.1. Workflow

deepQM is written in Python 3.9 (details of the functionality and different modules of deepQM are discussed elsewhere) and uses ASE libraries to calculate ANI and DFTD3 single-point energies of the given A–B complex structure (e.g, A = protein, P; B = ligand, L). As an end-to-end process with the DFT accuracy, binding free energy is predicted from the PL/LS complex, free protein, and ligand structures, all of which are extracted from the trajectories of classical MD simulations of the protein–ligand complex system in water.

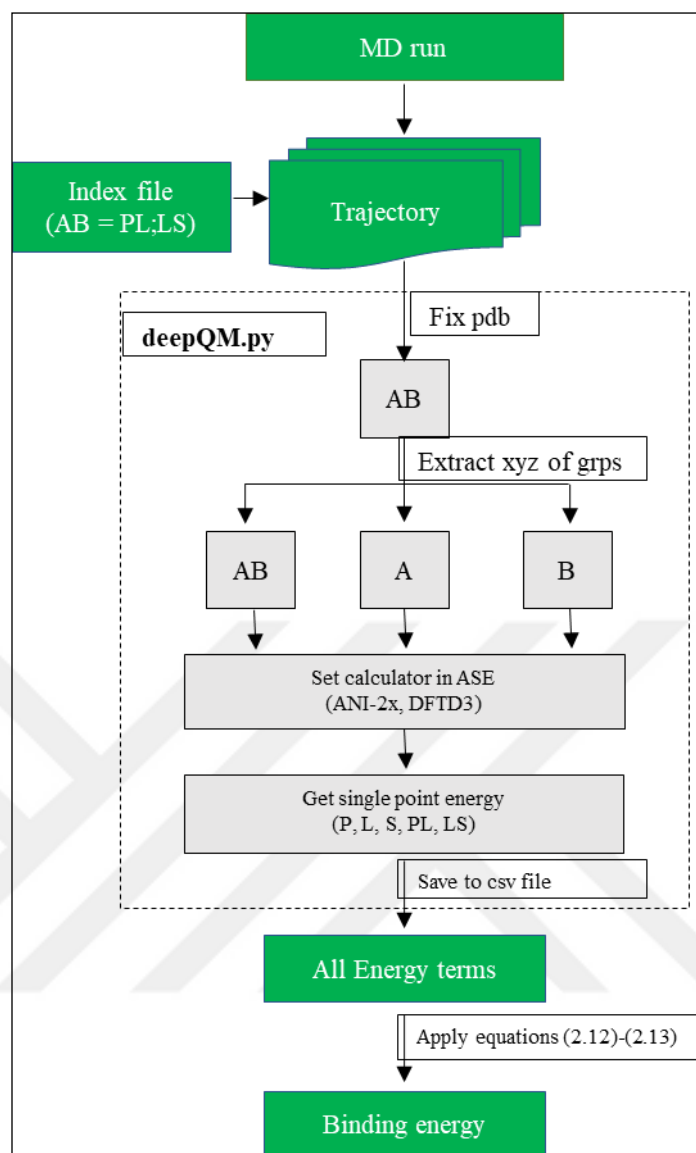


Figure 2.1: deepQM General Workflow for Performing EndState Binding Free Energy Calculations (The first step consists of extracting individual groups and converting to xyz file format. The second step is the calculator. Single or multicalculators can be selected. Finally, the results were printed and combined to predict binding free energy).

The workflow for the prediction of protein–ligand binding free energy from multiframe of an MD simulation is as follows: First, the atomic indices of the protein and ligand are read from the index groups of the index file and extracted into three separate xyz coordinates (PL, P, L or LS, L, S) after preprocessing the pdb files to make compatible for deepQM (Scheme 1). Next, ANI and DFT-D3 calculators are imported and set from the ASE platform, and then, each group is read by both calculators to predict total single-point energy (SPE) and D3 dispersion energy, respectively. Next, the interaction energy between the protein and ligand is written to

data files for each frame in both methods. In the final step, binding free energy is estimated using the equations discussed in the Computational Methods section. In principle, it can work with any MD simulation packages upon converting trajectories to separate pdb files that contain the protein–ligand–solvent complex and defining energy groups. Index file format must be Gromacs type.

2.4.2. ANI Results

2.4.2.1. Can ANI Reproduce DFT Interaction Energies?

Since ANI has been trained using atomic environments, it could still be scalable to the DFT calculations for larger systems such as protein–ligand interactions. It should be noted that DFT single-point energies of the protein and protein–ligand complex systems are in the order of hundreds of thousands of eVs, whereas the binding energy is only a few eVs. Therefore, it is important to have the total energies converged for accurate binding free energy calculations. Table 1.1 shows the absolute energies of truncated protein+ligand complex systems (HIV protease, PDB ID = 5IVS) calculated by ANI and G16 at the wb97x/6-31G* level. There are 55 residues around the ligand in the truncated protein with 920 atoms (Figure 2.1).

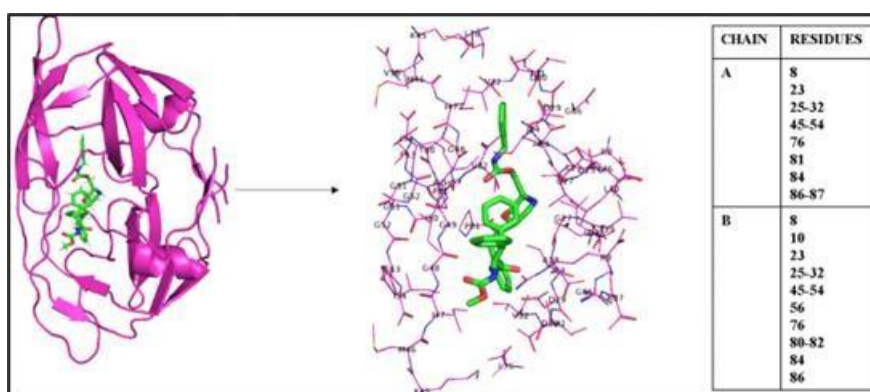


Figure 2.2: Truncated structure in the DFT calculation as a comparison to ANI, in which all the PL system can also be calculated. All the residues were saturated with H atoms to make the structure neutral. All H atoms were omitted from structures for better view. The table shows which residues included in the truncated structure.

ANI's prediction on ligand's energy is excellent. However, the ANI predicted absolute value of the protein–ligand complex deviates by ~2 eV from that of DFT

calculations. However, this energy is just a bias and canceled out by subtracting the energy of the protein in the interaction energy term. This bias cancellation will also occur on extended (untruncated) systems (i.e., a full protein + ligand complex). Therefore, the ANI-predicted interaction energies are in reasonable agreement with G16 calculations, with a shift by only ~ 0.252 eV (5.812 kcal/mol). This value can be minimized so as to converge by increasing the number of frames sampled by the MD simulation and increasing simulation time. It should also be noted that it has been trained for atomic environments with a cutoff (5.1 Å for radial and 4.5 Å for angular chemical environments), which limits its calculation to short-range interactions.

Table 2.1: Absolute Energies of Truncated Protein + Ligand Systems (HIV Protease, PDB ID = 5IVS) from MD Simulations (in eV).

	G16				ANI				
Frame	Complex PL	Receptor P	Ligand L	ΔE_{int}	Complex PL	Receptor P	Ligand L	ΔE_{int}	Abs. Err.
0	- 601409.137	- 543088.071	- 58316.912	- 4.154	- 601407.549	- 543086.919	- 58316.653	-3.977	0.177
1	- 611338.434	- 553017.357	- 58317.024	- 4.053	- 611337.259	- 553016.927	- 58316.904	-3.428	0.625
2	- 611343.896	- 553023.284	- 58316.737	- 3.876	- 611342.856	- 553022.874	- 58316.543	-3.439	0.436
3	- 611338.693	- 553021.427	- 58316.102	- 1.164	- 611337.181	- 553020.251	- 58315.918	-1.012	0.152
...									
100	- 611340.901	- 553023.152	- 58315.756	- 1.992	- 611340.670	- 553023.116	- 58315.703	-1.878	0.115
Mean	- 611366.228	- 553047.220	- 58316.584	- 2.424	- 611368.035	- 553046.484	- 58319.368	-2.183	0.252

To assess the convergence of the total energy, we have extracted snapshots from the trajectories of 10 ns MD simulations by retrieving the first frames in different time intervals (e.g, 10 ps for 1000 frames). Figure 2.2a shows the convergence of the average interaction energy when more snapshots are averaged, assuming 1000 frames in 10 ns MD simulations as the fully converged energy. The data suggests that using 100 frames in the average calculations is sufficient to have a convergence to a maximum of 0.04 eV (1.1 kcal/mol). This error drops to below 0.02 eV (0.46 kcal/mol) when 200 frames are used.

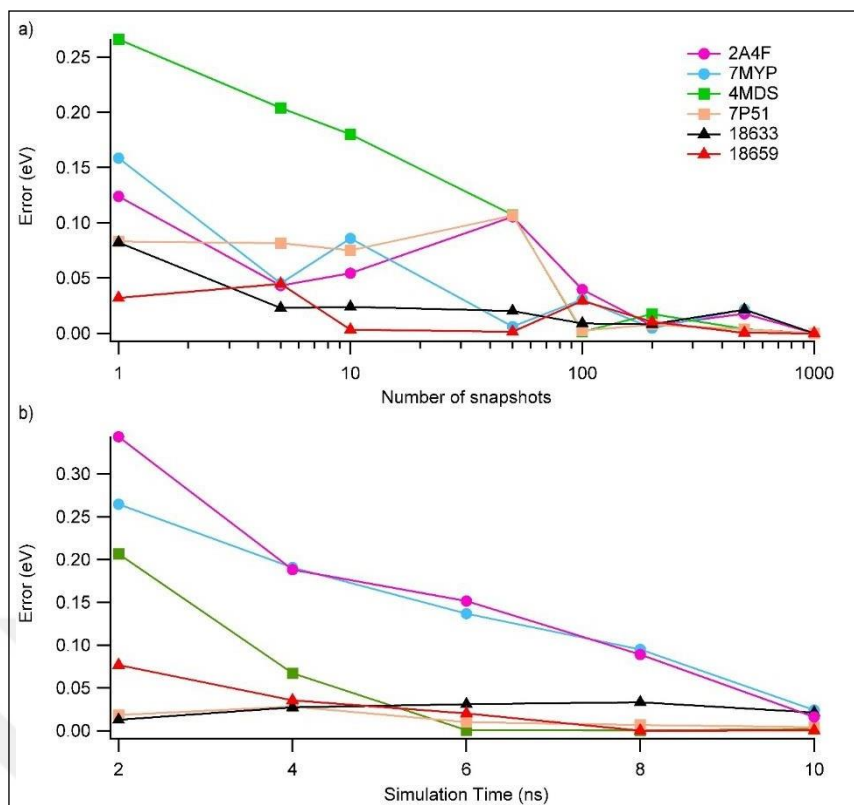


Figure 2.3: Absolute deviations of the average P-L interaction energies for selected complexes calculated by ANI, considering 1000 frames and 10 ns simulations as the converged value (a) as a function of the number of frames to average and (b) as a function of simulation time using only 100 frames in each time interval.

We have also investigated how MD simulation time affects the interaction energy (Figure 2.2b). It suggests that the simulation time does have more impact on the absolute errors and an 8 ns of simulation time can bring a maximum error of 0.09 eV (2.5 kcal/mol). This error reduces to below 0.5 kcal/mol in 10 ns simulations.

One of the most important features of the ANI calculations is that they can be run on GPU or CPU and either construction can be run in parallel. The computation time is million times faster than DFT calculations. On a 4x NVIDIA Tesla V100 construction, the 1000 frames of the P-L and L-S complexes can be calculated within hours if not minutes. A single frame in our truncated P-L structures with 55 residues and 920 atoms could take several hours at the wb97x/6-31G* level.

2.4.2.2. Can the Electrostatic Interaction Term Be Replaced by ANI?

The ANI and D3 energy terms are correlated and can be replaced with Coulombic and van der Waals terms, respectively. As both van der Waals interaction

terms in MM and Grimme's D3 functions originate from Lennard-Jones potentials, it would be trivial to correlate these two terms. Thus, the original LIE equation can easily be adopted by replacing D3 energies with MM-based van der Waals terms and by replacing MM-based electrostatic interaction energy terms with ANI-predicted interaction energies (Figure 2.3). ANI also shows great correlation to electrostatic terms, which allows us for the modification of the LIE equation.

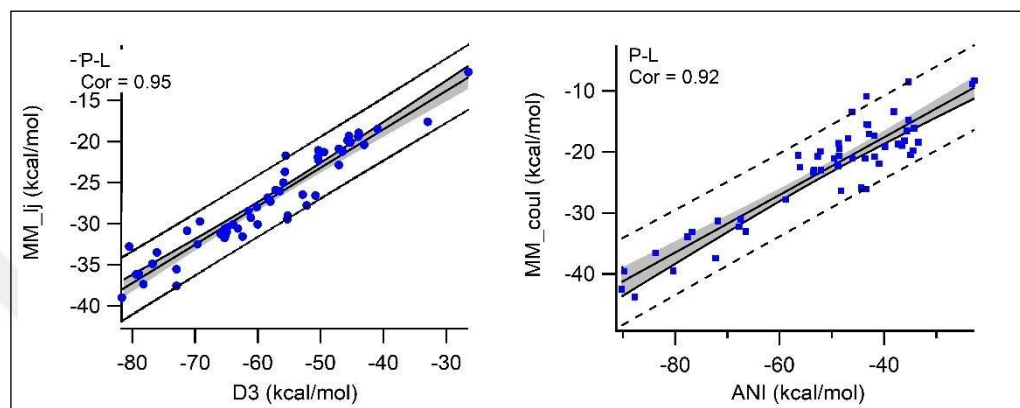


Figure 2.4: ANI-predicted average interaction energies and classical electrostatic interaction energies between the protein and ligand from PLS simulations are correlated while the energies from Grimme's D3 functions are correlated to van der Waals interactions.

Since ANI and MM energy terms are strongly correlated, one should expect LIE calculation by both methods to follow similar trends with respect to experimental free energies. Surprisingly, the correlation of ANI_LIE values to experimental free energies is greater by ~15% than that of MM_LIE.

2.4.2.3. BFE by Ignoring LS Interaction

Using the simplest modification of the LIE formula (i.e., eq 2.10), we have calculated the ANI interaction energies. The coefficients were determined from the fit to the experimental values (Figure 2.4). It should be emphasized that there are three different protein families with a total of 54 complexes and the ligands do not necessarily have similar scaffolds. We observed excellent agreement ($R = 0.87$) between the experimental and ANI/LIE-predicted absolute binding free energy values.

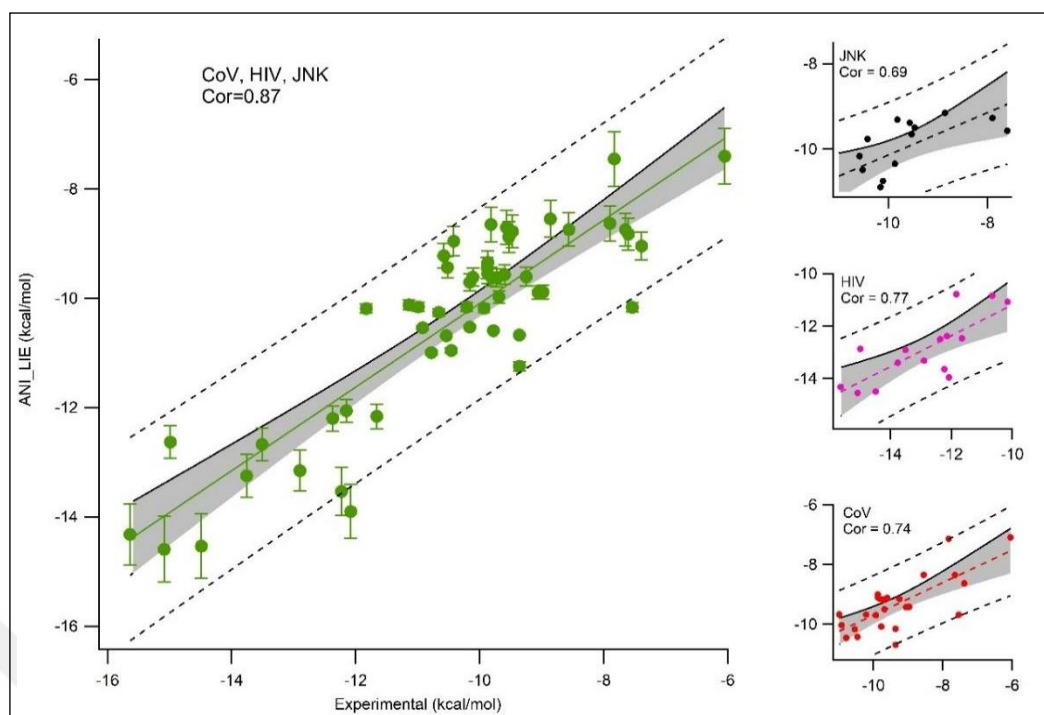


Figure 2.5: ANI interaction energies calculated from eq 2.10 with $\beta = 0.10639 \pm 0.0164$ and $\gamma = -4.9875 \pm 0.881$ using average of 100 snapshots of 10 ns MD simulations for 54 protein–ligand complexes, showing 0.87 correlation to the experimental values. The bar lines show uncertainties. The bottom plots show the fits of individual protein families as CoV, HIV-1, and JNK-1.

The mean absolute error (MAE) is just 1.76 kcal/mol. Similar to LIE calculations using only ANI potentials, we also investigated the effect of the inclusion of D3 terms in the prediction using eq 2.11. Although the correlation coefficient was not improved significantly, the MAE value decreased to 0.76 kcal/mol (Figure 2.5).

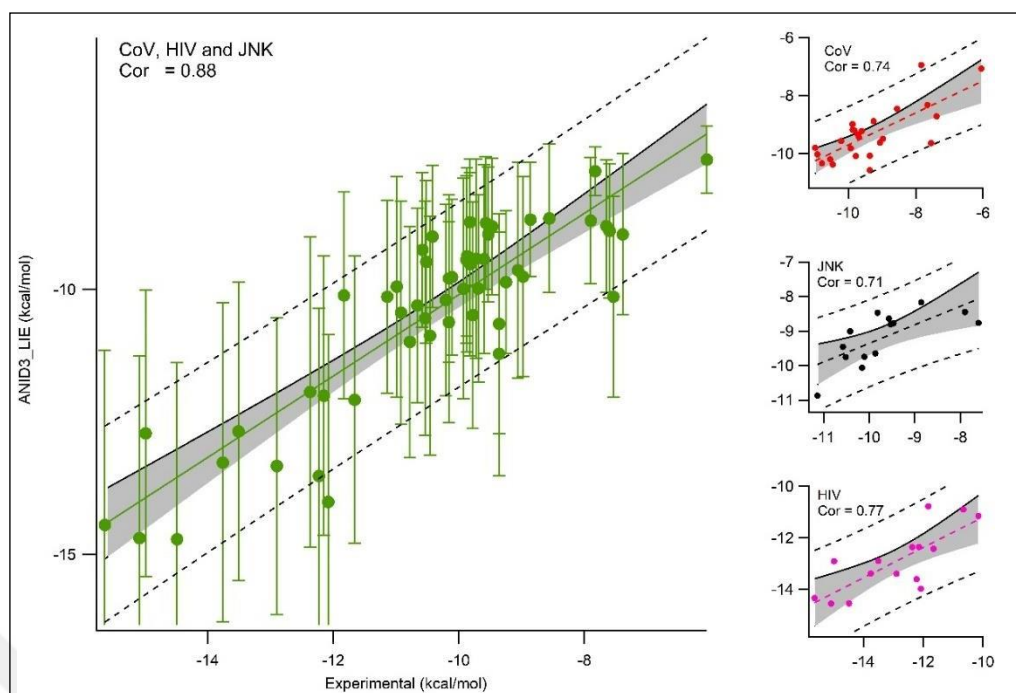


Figure 2.6: ANID3 interaction energies calculated from the equation (2.5) with $\alpha = -0.0254 \pm 0.0282$, $\beta = 0.1240 \pm 0.0212$ and $\gamma = -5.5692 \pm 0.7821$ using average of 1000 snapshots of 10 ns MD simulation for 54 protein ligand complexes, showing 0.88 correlation to the experimental values. The bar lines show the mean absolute error. The bottom plots show the fits of individual protein families as CoV, HIV-1 and JNK-1.

2.4.2.4. BFE by Including LS Interaction

After showing the success of ANI's prediction of interaction energies and binding free energies in the absence of solvents, we have also attempted to calculate the binding free energies using the most extensive LIE formalism in which the solvent effects are also included. The solvent requires a second simulation of ligand in water (LS) along with the contribution of L–S interactions in the PLS simulations as explained in the Theory section.

In principle, ligand–solvent interactions can be computed with implicit definitions as in the case of PBSA, GBSA, or any other continuum models. For instance, a recent work by Fox et al. introduced the so-called “ONETEP” program, in which they coupled the DFT energies with PBSA to calculate QM-PBSA energies [134]. However, here, we attempted the L–S interaction energies using explicit waters in the system sticking with the original formalism of the LIE method.

Although ANI enables us to calculate such large protein +ligand+water (PLS) systems, due to memory limitations, we could only perform L–S interaction energies

after reducing the PLS systems so as to include waters by only 4 Å around the PL complex. We achieved this by ordering the water in the trajectories and extracting the nearest solvents in the PLS system. Figure 2.6 shows examples of reduced vs complete water in PLS systems for each protein family.

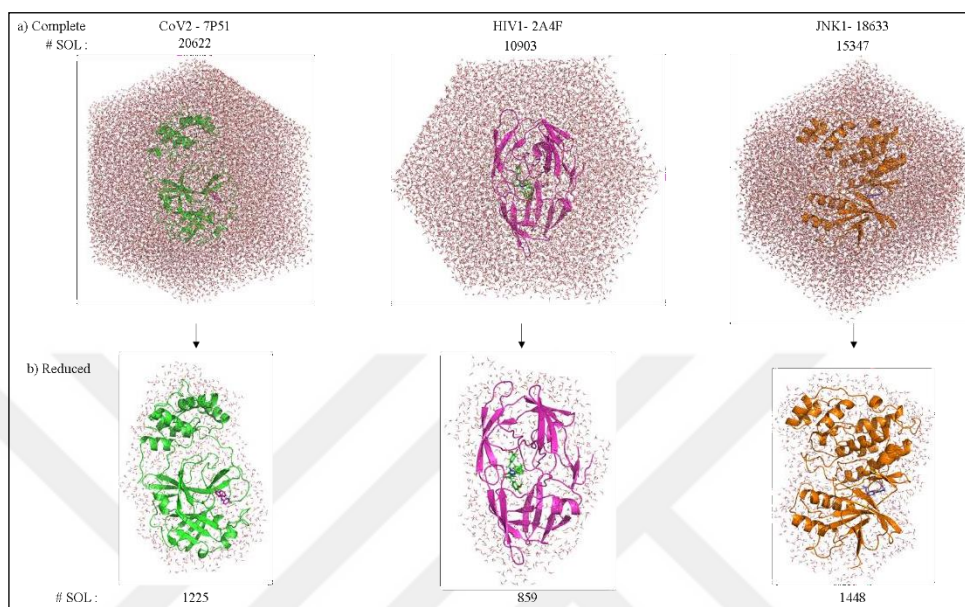


Figure 2.7. Illustration for the representative structures in protein + ligand + complete/reduced water (PLS) systems. The LIE calculations were performed using reduced water to ease memory cost of the computations. The results are almost identical in both cases.

Since the ligand is mostly buried in the protein, there are only a limited number of solvents around the ligand in the bound state. Thus, the effect of the solvents that are beyond 4 Å should not affect the calculations. Figure 2.7 shows that the solvent included ANI and ANID3 LIE calculations. We have summarized the correlation coefficients for different approaches in Table 2.3.

Table 2.2: MM energy terms and LIE for complete vs reduced systems, showing reducing system hardly affects the results.

	Complete System		$\alpha=0.181$; $\beta=0.5$	Reduced System		$\alpha=0.181$; $\beta=0.5$			
PDB ID	Lrest_coul (kcal/mol)	Lrest_LJ (kcal/mol)	LIE_whole systemt (kcal/mol)	L-rest_coul (kcal/mol)	L-rest_LJ (kcal/mol)	LIE reduced system only 4Åwaters (kcal/mol)	Diff_coul	Diff_lj	Diff_LIE
2GZ7	-10.76	-11.29	-3.23	-9.92	-9.94	-2.57	-0.84	-1.35	-0.67

Table 2.2: Continue

2GZ8	-18.02	-17.36	-2.09	-18.21	-14.33	-1.63	0.19	-3.03	-0.45
3V3M	-13.71	-11.04	-3.07	-10.51	-10.38	-1.35	-3.20	-0.67	-1.72
4MDS	-24.22	-17.62	0.54	-24.48	-15.78	0.75	0.26	-1.85	-0.20
4YOJ	-19.46	-10.19	-3.18	-19.61	-8.74	-2.99	0.15	-1.46	-0.19
7L10	-7.89	-13.27	-1.12	-8.59	-11.95	-1.23	0.69	-1.31	0.11
7L11	-4.85	-12.07	-2.11	-5.40	-11.01	-2.19	0.55	-1.06	0.08
7L12	-20.46	-12.98	1.01	-20.92	-11.38	1.07	0.46	-1.60	-0.06
7L13	-24.10	-13.36	-1.51	-24.65	-11.87	-1.51	0.55	-1.50	0.01
7L14	-4.96	-13.68	-2.45	-5.49	-12.44	-2.49	0.52	-1.23	0.04
7LMF	-12.31	-11.89	-3.15	-12.69	-10.76	-3.13	0.38	-1.13	-0.02
7LMG	-14.46	-11.42	-2.65	-14.69	-10.38	-2.57	0.23	-1.04	-0.07
7LMH	-15.27	-13.71	-1.78	-15.72	-12.38	-1.77	0.45	-1.34	-0.01
7LMI	-14.99	-11.89	-1.12	-15.26	-10.82	-1.06	0.28	-1.08	-0.06
7M8M	-25.97	-11.67	-2.64	-26.39	-10.35	-2.61	0.42	-1.33	-0.03
7M8N	-27.85	-13.39	-3.78	-28.22	-11.67	-3.65	0.36	-1.72	-0.13
7M8O	-19.07	-13.19	-0.79	-19.16	-11.69	-0.56	0.09	-1.50	-0.23
7M8P	-23.13	-14.29	-1.10	-23.96	-12.59	-1.20	0.82	-1.70	0.10
7M8X	-10.86	-13.16	-2.46	-11.54	-11.86	-2.56	0.68	-1.30	0.11
7M8Y	-25.35	-16.43	0.47	-25.82	-14.49	0.58	0.48	-1.95	-0.11
7M8Z	-28.66	-12.70	0.32	-28.94	-10.92	0.50	0.28	-1.78	-0.18
7M90	-33.05	-14.38	-1.91	-33.09	-12.64	-1.61	0.04	-1.73	-0.29
7M91	-18.63	-10.96	0.34	-18.98	-9.99	0.34	0.35	-0.97	0.00
7N44	-22.70	-12.97	0.20	-23.08	-11.48	0.28	0.38	-1.50	-0.08
7P51	-14.64	-9.94	-0.38	-14.95	-9.07	-0.38	0.31	-0.86	0.00
2A4F	-23.79	-9.82	0.20	-23.82	-9.01	0.33	0.02	-0.81	-0.13
3ROW	-55.87	-15.94	1.13	-55.73	-15.06	1.35	-0.13	-0.88	-0.22
3R0Y	-38.54	-16.56	-0.95	-38.43	-15.59	-0.72	-0.12	-0.98	-0.23
3SPK	-8.94	-8.21	-0.63	-8.83	-7.77	-0.49	-0.11	-0.44	-0.14
4KB9	-19.53	-6.86	-2.31	-19.52	-6.38	-2.22	-0.01	-0.49	-0.09
5DGW	-16.30	-10.96	0.39	-16.40	-10.19	0.48	0.10	-0.76	-0.09
5IVS	-31.12	-15.71	-9.37	-31.25	-14.20	-9.16	0.13	-1.51	-0.21
5IVT	-26.26	-10.23	-4.20	-26.14	-8.98	-3.91	-0.12	-1.25	-0.29
5TYS	-31.76	-11.71	-3.09	-31.50	-10.64	-2.76	-0.26	-1.07	-0.33
5ULT	-18.51	-9.58	-1.06	-18.77	-9.03	-1.10	0.26	-0.55	0.03
5UOV	-36.78	-25.32	-0.16	-37.10	-22.58	0.17	0.32	-2.73	-0.33
6CDJ	-32.87	-17.31	-6.78	-33.08	-15.98	-6.64	0.20	-1.33	-0.14
6CDL	-11.60	-13.63	-0.90	-11.72	-12.82	-0.82	0.12	-0.80	-0.09
6VOE	-31.69	-13.75	-7.78	-21.30	-13.81	-2.60	-10.38	0.06	-5.18
7MYP	-17.19	-9.22	-4.31	-17.19	-8.82	-4.24	0.00	-0.41	-0.07
18624	-6.77	-8.06	4.23	-6.96	-7.39	4.25	0.19	-0.67	-0.03
18629	-9.01	-8.22	3.77	-9.16	-7.50	3.83	0.15	-0.72	-0.05
18630	-6.19	-8.41	3.49	-6.49	-7.62	3.48	0.30	-0.79	0.01
18631	-4.78	-8.01	3.61	-5.02	-7.27	3.63	0.24	-0.74	-0.01
18632	-10.81	-8.45	3.13	-10.98	-7.68	3.18	0.17	-0.76	-0.06
18633	-6.03	-9.12	3.86	-6.16	-8.26	3.95	0.13	-0.86	-0.09
18634	-9.23	-9.02	2.98	-9.39	-8.17	3.05	0.16	-0.84	-0.07
18635	-5.55	-8.80	3.63	-5.69	-7.95	3.71	0.14	-0.84	-0.08
18637	-19.97	-11.21	4.25	-19.75	-9.80	4.61	-0.21	-1.40	-0.36
18638	-23.61	-10.50	3.63	-22.68	-9.04	4.36	-0.93	-1.46	-0.73
18639	-10.28	-10.38	2.11	-10.12	-9.19	2.40	-0.16	-1.19	-0.29

Table 2.2: Continue

18652	-17.17	-10.36	3.05	-16.87	-9.01	3.45	-0.30	-1.34	-0.39
18658	-17.49	-9.25	2.63	-17.49	-8.23	2.81	0.00	-1.02	-0.18
18659	-11.53	-10.09	3.48	-11.73	-9.10	3.56	0.20	-0.99	-0.08

Indeed, one can see the comparison of MM energy terms (Coulombic, van der Waals) along with MM_LIE calculations by taking only reduced water in the system rather than the complete water (Table 2.2).

By applying eqs 2.12 and 2.13, we could successfully produce the absolute binding free energies based on ANI/D3 interactions.

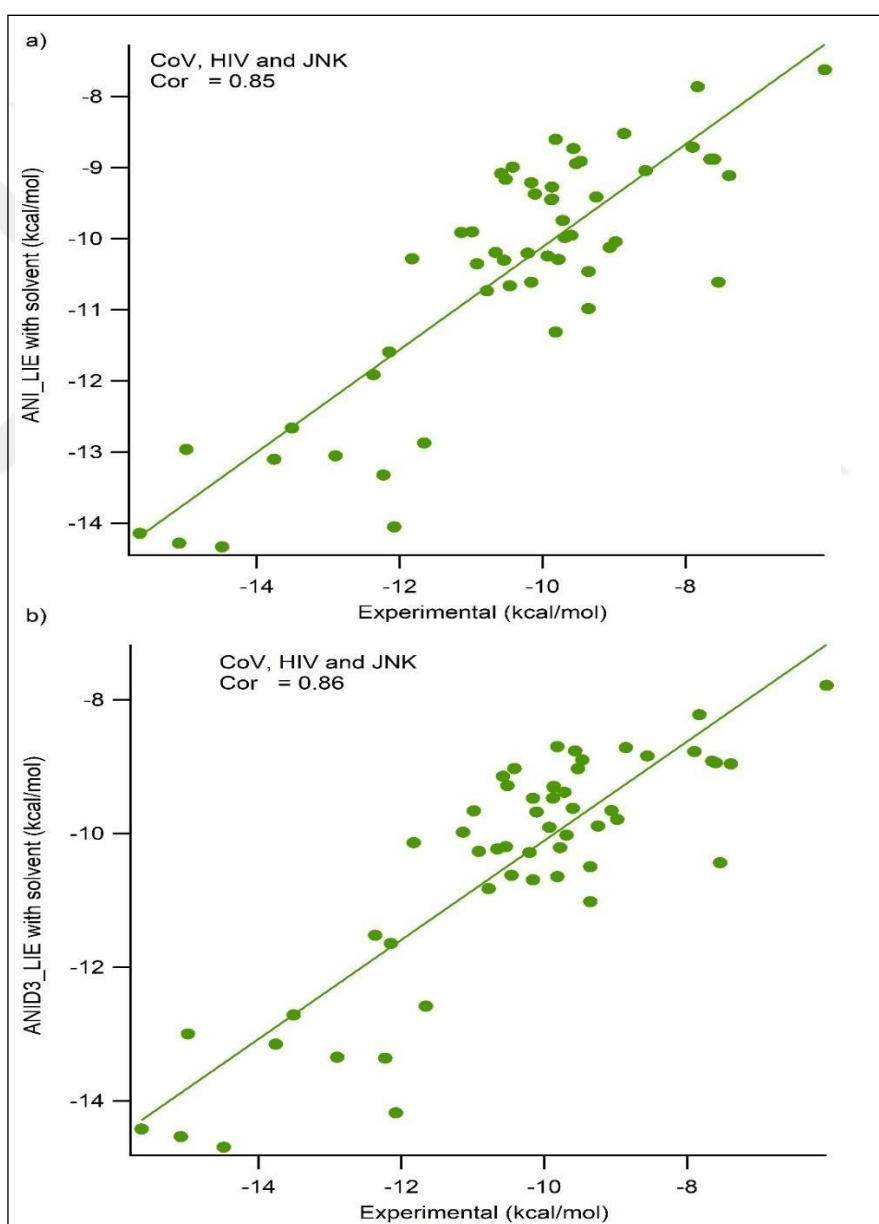


Figure 2.8: LIE calculated by ANI and ANID3 using a) eq 2.12 and b) eq 2.13, in which the solvent effects around the ligand were included in the calculations.

Table 2.3: Coefficient values derived from equations 2.10-2.13.

Equation number	Method	α	β	γ (kcal/mol)
4	ANI_LIE	0	0.106 ± 0.016	-4.987 ± 0.881
5	ANID3_LIE	-0.025 ± 0.028	0.124 ± 0.021	-5.569 ± 0.782
6	ANI_LIE	0	0.127 ± 0.011	-5.111 ± 0.480
7	ANID3_LIE	-0.035 ± 0.029	0.149 ± 0.021	-5.987 ± 0.806
7 (RS)	ANID3_LIE	-0.035 ± 0.008	0.148 ± 0.008	-5.978 ± 0.230
α : van der Waals terms (LJ or D3); β :electrostatic term (ele or ANI); γ : intercept RS: Random sampling as validation (90% training bootstrapped by 100 different selections) and test (10%)				

2.4.3. Benchmarking with Other Methods

2.4.3.1. Comparison to the Classical LIE Approach

In addition to the ANI method introduced here, we also performed calculations from several other methods for the studied protein–ligand complexes. The very first method was the classical LIE method, in which Coulombic and van der Waals interactions at the level of molecular mechanics were computed and fit to the experimental binding free energies to determine the coefficients.

The default parameters of LIE in Gromacs software use $\alpha = 0.181$, $\beta = 0.5$, and $\gamma = 0$ and yield very low correlation to the experimental values. The β coefficient for the electrostatic interactions was reported to have several values according to the charged vs uncharged ligands. Studying 18 PL complexes, Hansson et al. [135] concluded that $\beta = 0.5$ is a good approximation for charged ligands whereas lower values such as 0.43, 0.37, and 0.33 for neutral molecules with 0, 1, or >1 hydroxyl groups, respectively [55]. Since then, it has been used with care by fitting to the new sets of protein–ligand complexes [136]–[141].

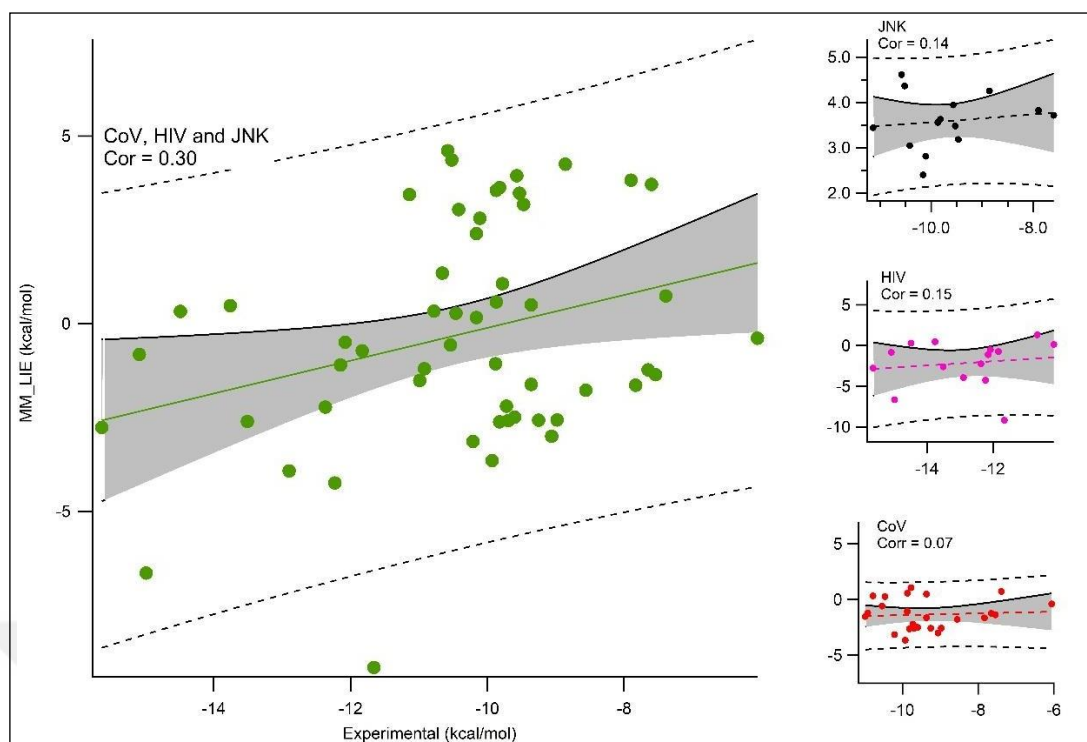


Figure 2.9: The LIE calculated by MM electrostatic and van der Waals interactions using equation 2.3. The preassigned parameters available in Gromacs software were used. The bottom plots show the correlation in each individual protein family.

With the default parameters ($\alpha = 0.181$, $\beta = 0.5$, and $\gamma = 0$), we have observed very low correlation between the experimental and MM_LIE prediction (Figure 2.8). On the other hand, when new parameters are assigned from our fit, we observe 0.70–0.73 correlation with $\alpha = 0.25$, $\beta = -0.06$, and $\gamma = -3.09$ (Figure 2.9). Here, all of the ligands studied are neutral and we did not classify them according to the presence of polar groups like hydroxyl groups. Note that MM_LIE is still quite successful ($R = 0.70$ – 0.73), and it is clear that ANI_LIE ($R = 0.85$ – 0.87) is much better in reproducing experimental values.

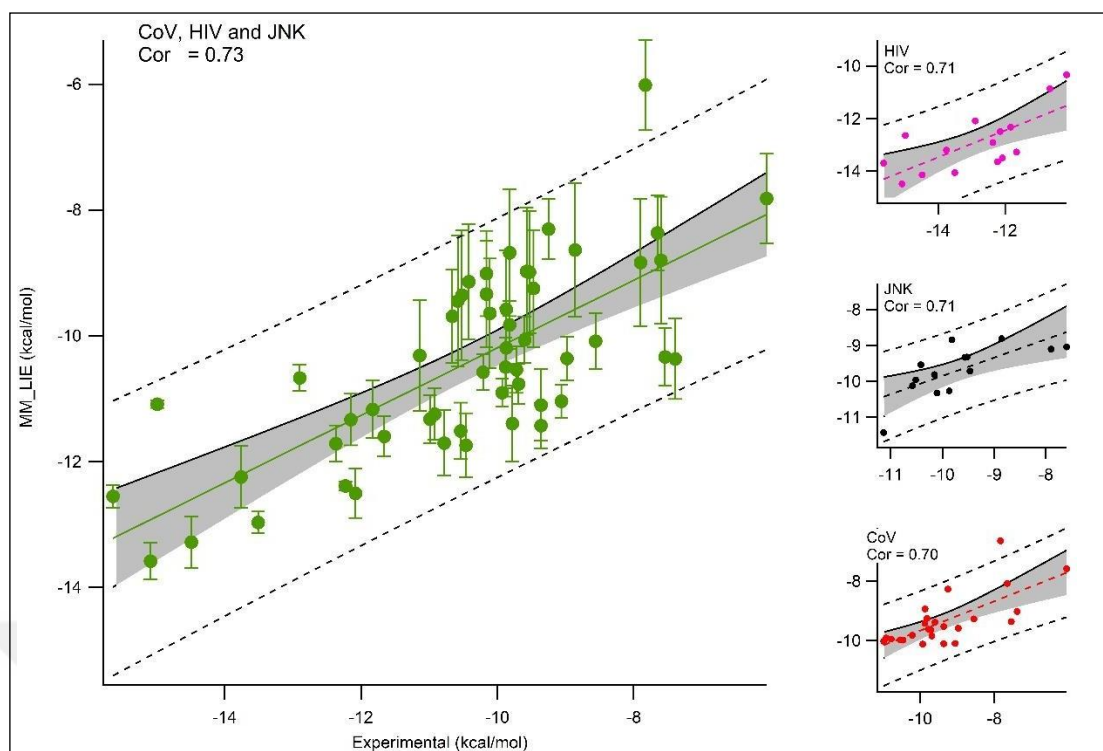


Figure 2.10: LIE calculated by MM electrostatic and van der Waals interactions using eq 2.9. The parameters were determined from the fit. The bottom plots show the correlation in each individual protein family.

In addition to the parameters calculated from all 54 data points of the dataset for both MM_LIE and ANID3_LIE, we have also analyzed the data by splitting training (90%, 48 points) and test sets (10%, 6 points). We have done this by randomly selecting 100 different trains and test sets. The coefficients were produced from the training sets and applied on test sets. The coefficients generated from random sampling and whole dataset are in a good agreement. We have reported the predicted values for the coefficients and their RMSE values on training and test sets.

2.4.3.2. Comparison to the MMPBSA

We have also performed MMPBSA and MMGBSA calculations as a comparison to ANI. The computational cost of these two methods is very similar to ANI since they are also end-state methods applied on bound-state (PLS) classical MD simulations. We should note that these methods do not use empirical fitting coefficients in the calculations. However, their success of predicting absolute binding energies can become very poor according to the system and they are mostly used to predict the relative binding free energies. In particular, when the ligands have similar

scaffolds, they can accurately predict relative binding free energies. Similarly, when the same ligand's binding to a protein and its mutations were compared, they can be quite successful. We have tested the performance of both MMPBSA and MMGBSA methods for the protein– ligand complexes studied here.

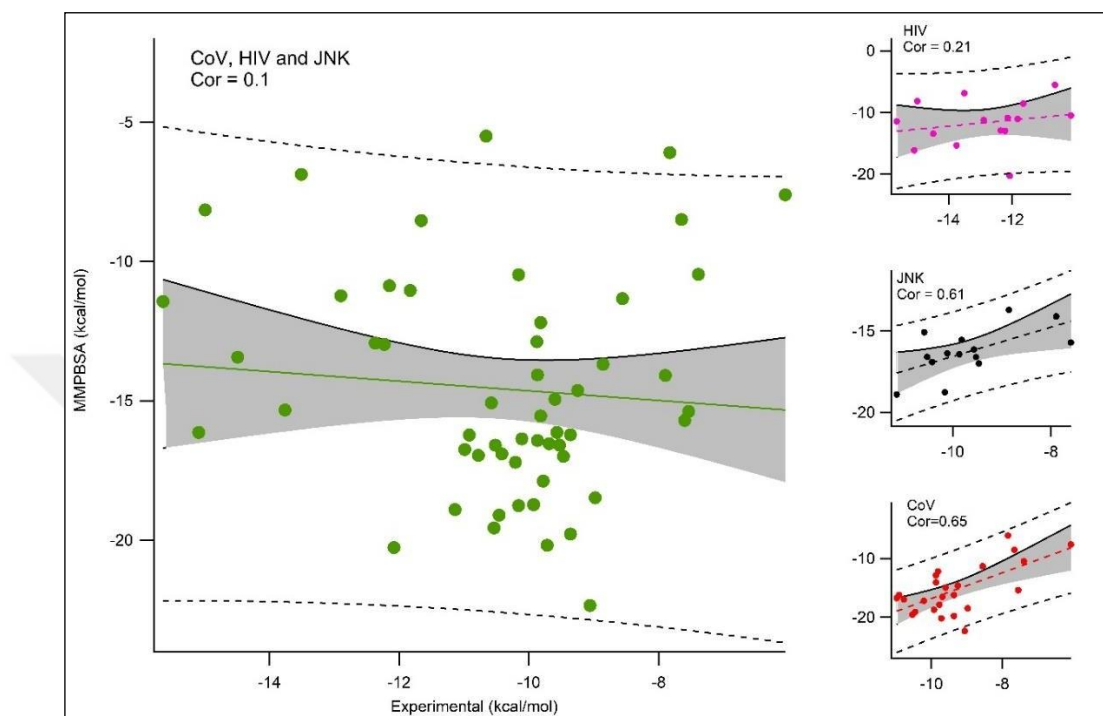


Figure 2.11: MMPBSA results show low correlation due to failure to predict BFEs in HIV-1 protease. The bottom plots show the correlation in each individual protein family.

Due to the fact that protein families of the ligands are very different, we could only observe somewhat meaningful correlations when the protein families were analyzed separately in the case of MMPBSA. When all proteins were treated with the same preassigned values of MMPBSA, we observed very poor correlation to the experimental values (Figure 2.10).

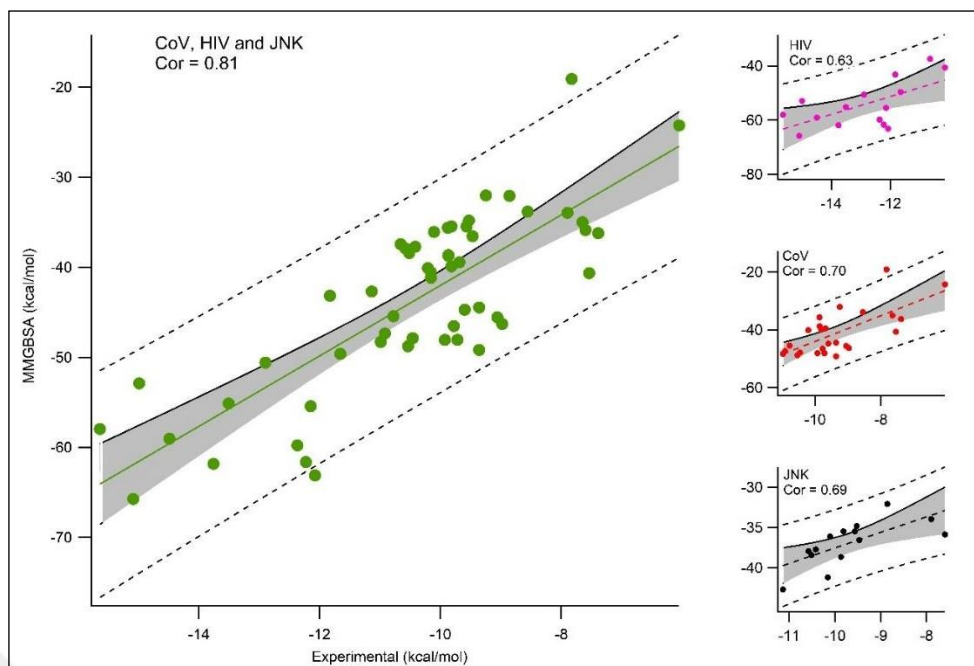


Figure 2.12: MMGBSA results showing higher correlation than LIE, but still lower than ANI_LIE. The bottom plots show the correlation in each individual protein family.

Surprisingly, MMGBSA showed a somewhat similar correlation ($R = 0.81$) with classical LIE to the experiments even when all 54 protein–ligand complexes were analyzed together (Figure 2.11). This is still less than the ANI-predicted values of $R = 0.87$ – 0.88 .

2.4.3.3. Comparison to alchemical calculations

We have also explored the binding energies by means of alchemical methods using a 10λ decoupling window for electrostatic interaction and 10λ for van der Waals interactions. During decoupling, no restraint was applied on the ligands; 10 ns from each window brings 20-fold computational cost to the simulations, whereas it still has a bare correlation to the experimental binding free energies (Figure 2.12). These low correlations may be due to the limited number of windows.

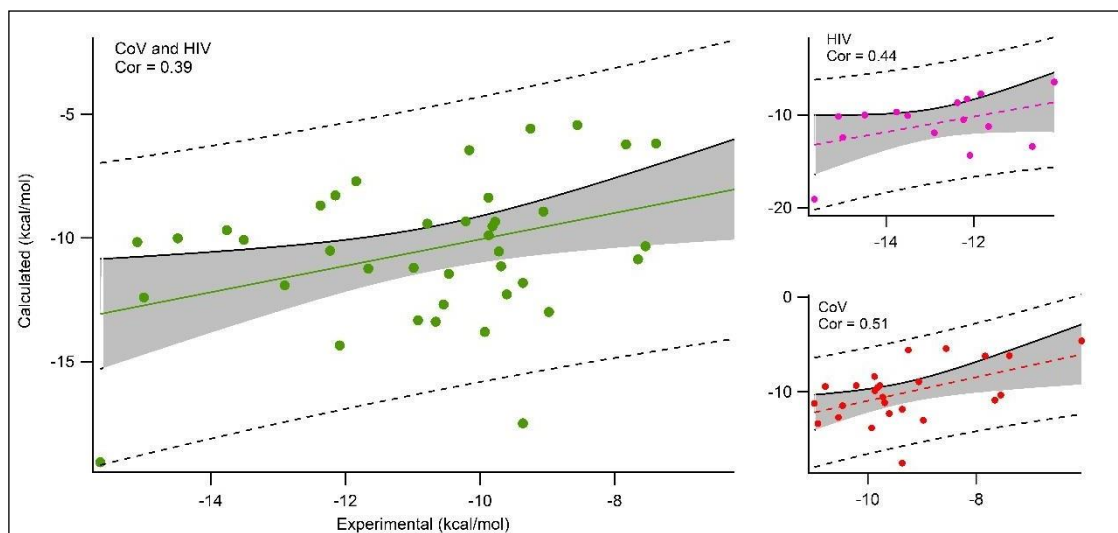


Figure 2.13: BAR results for CoV and HIV compared with experimental results when they are treated together or individually.

2.5. Summary and Conclusion

Here, we applied ANI-ML potentials, trained at the wb97x/6-31G* level, to calculate the interaction energies of protein–ligand and ligand–solvent pairs. The results show that the predicted interaction energies even when the solvent is totally ignored are highly correlated with experimental values. By modification of the LIE method, we have assigned coefficients for the ANI_LIE method. The ANI_LIE method outperforms conventional methods like LIE, BAR, and MMP(G)BSA in terms of accuracy and is comparable to MMP(B)SA in terms of computational cost. Our preassigned parameters determined for the method can be applied on any protein–ligand complex systems even when the protein(s)/ligand(s) have different scaffolds.

3. MODIFICATION OF MM/PBSA BY ANI: PERFORMANCE IMPROVEMENTS OF MM-PBSA BY ANI-ML POTENTIALS, MC BASED SURFACE AREA/VOLUME, AND TWO VALUED INTERIOR DIELECTRIC CONSTANT

3.1. Introduction

Accurate and fast prediction of free energy changes is a substantial research field in computational chemistry [142]. All chemical and biomolecular events are governed by the foundation of thermodynamic processes. Experimental techniques can become too costly, time consuming and not informative about the dynamic properties [143]–[145]. In particular, in biomolecular events such as binding free energy of a ligand to a macromolecule has been great deal in drug discovery. Computation of binding free energies (BFEs) and screening large libraries of potential drug candidates based on BFEs has been a routine process in rational drug design studies [146]. However, the computational studies require the most realistic modeling of the system. Due to the large size such that biomacromolecule-small ligand complex surrounded with explicit water molecules, the system cannot be simply simulated at the quantum mechanical (QM) level. At the current computational power, most parts of these systems, if not all, must be defined at the molecular mechanics (MM) level [59].

Molecular dynamics (MD) simulations have been widely used for BFE calculations. The methods to calculate the BFEs using MD simulations can be classified into two classes. The first, BFEs are predicted by using only bound and unbound states of the system so called end-state methods. Second, the system is evolved from bound to unbound state by physical or alchemical pathway via (ideally) infinitesimal changes of the energy function [41]. Although the pathway methods are more rigorous, they do not perform well in the trade-off between accuracy and computational cost [142]. On the contrary, the end-state methods, although more approximate methods, have been widely used in drug discovery [147], [148]. The most common end-state methods are linear interaction energy (LIE) [135], Molecular Mechanics-Poisson Boltzmann Surface Area (MMPBSA) and Generalized Born Surface Area (MMGBSA) [149], [150]. LIE uses averaged ensembles of interaction

energies between ligand and surroundings in both bound and unbound states and BFE is predicted by empirical coefficients produced by fitting to experimental data.

MM-P(G)BSA methods find BFE by composing contributions from gas phase binding energy (i.e. MM energy term), solvation, and entropy. The implicit solvation term is also assembled by contribution of polar (PB or GB) and non polar solvation based on surface area (SASA) or volume (SAV) energy terms [151]. These contributions to BFE can be collected by following two ways as separate trajectory protocol and single trajectory protocol.

In separate trajectory protocol (3A-MMP(G)BSA), independent MD simulations are performed for L in water (LS), apo-protein in water (PS) and protein-ligand complex in water (PLS). This protocol is more realistic approach to the experimental process since the conformational sampling of apo-protein and free ligand may differ from that of the complex. However, running separate MD simulations may bring additional errors and noise due to large internal energy fluctuations at different conformations. In most of the cases, these errors cannot be fixed by longer MD simulations. Indeed, studies have shown that longer MD simulations worsen the BFE prediction of the MMP(G)BSA calculations. On the other hand, these errors can be minimized by increasing the number of replicas in simulations at the end points. Su et al have shown that 4 ns of simulation time is sufficient along with 30-50 replica simulations [152].

In single trajectory protocol (MMP(G)BSA), only a single MD simulation of PLS is performed and the individual components of P and L are extracted from the same trajectory. Although this protocol suffers from the lack of conformational sampling of apo-protein and free ligand, the error due to the fluctuation of energy terms is much smaller since it is cancelled by the use of the same MD frames for each component. In addition, the computational cost is much less than the separate trajectory protocol. Therefore, the single trajectory protocol is more widespread in the field.

There have been several attempts to improve prediction power of MMP(G)BSA in literature [59], [153]–[155]. Wang et al have recently reviewed these studies in detail [142]. Most of the studies have been focused on correcting PB/GB and SA terms in the MMP(G)BSA energy formalism. Since efficiency and accuracy of these methods are system dependent, there is no consensus on the coefficients used to convert these terms in the energy equation. In Amber22

software, there are 6 regimes to calculate PBSA and regimes for GBSA contributions.

In the original MMPBSA calculation using SAV for the non-polar solvation term, the volume is calculated by integration of SASA [156], in which a spherical probe samples the accessible surface of the protein-ligand complex. Recently, Ongari et al [157] have introduced the use of accessible surface area (ASA), probe occupiable accessible volume (POAV) and probe centered accessible volume (PCAV) for the internal voids of microporous materials available in Zeo++ software [158]. In these strategies, the area and volume of cavities (or voids) are determined by Monte Carlo simulations of the spherical probe rather than integration of SASA. Ideally, ASA in Zeo++ should correspond to SASA in AMBER. Similarly, PCAV corresponds to SAV in protein-ligand complexes except for the calculation method. POAV differs from PCAV by including the additional pore volume due to the probe radius, which is not considered in PCAV (or SAV).

Here, we attempt to improve the BFE prediction by thoroughly assessing the performance of MM-P(G)BSA from four different perspectives. First, conventional MMPBSA is analyzed using different regimes covering SASA only, SAV only and SPT models in non polar solvation contribution part of the binding free energy. Second, the applicability of using POAV/PCAV rather than SAV in the MMP(G)BSA have been tested. Next, ANI-ML potentials are adopted as a replacement of MM terms in implicit solvent approach so called ANI_PBSA to predict BFEs with higher accuracy. In our previous work, we have introduced ANI_LIE method which uses ANI-2x machine learned (ML) potentials in the LIE formalism and showed significant improvements in BFE predictions over the classical LIE approach. Finally, the use of two different dielectric constants in the same calculation has been tested.

3.2. Theory

3.2.1. MM-P(G)BSA

In the single trajectory MM-PBSA or MM-GBSA, the free energy change for binding of a ligand (L) to a protein (P) to form a complex (PL) can be expressed as

$$\Delta G_{bind} = \langle G_{PL} - G_P - G_L \rangle_{PLS} \quad (3.1)$$

where $\langle \rangle_{PLS}$ shows the averaged ensemble from the simulation carried in protein-ligand-solvent (PLS) system. Here each individual component (PL, P, and L) are extracted from each of MD frames. The binding free energy can be decomposed into [159]:

$$\Delta G_{bind} = \Delta H - T\Delta S = \Delta E_{MM} + \Delta G_{solv} - T\Delta S \quad (3.2)$$

in which:

$$\Delta E_{MM} = \Delta E_{bonded} + \Delta E_{nb} = \Delta E_{bonded} + \Delta E_{ele} + \Delta E_{vdw} \quad (3.3)$$

where ΔE_{MM} , ΔG_{solv} and $-T\Delta S$ are the changes in the gas phase MM energy, solvation free energy and conformational entropy changes upon ligand binding.

ΔE_{MM} includes the energy from ΔE_{bonded} bonded terms (bond, angle, dihedral) and ΔE_{nb} non-bonding terms as ΔE_{ele} electrostatic and ΔE_{vdw} van der Waals interactions. For the non-covalent interaction between protein and ligand ΔE_{bonded} is basically zero and thus;

$$\Delta E_{MM} = \Delta E_{ele} + \Delta E_{vdw} \quad (3.4)$$

The solvation term is also decomposed as

$$\Delta G_{solv} = \Delta G_{polar} + \Delta G_{np} = \Delta G_{PB/GB} + \Delta G_{SA} \quad (3.5)$$

where the $\Delta G_{PB/GB}$ is the polar (electrostatic) contribution to the solvation free energy calculated by PB or GB model. Ideally, interior (solute) and exterior (solvent) dielectric constants are set to $\epsilon_{int} = 1$ and $\epsilon_{ext} = 80$, respectively [160]. Previous

studies have shown that the dielectric constant of the solute (interior dielectric constant) should be larger than 1 for highly charged solutes and thus need to be adjusted for better performance [161]–[164]. ΔG_{SA} is the non-polar contribution between solute and continuum solvent, which has been calculated either by solvent-accessible surface area (SASA) or solvent accessible volume (SAV) enclosed by SASA.

3.2.1.1. One Term (SA Only) Model for ΔG_{np}

The non-polar solvation term is expressed by an empirical function of SASA or SAV in this model [165], [166]:

$$\Delta G_{np}^{\Delta SA} = \gamma \cdot \Delta SASA + b \quad (3.6)$$

$$\Delta G_{np}^{SAV} = p \cdot \Delta SAV + b \quad (3.7)$$

where γ , p and b are surface tension constant, solvent pressure constant and correction terms, respectively ($\gamma = 0.00542 \text{ kcal} \cdot \text{mol}^{-1} \cdot \text{\AA}^{-2}$ and $b = 0.92 \text{ kcal/mol}$; $\gamma = 0.005 \text{ kcal} \cdot \text{mol}^{-1} \cdot \text{\AA}^{-2}$ and $b = 0$; or $\gamma = 0.0072 \text{ kcal} \cdot \text{mol}^{-1} \cdot \text{\AA}^{-2}$ and $b = 0$ in the AMBER package). The use of both SASA and SAV, which is called scaled particle theory (SPT) [166], has also been tested and showed better results

$$\Delta G_{np}^{\Delta SA} = \gamma \cdot \Delta SASA + p \cdot \Delta SAV + b \quad (3.8)$$

We note that the herein described SPT model, just like the other one-term models, corresponds to repulsive terms only, ΔG_{rep} , (i.e. ignores the attraction terms) in two-term model explained below. It has been shown that for small spherical cavities, ΔG_{rep} is more correlated with the cavity volume for the cavity radius below 10 Å whereas for larger cavity radius, it is more correlated with the cavity surface [167], [168]. The SPT model rather than SASA only and SAV only models could be more beneficial to reflect this crossover since it includes both terms in the equation.

3.2.1.2. Two Terms (Cavity-Dispersion) Model for ΔG_{np}

The true nature of non polar solvation term comes from repulsive and attractive components between solute and explicit solvents [165].

$$\Delta G_{np} = \Delta G_{rep} + \Delta G_{att} \quad (3.9)$$

where ΔG_{rep} is the solvation free energy contribution from solute-solvent repulsive interactions and the formation of solute cavity. ΔG_{att} is the solvent-solute attractive non-polar interactions and it also includes the solvent-solvent reorganization component. Therefore, in the two terms model, non-polar energy is further divided into cavity and dispersion contributions [165].

$$\Delta G_{cav} \approx \Delta G_{rep} \quad (3.10)$$

$$\Delta G_{vdW} \approx \Delta G_{disp} \approx \Delta G_{att} \quad (3.11)$$

$$\Delta G_{np} = \Delta G_{cav} + \Delta G_{disp} \quad (3.12)$$

where ΔG_{cav} is the cavity formation free energy corresponding to repulsive terms and ΔG_{disp} is the dispersion free energy corresponding to attractive non-polar interactions between solute and solvents. For the attractive term or dispersion contributions, several schemes have been tested such as using the attractive part of Lennard-Jones 6/12, Weeks-Chandler-Andersen (WCA), and σ -schemes [165]. For the repulsive part, SASA or SAV have been used in the same formula of one term approach. So,

$$\Delta G_{cav} = \gamma \cdot \Delta SA + b \quad (3.13)$$

$$\Delta G_{np}^{CD} = \gamma \cdot \Delta SA + b + \Delta G_{disp} \quad (3.14)$$

Similarly, the accessible volume instead of area can be used in these equations. The dispersion term, ΔG_{disp} , can also be calculated by solvent-accessible volume (or surface) integration (The scaling factors, $\gamma = 0.0378 \text{ kcal} \cdot \text{mol}^{-1} \cdot \text{\AA}^{-2}$ and $b = -0.5692 \text{ kcal/mol}$ in the AMBER package). Although it is the integration of the volume, we think (and show) it is still a direct function of solvent accessible volume and can be approximated to;

$$\Delta G_{disp} \cong m. \Delta SAV \quad (3.15)$$

Implementing this approximation in the two-term formula, the equation becomes similar to one-term approach of SASA-SAV model (i.e. SPT theory) or SAV-only model with new coefficients. If area as the repulsive (cavity) term is used;

$$\Delta G_{np}^{\Delta A} = \gamma. \Delta SASA + m. \Delta SAV + b \quad (3.16)$$

If the volume is used;

$$\Delta G_{np}^{\Delta A} = \gamma. \Delta SAV + b + m. \Delta SAV + n = (\gamma + m). \Delta SAV + b \quad (3.17)$$

If both area and volume are used:

$$\begin{aligned} \Delta G_{np}^{\Delta A} &= \gamma. \Delta SASA + p. \Delta SAV + b + m. \Delta SAV + n \\ &= \gamma. \Delta SASA + (p + m). \Delta SAV + b \end{aligned} \quad (3.18)$$

It is clear that if the dispersion energy term in the two term model is linear to accessible volume, one term model is reached in either case. The only change will be in the weight factors of SASA and SAV terms. Recently, Zhang et al. [45] have proposed a new approach for energy estimator, PBSA_E, which is defined as;

$$\begin{aligned} MMPBSA_E &= \alpha_1. (\Delta E_{ele} + \Delta G_{PB}) + \alpha_2. \Delta E_{vdW} + \alpha_3. \Delta SASA \\ &\quad + \alpha_4 N_{rot} \end{aligned} \quad (3.19)$$

where $\gamma=0.0072 \text{ kcal.mol}^{-1}.\text{\AA}^{-2}$ and $\alpha_1, \alpha_2, \alpha_3$ and α_4 coefficients are 0.03037, 0.07791, 1.2193 and 0.1854, respectively. Here, N_{rot} was introduced as rotatable bonds to account for entropic contribution. This formula can be extended to entirely empirical parameters of all energy components and can be even represented by;

$$\begin{aligned} MMPBSA_{Ex} = & \alpha_1. \Delta E_{ele} + \alpha_2. \Delta E_{vdW} + \alpha_3. \Delta G_{PB} \\ & + \alpha_4. \Delta SASA + \alpha_5. \Delta SAV + b \end{aligned} \quad (3.20)$$

3.2.2. ANI Replacement with MM Terms

Previously, we have shown that ANI-ML potentials can be utilized in the linear interaction energy (LIE) formalism by replacing of ANI with MM term [169]. Herein, we borrow the same change in the MM/PBSA formalism so as to test ANI_PBSA approach. The updated formula for the free energy of binding would be;

$$a. \Delta E_{ANI} + \Delta G_{PB} + \gamma. \Delta SA + b \quad (3.21)$$

one-term (SA only) model

$$a. \Delta E_{ANI} + \Delta G_{PB} + \gamma. \Delta SA + b + \Delta E_{Disp} \quad (3.22)$$

two-term (SAV+Disp) model

$$a. \Delta E_{ANI} + \Delta G_{PB} + b. \Delta SASA + c. \Delta SAV + d \quad (3.23)$$

SPT model

$$a. \Delta E_{ANI} + b. \Delta G_{PB} + c. \Delta SASA + d. \Delta SAV + d \quad (3.24)$$

PBSA_E model

In order to calculate the non-polar contribution to the solvation free energy, we have assessed the use of ASA instead of SASA, PCAV or POAV instead of SAV in all the equations.

3.3. Computational Methods

3.3.1. System Preparation

For the binding studies, the crystal structures of non-covalent ligands complexed with 3CLpro of SARS-CoV-1, SARS-CoV-2 and HIV-1 proteases, a total of 54 protein-ligand complex with known IC₅₀ values, were retrieved from Protein Data Bank (PDB) except for c-Jun N-terminal kinases, which were retrieved from the work by Khalak et al. [36]. Using Gaussian 16 software, the model ligands were first optimized at the B3LYP/6-31G* and ESP charges were generated at the HF/6-31G* level. Using Antechamber module in AmberTools 2021, RESP charges and GAFF force field atom types were generated. Amber2gmx module from AmberTools 2021 was used to convert amber type input files to Gromacs.

3.3.2. MD Simulations

The molecular dynamics simulations were carried out using Gromacs 2018+ software package [125] with all-atom model of Amber ff99SB-ILDN [131]–[133] force field implemented in Gromacs. The protein-ligand complex (~6000 atoms) was placed in the center of a dodecahedron box. Each system was solvated in the TIP3P model type water [126] with a cell margin distance of 10 Å for each dimension. The protein-ligand systems with ~70,000 atoms were neutralized in 0.15 M NaCl. Classical harmonic motions of all bonds were constrained to their equilibrium values with LINCS algorithm.

Energy minimization was carried out to a maximum 100 kJ.mol⁻¹.nm⁻¹ force using Verlet cutoff scheme. For both long range electrostatic and Van der Waals interactions, a cutoff length of 12 Å was used with the Particle Mesh Ewald method (PME) (4th order interpolation). The neighbor list update frequency was set

to 20 ps⁻¹. Two-steps energy minimization and equilibration schemes were used. Each minimization step consisted of up to a 50000 cycle of Steepest Descent and a subsequent 50000 cycles of l-bfgs integrators.

After minimization, each system was equilibrated within three steps using Langevin Dynamics. The first step consisted of a 1 ns of NVT ensemble. The protein-ligand and rest of the system were defined as two temperature groups at 310 K. The next steps consisted of NPT ensembles in which the systems were equilibrated to the 1 atm pressure by Berendsen for 200 ps and followed by Parrinello-Rahman isotropic pressure coupling for 1 ns to a reference pressure of 1 atm. When systems reached to equilibrium, an MD simulation of 10 ns were carried out at NPT ensemble.

3.3.3. PBSA Calculations

MMPBSA calculations were performed using the trajectories of Gromacs MD simulations via gmx_MMPBSA package [170], [171]. Our custom scripts were used to combine ANI and PBSA calculations as ANI_PBSA.

Statistical analysis were performed by using Igor Pro and custom python scripts for predictive index (PI) as outlined in [172], Pearson R, Spearman R, MUE, MUEtr and MUEsc [147].

3.4. Results and Discussion

3.4.1. MMPBSA by SASA and/or SAV

Single trajectory approach MMP(G)BSA calculations can be carried out using several schemes as outlined in the Theory section at this chapter. The most popular methods, which are also embedded in Amber22 software, are SASA only and SAV+Disp models. In addition, SAV only model can be modulated, but it mostly mimics the results of SASA only model. Although SASA is in area unit with numeric values ranging from ~10000 to 15000 Å² and SAV is in volume unit ranging from ~40000 to 80000 Å³ for protein-ligand complexes (PL) for our case, the difference values (PL-P-L) on both SASA and SAV are very similar in the range of -500 to -1000 Å² or Å³ and thus the coefficient values of γ and p in equations 3.1-3.2 are very similar, thus they can be simply represented with the same coefficient. Therefore,

SASA only model and SAV only model yield almost identical results in reproducing experimental relative binding free energies (RBFE).

We have first tested the performance of predefined coefficients using SASA (or SAV) only model (Table 3.1). Setting the interior dielectric constant to 1 ($\epsilon_{\text{int}}=1$), we could only get Pearson's correlation coefficient, $R=0.42-0.47$. However, when $\epsilon_{\text{int}}=2.1$, we observed dramatic increase in the correlation coefficient to $R=0.78-0.79$. We have also tested different ϵ_{int} values (up to 10) and got no further significant improvement. In addition to single term non-polar model, we have also tested the two-term approach in which SASA (or SAV) combined with dispersion (Disp) using predefined values of $\gamma=0.0378$ and $b=-0.5692$. Unfortunately, this yielded very poor correlation (even reverse correlation) to the experimental values with $\epsilon_{\text{int}}=1$. By increasing ϵ_{int} (to 10), we could produce $R=0.50$ correlation. More of ϵ_{int} will be discussed later in the chapter. Instead of using predefined coefficients, we have also fitted the SASA (or SAV) in both one and two term models using the experimental BFE of the 54 PL complexes. For one term model, we improved the correlation by 5% (using $\gamma=0.08$) with $\epsilon_{\text{int}}=1$ or 2.1. This correlation could not be further improved by changing ϵ_{int} . In the SPT model, which is not a default model in Amber22, we produced slightly better fit to the experimental binding free energies ($R=0.84$). Figure 1.1 shows correlation between experimental binding free energy values of PL complexes and SASA only/SPT regimes by setting $\epsilon_{\text{int}}=2.1$. Interestingly, the success of this model was not affected by different ϵ_{int} values. Table 3.1 shows the summary of Pearson coefficients produced by original and our fit parameters.

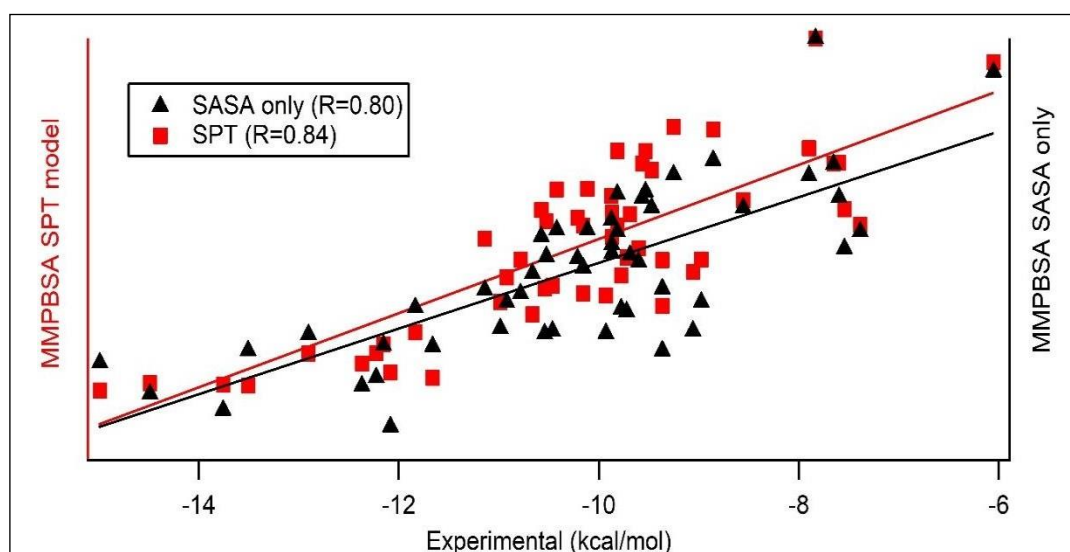


Figure 3.1: The experimental and calculated binding free energies by MMPBSA using SASA only/SPT regimes for non-polar contribution where $\epsilon_{int}=2.1$ for electrostatic and polar solvation terms.

Table 3.1: MMPBSA calculated by one term, two term and SPT approaches for non-polar contribution to solvation free energy.

	SASA_only			SAV_only			SAV+Disp		SPT
$\epsilon_{int}=1$	GB1=PB4 $\gamma=0.0072$	GB2=PB5 $\gamma=0.005$	Our $\gamma=0.0836$	GB1=PB4 $\gamma=0.0072$	GB2=PB5 $\gamma=0.005$	Our $\gamma=0.0785$	PB1 $\gamma=0.0378$, $b=-0.5692$	Our $\gamma=0.1481$	Our $\gamma=0.139$ 4, $p=0.016$ 3
PI	0.42	0.36	0.86	0.43	0.36	0.85	-0.51	0.83	0.87
Pearson R	0.47	0.42	0.78	0.47	0.42	0.82	-0.54	0.80	0.84
Spearman R	0.39	0.34	0.72	0.40	0.33	0.77	-0.45	0.76	0.80
MUE	24.52	22.31	101.35	24.46	22.26	95.51	10.70	99.92	101.86
MUEtr	3.79	3.84	13.34	3.85	3.87	13.00	7.40	13.24	13.32
MUEsc	2.93	3.44	1.04	2.93	3.43	1.10	2.61	1.20	1.03
$\epsilon_{int}=2.1$	$\gamma=0.0582$			$\gamma=0.0538$			$\gamma=0.1234$		$\gamma=0.09$ 33, $p=0.03$ 43
PI	0.83	0.83	0.85	0.82	0.83	0.84	-0.13	0.82	0.86
Pearson R	0.79	0.79	0.80	0.79	0.78	0.83	-0.11	0.80	0.84
Spearman R	0.76	0.75	0.78	0.74	0.75	0.76	-0.09	0.75	0.78
MUE	39.20	36.99	90.46	39.14	36.95	85.62	5.39	90.03	91.58
MUEtr	5.31	5.00	13.26	5.42	5.08	12.92	3.88	13.26	13.17
MUEsc	1.22	1.25	1.07	1.26	1.28	1.11	15.11	1.23	1.06
PI: predictive index, MUE: Mean unsigned error; MUEtr: Mean unsigned error after subtraction of average signed error, MUEsc: MUE rescaled by slope and intercept of the predicted vs experimental results									

As discussed in the Theory section, when the dispersion energies are a linear function of SASA, SAV+Disp model should converge to the same equation with the SPT model (with new weight factors). Figure 3.2 shows the correlation between dispersion energy and SASA (or SAV enclosed by SASA). Therefore, using SASA+SAV terms (SPT model) instead of SAV+Disp (two term model) should produce very similar results. We find the SPT model slightly better results than SAV+Disp model since the coefficient for Disp is set to 1 (i.e., not a fitting parameter) in the case of SAV+Disp whereas it is also fitted in the SPT model.

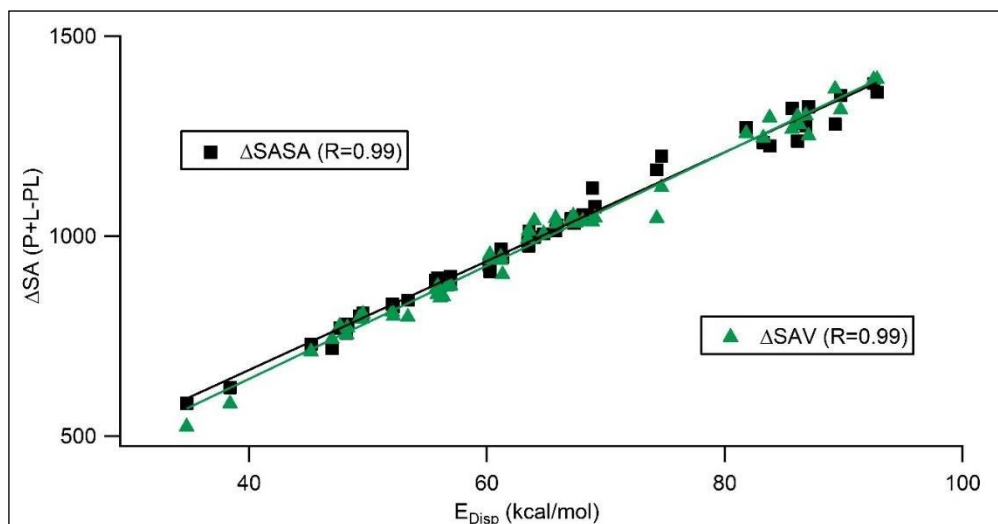


Figure 3.2: The relationship between SASA (or SAV enclosed by SASA) and E_{disp} (calculated by integration of SASA). The values are of averages of PL-P-L differences over 100 MD snapshots.

When all energy terms are defined with empirical parameters as in the MMPBSA_E formalism in equation [19] using original values of coefficients, we observed the Pearson correlation of 0.84, which is almost the same as SPT model. However, when we used our fitting parameters, we could slightly improve the correlation to 0.85-0.86. In this model, electrostatic MM energy and polar solvation energy difference (ΔE_{ele} and ΔG_{PB}) are scaled with the same scaling factor. The logical idea is that since both terms are electrostatic and use the simple Coulombic potential, they must be scaled with the same coefficient, which corresponds the effect of interior dielectric constant. Using the extended version of MMPBSA_Ex in equation [20], the correlation can be improved up to 0.91. Table 3.2 shows Pearson correlation coefficients using all these different regimes.

Table 3.2: MMPBSA_E by SASA only and MMPBSA_Ex (extended) by SPT models for non-polar contribution to solvation free energy.

	$\epsilon_{int}=1$			$\epsilon_{int}=2.1$		
	MMPBSA_E by SASA only		MMPBSA_Ex by SPT	MMPBSA_E by SASA only		MMPBSA_Ex by SPT
	Original $\alpha_1=0.0303$ $\alpha_2=0.0779$ $\alpha_3=0.0072$	Our fit $\alpha_1=0.5566$ $\alpha_2=0.1969$ $\alpha_3=0.1046$	Our fit $\alpha_1=1.0265$ $\alpha_2=0.8330$ $\alpha_3=0.0894$ $\alpha_4=0.0870$ $\alpha_5=-0.0838$	Original $\alpha_1=0.0303$ $\alpha_2=0.0779$ $\alpha_3=0.0072$	Our fit $\alpha_1=1.2792$ $\alpha_2=0.2029$ $\alpha_3=0.1057$	Our fit $\alpha_1=2.1878$ $\alpha_2=0.8444$ $\alpha_3=0.2315$ $\alpha_4=0.0878$ $\alpha_5=-0.0841$
PI	0.85	0.86	0.90	0.84	0.86	0.90
Pearson R	0.84	0.85	0.91	0.84	0.85	0.91
Spearman R	0.77	0.79	0.86	0.77	0.79	0.86
MUE	0.98	90.46	62.38	1.16	91.13	62.81
MUEtr	0.94	12.98	12.26	0.99	12.98	12.27
MUEsc	1.08	1.03	0.77	1.10	1.02	0.77
PI: predictive index, MUE: Mean unsigned error; MUEtr: Mean unsigned error after subtraction of average signed error, MUEsc: MUE rescaled by slope and intercept of the predicted vs experimental results						

3.4.2. ASA/PCAV(POAV) in Replacement of SASA/SAV

In addition to the results of superiority of SPT model over the other tested schemes, we also investigated the area and volume calculations using a different approach. In Zeo++, the accessible surface area (ASA), probe centered accessible volume (PCAV) and probe occupiable accessible volume (POAV) could be utilized for the effect of area/volume on the success of binding free estimation in MMPBSA calculations. Ideally, ASA corresponds to SASA and PCAV corresponds to SAV. The only difference is that Zeo++ uses Monte Carlo sampling of the probe in calculation of void of the bulk structure. Since we are interested in molecular surface rather than pore volume, we have subtracted Zeo++ calculated pore volumes from the total volume of the cell (Figure 3.3). On the other hand, POAV eliminates the volume which is over calculated due to the probe radius.

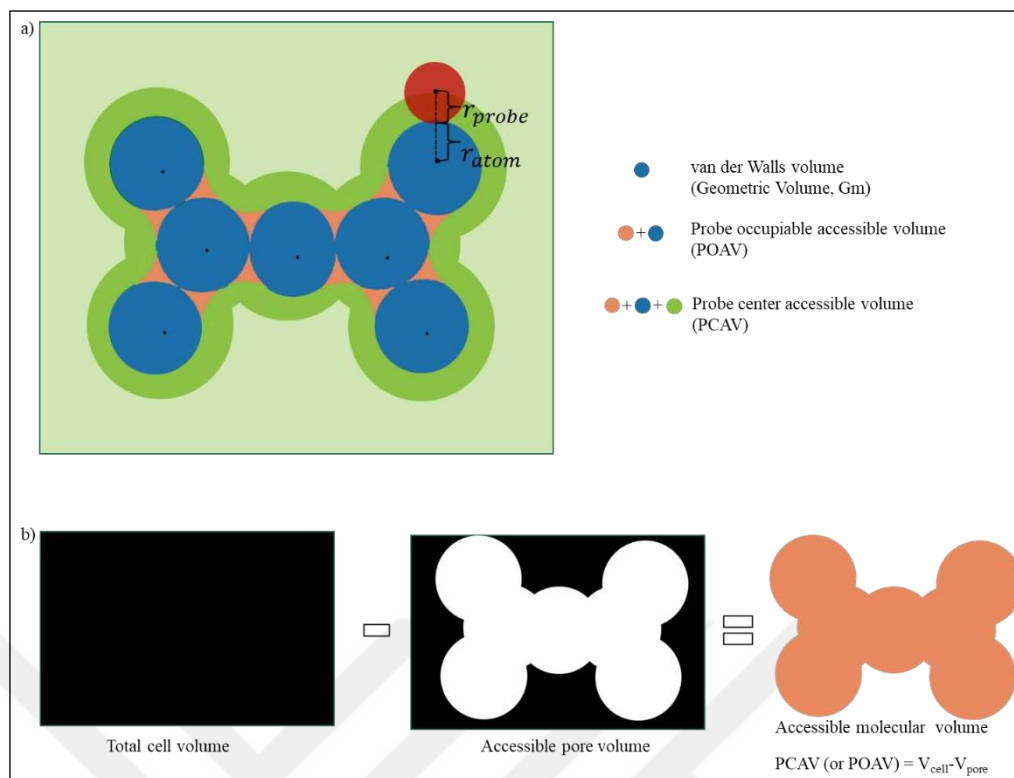


Figure 3.3: Illustration of pore volume produced by Zeo++ showing a) covered regions in PCAV and POAV definition b) how we calculated the molecular surface from pore volume.

We first compared the values of ASA generated by Zeo++ instead of SASA from the Amber software. They are almost identical and as expected there is no improvement in the MMPBSA calculations when SASA is replaced by ASA. On the other hand, PCAV generated by Zeo++ and SAV from the Amber software produced slightly different results (Table 3.3). When the MMPBSA equations are modified to include ASA/PCAV rather than SASA/SAV in SPT model, the R values are slightly improved from 0.84 to 0.87 whereas the use of ASA/POAV has yielded very similar results with the original SPT model (i.e., SASA/SAV). Similarly, MMPBSA_Ex is also improved by ~3% (to R=0.91) by using ASA/PCAV or ASA/POAV instead of SASA/SAV.

Table 3.3: MMPBSA calculated by Zeo++ produced SA terms for non-polar contribution to solvation free energy.

	One-term			Two-term		SPT	
$\epsilon_{int}=1$	ASA only $\gamma=0.08525$	PCAV only $\gamma=0.06748$	POAV only $\gamma=-0.1150$	PCAV+Disp $\gamma=0.1327$	POAV+Disp $\gamma=-0.1907$	ASA+PCAV $\gamma=0.1151,$ $p=-0.0326$	ASA+POAV $\gamma=0.0797,$ $p=-0.0299$
PI	0.87	0.74	0.55	0.45	-0.36	0.88	0.87
Pearson R	0.85	0.70	0.53	0.45	-0.31	0.86	0.85
Spearman R	0.81	0.66	0.47	0.38	-0.34	0.83	0.80
MUE	101.87	70.75	24.77	57.44	35.28	105.67	98.29
MUEtr	13.36	11.38	6.42	13.03	13.19	13.49	13.25
MUEsc	1.02	1.63	2.28	3.35	4.85	0.94	0.97
$\epsilon_{int}=2.1$	$\gamma=0.0594$	$\gamma=0.0414$	$\gamma=-0.0925$	$\gamma=0.1067$	$\gamma=-0.1682$	$\gamma=0.1052,$ $p=-0.05$	$\gamma=0.0527,$ $p=-0.0362$
PI	0.86	0.80	0.73	0.52	-0.23	0.88	0.85
Pearson R	0.84	0.75	0.77	0.50	-0.13	0.87	0.85
Spearman R	0.78	0.71	0.64	0.45	-0.22	0.82	0.77
MUE	90.92	64.82	37.98	51.50	22.38	96.73	86.58
MUEtr	13.31	11.17	7.57	12.28	8.91	13.12	13.06
MUEsc	1.05	1.46	1.34	2.98	11.07	0.91	1.01
PI: predictive index, MUE: Mean unsigned error; MUEtr: Mean unsigned error after subtraction of average signed error, MUEsc: MUE rescaled by slope and intercept of the predicted vs experimental results							

3.4.3. ANI in Replacement of MM Terms (ANI_PBSA)

As discussed in the theory section, ANI has been tested using several schemes in MMPBSA calculations. Instead of MM energy terms so as to estimate protein-ligand gas phase binding free energy contribution to the total binding free energy of solvated systems, ANI could be used referred as “ANI_PBSA”. Introducing ANI, we have observed dramatic increase in the correlation plots in almost all the different regimes (Table 3.4-3.5) of PBSA calculations with original parameters.

In addition, ANI_PBSA calculations were compared to the MMPBSA calculations with our fit parameters. The correlation coefficient by ANI_PBSA calculations is either comparable to or even slightly better than MMPBSA calculations with our fit parameters when $\epsilon_{int}=1$ or 2.1. In particular, when ANI is implemented in the SPT formalism using in either ASA+PCAV or SASA+SAV (i.e., equation 2.3), we produced the highest correlation as much as $R=0.89$ and $R=0.90$, respectively. This is almost 5% improvement over MMPBSA using the same formalism (SPT).

ANI has been also implemented in the extended form of PBSA_E formalism as named “ANI_PBSAe”. In almost all the different regimes, introducing ANI as a

replacement of MM term increase the correlation coefficient. Using SASA+SAV or ASA+PCAV in SPT scheme, ANI_PBSAe yields 0.90 and 0.91, respectively (Table 3.6). The correlation coefficient was not affected by changing ϵ_{int} , but the fitting parameter of PB term in the equation is linearly dependent to ϵ_{int} as expected. Figure 3.4 shows comparison of MMPBSA_E and ANI_PBSAe by using SASA only and SPT regimes respectively.

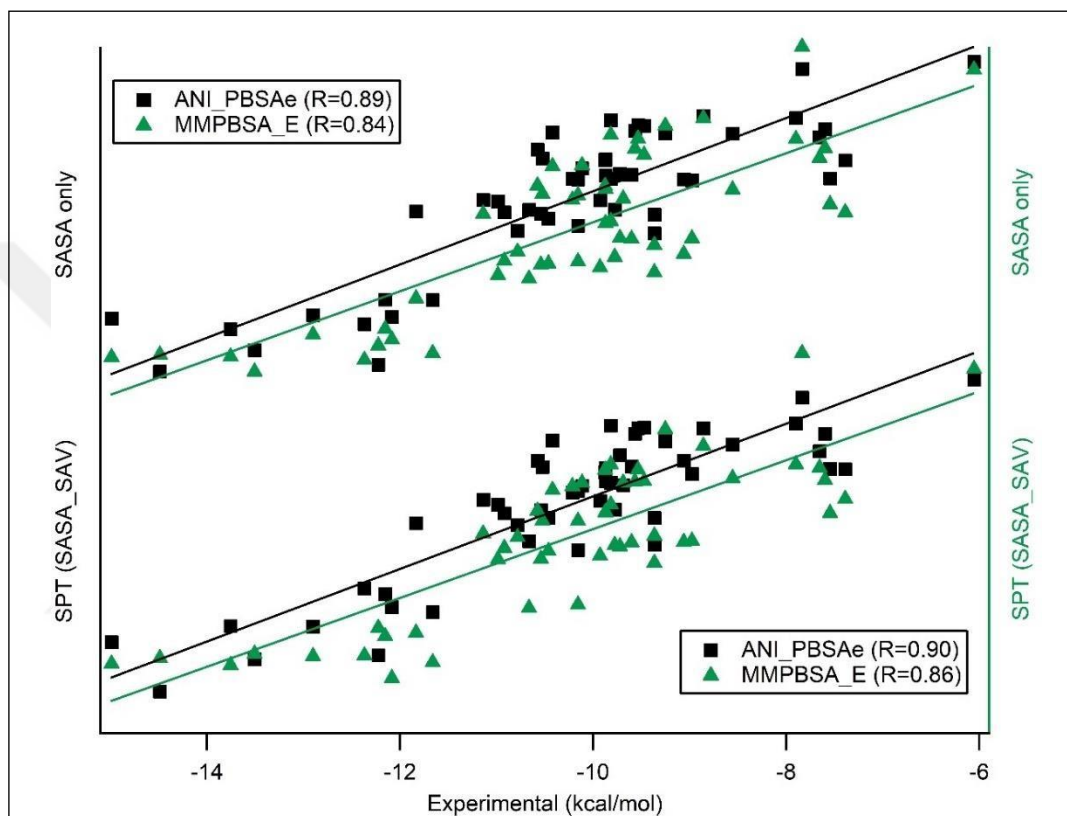


Figure 3.4: The comparison of MMPBSA_E and ANI_PBSAe by using SASA only and SPT regimes.

Table 3.4: ANI_PBSA calculated by AMBER produced SA terms for non-polar contribution to solvation free energy.

	One-term		Two-term	SPT
$\epsilon_{int}=1$	SASA only $a=1.5508$, $\gamma=0.0609$	SAV only $a=1.6187$, $\gamma=0.0511$	SAV+Disp $a=1.6497$, $\gamma=0.1184$	SASA+SAV $a=1.7478$, $\gamma=0.1112$, $p=-0.0636$
PI	0.82	0.81	0.79	0.80
Pearson R	0.79	0.78	0.76	0.80
Spearman R	0.74	0.73	0.69	0.73
MUE	76.05	69.17	72.87	73.28
MUEtr	13.91	13.54	13.70	13.99
MUEsc	1.17	1.24	1.34	1.16

Table 3.4: Continue

$\epsilon_{int}=2.1$	a=1.1425, $\gamma=0.0442$	a=1.2554, $\gamma=0.0316$	a=1.2863, $\gamma=0.098937$	a=1.3757, $\gamma=0.10362$, p=0.075252
PI	0.84	0.83	0.82	0.85
Pearson R	0.85	0.84	0.82	0.86
Spearman R	0.76	0.75	0.72	0.77
MUE	67.19	60.09	63.78	213.96
MUEtr	13.44	13.07	13.30	38.04
MUEsc	0.98	1.02	1.12	0.98
PI: predictive index, MUE: Mean unsigned error; MUEtr: Mean unsigned error after subtraction of average signed error, MUEsc: MUE rescaled by slope and intercept of the predicted vs experimental results				

Table 3.5: ANI_PBSA calculated by Zeo++ produced SA terms for non-polar contribution to solvation free energy.

	One-term			Two-term		SPT	
$\epsilon_{int}=1$	ASA only a=1.4993, $\gamma=0.0662$	PCAV only a=2.4414, $\gamma=-0.0212$	POAVonly a=2.1347, $\gamma=-0.0498$	PCAV+EDisp a=2.9684, $\gamma=0.0072$	POAV+EDisp a=2.922, $\gamma=-0.07088$	ASA_PCAV a=1.6755, $\gamma=0.1139$, p=-0.0669	ASA_POAV a=1.5568, $\gamma=0.0562$, p=-0.0300
PI	0.83	0.79	0.78	0.70	0.70	0.86	0.81
Pearson R	0.80	0.77	0.78	0.69	0.72	0.84	0.80
Spearman R	0.75	0.70	0.70	0.60	0.60	0.81	0.73
MUE	77.77	43.63	47.96	28.26	24.76	81.15	72.74
MUEtr	13.93	13.41	13.92	15.04	15.05	14.22	13.87
MUEsc	1.15	1.29	1.31	1.67	1.51	1.02	1.16
$\epsilon_{int}=2.1$	a=1.1029, $\gamma=0.0482$	a=1.8698, $\gamma=-0.0232$	a=1.5612, $\gamma=0.0389$	a=2.3969, $\gamma=0.0052$,	a=2.3485, $\gamma=-0.0599$	a=1.2596, $\gamma=0.0907$, p=-0.0596	a=1.1504, $\gamma=0.0399$, p=-0.0248
PI	0.84	0.83	0.82	0.76	0.77	0.89	0.84
Pearson R	0.85	0.84	0.84	0.76	0.79	0.89	0.86
Spearman R	0.76	0.76	0.74	0.66	0.68	0.84	0.77
MUE	68.53	41.66	46.77	26.30	23.56	71.56	64.38
MUEtr	13.47	12.90	13.41	14.02	14.14	14.02	13.58
MUEsc	0.97	1.04	1.05	1.35	1.27	0.86	0.98
PI: predictive index, MUE: Mean unsigned error; MUEtr: Mean unsigned error after subtraction of average signed error, MUEsc: MUE rescaled by slope and intercept of the predicted vs experimental results							

Table 3.6: ANI_PBSAe calculated by SPT model for non-polar contribution to solvation free energy.

$\epsilon_{int}=1$	SASA+SAV $\alpha_1=0.7568,$ $\alpha_2=0.4126,$ $\alpha_3=0.0850,$ $\alpha_4=0.0882$	ASA+PCAV $\alpha_1=0.7294,$ $\alpha_2=0.2250,$ $\alpha_3=0.0589,$ $\alpha_4=0.0492$	ASA+POAV $\alpha_1=0.5403,$ $\alpha_2=-0.3433,$ $\alpha_3=0.0150,$ $\alpha_4=0.0144$
PI	0.87	0.90	0.86
Pearson R	0.90	0.91	0.89
Spearman R	0.81	0.85	0.79
MUE	13.50	13.89	10.65
MUEtr	0.80	0.79	0.84
MUEsc	0.87	0.90	0.86
$\epsilon_{int}=2.1$	$\alpha_1=0.7633,$ $\alpha_2=0.8606,$ $\alpha_3=0.0834,$ $\alpha_4=0.0863$	$\alpha_1=0.7349,$ $\alpha_2=0.4710,$ $\alpha_3=0.0589,$ $\alpha_4=0.0491$	$\alpha_1=0.5433,$ $\alpha_2=0.7327,$ $\alpha_3=0.0150,$ $\alpha_4=0.0137$
PI	0.87	0.90	0.85
Pearson R	0.90	0.91	0.89
Spearman R	0.81	0.85	0.78
MUE	13.49	13.89	13.34
MUEtr	0.80	0.79	0.85
MUEsc	0.87	0.90	0.85
PI: predictive index, MUE: Mean unsigned error; MUEtr: Mean unsigned error after subtraction of average signed error, MUEsc: MUE rescaled by slope and intercept of the predicted vs experimental results			

3.4.4. Bivalent ϵ_{int} in MMPBSA/ANIPBSA Calculations

In the original MMPBSA formalism, ϵ_{int} , which corresponds to the solute dielectric constant is set to 1. The use of a single solute dielectric constant for ligand-receptor complexes, which are not uniform dielectric media, is controversial and can cause significant errors [147], [161], [173], [174]. There have been numerous attempts to correct the MMPBSA calculations by adjusting ϵ_{int} [147], [148], [156], [175], [176]. One of the most promising methods is to use variable dielectric constant, which take different values on acidic basic and neutral amino acids in the protein [65]. However, most calculations still use single value of ϵ_{int} which is set to 2-4 for larger datasets of diverse proteins [100], [177].

Although the attempts were focused on correcting the solvation free energy part of the MMPBSA calculations by adjusting the dielectric constant of the solute, it is appeared in two terms and in both it is reversely proportional to energy. The first is the coulombic energy term of the MM part of P-L and the second is the polar solvation term of the PB/GB calculations. Thus, the total binding free energy is affected by these two altered values.

Our MMPBSA calculations using default parameters in Amber software yielded comparatively poorer results with $\epsilon_{\text{int}}=1$. When this value is increased to 2.1, we have observed significant improvement in correlation coefficient with greater MUEs. However, when further increased up to 10, negligible improvements were observed.

Instead of using single dielectric constant for both protein-ligand electrostatic term and polar solvation free energy change upon complexation, we considered a different approach and defined two ϵ_{int} values (Figure 3.5), one for coulombic term of P-L, the other for PB term of (PL-P-L).

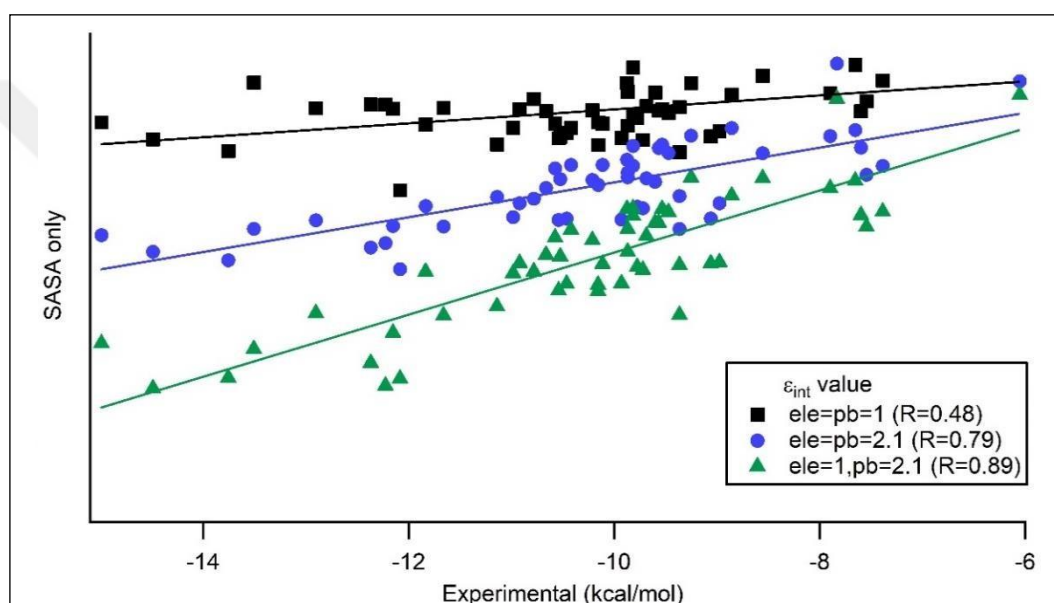


Figure 3.5: The comparison of the single valued vs two valued ϵ_{int} for coulombic (P-L) and polar solvation term, respectively.

The rationale behind this is that the neutral ligand is mostly buried in the binding pocket whose amino acids are mostly neutral. Therefore, the ϵ_{int} of the coulombic term appearing in the MM term must not be changed. On the other hand, amino acids that are surface exposed have direct interaction with solvent. Thus, ϵ_{int} in the PB term must be optimized to better estimate solvation energies. Using this idea, we defined divalent ϵ_{int} in which it is fixed at 1 for P-L coulombic term whereas various values of ϵ_{int} was used for polar solvation term (from 2 to 10). Our results showed an improvement of $\sim 10\%$ in the correlation coefficient in all cases. Table 3.7-3.8 shows the comparison between single valued and two valued ϵ_{int} results.

Table 3.7: Bivalent epsilon calculated by MMPBSA produced SA terms for non polar contribution to solvation free energy.

	One-term		Two-term		SPT
$\epsilon_{ele}=1, pb=2.1$	SASA only $\gamma=0.0360,$ $b=0$	SAV only $\gamma=0.0319,$ $b=0$	SASA+Disp $\gamma=0.1085,$ $b=0$	SAV+Disp $\gamma=0.1015,$ $b=0$	SASA+SAV $\gamma_1=0.10301,$ $p=-0.0655,$ $b=0$
PI	0.88	0.87	0.90	0.85	0.90
Pearson R	0.89	0.88	0.89	0.85	0.90
Spearman R	0.83	0.82	0.84	0.79	0.84
MUE	82.44	78.01	90.42	82.42	84.58
MUEtr	12.98	12.66	13.03	12.95	13.10
MUEsc	0.84	0.89	0.82	0.99	0.79

Table 3.8: Bivalent epsilon calculated by Zeo++ produced SA terms for non-polar contribution to solvation free energy.

	One-term			Two-term			SPT	
$\epsilon_{ele}=1, pb=2.1$	ASA only $\gamma=0.0368,$ $b=0$	PCAV only $\gamma=0.0256,$ $b=0$	POAVonly $\gamma=-0.0642,$ $b=0$	ASA+EDisp $\gamma=0.1103,$ $b=0$	PCAV+EDisp $\gamma=0.0909,$ $b=0$	POAV+EDisp $\gamma=-0.1399,$ $b=0$	ASA_PCAV $\gamma_1=0.0653,$ $\gamma_2=0.0311,$ $b=0$	ASA_POAV $\gamma_1=0.0310,$ $\gamma_2=0.0311,$ $b=0$
PI	0.89	0.86	0.26	0.90	0.70	-0.74	0.90	0.87
Pearson R	0.89	0.86	0.21	0.89	0.70	-0.64	0.90	0.88
Spearman R	0.84	0.79	0.21	0.84	0.60	-0.66	0.85	0.81
MUE	82.76	66.57	8.63	90.74	53.26	129.61	86.38	52.36
MUEtr	13.04	11.34	6.94	13.15	11.48	28.87	13.42	8.72
MUEsc	0.83	0.95	7.51	0.79	1.73	2.03	0.81	0.91

3.5. Summary and Conclusion

We have thoroughly revisited single trajectory approach MMPBSA and MMPBSA_E calculations for selected 54 protein-ligand complex systems, of which experimental values are determined with high accuracy. We have tested predefined coefficients in the non-polar solvation term as well as our own fitting parameters. Our results have shown that SPT model superior to other regimes. We have also investigated the success of ASA/PCAV over SASA/SAV and shown slight improvement at the SPT regime. In addition, we have implemented ANI-ML potentials in the MM terms of the MMPBSA calculations, which lead to significant

improvements over the single and two term approaches and somewhat similar to SPT model. Finally, we have also investigated the use of two valued dielectric constant and shown superiority over the single valued one.



4. CONCLUSION

In this thesis, firstly ANI-ML potentials, trained at the wb97x/6-31G* level, are utilized to compute interaction energies for both protein-ligand and ligand-solvent complexes. Our findings unravel a strong correlation between predicted interaction energies and experimental values, even when the solvent is disregarded. We have modified the conventional LIE method by replacing MM interaction energy terms (coulombic and van der Waals) with ANI predicted interaction energy terms in our so-called ANI_LIE method. We extensively benchmark our ANI_LIE method with other existing methods such as BAR, and LIE, MMP(G)BSA. The ANI_LIE approach shows superior accuracy while maintaining similar computational cost to MMP(B)SA. Additionally, the preassigned parameters determined for the ANI_LIE method are transferable and can be applied to any protein–ligand complex system, regardless of the protein(s)/ligand(s) scaffolds.

As an extension to ANI_LIE method, we have implemented the same approach of replacing MM terms with ANI potentials in MM-PBSA method. Due to the complex nature of decomposed energy components of MMPBSA approach, we have thoroughly revisited each component in these calculations and suggested several new perspectives to this field. We have focused on the electrostatic P-L energy terms along with non-polar solvation free energy change. Our findings indicate that the SPT model for prediction of non-polar contribution to the solvation free energy outperforms to other regimes. We also examined the success of ASA/PCAV compared to SASA/SAV and noted a slight improvement with the SPT regime. Finally, we investigated the effect of using a two-valued dielectric constant for electrostatic terms, which demonstrated superiority over the single-valued approach.

The present study evaluated the performance of ANI_LIE and ANI_PBSA methods on a total of 54 proteins from three different families. The correlations showed that both methods yielded excellent outcomes when compared to experimental binding free energies (BFEs). Moreover, we demonstrated that ANI_LIE and ANI_PBSA are applicable to any protein-ligand pairs that have atoms within the ANI-2x capacity. As a future direction of the study, it would be beneficial to extend the range of protein families and ligand types tested to further validate the generalizability of ANI_LIE and ANI_PBSA. To achieve this, these methods could be applied to a larger and more diverse set of protein-ligand systems, and the results

compared against experimental data. In addition, protein-protein interactions are also so important at computational chemistry studies, these methods would also be used to predict protein-protein interactions by beginning with studies on protein-peptide interactions. One of the fundamental methods for studying protein-substrate interactions is to determine the BFE between the protein and ligand. However, both the macro-molecular structure of the protein and the fact that the substrate is a peptide structure containing multiple amino acids make it difficult to calculate the BFE. Therefore, there are very limited methods available for calculating the protein-peptide BFE, with the most common being methods such as MM/PBSA and MM/GBSA. These methods, which are referred to as "end-state" methods, are attractive because of their low computational cost, but their accuracy is low, they exhibit high variability from system to system, and they can only calculate relative binding energy, which creates a need for alternative strategies.

REFERENCES

- [1] Du X.,Li Y.,Xia Y.L.,Ai S.M.,Liang J.,Sang P.,Ji X.L.,Liu S.Q., (2016), “Insights Into Protein–Ligand Interactions: Mechanisms, Models, And Methods”, *International Journal of Molecular Sciences*, 17 (2), 1–34.
- [2] Friedman R., (2022), “Computational Studies Of Protein–Drug Binding Affinity Changes Upon Mutations In The Drug Target”, *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 12 (1), 1–31.
- [3] Burlingham B.T.,Widlanski T.S., (2003), “An Intuitive Look At The Relationship Of K_i And IC_{50} : A More General Use For The Dixon Plot”, *Journal of Chemical Education*, 80 (2), 214–218.
- [4] Warshel A.,Levitt M., (1976), “Theoretical Studies Of Enzymic Reactions: Dielectric, Electrostatic And Steric Stabilization Of The Carbonium Ion In The Reaction Of Lysozyme”, *Journal of Molecular Biology*, 103 (2), 227–249.
- [5] Field M.J.,Bash P.A.,Karplus M., (1990), “A Combined Quantum Mechanical And Molecular Mechanical Potential For Molecular Dynamics Simulations”, *Journal of Computational Chemistry*, 11 (6), 700–733.
- [6] Chang C.E.A.,Huang Y.M.M.,Mueller L.J.,You W., (2016), “Investigation Of Structural Dynamics Of Enzymes And Protonation States Of Substrates Using Computational Tools”, *Catalysts*, 6 (6).
- [7] Kontoyianni M.,McClellan L.M.,Sokol G.S., (2004), “Evaluation Of Docking Performance: Comparative Data On Docking Algorithms”, *Journal of Medicinal Chemistry*, 47 (3), 558–565.
- [8] Friesner R.A.,Banks J.L.,Murphy R.B.,Halgren T.A.,Klicic J.J.,Mainz D.T.,Repasky M.P.,Knoll E.H.,Shelley M.,Perry J.K.,Shaw D.E.,Francis P.,Shenkin P.S., (2004), “Glide: A New Approach For Rapid, Accurate Docking And Scoring. 1. Method And Assessment Of Docking Accuracy”, *Journal of Medicinal Chemistry*, 47 (7), 1739–1749.
- [9] Morris G.M.,Goodsell D.S.,Halliday R.S.,Huey R.,Hart W.E.,Belew R.K.,Olson A.J., (1998), “Automated Docking Using A Lamarckian Genetic Algorithm And An Empirical Binding Free Energy Function”, *Journal of Computational Chemistry*, 19 (14), 1639–1662.
- [10] Shoichet B.K., (2004), “Successes From Virtual Screening”, 432 (December).
- [11] Choi S.B.,Yap B.K.,Choong Y.S.,Wahab H., (2018), “Molecular Dynamics Simulations In Drug Discovery”, *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, 1–3, 652–665.
- [12] Spreiter Q.,Walter M., (1999), “Classical Molecular Dynamics Simulation With The Velocity Verlet Algorithm At Strong External Magnetic Fields”, *Journal of Computational Physics*, 152 (1), 102–119.

- [13] Bayly C.I., Merz K.M., Ferguson D.M., Cornell W.D., Fox T., Caldwell J.W., Kollman P.A., Cieplak P., Gould I.R., Spellmeyer D.C., (1995), "A Second Generation Force Field For The Simulation Of Proteins, Nucleic Acids, And Organic Molecules", *Journal of the American Chemical Society*, 117 (19), 5179–5197.
- [14] Marais A., Adams B., Ringsmuth A.K., Ferretti M., Gruber J.M., Hendrikx R., Schuld M., Smith S.L., Sinayskiy I., Krüger T.P.J., Petruccione F., van Grondelle R., (2018), "The Future Of Quantum Biology", *Journal of the Royal Society Interface*, 15 (148).
- [15] Tully J.C., (2000), "Perspective On 'Zur Quantentheorie Der Molekeln'", *Theoretical Chemistry Accounts*, 103 (3–4), 173–176.
- [16] Hotokka M., (2006), "Calculation Of Vibrational Frequencies By Molecular Mechanics", *Handbook of Vibrational Spectroscopy*.
- [17] Arodola O.A., Soliman M.E.S., (2017), "Quantum Mechanics Implementation In Drug-Design Workflows: Does It Really Help?", *Drug Design, Development and Therapy*, 11, 2551–2564.
- [18] Burke K., (2012), "Perspective On Density Functional Theory", *Journal of Chemical Physics*, 136 (15).
- [19] Grimme S., Antony J., Schwabe T., Mück-Lichtenfeld C., (2007), "Density Functional Theory With Dispersion Corrections For Supramolecular Structures, Aggregates, And Complexes Of (Bio)Organic Molecules", *Organic and Biomolecular Chemistry*, 5 (5), 741–758.
- [20] Car R., (2002), "Introduction To Density-Functional Theory And Ab-Initio Molecular Dynamics", *Quantitative Structure-Activity Relationships*, 21 (2), 97–104.
- [21] Senn H.M., Thiel W., (2009), "QM/MM Methods For Biomolecular Systems", *Angewandte Chemie - International Edition*, 48 (7), 1198–1229.
- [22] Rupp M., Tkatchenko A., Müller K.R., Von Lilienfeld O.A., (2012), "Fast And Accurate Modeling Of Molecular Atomization Energies With Machine Learning", *Physical Review Letters*, 108 (5), 1–5.
- [23] Manzhos S., Carrington T., (2006), "Using Neural Networks To Represent Potential Surfaces As Sums Of Products", *Journal of Chemical Physics*, 125 (19).
- [24] Schütt K.T., Arbabzadah F., Chmiela S., Müller K.R., Tkatchenko A., (2017), "Quantum-Chemical Insights From Deep Tensor Neural Networks", *Nature Communications*, 8, 6–13.
- [25] Ho T.H., Pham-Tran N.N., Kawazoe Y., Le H.M., (2016), "Ab Initio Investigation Of O-H Dissociation From The Al-OH₂ Complex Using Molecular Dynamics And Neural Network Fitting", *Journal of Physical Chemistry A*, 120 (3), 346–355.

- [26] Kolb B., Zhao B., Li J., Jiang B., Guo H., (2016), "Permutation Invariant Potential Energy Surfaces For Polyatomic Reactions Using Atomistic Neural Networks", *Journal of Chemical Physics*, 144 (22).
- [27] Jiang B., Li J., Guo H., (2016), "Potential Energy Surfaces From High Fidelity Fitting Of Ab Initio Points: The Permutation Invariant Polynomial - Neural Network Approach", *International Reviews in Physical Chemistry*, 35 (3), 479–506.
- [28] Häse F., Valleau S., Pyzer-Knapp E., Aspuru-Guzik A., (2016), "Machine Learning Exciton Dynamics", *Chemical Science*, 7 (8), 5139–5147.
- [29] Faber F.A., Lindmaa A., Von Lilienfeld O.A., Armiento R., (2016), "Machine Learning Energies Of 2 Million Elpasolite (ABC2D6) Crystals", *Physical Review Letters*, 117 (13), 2–7.
- [30] Rupp M., Ramakrishnan R., Von Lilienfeld O.A., (2015), "Machine Learning For Quantum Mechanical Properties Of Atoms In Molecules", *Journal of Physical Chemistry Letters*, 6 (16), 3309–3313.
- [31] Behler J., (2015), "Constructing High-Dimensional Neural Network Potentials: A Tutorial Review", *International Journal of Quantum Chemistry*, 115 (16), 1032–1050.
- [32] Jasrasaria D., Pyzer-Knapp E.O., Rappoport D., Aspuru-Guzik A., (2016), "Space-Filling Curves As A Novel Crystal Structure Representation For Machine Learning Models", (December). [Online]. Available: <http://arxiv.org/abs/1608.05747>
- [33] Behler J., Parrinello M., (2007), "Generalized Neural-Network Representation Of High-Dimensional Potential-Energy Surfaces", *Physical Review Letters*, 98 (14), 1–4.
- [34] Von Lilienfeld O.A., Ramakrishnan R., Rupp M., Knoll A., (2015), "Fourier Series Of Atomic Radial Distribution Functions: A Molecular Fingerprint For Machine Learning Models Of Quantum Chemical Properties", *International Journal of Quantum Chemistry*, 115 (16), 1084–1093.
- [35] Manzhos S., Dawes R., Carrington T., (2015), "Neural Network-Based Approaches For Building High Dimensional And Quantum Dynamics-Friendly Potential Energy Surfaces", *International Journal of Quantum Chemistry*, 115 (16), 1012–1020.
- [36] Smith J.S., Isayev O., Roitberg A.E., (2017), "ANI-1: An Extensible Neural Network Potential With DFT Accuracy At Force Field Computational Cost", *Chemical Science*, 8 (4), 3192–3203.
- [37] Smith J.S., Isayev O., Roitberg A.E., (2017), "Data Descriptor: ANI-1, A Data Set Of 20 Million Calculated Off-Equilibrium Conformations For Organic Molecules", *Scientific Data*, 4, 1–8.
- [38] Gastegger M., Behler J., Marquetand P., (2017), "Machine Learning Molecular

Dynamics For The Simulation Of Infrared Spectra”, *Chemical Science*, 8 (10), 6924–6935.

- [39] Artrith N.,Behler J., (2012), “High-Dimensional Neural Network Potentials For Metal Surfaces: A Prototype Study For Copper”, *Physical Review B - Condensed Matter and Materials Physics*, 85 (4), 1–13.
- [40] Devereux C.,Smith J.S.,Davis K.K.,Barros K.,Zubatyuk R.,Isayev O.,Roitberg A.E., (2020), “Extending The Applicability Of The ANI Deep Learning Molecular Potential To Sulfur And Halogens”, *Journal of Chemical Theory and Computation*, 16 (7), 4192–4202.
- [41] Limongelli V., (2020), “Ligand Binding Free Energy And Kinetics Calculation In 2020”, *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 10 (4).
- [42] Åqvist J.,Medina C.,Samuelsson J.E., (1994), “A New Method For Predicting Binding Affinity In Computer-Aided Drug Design”, *Protein Engineering, Design and Selection*, 7 (3), 385–391.
- [43] Åqvist J.,Luzhkov V.B.,Brandsdal B.O., (2002), “Ligand Binding Affinities From MD Simulations”, *Accounts of Chemical Research*, 35 (6), 358–365.
- [44] Wang R.,Lai L.,Wang S., (2002), “Further Development And Validation Of Empirical Scoring Functions For Structure-Based Binding Affinity Prediction”, *Journal of Computer-Aided Molecular Design*, 16 (1), 11–26.
- [45] Liu X.,Liu J.,Zhu T.,Zhang L.,He X.,Zhang J.Z.H., (2016), “PBSA-E: A PBSA-Based Free Energy Estimator For Protein-Ligand Binding Affinity”, *Journal of Chemical Information and Modeling*, 56 (5), 854–861.
- [46] Method P.,Gases I.N.,Zwanzig R.W., (2004), “High-Temperature Equation Of State By A”, 1420 (March 1954).
- [47] Fratev F.,Steinbrecher T.,Jónsdóttir S.Ó., (2018), “Prediction Of Accurate Binding Modes Using Combination Of Classical And Accelerated Molecular Dynamics And Free-Energy Perturbation Calculations: An Application To Toxicity Studies”, *ACS Omega*, 3 (4), 4357–4371.
- [48] Gilson M.K.,Given J.A.,Bush B.L.,McCammon J.A., (1997), “The Statistical-Thermodynamic Basis For Computation Of Binding Affinities: A Critical Review”, *Biophysical Journal*, 72 (3), 1047–1069.
- [49] Jorgensen W.L.,Buckner J.K.,Boudon S.,Tirado-Rives J., (1988), “Efficient Computation Of Absolute Free Energies Of Binding By Computer Simulations. Application To The Methane Dimer In Water”, *The Journal of Chemical Physics*, 89 (6), 3742–3746.
- [50] Helms V.,Wade R.C., (1998), “Computational Alchemy To Calculate Absolute Protein-Ligand Binding Free Energy”, *Journal of the American Chemical Society*, 120 (12), 2710–2713.

- [51] Lawrenz M., Wereszczynski J., Ortiz-Sánchez J.M., Nichols S.E., McCammon J.A., (2012), "Thermodynamic Integration To Predict Host-Guest Binding Affinities", *Journal of Computer-Aided Molecular Design*, 26 (5), 569–576.
- [52] Bennett C.H., (1976), "Efficient Estimation Of Free Energy Differences From Monte Carlo Data",.
- [53] Shirts M.R., Bair E., Hooker G., Pande V.S., (2003), "Equilibrium Free Energies From Nonequilibrium Measurements Using Maximum-Likelihood Methods", *Physical Review Letters*, 91 (14), 1–4.
- [54] Lapelosa M., Gallicchio E., Levy R.M., (2012), "Conformational Transitions And Convergence Of Absolute Binding Free Energy Calculations", *Journal of Chemical Theory and Computation*, 8 (1), 47–60.
- [55] King E., Aitchison E., Li H., Luo R., (2021), "Recent Developments In Free Energy Calculations For Drug Discovery", *Frontiers in Molecular Biosciences*, 8 (August), 1–24.
- [56] De Ruiter A., Boresch S., Oostenbrink C., (2013), "Comparison Of Thermodynamic Integration And Bennett's Acceptance Ratio For Calculating Relative Protein-Ligand Binding Free Energies", *Journal of Computational Chemistry*, 34 (12), 1024–1034.
- [57] Izrailev S., Stepaniants S., Balsera M., Oono Y., Schulten K., (1997), "Molecular Dynamics Study Of Unbinding Of The Avidin-Biotin Complex", *Biophysical Journal*, 72 (4), 1568–1581.
- [58] Yung-Chi C., Prusoff W.H., (1973), "Relationship Between The Inhibition Constant (KI) And The Concentration Of Inhibitor Which Causes 50 Per Cent Inhibition (I50) Of An Enzymatic Reaction", *Biochemical Pharmacology*, 22 (23), 3099–3108.
- [59] Genheden S., Ryde U., (2015), "The MM/PBSA And MM/GBSA Methods To Estimate Ligand-Binding Affinities", *Expert Opinion on Drug Discovery*, 10 (5), 449–461.
- [60] Kumari R., Kumar R., Lynn A., (2014), "G-Mmpbsa -A GROMACS Tool For High-Throughput MM-PBSA Calculations", *Journal of Chemical Information and Modeling*, 54 (7), 1951–1962.
- [61] Rifai E.A., Van Dijk M., Vermeulen N.P.E., Yanuar A., Geerke D.P., (2019), "A Comparative Linear Interaction Energy And MM/PBSA Study On SIRT1-Ligand Binding Free Energy Calculation", *Journal of Chemical Information and Modeling*, 59 (9), 4018–4033.
- [62] Swanson J.M.J., Henchman R.H., McCammon J.A., (2004), "Revisiting Free Energy Calculations: A Theoretical Connection To MM/PBSA And Direct Calculation Of The Association Free Energy", *Biophysical Journal*, 86 (1 I), 67–74.
- [63] Steinbrecher T., Joung I., Case D.A. (2011), "Soft-Core Potentials In

Thermodynamic Integration: Comparing One-And Two-Step Transformations”, *Journal of Computational Chemistry*, 32 (15), 3253–3263.

- [64] König G.,Hudson P.S.,Boresch S.,Woodcock H.L., (2014), “Multiscale Free Energy Simulations: An Efficient Method For Connecting Classical MD Simulations To QM Or QM/MM Free Energies Using Non-Boltzmann Bennett Reweighting Schemes”, *Journal of Chemical Theory and Computation*, 10 (4), 1406–1419.
- [65] Allouche A., (2012), “Software News And Updates Gabedit — A Graphical User Interface For Computational Chemistry Softwares”, *Journal of computational chemistry*, 32, 174–182.
- [66] Salmaso V.,Moro S., (2018), “Bridging Molecular Docking To Molecular Dynamics In Exploring Ligand-Protein Recognition Process: An Overview”, *Frontiers in Pharmacology*, 9 (AUG), 1–16.
- [67] Kästner J.,Senn H.M.,Thiel S.,Otte N.,Thiel W., (2006), “QM/MM Free-Energy Perturbation Compared To Thermodynamic Integration And Umbrella Sampling: Application To An Enzymatic Reaction”, *Journal of Chemical Theory and Computation*, 2 (2), 452–461.
- [68] Shirts M.R.,Pande V.S., (2005), “Comparison Of Efficiency And Bias Of Free Energies Computed By Exponential Averaging, The Bennett Acceptance Ratio, And Thermodynamic Integration”, *Journal of Chemical Physics*, 122 (14).
- [69] Zacharias M.,Straatsma T.P.,McCammon J.A., (1994), “Separation-Shifted Scaling, A New Scaling Method For Lennard-Jones Interactions In Thermodynamic Integration”, *The Journal of Chemical Physics*, 100 (12), 9025–9031.
- [70] Lee T.S.,Allen B.K.,Giese T.J.,Guo Z.,Li P.,Lin C.,Dwight McGee T.,Pearlman D.A.,Radak B.K.,Tao Y.,Tsai H.C.,Xu H.,Sherman W.,York D.M., (2020), “Alchemical Binding Free Energy Calculations In AMBER20: Advances And Best Practices For Drug Discovery”, *Journal of Chemical Information and Modeling*, 60 (11), 5595–5623.
- [71] Kim I.,Allen T.W., (2012), “Bennetts Acceptance Ratio And Histogram Analysis Methods Enhanced By Umbrella Sampling Along A Reaction Coordinate In Configurational Space”, *Journal of Chemical Physics*, 136 (16).
- [72] Giese T.J.,York D.M., (2019), “Development Of A Robust Indirect Approach For MM → QM Free Energy Calculations That Combines Force-Matched Reference Potential And Bennett’s Acceptance Ratio Methods”, *Journal of Chemical Theory and Computation*, 15 (10), 5543–5562.
- [73] Jia X.,Wang M.,Shao Y.,König G.,Brooks B.R.,Zhang J.Z.H.,Mei Y., (2016), “Calculations Of Solvation Free Energy Through Energy Reweighting From Molecular Mechanics To Quantum Mechanics”, *Journal of Chemical Theory and Computation*, 12 (2), 499–511.

- [74] Li Y., Nam K., (2020), “Repulsive Soft-Core Potentials For Efficient Alchemical Free Energy Calculations”, *Journal of Chemical Theory and Computation*, 16 (8), 4776–4789.
- [75] Li P., Jia X., Pan X., Shao Y., Mei Y., (2018), “Accelerated Computation Of Free Energy Profile At Ab Initio Quantum Mechanical/Molecular Mechanics Accuracy Via A Semi-Empirical Reference Potential. I. Weighted Thermodynamics Perturbation”, *Journal of Chemical Theory and Computation*, 14 (11), 5583–5596.
- [76] Gumbart J.C., Roux B., Chipot C., (2013), “Efficient Determination Of Protein-Protein Standard Binding Free Energies From First Principles”, *Journal of Chemical Theory and Computation*, 9 (8), 3789–3798.
- [77] Gumbart J., Chipot C., Schulten K., (2011), “Free-Energy Cost For Translocon-Assisted Insertion Of Membrane Proteins”, *Proceedings of the National Academy of Sciences of the United States of America*, 108 (9), 3596–3601.
- [78] Giese T.J., York D.M., (2021), “Variational Method For Networkwide Analysis Of Relative Ligand Binding Free Energies With Loop Closure And Experimental Constraints”, *Journal of Chemical Theory and Computation*, 17 (3), 1326–1336.
- [79] Jarzynski C., (1997), “Nonequilibrium Equality For Free Energy Differences”, *Physical Review Letters*, 78 (14), 2690–2693.
- [80] Boresch S., Woodcock H.L., (2017), “Convergence Of Single-Step Free Energy Perturbation”, *Molecular Physics*, 115 (9–12), 1200–1213.
- [81] Crooks G.E., (1999), “Entropy Production Fluctuation Theorem And The Nonequilibrium Work Relation”, *Physical Review E*, 60 (3), 2721–2726.
- [82] Case D.A., Aktulga H.M., Belfon K., Ben-Shalom I.Y., Brozell S.R., Cerutti D.S., Cheatham T.E., Cisneros G.A., Cruzeiro V.W.D., Darden T.A., Duke R.E., Giambasu G., Gilson M.K. *et al.*, (2021), “AmberTools20”, University of California, San Francisco, 2021.
- [83] Cournia Z., Chipot C., Roux B., York D.M., Sherman W., (2021), “Free Energy Methods In Drug Discovery - Introduction”, *ACS Symposium Series*, 1397, 1–38.
- [84] Gundelach L., Fox T., Tautermann C.S., Skylaris C.K., (2021), “Protein-Ligand Free Energies Of Binding From Full-Protein DFT Calculations: Convergence And Choice Of Exchange-Correlation Functional”, *Physical Chemistry Chemical Physics*, 23 (15), 9381–9393.
- [85] Gräter F., Schwarzl S.M., Dejaegere A., Fischer S., Smith J.C., (2005), “Protein/Ligand Binding Free Energies Calculated With Quantum Mechanics/Molecular Mechanics”, *Journal of Physical Chemistry B*, 109 (20), 10474–10483.
- [86] Ibrahim M.A.A., (2011), “Performance Assessment Of Semiempirical

- Molecular Orbital Methods In Describing Halogen Bonding: Quantum Mechanical And Quantum Mechanical/Molecular Mechanical-Molecular Dynamics Study”, *Journal of Chemical Information and Modeling*, 51 (10), 2549–2559.
- [87] Retegan M.,Milet A.,Jamet H., (2009), “Exploring The Binding Of Inhibitors Derived From Tetrabromobenzimidazole To The CK2 Protein Using A QM/MM-PB/SA Approach”, *Journal of Chemical Information and Modeling*, 49 (4), 963–971.
- [88] Wichapong K.,Rohe A.,Platzer C.,Slynko I.,Erdmann F.,Schmidt M.,Sippl W., (2014), “Application Of Docking And QM/MM-GBSA Rescoring To Screen For Novel Myt1 Kinase Inhibitors”, *Journal of Chemical Information and Modeling*, 54 (3), 881–893.
- [89] Díaz N.,Suárez D.,Merz K.M.,Sordo T.L., (2005), “Molecular Dynamics Simulations Of The TEM-1 β -Lactamase Complexed With Cephalothin”, *Journal of Medicinal Chemistry*, 48 (3), 780–791.
- [90] Wang Y.T.,Chen Y.C., (2014), “Insights From QM/MM Modeling The 3D Structure Of The 2009 H1N1 Influenza A Virus Neuraminidase And Its Binding Interactions With Antiviral Drugs”, *Molecular Informatics*, 33 (3), 240–249.
- [91] Barbault F.,Maurel F., (2012), “Is Inhibition Process Better Described With MD(QM/MM) Simulations? The Case Of Urokinase Type Plasminogen Activator Inhibitors”, *Journal of Computational Chemistry*, 33 (6), 607–616.
- [92] Tsitsanou K.E.,Hayes J.M.,Keramioti M.,Mamais M.,Oikonomakos N.G.,Kato A.,Leonidas D.D.,Zographos S.E., (2013), “Sourcing The Affinity Of Flavonoids For The Glycogen Phosphorylase Inhibitor Site Via Crystallography, Kinetics And QM/MM-PBSA Binding Studies: Comparison Of Chrysin And Flavopiridol”, *Food and Chemical Toxicology*, 61, 14–27.
- [93] Fedorov D.G., (2017), “The Fragment Molecular Orbital Method: Theoretical Development, Implementation In GAMESS, And Applications”, *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 7 (6), 1–17.
- [94] Kurauchi R.,Watanabe C.,Fukuzawa K.,Tanaka S., (2015), “Novel Type Of Virtual Ligand Screening On The Basis Of Quantum-Chemical Calculations For Protein-Ligand Complexes And Extended Clustering Techniques”, *Computational and Theoretical Chemistry*, 1061, 12–22.
- [95] Tagami U.,Takahashi K.,Igarashi S.,Ejima C.,Yoshida T.,Takeshita S.,Miyanaga W.,Sugiki M.,Tokumasu M.,Hatanaka T.,Kashiwagi T.,Ishikawa K.,Miyano H.,Mizukoshi T., (2016), “Interaction Analysis Of FABP4 Inhibitors By X-Ray Crystallography And Fragment Molecular Orbital Analysis”, *ACS Medicinal Chemistry Letters*, 7 (4), 435–439.
- [96] Heifetz A.,Chudyk E.I.,Gleave L.,Aldeghi M.,Cherezov V.,Fedorov D.G.,Biggin P.C.,Bodkin M.J., (2016), “The Fragment Molecular Orbital Method Reveals New Insight Into The Chemical Nature Of GPCR-Ligand

Interactions”, *Journal of Chemical Information and Modeling*, 56 (1), 159–172.

- [97] Chen X., Zhao X., Xiong Y., Liu J., Zhan C.G., (2011), “Fundamental Reaction Pathway And Free Energy Profile For Hydrolysis Of Intracellular Second Messenger Adenosine 3',5'-Cyclic Monophosphate (cAMP) Catalyzed By Phosphodiesterase-4”, *Journal of Physical Chemistry B*, 115 (42), 12208–12219.
- [98] Kaukonen M., Söderhjelm P., Heimdal J., Ryde U., (2008), “QM/MM-PBSA Method To Estimate Free Energies For Reactions In Proteins”, *Journal of Physical Chemistry B*, 112 (39), 12537–12548.
- [99] Wang M., Wong C.F., (2007), “Rank-Ordering Protein-Ligand Binding Affinity By A Quantum Mechanics/Molecular Mechanics/Poisson-Boltzmann-Surface Area Model”, *Journal of Chemical Physics*, 126 (2), 5–8.
- [100] Söderhjelm P., Kongsted J., Ryde U., (2010), “Ligand Affinities Estimated By Quantum Chemical Calculations”, *Journal of Chemical Theory and Computation*, 6 (5), 1726–1737.
- [101] Manta S., Xipnitou A., Kiritsis C., Kantsadi A.L., Hayes J.M., Skamnaki V.T., Lamprakis C., Kontou M., Zoumpoulakis P., Zographos S.E., Leonidas D.D., Komiotis D., (2012), “3'-Axial CH 2OH Substitution On Glucopyranose Does Not Increase Glycogen Phosphorylase Inhibitory Potency. QM/MM-PBSA Calculations Suggest Why”, *Chemical Biology and Drug Design*, 79 (5), 663–673.
- [102] Söderhjelm P., Kongsted J., Genheden S., Ryde U., (2010), “Estimates Of Ligand-Binding Affinities Supported By Quantum Mechanical Methods”, *Interdisciplinary Sciences – Computational Life Sciences*, 2 (1), 21–37.
- [103] Sawada T., Fedorov D.G., Kitauro K., (2010), “Role Of The Key Mutation In The Selective Binding Of Avian And Human Influenza Hemagglutinin To Sialosides Revealed By Quantum-Mechanical Calculations”, *Journal of the American Chemical Society*, 132 (47), 16862–16872.
- [104] Lu H., Huang X., Abdulhameed M.D.M., Zhan C.G., (2014), “Binding Free Energies For Nicotine Analogs Inhibiting Cytochrome P450 2A6 By A Combined Use Of Molecular Dynamics Simulations And QM/MM-PBSA Calculations”, *Bioorganic and Medicinal Chemistry*, 22 (7), 2149–2156.
- [105] Smith J.S., Nebgen B.T., Zubatyuk R., Lubbers N., Devereux C., Barros K., Tretiak S., Isayev O., Roitberg A., (2018), “Outsmarting Quantum Chemistry Through Transfer Learning”, *ChemRxiv*.
- [106] Ahneman D.T., Estrada J.G., Lin S., Dreher S.D., Doyle A.G., (2018), “Supplementary Material For Predicting Reaction Performance In C–N Cross-Coupling Using Machine Learning Predicting Reaction Performance In C–N Cross-Coupling Using Machine Learning”, *Science*, 190 (April), 5169. [Online]. Available: <http://science.sciencemag.org/%0Awww.sciencemag.org/cgi/content/full/scien>

- [107] Chmiela S.,Tkatchenko A.,Sauceda H.E.,Poltavsky I.,Schütt K.T.,Müller K.R., (2017), “Machine Learning Of Accurate Energy-Conserving Molecular Force Fields”, *Science Advances*, 3 (5).
- [108] Behler J., (2017), “First Principles Neural Network Potentials For Reactive Simulations Of Large Molecular And Condensed Systems”, *Angewandte Chemie - International Edition*, 56 (42), 12828–12840.
- [109] Li Z.,Kermode J.R.,De Vita A., (2015), “Molecular Dynamics With On-The-Fly Machine Learning Of Quantum-Mechanical Forces”, *Physical Review Letters*, 114 (9), 1–5.
- [110] Glielmo A.,Sollich P.,De Vita A., (2017), “Accurate Interatomic Force Fields Via Machine Learning With Covariant Kernels”, *Physical Review B*, 95 (21), 1–10.
- [111] Kruglov I.,Sergeev O.,Yanilkin A.,Oganov A.R., (2017), “Energy-Free Machine Learning Force Field For Aluminum”, *Scientific Reports*, 7 (1), 1–7.
- [112] Faber F.A.,Hutchison L.,Huang B.,Gilmer J.,Schoenholz S.S.,Dahl G.E.,Vinyals O.,Kearnes S.,Riley P.F.,Von Lilienfeld O.A., (2017), “Prediction Errors Of Molecular Machine Learning Models Lower Than Hybrid DFT Error”, *Journal of Chemical Theory and Computation*, 13 (11), 5255–5264.
- [113] Ragoza M.,Hochuli J.,Idrobo E.,Sunseri J.,Koes D.R., (2017), “Protein-Ligand Scoring With Convolutional Neural Networks”, *Journal of Chemical Information and Modeling*, 57 (4), 942–957.
- [114] Zubatyuk R.,Smith J.S.,Leszczynski J.,Isayev O., (2019), “Accurate And Transferable Multitask Prediction Of Chemical Properties With An Atoms-In-Molecules Neural Network”, *Science Advances*, 5 (8).
- [115] Unke O.T.,Meuwly M., (2019), “PhysNet: A Neural Network For Predicting Energies, Forces, Dipole Moments, And Partial Charges”, *Journal of Chemical Theory and Computation*, 15 (6), 3678–3693.
- [116] Mascotti M.L.,Juri Ayub M.,Furnham N.,Thornton J.M.,Laskowski R.A., (2016), “Chopping And Changing: The Evolution Of The Flavin-Dependent Monooxygenases”, *Journal of Molecular Biology*, 428 (15), 3131–3146.
- [117] Smith J.S.,Roitberg A.E.,Isayev O., (2018), “Transforming Computational Drug Discovery With Machine Learning And AI”, *ACS Medicinal Chemistry Letters*, 9 (11), 1065–1069.
- [118] Smith J.S.,Nebgen B.,Lubbers N.,Isayev O.,Roitberg A.E., (2018), “Less Is More: Sampling Chemical Space With Active Learning”, *Journal of Chemical Physics*, 148 (24).
- [119] Breidenbach J.,Lemke C.,Pillaiyar T.,Schäkel L.,Al Hamwi G.,Dietl

- M.,Gedschold R.,Geiger N.,Lopez V.,Mirza S.,Namasivayam V.,Schiedel A.C.,Sylvester K.*et al.*, (2021), “Targeting The Main Protease Of SARS-CoV-2: From The Establishment Of High Throughput Screening To The Design Of Tailored Inhibitors”, *Angewandte Chemie - International Edition*, 60 (18), 10423–10429.
- [120] Trougakos I.P.,Stamatelopoulou K.,Terpos E.,Tsitsilonis O.E.,Aivalioti E.,Paraskevis D.,Kastritis E.,Pavlakakis G.N.,Dimopoulos M.A., (2021), “Insights To SARS-CoV-2 Life Cycle, Pathophysiology, And Rationalized Treatments That Target COVID-19 Clinical Complications”, *Journal of Biomedical Science*, 28 (1), 1–18.
- [121] Hu Q.,Xiong Y.,Zhu G.H.,Zhang Y.N.,Zhang Y.W.,Huang .P,Ge G.B., (2022), “The SARS-CoV-2 Main Protease (Mpro): Structure, Function, And Emerging Therapies For COVID-19”, *MedComm*, 3 (3), 1–27.
- [122] Ghosh A.K.,Osswald H.L.,Prato G., (2016), “Recent Progress In The Development Of HIV-1 Protease Inhibitors For The Treatment Of HIV/AIDS”, *Journal of Medicinal Chemistry*, 59 (11), 5172–5208.
- [123] Liu J.,Minemoto Y.,Lin A., (2004), “C-Jun N-Terminal Protein Kinase 1 (JNK1), But Not JNK2, Is Essential For Tumor Necrosis Factor Alpha-Induced C-Jun Kinase Activation And Apoptosis”, *Molecular and Cellular Biology*, 24 (24), 10844–10856.
- [124] Khalak Y.,Tresadern G.,Aldeghi M.,Baumann H.M.,Mobley D.L.,de Groot B.L.,Gapsys V., (2021), “Alchemical Absolute Protein-Ligand Binding Free Energies For Drug Design”, *Chemical Science*, 12 (41), 13958–13971.
- [125] Abraham M.J.,Murtola T.,Schulz R.,Páll S.,Smith J.C.,Hess B.,Lindah E., (2015), “Gromacs: High Performance Molecular Simulations Through Multi-Level Parallelism From Laptops To Supercomputers”, *SoftwareX*, 1–2, 19–25.
- [126] Toukan K.,Rahman A., (1985), “Molecular-Dynamics Study Of Atomic Motions In Water”, *Physical Review B*, 31 (5), 2643–2648.
- [127] Darden T.,York D.,Pedersen L., (1993), “Particle Mesh Ewald: An $N \cdot \log(N)$ Method For Ewald Sums In Large Systems”, *The Journal of Chemical Physics*, 98 (12), 10089–10092.
- [128] Kocak A.,Erol I.,Yildiz M.,Can H., (2016), “Computational Insights Into The Protonation States Of Catalytic Dyad In BACE1–Acyl Guanidine Based Inhibitor Complex”, *Journal of Molecular Graphics and Modelling*, 70, 226–235.
- [129] Kocak A., (2019), “HBGA Binding Modes And Selectivity In Noroviruses Upon Mutation: A Docking And Molecular Dynamics Study”, *Journal of Molecular Modeling*, 25 (12), 1–11.
- [130] Kocak A.,Yildiz M., (2017), “Docking, Molecular Dynamics And Free Energy Studies On Aspartoacylase Mutations Involved In Canavan Disease”, *Journal of Molecular Graphics and Modelling*, 74, 44–53.

- [131] Caleman C., Van Maaren P.J., Hong M., Hub J.S., Costa L.T., Van Der Spoel D., (2012), “Force Field Benchmark Of Organic Liquids: Density, Enthalpy Of Vaporization, Heat Capacities, Surface Tension, Isothermal Compressibility, Volumetric Expansion Coefficient, And Dielectric Constant”, *Journal of Chemical Theory and Computation*, 8 (1), 61–74.
- [132] Wang J., Wolf R.M., Caldwell J.W., Kollman P.A., Case D.A., (2004), “Development And Testing Of A General Amber Force Field”, *Journal of Computational Chemistry*, 25 (9), 1157–1174.
- [133] Lindorff-Larsen K., Piana S., Palmo K., Maragakis P., Klepeis J.L., Dror R.O., Shaw D.E., (2010), “Improved Side-Chain Torsion Potentials For The Amber Ff99SB Protein Force Field”, *Proteins: Structure, Function and Bioinformatics*, 78 (8), 1950–1958.
- [134] Fox S.J., Dziedzic J., Fox T., Tautermann C.S., Skylaris C.K., (2014), “Density Functional Theory Calculations On Entire Proteins For Free Energies Of Binding: Application To A Model Polar Binding Site”, *Proteins: Structure, Function and Bioinformatics*, 82 (12), 3335–3346.
- [135] Hansson T., Marelus J., Åqvist J., (1998), “Ligand Binding Affinity Prediction By Linear Interaction Energy Methods”, *Journal of Computer-Aided Molecular Design*, 12 (1), 27–35.
- [136] Åqvist J., Hansson T., (1996), “On The Validity Of Electrostatic Linear Response In Polar Solvents”, *Journal of Physical Chemistry*, 100 (22), 9512–9521.
- [137] Hansson T., Åqvist J., (1995), “By Molecular Dynamics Simulations”, 8 (11), 1137–1144.
- [138] Shamsudin Khan Y., Gutiérrez-De-Terán H., Boukharta L., Åqvist J., (2014), “Toward An Optimal Docking And Free Energy Calculation Scheme In Ligand Design With Application To COX-1 Inhibitors”, *Journal of Chemical Information and Modeling*, 54 (5), 1488–1499.
- [139] Almlöf M., Carlsson J., Åqvist J., (2007), “Improving The Accuracy Of The Linear Interaction Energy Method For Solvation Free Energies”, *Journal of Chemical Theory and Computation*, 3 (6), 2162–2175.
- [140] Suparyanto dan Rosad (2015, (2020), “~~濟無~~No Title No Title No Title”, Suparyanto dan Rosad (2015, 5 (3), 248–253.
- [141] Wall I.D., Leach A.R., Salt D.W., Ford M.G., Essex J.W., (1999), “Binding Constants Of Neuraminidase Inhibitors: An Investigation Of The Linear Interaction Energy Method”, *Journal of Medicinal Chemistry*, 42 (25), 5142–5152.
- [142] Wang E., Sun H., Wang J., Wang Z., Liu H., Zhang J.Z.H., Hou T., (2019), “End-Point Binding Free Energy Calculation With MM/PBSA And MM/GBSA: Strategies And Applications In Drug Design”, *Chemical Reviews*, 119 (16), 9478–9508.

- [143] van Gunsteren W.F., Berendsen H.J.C., (1990), "Computer Simulation Of Molecular Dynamics: Methodology, Applications, And Perspectives In Chemistry", *Angewandte Chemie International Edition in English*, 29 (9), 992–1023.
- [144] Filipe H.A.L., Loura L.M.S., (2022), "Molecular Dynamics Simulations: Advances And Applications", *Molecules*, 27 (7), 37–47.
- [145] Plattner N., Doerr S., De Fabritiis G., Noé F., (2017), "Complete Protein-Protein Association Kinetics In Atomic Detail Revealed By Molecular Dynamics Simulations And Markov Modelling", *Nature Chemistry*, 9 (10), 1005–1011.
- [146] Ytreberg F.M., Swendsen R.H., Zuckerman D.M., (2006), "Comparison Of Free Energy Methods For Molecular Systems", *Journal of Chemical Physics*, 125 (18).
- [147] Genheden S., Ryde U., (2012), "Comparison Of End-Point Continuum-Solvation Methods For The Calculation Of Protein-Ligand Binding Free Energies", *Proteins: Structure, Function and Bioinformatics*, 80 (5), 1326–1342.
- [148] Adekoya O.A., Willassen N.P., Sylte I., (2006), "Molecular Insight Into Pseudolysin Inhibition Using The MM-PBSA And LIE Methods", *Journal of Structural Biology*, 153 (2), 129–144.
- [149] Kollman P.A., Massova I., Reyes C., Kuhn B., Huo S., Chong L., Lee M., Lee T., Duan Y., Wang W., Donini O., Cieplak P., Srinivasan J., Case D.A., Cheatham T.E., (2000), "Calculating Structures And Free Energies Of Complex Molecules: Combining Molecular Mechanics And Continuum Models", *Accounts of Chemical Research*, 33 (12), 889–897.
- [150] Srinivasan J., Miller J., Kollman P.A., Case D.A., (1998), "Continuum Solvent Studies Of The Stability Of Rna Hairpin Loops And Helices", *Journal of Biomolecular Structure and Dynamics*, 16 (3), 671–682.
- [151] Hermann R.B., (1972), "Theory Of Hydrophobic Bonding. II. The Correlation Of Hydrocarbon Solubility In Water With Solvent Cavity Surface Area", *Journal of Physical Chemistry*, 76 (19), 2754–2759.
- [152] Su P.C., Tsai C.C., Mehboob S., Hevener K.E., Johnson M.E., (2015), "Comparison Of Radii Sets, Entropy, QM Methods, And Sampling On MM-PBSA, MM-GBSA, And QM/MM-GBSA Ligand Binding Energies Of *F. Tularensis* Enoyl-ACP Reductase (FabI)", *Journal of Computational Chemistry*, 36 (25), 1859–1873.
- [153] Homeyer N., Stoll F., Hillisch A., Gohlke H., (2014), "Binding Free Energy Calculations For Lead Optimization: Assessment Of Their Accuracy In An Industrial Drug Design Context", *Journal of Chemical Theory and Computation*, 10 (8), 3331–3344.
- [154] Hansen N., Van Gunsteren W.F., (2014), "Practical Aspects Of Free-Energy Calculations: A Review", *Journal of Chemical Theory and Computation*, 10

(7), 2632–2647.

- [155] Ashida T., Kikuchi T., (2015), “Overview Of Binding Free Energy Calculation Techniques For Elucidation Of Biological Processes And For Drug Discovery”, *Medicinal Chemistry*, 11 (3), 248–253.
- [156] Wang J., Hou T., Xu X., (2006), “Recent Advances In Free Energy Calculations With A Combination Of Molecular Mechanics And Continuum Models”, *Current Computer Aided-Drug Design*, 2 (3), 287–306.
- [157] Ongari D., Boyd P.G., Barthel S., Witman M., Haranczyk M., Smit B., (2017), “Accurate Characterization Of The Pore Volume In Microporous Crystalline Materials”, *Langmuir*, 33 (51), 14529–14538.
- [158] Willems T.F., Rycroft C.H., Kazi M., Meza J.C., Haranczyk M., (2012), “Algorithms And Tools For High-Throughput Geometry-Based Analysis Of Crystalline Porous Materials”, *Microporous and Mesoporous Materials*, 149 (1), 134–141.
- [159] Srinivasan J., Cheatham T.E., Cieplak P., Kollman P.A., Case D.A., (1998), “Continuum Solvent Studies Of The Stability Of DNA, RNA, And Phosphoramidate-DNA Helices”, *Journal of the American Chemical Society*, 120 (37), 9401–9409.
- [160] Massova I., Kollman P.A., (2000), “Combined Molecular Mechanical And Continuum Solvent Approach (MM- PBSA/GBSA) To Predict Ligand Binding”, *Perspectives in Drug Discovery and Design*, 18 (i), 113–135.
- [161] Sun H., Li Y., Shen M., Tian S., Xu L., Pan P., Guan Y., Hou T., (2014), “Assessing The Performance Of MM/PBSA And MM/GBSA Methods. 5. Improved Docking Performance Using High Solute Dielectric Constant MM/GBSA And MM/PBSA Rescoring”, *Physical Chemistry Chemical Physics*, 16 (40), 22035–22045.
- [162] Chen F., Sun H., Wang J., Zhu F., Liu H., Wang Z., Lei T., Li Y., Hou T., (2018), “Assessing The Performance Of MM/PBSA And MM/GBSA Methods. 8. Predicting Binding Free Energies And Poses Of Protein-RNA Complexes”, *Rna*, 24 (9), 1183–1194.
- [163] Chen F., Liu H., Sun H., Pan P., Li Y., Li D., Hou T., (2016), “Assessing The Performance Of The MM/PBSA And MM/GBSA Methods. 6. Capability To Predict Protein-Protein Binding Free Energies And Re-Rank Binding Poses Generated By Protein-Protein Docking”, *Physical Chemistry Chemical Physics*, 18 (32), 22129–22139.
- [164] Sun H., Li Y., Tian S., Xu L., Hou T., (2014), “Assessing The Performance Of MM/PBSA And MM/GBSA Methods. 4. Accuracies Of MM/PBSA And MM/GBSA Methodologies Evaluated By Various Simulation Protocols Using PDBbind Data Set”, *Physical Chemistry Chemical Physics*, 16 (31), 16719–16729.
- [165] Tan C., Tan Y.H., Luo R., (2007), “Implicit Nonpolar Solvent Models”, *Journal*

of Physical Chemistry B, 111 (42), 12263–12274.

- [166] Wagoner J.A., Baker N.A., (2006), “Assessing Implicit Models For Nonpolar Mean Solvation Forces: The Importance Of Dispersion And Volume Terms”, *Proceedings of the National Academy of Sciences of the United States of America*, 103 (22), 8331–8336.
- [167] Strömberg A., Ryttsén F., Chiu D.T., Davidson M., Eriksson P.S., Wilson C.F., Orwar O., Zare R.N., (2000), “Manipulating The Genetic Identity And Biochemical Surface Properties Of Individual Cells With Electric-Field-Induced Fusion”, *Proceedings of the National Academy of Sciences of the United States of America*, 97 (1), 7–11.
- [168] Ou S.C., Drake J.A., Pettitt B.M., (2017), “Nonpolar Solvation Free Energy From Proximal Distribution Functions”, *Journal of Physical Chemistry B*, 121 (15), 3555–3564.
- [169] Akkus E., Tayfuroglu O., Yildiz M., Kocak A., (2022), “Accurate Binding Free Energy Method From End-State MD Simulations”, *Journal of Chemical Information and Modeling*, 62 (17), 4095–4106.
- [170] Valdés-Tresanco M.S., Valdés-Tresanco M.E., Valiente P.A., Moreno E., (2021), “Gmx_MMPBSA: A New Tool To Perform End-State Free Energy Calculations With GROMACS”, *Journal of Chemical Theory and Computation*, 17 (10), 6281–6291.
- [171] Miller B.R., Mcgee T.D., Swails J.M., Homeyer N., Gohlke H., Roitberg A.E., (2012), “MMPBSA . Py : An Efficient Program For End-State Free Energy Calculations”,.
- [172] Pearlman D.A., Charifson P.S., (2001), “Scoring Functions For The P38 MAP Kinase Protein System †”, *Journal of medicinal chemistry*, 44 (21), 3417–3423.
- [173] Venken T., Krnavek D., Münch J., Kirchhoff F., Henklein P., de Maeyer M., Voet A., (2011), “An Optimized MM/PBSA Virtual Screening Approach Applied To An HIV-1 Gp41 Fusion Peptide Inhibitor”, *Proteins: Structure, Function and Bioinformatics*, 79 (11), 3221–3235.
- [174] Mikulskis P., Genheden S., Rydberg P., Sandberg L., Olsen L., Ryde U., (2012), “Binding Affinities In The SAMPL3 Trypsin And Host-Guest Blind Tests Estimated With The MM/PBSA And LIE Methods”, *Journal of Computer-Aided Molecular Design*, 26 (5), 527–541.
- [175] Foloppe N., Hubbard R., (2006), “Towards Predictive Ligand Design With Free-Energy Based Computational Methods?”, *Current Medicinal Chemistry*, 13 (29), 3583–3608.
- [176] Singh N., Warshel A., (2010), “Absolute Binding Free Energy Calculations: On The Accuracy Of Computational Scoring Of Protein-Ligand Interactions”, *Proteins: Structure, Function and Bioinformatics*, 78 (7), 1705–1723.

- [177] Hou T., Wang J., Li Y., Wang W., (2011), “Assessing The Performance Of The MM/PBSA And MM/GBSA Methods. 1. The Accuracy Of Binding Free Energy Calculations Based On Molecular Dynamics Simulations”, *Journal of Chemical Information and Modeling*, 51 (1), 69–82.



BIOGRAPHY

Ebru AKKUŞ finished Burdur Anatolian High School at 2013. She continued her education life at the department of Molecular Biology and Genetics at Gebze Technical University and graduated at 2019. With following times, Ebru applied for Master Degree to Gebze Technical University for Bioinformatics and Systems Biology program at Bioengineering Department and is now preparing to have diploma form there. She gained a lot of knowledge about computational methods for drug discovery, particularly protein-ligand interactions, during her master's studies and had the opportunity to participate in a variety of projects-studies at the computational chemistry field. She also had the ability to combine different techniques to use and produce results that were more advantageous for the field.

APPENDICES

Appendix A: Thesis Supporting Project

This thesis and its calculations were supported by Tubitak ULAKBİM and numerical calculations reported in this study were partially performed at TUBITAK ULAKBİM, High Performance and Grid Computing Center (TRUBA resources). In addition, This study is founded by Scientific and Technological Research Council of Turkey-TÜB İTAK project (Project Number 120Z732).

