

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

**ACCURATE FREE ENERGY CALCULATIONS FOR  
SOLVATION OF SMALL MOLECULES; AND HBGA<sub>s</sub> BINDING  
TO HUMAN NOROVIRUS-VP1 PROTEIN USING QM, MD AND  
ML-MD METHODS**

**MÜTESİR TEMEL**  
**A THESIS SUBMITTED FOR THE DEGREE OF**  
**MASTER OF SCIENCE**  
**DEPARTMENT OF CHEMISTRY**

**GEBZE**  
**2022**

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

**ACCURATE FREE ENERGY  
CALCULATIONS FOR SOLVATION OF  
SMALL MOLECULES; AND HBGA<sub>s</sub>  
BINDING TO HUMAN NOROVIRUS-VP1  
USING QM, MD AND ML-MD METHODS**

**MÜTESİR TEMEL**  
**A THESIS SUBMITTED FOR THE DEGREE OF**  
**MASTER OF SCIENCE**  
**DEPARTMENT OF CHEMISTRY**

**THESIS SUPERVISOR**  
**PROF. DR. AZİZ TANRISEVEN**  
**II. THESIS SUPERVISOR**  
**ASSOC. PROF. DR. ABDÜLKADİR KOÇAK**

**GEBZE**  
**2022**

**T.C.**  
**GEBZE TEKNİK ÜNİVERSİTESİ**  
**FEN BİLİMLERİ ENSTİTÜSÜ**

**QM, MD VE ML-MD METOTLARINI**  
**KULLANARAK KÜÇÜK**  
**MOLEKÜLLERİN HASSAS ÇÖZÜNME**  
**SERBEST ENERJİ VE HBGA'LARIN**  
**HUNOV-VP1 PROTEİNİNE BAĞLANMA**  
**HESAPLAMALARI**

**MÜTESİR TEMEL**  
**YÜKSEK LİSANS TEZİ**  
**KİMYA ANABİLİM DALI**

DANIŞMANI  
PROF. DR. AZİZ TANRISEVEN  
II. DANIŞMANI  
DOÇ. DR. ABDÜLKADİR KOÇAK

**GEBZE**  
**2022**



## YÜKSEK LİSANS JÜRİ ONAY FORMU

GTÜ Fen Bilimleri Enstitüsü Yönetim Kurulu'nun 28/07/2022 tarih ve 2022/36 sayılı kararıyla oluşturulan jüri tarafından 11/08/2022 tarihinde tez savunma sınavı yapılan Mütessir TEMEL'in tez çalışması Kimya Anabilim Dalında YÜKSEK LİSANS tezi olarak kabul edilmiştir.

### JÜRİ

ÜYE

(TEZ DANIŞMANI) : Prof. Dr. Aziz TANRISEVEN

ÜYE

: Prof. Dr. Viktorya AVİYENTE

ÜYE

: Prof. Dr. Bünyemin ÇOŞUT

### ONAY

Gebze Teknik Üniversitesi Fen Bilimleri Enstitüsü Yönetim Kurulu'nun

...../...../..... tarih ve ...../..... sayılı kararı.

İMZA/MÜHÜR

## SUMMARY

Recent studies have showed that machine learning (ML ) methods can be used to predict quantum mechanics (QM) energy at a given non-equilibrium geometry for small organic compounds. Several models using active learning such as “Accurate Neural network engine for Molecular Energies” (ANI) have been trained to calculate energies of small organic compounds containing C, H, O, N, F, Cl and S atoms in non-equilibrium conformations at the Density Functional Theory (DFT) level, which can be utilized to solve problems in the field of biochemistry and biophysics.

In this thesis, ANI-ML potentials have been utilized as a new method to study the solvation free energy calculations and interaction energy. Here, the method which uses ANI-2x to calculate the single point energies the trajectory of an MD simulation at the wb97x/6-31G\* level for the molecules (~640) in FreeSolv Database was introduced. Then, the ensemble averages from our method were used to estimate free energies of solvation of small organic compounds in the presence of explicit solvent. In the final part, newly introduced method was applied to analyze torsions between glycan subunits in the existence and absence of HuNoV-VP1 protein to understand details of the interaction by means of generation of multi-dimensional potential energy surface via calculating SPE with ANI-2X. As a result, after all calculations, a good correlation value between calculated and experimental SFEs was obtained. Therefore, all results indicate that ANI potentials may be a useful and robust candidate as a post-MD method for SFE and BFE calculations.

**Keywords:** Solvation Free Energy, Machine Learning, Molecular Dynamics, Quantum Mechanics, Small Ligand, Human Norovirus.

## ÖZET

Son çalışmalar, küçük moleküller için belirli bir denge dışı geometride kuantum mekanik (QM) enerjisini tahmin etmek için makine öğrenimi (ML) yöntemlerinin kullanılabileceğini göstermiştir. “Moleküler Enerjiler için Doğru Nöral ağ motoru” (ANI) gibi birkaç model, C, H, O, N, F, Cl ve S içeren küçük moleküllerin enerjilerini, biyokimya ve biyofizik alanındaki problemleri çözmede faydalanan Yoğunluk Fonksiyonel Teorisi (DFT) seviyesinde hesaplamak için eğitilmiştir.

Bu tezde, ANI-ML potansiyelleri, çözünme serbest enerji hesaplamalarını ve etkileşim enerjisini incelemek için yeni bir yöntem olarak kullanılmıştır. Burada, FreeSolv Veritabanındaki moleküller (~640) için wb97x/6-31G\* düzeyinde bir MD simülasyonunun yörüngesini tek nokta enerjilerini hesaplamak için ANI-2x kullanan yöntem tanıtıldı. Daha sonra, yöntemimizden elde edilen topluluk ortalamaları, açık çözücü varlığında küçük organik bileşiklerin serbest çözünme enerjilerini tahmin etmek için kullanılır. Son bölümde, HuNoV-VP1 proteininin varlığında ve yokluğunda glikan alt birimleri arasındaki torsiyonları analiz etmek için yeni tanıtılan yöntem, ANI-2X ile SPE'yi hesaplayarak çok boyutlu potansiyel enerji yüzeyinin üretilmesi yoluyla etkileşimin ayrıntılarını anlamak için uygulanır. Sonuç olarak, tüm hesaplamalardan sonra, hesaplanan ve deneysel SFE'ler arasında iyi bir korelasyon değeri elde edilmiştir. Bu nedenle, tüm sonuçlar, ANI potansiyellerinin SFE ve BFE hesaplamaları için MD sonrası bir yöntem olarak yararlı ve sağlam bir aday olabileceğini göstermiştir.

**Anahtar Kelimeler:** Serbest Çözünme Enerjisi, Makine Öğrenimi, Moleküler Dinamic, Quantum Mekanik, Küçük Ligandlar, İnsan Norovirüsü.

## ACKNOWLEDGEMENTS

I would like to express my deep and sincere gratitude to my supervisor, Prof. Dr. Aziz TANRISEVEN, Doç. Dr. Abdülkadir KOÇAK, and Dr. Müslüm YILDIZ who not only shared their profound scientific knowledge with me but also taught me great lessons of life. Their support, suggestions and encouragement gave me the drive and will to complete this work.

I wish to express my warm and sincere thanks to Ömer TAYFUROĞLU for his collaboration to this research.

I am grateful to my all-family members -especially my mother- for their love and support along my master's degree education and I am really pleased to have a family like that.

I want to deeply thank my lab collaborators Ebru AKKUŞ and Mehmet Serdar KOCA, who helped me about everything during my study. In addition, I want to thank to Fatma ERTAN for her support and wishes.

Finally, I want to thank my friends Halilcan DEMİRCAN and Rıdvan ORMAN for their help and support in all matters with all my heart. Also, I am really happy to have such an everlasting friendship.

I thank to Tubitak ULAKBİM because numerical calculations reported in this study were partially performed at TUBITAK ULAKBİM, High Performance and Grid Computing Center (TRUBA resources). This work supported by Scientific and Technological Research Council of Turkey-TÜBİTAK (project number 120Z732).

# TABLE of CONTENTS

|                                                                  | <b><u>Page</u></b> |
|------------------------------------------------------------------|--------------------|
| SUMMARY                                                          | v                  |
| ÖZET                                                             | vi                 |
| ACKNOWLEDGEMENTS                                                 | vii                |
| TABLE of CONTENTS                                                | viii               |
| LIST OF ABBREVIATIONS AND ACRONYMS                               | x                  |
| LIST OF FIGURES                                                  | xii                |
| LIST OF TABLES                                                   | xiv                |
| <br>                                                             |                    |
| 1. INTRODUCTION                                                  | 1                  |
| 1.1. The Aim of This Study                                       | 4                  |
| 1.2. Quantum Chemical/Mechanical (QC/M) Methods                  | 5                  |
| 1.2.1. Wave Function-Dependent Methods                           | 5                  |
| 1.2.2. Density Functional Theory (DFT) Based Methods             | 6                  |
| 1.3. Molecular Mechanics (MM) And Molecular Dynamic (MD) Methods | 8                  |
| 1.3.1. Free Energy Calculations Using MD Simulations             | 14                 |
| 1.4. Quantum Mechanics/Molecular Mechanics Methods (QM/MM)       | 21                 |
| 1.5. Machine Learning-Based (ML) Methods                         | 22                 |
| 1.6. Human Norovirus (HuNoV)                                     | 26                 |
| 2. EXPERIMENT                                                    | 32                 |
| 2.1. Theory                                                      | 32                 |
| 2.1.1. Average interaction energy (“AVG”)                        | 32                 |
| 2.1.2. Exponential average with numerical correction (“nc-EXP”)  | 33                 |
| 2.1.3. Exponential average method (“EXP”)                        | 33                 |
| 2.1.4. Cumulant expansion (“cu-EXP”)                             | 34                 |
| 2.2. Computational Methods                                       | 35                 |
| 2.2.1. Test Set for MD Simulations                               | 35                 |
| 2.2.2. Classical MD and Generating Input Files                   | 35                 |
| 2.2.3. Different Implicit Water Methods in G16                   | 37                 |
| 2.2.4. ANI as Utilizable ML Potentials and Post-MD Processor     | 38                 |
| 2.2.5. Conformer Generation and SPE of HBGs                      | 40                 |

|                                                                 |    |
|-----------------------------------------------------------------|----|
| 3. RESULT and DISCUSSION                                        | 42 |
| 3.1. Testing Database and Defining Input Parameters             | 42 |
| 3.2. ANI's performance on reproducing single point DFT energies | 46 |
| 3.3. ANI's performance on reproducing QM interaction energies   | 48 |
| 3.4. Prediction of solvation free energies                      | 60 |
| 3.5. Conformer Search of HBGAs                                  | 65 |
| 4. CONCLUSION                                                   | 83 |
| REFERENCES                                                      | 84 |
| BIOGRAPHY                                                       | 98 |
| APPENDICES                                                      | 99 |

## LIST OF ABBREVIATIONS AND ACRONYMS

| <b><u>Abbreviations<br/>and Acronyms</u></b> | <b><u>Explanation</u></b>                               |
|----------------------------------------------|---------------------------------------------------------|
| $\lambda$                                    | : Number of state                                       |
| q                                            | : Atomic charge                                         |
| $\Phi$                                       | : Psi                                                   |
| $\Psi$                                       | : Wavefunction and phi                                  |
| $\Delta$                                     | : Delta                                                 |
| $\hat{H}$                                    | : Hamiltonian                                           |
| m                                            | : Mass                                                  |
| F                                            | : Force                                                 |
| k                                            | : Spring constant                                       |
| v                                            | : Velocity                                              |
| a                                            | : Acceleration                                          |
| U                                            | : Total energy                                          |
| V                                            | : Potential energy                                      |
| K                                            | : Kinetic energy                                        |
| AAMD                                         | : Viral Protein                                         |
| AI                                           | : Artificial Intelligence                               |
| ANI                                          | : Accurate Neural Network Engine for Molecular Energies |
| BAR                                          | : Bennet Acceptance Ratio                               |
| CC                                           | : Coupled Cluster                                       |
| CCSD(T)                                      | : Coupled cluster single double triple                  |
| CI                                           | : Configuration interaction                             |
| D                                            | : Dispersion                                            |
| DFT                                          | : Density Functional Theory                             |
| DNA                                          | : Deoxyribonucleic acid                                 |
| EP                                           | : Empirical interatomic particles                       |
| FEP                                          | : Free Energy Perturbation                              |
| FF                                           | : Force Field                                           |
| HBGA                                         | : Histo-blood group antigens                            |
| HF                                           | : Hartree-Fock                                          |

|        |   |                                                         |
|--------|---|---------------------------------------------------------|
| HuNoV  | : | Human Norovirus                                         |
| IEFPCM | : | Integral Equation Formalism Polarizable Continuum Model |
| IP     | : | Interatomic Potential                                   |
| LIE    | : | Linear Interaction Energy                               |
| LRA    | : | Linear Response Approximation                           |
| MBAR   | : | Multiple Bennet Acceptance Ratio                        |
| MC     | : | Monte Carlo                                             |
| MD     | : | Molecular Dynamics                                      |
| MM     | : | Molecular Mechanics                                     |
| ML     | : | Machine Learning                                        |
| MP     | : | Møller–Plesset                                          |
| NN     | : | Neural network                                          |
| NPE    | : | nonhomogeneous Poisson equation                         |
| NPT    | : | Isothermal-isobaric ensemble                            |
| NVT    | : | Canonical ensemble                                      |
| ORF    | : | Open Reading Frame                                      |
| PBC    | : | Periodic boundary conditions                            |
| PCM    | : | Polarizable Continuum Model                             |
| QC     | : | Quantum Chemical                                        |
| QM     | : | Quantum Mechanical                                      |
| QSPR   | : | Quantitative structure-property relationship            |
| RNA    | : | Ribonucleic acid                                        |
| SE     | : | Semi empirical                                          |
| SFE    | : | Solvation Free Energy                                   |
| SMD    | : | Solvation Model Based on Density                        |
| SPE    | : | Single Point energy                                     |
| TI     | : | Thermodynamic Integration                               |
| vdW    | : | Van der Waals                                           |
| VP     | : | Viral Protein                                           |

## LIST OF FIGURES

| <b><u>Figure No:</u></b>                                                                                                                                                                    | <b><u>Page</u></b> |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------|
| 1.1: The schematic representation of MD working mechanism.                                                                                                                                  | 13                 |
| 1.2: Mathematical definition of different MD methods.                                                                                                                                       | 20                 |
| 1.3: The view of QM/MM region in a solvated ligand system.                                                                                                                                  | 22                 |
| 1.4: Different Types of HBGA based on the type of sugar unit and connection type.                                                                                                           | 28                 |
| 1.5: A general genome view of VP1.                                                                                                                                                          | 29                 |
| 1.6: HBGA interaction mechanism and residues.                                                                                                                                               | 30                 |
| 2.7: The view of Norovirus G.II-4 with its ligand (PDB ID:4X07).                                                                                                                            | 41                 |
| 3.8: Testing data selected based on FreeSolv Database.                                                                                                                                      | 42                 |
| 3.9: The BAR results for selected 28 molecules at NVT and NPT ensemble for 5 ns of simulation.                                                                                              | 44                 |
| 3.10: Workflow for ANI prediction of SFE by applying on single step MD trajectory.                                                                                                          | 46                 |
| 3.11: SPE of isolated molecules in FreeSolv predicted by ANI2x vs DFT.                                                                                                                      | 47                 |
| 3.12: The SPE of the ligand (mobley_1636752)-nH <sub>2</sub> O (n=10,20..100) systems calculated by ANI and DFT obtained MD trajectories.                                                   | 50                 |
| 3.13: The SPE of the different ligands with different nH <sub>2</sub> O systems calculated by ANI and DFT obtained MD trajectories.                                                         | 51                 |
| 3.14: Interaction energies calculated by ANI and G16 for C <sub>6</sub> H <sub>6</sub> -nH <sub>2</sub> O (n=1-10), $\Delta U = U_{\text{complex}} - (U_{\text{lig}} + U_{\text{water}})$ . | 52                 |
| 3.15: Comparison of interaction energies predicted by ANI and G16 throughout MD trajectories in the presence of 70-90 (a, b, and c, respectively) water for a selected ligand.              | 53                 |
| 3.16: DFT and ANI calculated interaction energies for methanol (mobley_16366752)268H <sub>2</sub> O system along 100 frames of a 40 ns MD simulation.                                       | 55                 |
| 3.17: DFT and ANI calculated interaction energies for methanol (mobley_1636752)-100H <sub>2</sub> O system along 500 frames of a 5 ns MD simulation.                                        | 56                 |

|       |                                                                                                                |    |
|-------|----------------------------------------------------------------------------------------------------------------|----|
| 3.18: | Comparison of scaled ANI and DFT interaction energy calculation for different molecules.                       | 57 |
| 3.19: | Average interaction energy of all 500 frames from MD simulations.                                              | 58 |
| 3.20: | The view of interaction energy calculated by ANI and G16 for different molecules in different solvent volumes. | 59 |
| 3.21: | Average ANI and D3 interaction energies.                                                                       | 60 |
| 3.22: | Three different approaches to predict solvation free energies.                                                 | 62 |
| 3.23: | The other continuum model for solvation free energy calculation.                                               | 63 |
| 3.24: | ANI predicted solvation free energies shows more accuracy than LIE and less computational cost than BAR.       | 64 |
| 3.25: | ANI's success of reproducing experimental SFEs using equations (2.7)-(2.11).                                   | 65 |
| 3.26: | HBGA A Type 1 for only HBGA's SPE.                                                                             | 66 |
| 3.27: | HBGA A Type 1 for HBGA-protein complex.                                                                        | 68 |
| 3.28: | HBGA A Type 2 for only HBGA's SPE.                                                                             | 69 |
| 3.29: | HBGA A Type 2 for HBGA-protein complex.                                                                        | 70 |
| 3.30: | HBGA B Type 1 for only HBGA's SPE.                                                                             | 71 |
| 3.31: | HBGA B Type 1 for HBGA-protein complex.                                                                        | 73 |
| 3.32: | HBGA B Type 2 for only HBGA's SPE.                                                                             | 74 |
| 3.33: | HBGA B Type 2 for HBGA-protein complex.                                                                        | 75 |
| 3.34: | HBGA H Type 1 for only HBGA's SPE.                                                                             | 76 |
| 3.35: | HBGA H Type 1 for HBGA-protein complex.                                                                        | 77 |
| 3.36: | HBGA H Type 2 for only HBGA's SPE.                                                                             | 78 |
| 3.37: | HBGA H Type 2 for HBGA-protein complex.                                                                        | 79 |
| 3.38: | HBGA Precursor Type 1 for only HBGA SPE and ligand-protein complex, respectively.                              | 80 |
| 3.39: | HBGA Precursor Type 1 for only HBGA SPE and ligand-protein complex, respectively.                              | 81 |

## LIST OF TABLES

| <b><u>Table No:</u></b> |                                                                           | <b><u>Page</u></b> |
|-------------------------|---------------------------------------------------------------------------|--------------------|
| 3.1:                    | The interaction energy of benzene with 1 to 10 H <sub>2</sub> O molecules | 48                 |

# 1. INTRODUCTION

Getting mechanistic insights into how biomolecules function is one of the fundamental goals of molecular biochemistry. To do this, scientists must devise tests utilizing a variety of biochemical and biophysical methods that provide novel information about the target molecule. The experimental findings must then relate to the unique properties of the molecule under investigation. In some instances, this procedure is a simple interpretation, but in many others, it is challenging to understand the molecular significance of the data. As a result, one of an experimentalist's primary responsibilities is to analyze the data and draw new conclusions about a particular molecular mechanism from the findings [1], [2].

In recent years, the rate of use of computer science and related fields has shown a rising trend and developments in this area affect directly areas showing highly dependencies to them. Computational chemistry, a branch of chemistry that uses models to aid in the operation of chemical issues, is one of the areas of usage that is expanding quickly. To analyze the structures and portions of molecules, groups of materials and molecules, and solids, theoretical chemistry techniques are combined into computer algorithms. [3]. It's essential because of, piecemeal from fairly recent results regarding the chemical element molecular ion the quantity numerous- body downside cannot be answered analytically, a lot of less in unrestricted kind. Whereas procedure results usually round the info earned by chemical trials, it will in some cases prognosticate hitherto unobserved chemical marvels. It's extensively utilized within the designing of latest medicines and accoutrements. Structure, binding free energy, interaction energy, hydration free energy, electronic charge density distributions, dipoles and advanced multipole moments, undulation frequency, reactivity, or varied chemical analysis quantities, and cross sections for impact with distinct patches are examples of comparable properties [4]. One typical downside in cheminformatics is to predict the binding affinity of medication molecules to a given target. different issues embody estimating binding affinity, off- target product, toxin, and pharmacokinetic parcels [5].

In general, the aim of computational chemistry methods is to have the knowledge about thermodynamic and electronic properties of molecules of interest. Solvation free energy and binding free energy are the ones of the most important properties. Also,

they are utilizable in a lot of fields of natural science, especially drug development, biochemistry and pharmacokinetics [3].

In the fields of material, biological, and pharmacological sciences, solvation free energy is used to describe the distinctive characteristics of molecules. For instance, since the desolvation costs for complexation can significantly affect the total binding free energy, knowledge of solvation free energy is necessary to determine the equilibrium constant for protein-ligand association [6]. Due to their impact on the bioactivity of drug candidates at the site of action, solvation characteristics are also significant in the drug discovery process. This necessitates the early stage of drug discovery to determine the molecular solvation free energy or solubility [7]. However, it is exceedingly challenging to screen a sizable chemical library from which physically or physiologically active molecules can be found because the experimental measurement of solvation free energy is a time-consuming operation. With the advent of combinatorial chemistry in recent years, the need to estimate the differences between solvation free energies of structurally related compounds has grown more urgent, further necessitating the development of a trustworthy computational method to predict solvation free energies of organic molecules [8].

Due to the intricate nature of the interactions between solutes and solvents, the solvation free energy has, nonetheless, been regarded as one of the most calculably challenging energy concepts [9]. Nevertheless, since the early work of Onsager [10], other computational techniques for the prediction of molecular solvation free energies have been investigated. To replicate the electrostatic screening by solvent, the dielectric constant might be adjusted in a distant-dependent manner to approximate the simplest solvation model [11]. On the basis of the Poisson-Boltzmann equation to determine the electrostatic potentials surrounding the solute molecule, more accurate methods for forecasting molecular solvation free energies were also proposed [12]. This has the effect of restricting the reliability of the calculated molecular solvation free energy because these implicit solvation models were unable to depict the solute-solvent interactions on an atomic scale. The accurate determination of molecular solvation free energy, however, has been demonstrated to be possible even in the case of protein-ligand complexes using all-atom model computations based on molecular dynamics and Monte Carlo (MC) simulations [13]. High-level quantum chemistry computations have also been used to accurately estimate solvation free energy for molecules [14], [15]. The quantum mechanical and all-atom models for molecular

solvation have not been widely used in practical applications, despite the increased accuracy [16], due to the high computing costs. In order to strike a balance between computational cost and accuracy, several effective computational techniques have been suggested. These include the solvent-accessible surface area model [17], [18], the 3-D reference interaction site model [19], the cellular automata based approach [20], the quantitative structure-property relationship (QSPR) model [21], the linear interaction energy method [22], and the quantum mechanical continuum solvation [23], [24].

In the scientific literature, solvent models have undergone thorough testing and analysis. There are two sorts of models: explicit and implicit models, each with its own benefits and drawbacks. A solute molecule's immediate environment is subject to local changes in solvent density, which are not taken into account by implicit models, despite the fact that they are typically computationally efficient and can provide a realistic description of the behavior of the solvent [25]. When one considers water as the solvent, the density fluctuation behavior, which is caused by solvent ordering around a solute, is more common. Although explicit models can offer a physical, spatially detailed description of the solvent, they are frequently less efficient computationally. Although many of these explicit models can fail to reproduce experimental results, this is frequently because of high entropic contribution in the partition function due to the mobility of water molecules. Another choice is hybrid approaches. These techniques combine implicit and explicit elements to reduce computing expense while maintaining at least some spatial resolution of the solvent. These techniques frequently include post-calculation adjustment phrases and can be more difficult to employ appropriately [26].

In the earlier of 21<sup>st</sup> and latest of 20<sup>th</sup> century, different kind of implicit solvation models as example of integral equation formalism polarizable continuum model (IEF-PCM) [27], conductor-like polarizable model (C-PCM) [28], were devised based on the polarizable continuum model (PCM) of Tomasi that solves the nonhomogeneous Poisson equation (NPE) to simulate bulk electrostatic effects of the solvent with the development in density function theory (DFT) and progress in computing strength [29]. Alternatively, some other implicit solvation models such as SM12 and SM8 are designed in a way to represent the solute as a collection of point charges [30], [31]. These models use Poisson Boltzmann or the generalized Born approximation for the polar contribution and solvent accessible surface area or volume for the non-polar

contribution to the solvation[32]. For correct prediction of solvation free energies, especially in the case of charged solutes, thorough simulation of bulk electrostatic interactions is essential. With the use of a cavity that simulates a solvated molecule by a dielectric medium, these bulk electrostatic contributions are considered when calculating the final solvation free energy [33], [34].

The computation of condensed phase parameters like SFE has benefited greatly from the use of an implicit solvation model based on density (SMD) [35], which was created as a universal solvation model where "universal" refers to its application to both ionic and neutral solutes in any solvent. Schlegel and colleagues have demonstrated that using the implicit SMD solvation model in combination with up to three explicit water molecules improves SFE values for a number of functional groups compared to implicit-only simulations [36], [37].

The aquatic environment has received only sporadic attention in the quest for a force field that is sufficiently realistic. The hydrophobic effect and Coulomb interaction screening, however, disappear if water molecules are not present around a biomolecule since the water environment is crucial to biomolecular functions. For a description of at least solute–solvent interactions to be reasonably realistic, water molecules must therefore be explicitly taken into account [38]. In an effort to account for all of the physicochemical characteristics of water, numerous explicit water models have been created [39]. The two that are most frequently used for biomolecular simulations are Jorgensen's TIPnP [40] and Berendsen's SPC/E [41] models. However, it should be mentioned that while the balance of solute–solvent interactions in these models is still debatable, the explicit solvent models have been developed to accurately capture the physicochemical characteristics of bulk water, i.e., their solvent–solvent interactions [42].

Herein, we used different methods based on the MM, QM and ML for estimation of small organic moieties and HBGA interaction mechanism. Also, various solvent models were used as implicit and explicit while performing calculations.

## **1.1. The Aim of This Study**

There are many common methods to calculate the solvation free energy of small organic molecules or binding free energy of organic complexes to biological macromolecules such as protein and lipid. Besides this, the effect of Artificial

Intelligence (AI) and ML has increased day by day because they provide a lot of innovations for many various fields with example of Biology, Data analyze, Physics, Computational Chemistry and so forth. The present methods in Computational Chemistry have different scopes, abilities, advantages, and drawback. The use of Machine Learning based methods combining with those make a new attempt in this field. To put it briefly, the goal of this study is to use a combination of traditional methods and ML-based ANI to calculate the solvation free energies and single point energies of small molecules and to apply this method to massive structures and complexes to determine the interaction mechanism in human norovirus.

## 1.2. Quantum Chemical/Mechanical (QC/M) Methods

Electronic structure methods, such as Quantum Chemical/Mechanical (QM) methods, utilize the time-independent Schrödinger equation (Equation (1.1)) covered by Born–Oppenheimer approximation for calculation of interaction mechanism between electron and atom core. This approximation is often used in chemical and physical theories. There are two types QM methods: those that use wave function methods, and those that use the density functional theory [43].

$$\left[-\frac{1}{2}\nabla^2 - \frac{Z}{r}\right] \psi(r) = E\psi(r) \quad (1.1)$$

A charge-carrying nucleus  $Z$  is indicated by the amount enclosed in square brackets as well as the kinetic and potential energy of an electron. Electronic energy ( $E$ ) is measured in atomic units, and the electron coordinates are described by a wavefunction,  $r$ , which the fullest extent possible, the electron's velocity.

### 1.2.1. Wave Function-Dependent Methods

Methods based wave function, by dealing with Schrodinger wave equation, are used to calculate the expressly related electronic wave functions. The simplest one of the wave function-based techniques is that the Hartree–Fock (HF), inside that the multielectron wave perform is approximated to at least one Slater determinant but this approximation leads to the biggest HF mistake: it neglects the immediate electron correlation. In other words, it assumes that one electron moves in an averaged field

because of the other electrons when the motion of the electrons is correlated (the motion of one electron depends on the instant, mutual interaction with the other electrons). A flaw in the HF wave perform and energy is introduced by the lack of immediate electron correlation, substantially impairing predictions of kinetic barriers or the definition of London dispersion forces [44].

Despite this imprecision, the HF solution are typically consistently improved. A way is through many body perturbation theory, culminated inside the Møller–Plesset (MPn) strategies, or by expressing the wave operate as a linear combination of slater determinants, i.e., the HF one and those representing single, double, triple, etc., excitations, giving rise to the configuration interaction (CI) approaches. The coupled cluster (CC) theory additionally introduces excited state configurations to the wave operate however adopting associate degree exponential enlargement (through the cluster operator) on the HF one. The foremost fashionable CC derivation is that the CCSD(T) methodology, inside that single (S) and double (D) excitations are enclosed through the cluster operator and triple (T) excitations by utilizing the perturbation theory [45]. In addition to using primarily expanded basis sets, CCSD(T) incorporates electron correlation in great detail and is regarded as the "gold standard" in QM [46].

The scale of the systems limits the applicability of the post-HF techniques because, despite their higher quality, they are computationally expensive. In practice, these calculations in potential energy surface (PES) modeling are primarily used for calibrating additional approximate (but less expensive) strategies to make sure their quality inside the outline of the electronic structure of the simulated surfaces. This is because CCSD(T) calculations for quite 25–30 atoms are impractical [47].

### **1.2.2. Density Functional Theory (DFT) Based Methods**

The mathematical formulation created by Korn and Sham in 1965 serves as the foundation for the density functional theory (DFT) [48]. According to the central theorem, a non-reduced electronic system's base state energy is precisely and universally determined by the total electron density ( $r$ ) using a general mathematical function (the general function is a function).

The energy of the system can be written as the function  $E$  of  $r$ , that is,  $E[r]$ , by taking a single number from it and defining it, such as the definite integral. Since the  $\rho(r)$  for a system of  $N$  electrons in DFT only depends on three spatial coordinates,

compared to the corresponding wave function's dependence on  $3N$  spatial and  $N$  spin variables, this offers a significant advantage over wave function-based approaches. Because of this, the DFT technique can be even more affordable than HF (depending on the function  $E[r]$ ) while still including a sizable amount of the missing instantaneous electron correlation. The drawback of DFT is that, although it is mathematically conceivable to demonstrate the existence of a universal function connecting electron density to system energy, the precise form of  $E[r]$  is still unknown. There are numerous DFT techniques because of the pursuit of the exchange-correlated functional  $E_{xc}[r]$ , which connects these two values.

Numerous DFT techniques have been created, as was previously noted. There are many different molecular and solid system systems, but only a few are most frequently employed to give a good account of the electrical structure and related features of these systems. PBE, OPBE, PW91, and BLYP are popular GGA techniques. B3LYP, BHLYP, PBE0, and WB97X are examples of common hybrid functionals. Truhlar and his associates created the M06L meta-GGA 20 and M06 and M06-2X hybrid meta-GGA 20, which are functionals of the M06 family and the most widely used meta-GGA and hybrid meta-GGA. It serves its purpose [49].

The DFT approach's inability to handle long-distance non-shared (distributed or London) interactions is one of its fundamental flaws, especially when compared to the older method. Grimme put forth a practical approach to accounting for dispersion interactions in DFT methods, in which the DFT energy is supplemented with an a posteriori correction factor based on an atom-atom additive London-type empirical potential (D) [50]. Iteratively proposed and refined corrective words include (D, D1, D2, D3 and D4). The primary benefit of the DFT-D scheme is that it incorporates the dispersion interactions at a low computing cost while preserving the original accuracy of the pure DFT method for short-range interactions [51].

It is hard to make calculation of solvation free energy of small molecules in an explicit water environment at the level of QM because of high degree of freedom in QM calculations. Thus, implicit solvent models were developed to calculate solvation free energy of small molecules rapidly and (modeling) the solvent applying as a continuum, in oppose to individual molecules, enable ab initio computation feasible for SFE calculation. Thus, there are several implicit solvent models like CPCM, IFPCM, and SMD that are developed and combined with DFT to make SFE calculations with a high rapidity. Instead of using molecules in explicit water medium

to calculate SFE, the methods known as implicit water methods can be used to get SFE of small molecules at DFT level [52].

### **1.3. Molecular Mechanics (MM) and Molecular Dynamic (MD) Methods**

The electrons' velocity is disregarded by force field approaches, which are based on classical molecular mechanics (MM). A "ball and spring" model is used to describe chemical systems, in which atoms are represented by balls of varying sizes and softness and bonds by springs of varying stiffness. Only the nuclear locations are used to compute the system's energy. To this aim, molecular mechanics energy, which gauges the degree of mechanical strain inside the system, is defined using force fields (FFs), a collection of interatomic potentials (IPs) including energy functions and parameters. Very large systems (thousands of atoms) can be simulated to forecast conformational flexibility and relative stability because electrons are not explicitly considered in the standard MM [53].

According to molecular mechanics, molecules are described as "bound atoms," which are deformed from an idealized shape by nonbonded van der Waals and Coulombic interactions [1]. Compared to quantum chemical models, which have absolutely no mention of chemical bonding, this is significantly different. The effectiveness of molecular mechanics models relies on the geometrical parameters' high degree of transferability from one molecule to another and their predicted influence on atomic hybridization [54]. If the molecule has already been represented in terms of a specific valence structure, it is possible to make a pretty accurate "guess" about the molecular geometry in terms of bond lengths, bond angles, and torsion angles. This category encompasses the vast majority of organic compounds [55].

The intra- (i.e., covalent) and inter- (i.e., non-covalent) molecular forces inside the system are represented by FFs, which contain energy functions with bonded and non-bonded terms. In general, the various contributions present in the IPs should simulate: (i) the interaction between pairs of bound atoms (covalent bond stretching), (ii) the energy needed to bend an angle created by three atoms (angle bending), (iii) the energy change associated with a bond rotation (torsional or dihedral), and (iv) the non-bonded pairwise interactions, typically utilizing a Coulomb potential for electrostatic interactions and a Lennard-Jones for van der Waals interactions [56].

The "energy" of a molecule as it relates to molecular mechanics is represented as the sum of contributions from distortions of "ideal" bond distances, bond angles, and torsion angles, as well as contributions from "non-bonded" (van der Waals and Coulombic) interactions. It is frequently referred to as a "strain energy," which denotes that it represents the "strain" a "actual" molecule experiences in contrast to an idealized form. For the following equation, each item that makes up the force field energy describes the amount of energy necessary to deform a molecule in a particular way.

$$E_{\text{ForceField}} = E_{\text{stretch}} + E_{\text{bend}} + E_{\text{torsion}} + E_{\text{non-bonded}} \quad (1.2)$$

Equation 1's first three summations, which cover all "bonds stretching" all "bond angles," and all "torsion angles," are done in that order. Thus, bonding knowledge is "part of the input" to a molecular mechanics calculation as opposed to "part of the output" in a quantum chemistry calculation. All atom pair combinations that are not bound are the subject of equation 1's final summation.

Different force fields have different mathematical representations for the energy terms. Based on Hook's law, we may calculate the stretching energy. The bond spring's stiffness is determined by the  $k^{\text{stretch}}$  parameter (force constant). The two atoms' equilibrium separation is denoted by the symbol  $r^{\text{eq}}$ . That variations from the equilibrium length would be linked to increased energy should make sense. Hook's law is also the foundation for the bending energy equation. The angle spring's stiffness is controlled by the  $k^{\text{bend}}$  parameter, while the equilibrium angle is defined by the  $\alpha^{\text{eq}}$  parameter. The following equations are for stretching and bending, respectively:

$$E^{\text{stretch}}(r) = \frac{1}{2} k^{\text{stretch}} (r - r^{\text{eq}})^2 \quad (1.3)$$

$$E^{\text{bending}}(\alpha) = \frac{1}{2} k^{\text{bend}} (\alpha - \alpha^{\text{eq}})^2 \quad (1.4)$$

The bond length and angle are denoted by  $r$  and  $\alpha$ , respectively. The "ideal" (equilibrium) bond length and angle are denoted by  $r^{\text{eq}}$  and  $\alpha^{\text{eq}}$ , and they are obtained either through experimentation or precise quantum chemical calculations and the

parameters  $k^{\text{stretch}}$  and  $k^{\text{bend}}$ , which are referred to as the stretch and bend "force constants," respectively.

The torsional potential must be described in a way that accurately captures its innate periodicity. The basic functional form, for instance, may be used to explain the three-fold periodicity of rotation about the carbon-carbon bond in ethane.

$$E^{\text{torsion}}(\omega) = k^{\text{torsion}3} [1 - \cos 3 (\omega - \omega^{\text{eq}})] \quad (1.5)$$

$\omega^{\text{eq}}$  is the ideal torsion angle,  $k^{\text{torsion}3}$  is handled as a parameter, and  $\omega$  is the torsion angle. One-fold and two-fold periodic terms may also need to be included in the bond torsion contributions to the total energy.

Van der Waals (vdW) interactions and coulombic interactions make up the majority of non-bonded interactions.

$$E^{\text{non-bonded}}(r) = E^{\text{vdW}}(r) + E^{\text{coulombic}}(r) \quad (1.6)$$

To explicitly take into consideration interactions like hydrogen bonding, more non-bonded terms may be included. van der Waals interactions are typically depicted as the sum of an attracting and repulsive term.

$$E^{\text{vdW}}(r) = \varepsilon \left[ \left( \frac{r^0}{r} \right)^{12} - 2 \left( \frac{r^0}{r} \right)^6 \right] \quad (1.7)$$

$\varepsilon$  and  $r^0$  are variables, while  $r$  represents the non-bonded distance. For the two atoms involved, this functional form offers a highly steep energy barrier inside the sum of van der Waals radii and a shallow energy well at larger separations. As a result, it compensates for both the fundamental size needs of atoms and the weak attraction forces that exist between separated atoms.

The interaction of charges is considered in the Coulombic term.

$$E^{\text{coulombic}}(r) = \frac{qq'}{r} \quad (1.8)$$

Atomic partial charges,  $q$ , may either be used as parameters or as a result of quantum chemical computations, with  $r$  being the non-bonded distance. The total

molecular charge, or 0 in the case of a neutral molecule, must match the sum of the atomic partial charges.

The parameters that characterize the various types of atoms, chemical bonds, angles torsions, non-bonded interactions, and other terms are contained in the energy functions, such as equilibrium values for bond lengths, angles, dihedrals, and atomic charges. Historically, the FF parametrization (i.e., the numerical assignment of the parameters) was determined from experimental data; however, today, QC calculations have taken its place [57].

Due to their low cost, molecular dynamic (MD) simulations are used to simulate the dynamics of the nuclear motion within a molecule using MM approaches. In MD, the system's subsequent configurations are created by integrating Newton's equations of motion.

$$F = ma = -\nabla V \quad (1.9)$$

This term corresponds to all nuclei in x, y, z 3D space.

$$F_i = -\nabla_{r_i} V(r_1..r_i..r_N) \quad (1.10)$$

Therefore Simplest choice for potential is a sum of pairwise interactions

$$V(r_1....r_N) = \sum_i \sum_{j>i} V(|r_i - r_j|) \quad (1.11)$$

The end result is a trajectory that describes how the nuclei's positions and velocities change over time and includes measurements of the system's molecular mechanics energy and nuclear forces at each nuclear position. The average kinetic energy associated with nuclear motion provides the temperature of the system, making MD simulations an excellent choice for studying the evolution in time-space phase of the atomic locations due to internal chemical forces. The ergodicity of the system can be attained by using traditional MD simulations, which can achieve time scales of hundreds of picoseconds to microseconds [55].

$$\bar{K} = \frac{1}{2}m\bar{v}^2 = \frac{3}{2}RT \quad (1.12)$$

One such simulation method is molecular dynamics (MD). It seeks to derive claims about a molecular system's structural, dynamical, and thermodynamic characteristics of a system. The present drug development process has benefited more and more from MD simulations. For instance, MD enables the assessment of the binding energetics and kinetics of the ligand-receptor interactions during the lead discovery and lead optimization phases, assisting in the selection of the best candidate compounds for continued development [58]. The latter is often a biomolecule (solute) that is submerged in an aqueous solvent and includes small molecules, enzymes, sugar units with lipid that make up a membrane (water or electrolyte). In atomistic "all-atom" MD (AAMD), the model system is made up of a number of interacting particles that are represented as atoms and describe both the solute and the solvent. These particles are placed inside a sufficiently large simulation box, where their motions are described by Newton's laws of motion. The state of the model system as a function of time is advanced across a large number of time steps using an algorithm like velocity-Verlet or leap-frog [59]. In the velocity verlet algorithm, velocity, distance and accelaretion of atoms in the systems are updated by the following formula:

$$\begin{aligned} \mathbf{v}(t + \Delta t) &= \mathbf{v}(t + \Delta t/2) + \mathbf{a}(t + \Delta t)\Delta t \\ \mathbf{v}(t + \Delta t/2) &= \mathbf{v}(t) + \mathbf{a}(t)\Delta t \\ \mathbf{r}(t + \Delta t) &= \mathbf{r}(t) + \mathbf{v}(t + \Delta t/2)\Delta t \\ \mathbf{a}(t + \Delta t) &= (-1/m) \nabla V(\mathbf{r}(t + \Delta t)) \end{aligned} \quad (1.13)$$

The sum of the locations and velocities of the atoms constitutes one state. For the purpose of advancing the state, forces acting on particles are calculated from a model or empirical potential energy ("force field"), a function of particle locations. This potential energy contains all sorts of "non-bonded" interactions, electrostatic and Lennard-Jones forces, as well as many types of "bonded" potentials for maintaining the structural integrity of the specific biomolecular system. The latter include words

for dihedrals as well as harmonic potentials for maintaining bonds, bond angles, and "improper" dihedrals. It is noteworthy that distinct force fields have been created for diverse kinds of molecules. Particular importance pertains to force fields for small moieties, enzymes, glycans [60], [61]. In order to simulate laboratory settings for which specialized methods are available, such simulations are typically performed under circumstances of constant temperature and pressure. Additionally, a number of methods for artificially enlarging the scale of the model system have been devised in order to mimic a very big molecular system. The periodic boundary condition is the most prevalent (PBC). In this case, the main simulation box is surrounded by an endless number of copies of it. Electrostatic interactions have a long range, hence specific approaches, like Ewald-based methods, are needed to account for the interactions between the particles in the central box and their duplicates [62]. During a simulation, successive states occurring at regular intervals are recorded in a trajectory for study at a later time [58]. There are other software programs available for MD of biomolecules, including GROMACS [63], AMBER [64], and NAMD [65].

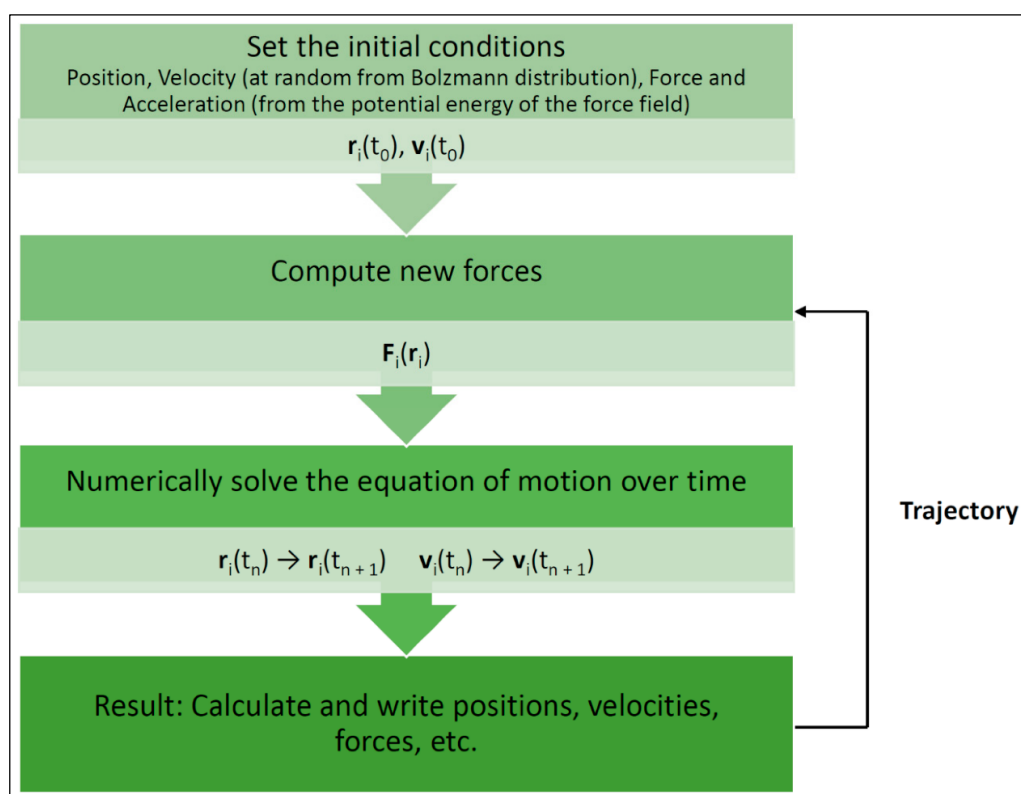


Figure 1.1: The schematic representation of MD working mechanism.

MD simulations and Methods for MD modeling have proven to be helpful in contemporary drug research, development processes, pharmaceutical development and formulations studies in addition to the initial stages of the drug discovery process during the past 50 years. For instance, drug-polymer formulations, drug-loaded nanoparticles, and crystalline and amorphous medicines can all be examined by MD simulations to supplement experimental studies. The solubility, stability, or other qualities of drug formulations can be improved with the aid of molecular-level understanding of such systems. It is now possible to model even large and complicated (bio)molecular systems on time scales that can provide significant insights into actual molecular events and interactions thanks to advances in computing power, novel force fields, and improved sampling techniques. The conformational dynamics of extremely flexible target proteins can be successfully uncovered by MD simulations, offering important insights for drug discovery and design. Additionally, the design of ligands for extremely flexible, intrinsically disordered proteins may benefit from the use of conformational ensembles produced by MD. In order to take on such difficult objectives, improved MD methods and force field development are essential. Additionally, MD simulations can provide crucial insights into the workings of antibody-antigen interactions and aid in the development of new, more effective antibody therapies.

To increase the precision of ranking the putative hits, MD-based binding free energy calculations are frequently performed in the hit identification phase. Additionally, a number of enhanced MD approaches that examine the drug residence durations at the target site are helpful in the creation of medications with improved binding kinetics [66].

Finally, MD can be an effective tool for accelerating the early stages of contemporary drug discovery and development. Given the theoretical and technological developments in the field, MD simulations are likely to become even more significant in the future.

### **1.3.1. Free Energy Calculations Using MD Simulations**

MD simulations have long been utilized to predict binding of a small drug like molecule non-covalently interacting with a larger biological macromolecules as well as solvation free energies of small ligands.

Methods for calculating ligand solvation free energies can be divided into 2 categories [67].

The first one is known as end-state methods. These are the methods that only consider the bound and/or unbound end states of the ligand. Classically, the system belonging to the ligand-solvent complex is simulated, and then the average properties (averaged ensemble) is obtained by calculating the free energy of all conformations of the trajectory obtained during the simulation. One of the common end-state methods is linearinteraction energy(LIE) [68].

In LIE method, the simulation of ligand in solvent medium in bound state is performed. Thus, the ligand environment interaction energies of the bound (fully interacting) and unbound (non-interacting) states of the ligand are calculated as Boltzmann averages. It assumes that the free energy varies linearly with these interaction energies. The interaction energies are completely at Molecular Mechanical (MM) level and are handled in two parts as electrostatic and van der Waals. It is a great advantage that it does not add any computational cost to a normal MD simulation. It can be easily obtained when ligand and solvent are introduced as different energy groups during MD. In addition, the accuracy level is very low because it only considers the final states of the system and is completely MM-level calculation. Solvation free energy for non-covalent interaction is given by:

$$\Delta G_{SFE} = \alpha \langle \Delta U_{el} \rangle_1 + \beta \langle \Delta U_{vdW} \rangle_1 + \gamma \langle Y \rangle_1 + \delta \quad (1.14)$$

Here,  $\langle \Delta U_x \rangle$  terms correspond to the mean electrostatic and van der Waals interaction energy difference (interaction energy) of the ligand in the solvent calculated at the MM level, and  $\langle Y \rangle$  corresponds to the solvent-accessible surface area of the ligand (solvent accessible surface area) and other parameters ( $\alpha$ ,  $\beta$ ,  $\gamma$ ,  $\delta$ ) are empirical coefficients that describe the linear relationship between solvation energy and interaction energy and can be obtained by fitting the experimental results [69], [70].

Linear Interaction energy (LIE) is an extension of the linear response approach (Linear Response Approximation, LRA) developed as an electrostatic term [70]. In this older approach, the free energy change of the system is:

$$\Delta G_{solv} = \alpha \langle V \rangle_1 \text{ and } \Delta G_{solv} = 1/2 \langle V \rangle_1 \quad \text{if } \langle V \rangle_0 \ll \frac{\partial^2 G}{\partial \lambda^2} \quad (1.15)$$

Here,  $\langle V_1 \rangle$  indicates the interaction energy of the ligand when the environment is fully interacting (bound state), and  $\langle V_0 \rangle$  when the ligand is not interacting with the environment at all (unbound state), which corresponds to the MM energy differences in the original LIE. This equation shows that in this approach, the free energy is equal to half the potential energy of the ligand-environment interaction.

Alchemical methods are the second part in the calculation of ligand-solvent free energies methods. These methods provide step-by-step transformation of the system from the initial state to the final state by drawing an alchemical thermodynamic path. Since the free energy change is a state function, it does not depend on path, which can be real or hypothetical. The hypothetical (unphysical, alchemical) intermediate states in which the ligand goes from being bound to unbound state are defined by values between 0 and 1 for lambda ( $\lambda$ ), a partial interaction parameter [71]. The total free energy change is equal to the sum of the energy changes in these intermediate lambda states. For example, the bound state of the solute-solvent system in the solvent is defined as the initial state with  $\lambda=0$ , and when the ligand is gradually deleted/destroyed from the system (annihilation), only the solvent molecules remain in the medium in the final state ( $\lambda=1$ ). As another example, lambda is defined as the parameter that creates the ligand-surroundings interaction energy, and it can be obtained by summing up the interaction energies in each step by gradually decoupling these interactions while providing the transition from the initial state of  $\lambda=0$  to the state of  $\lambda=1$ . This method can be used to calculate both the hydration free energy and the absolute or relative binding energy. It can also be designed as a parameter for the conversion of one ligand to another ligand. Thus, the interaction and free energy can be calculated relatively [72].

The important aspect of these techniques is that there must be a string of intermediates with mutual phase-space overlap linking the two physical end states for practically all alchemical conversions between the initial and final Hamiltonians. The linear route is the most straightforward option for most transformations between the two Hamiltonians  $U_0$  and  $U_1$ :

$$U(x) = (1 - \lambda)U_0(x) + \lambda U_1(x) \quad (1.16)$$

The most common alchemical methods are Thermodynamic Integration (TI) [73], Free Energy Perturbation (FEP) [74], and Bennett Acceptance Ratio (BAR) methods [75] and Jarzynski's relationship [76]. All are based on the alchemical change of the parameter  $\lambda$  from 0 to 1 .

The FEP method, also known as exponential averaging and using the Zwanzig relationship, is the origin of all alchemical free energy methods and is seen as the first method to calculate the absolute free energy change [74]. In FEP theory, the energy of the final state of the system is expressed as the energy of the initial state and an additional perturbation. Thus, the energy change between the two states (start,0 and end,1) is given as the exponential mean of the deviation of the initial state:

$$U_i(\lambda) = U_0 + \lambda \Delta U \quad (1.17)$$

$$Z_i(N, V, T, \lambda) = \sum_i e^{-U_i(\lambda)/kT} \quad (1.18)$$

$$A(N, V, T, \lambda) = -kT \ln[Z(N, V, T, \lambda)] \quad (1.19)$$

$$\Delta A(0 \rightarrow 1) = -kT \ln \langle e^{-\Delta U/kT} \rangle_0 \quad (1.20)$$

Here,  $\Delta U = U_1 - U_0$  represents the potential energy difference of the initial and final states, T temperature, k Boltzmann constant, and  $\Delta A$  Helmholtz free energy (NVT ensemble). When this equation is revised again using the NPT ensemble, it should include enthalpy instead of potential energy and Gibbs free energy terms instead of Helmholtz [77]. However, since the PV term is constant in the initial and final states, it falls out of the equations and the general formula for FEP is given as follows:

$$\Delta G(0 \rightarrow 1) = -kT \ln \langle e^{-\Delta U/kT} \rangle_0 \quad (1.21)$$

This equation gives the transition from the initial state to the final state, if the energy of the system is defined as a function of  $\lambda$  with small intermediate steps  $\lambda$ ;

$$\Delta G = \sum_{\lambda_i=0}^1 \Delta G_{\lambda_i \rightarrow \lambda_{i+1}} = -kT \sum_{\lambda_i=0}^1 \ln \langle e^{-\Delta U_{\lambda_i \rightarrow \lambda_{i+1}}/kT} \rangle_{\lambda_i} \quad (1.22)$$

Here,  $\lambda_i$  ligand is the environment (environment=ligand, solvent) interaction energy, and the free energy changes in each step are obtained by changing it from 0 to 1 in small steps. If the ligand is defined in such a way that it gradually separates from the surrounding complex (decoupling)  $\lambda_i = 0$  there is one hundred percent interaction;  $\lambda_i = 1$  also refers to situations where there is no interaction. Each of these steps is defined as the  $\lambda$  window and introduces an additional MD simulation into the calculation. The important thing to note here is that for each  $\lambda$  value, the simulated configurations are averaged with the energy difference in an upper window (neighborhood). Therefore, the energy differences are unidirectional and always belong to the lower  $\lambda$  configurations [78].

If the derivative of the free energy is taken against  $\lambda$ , the following general formula of Thermodynamic Integration (TI) is obtained as:

$$\begin{aligned} \Delta A(0 \rightarrow 1) &= \int_0^1 \frac{\partial A}{\partial \lambda} d\lambda = \\ &= - \int_0^1 \frac{kT}{Z} \frac{\partial Z}{\partial \lambda} d\lambda \\ &= \int_0^1 \frac{kT}{Z} \sum_{\lambda_i} \frac{1}{kT} \cdot e^{-U(\lambda_i)/kT} \cdot \frac{\partial U(\lambda_i)}{\partial \lambda} \cdot d\lambda \end{aligned} \quad (1.23)$$

$$\Delta A(0 \rightarrow 1) = \int_0^1 \langle \frac{\partial U(\lambda_i)}{\partial \lambda} \rangle_{\lambda} d\lambda \quad (1.24)$$

The TI method, which can give more accurate results than FEP in cases where each lambda step is quite small and equal, can be almost useless, especially at different step sizes [79].

In the FEP method, the simulated configurations for each  $\lambda$  value are averaged with the energy difference in an upper window (neighborhood) and the system's transition from the initial state to the final state is calculated unidirectionally, while

Bennett actually revealed that the two-way configurations of the system should be used and obtained the following BAR general formula:

$$\Delta A(0 \rightarrow 1) = \frac{1}{\beta} \ln \left[ \frac{\sum_1 f(U_0 - U_1 + C)}{\sum_0 f(U_1 - U_0 - C)} \right] + C - \frac{1}{\beta} \ln \frac{n_1}{n_0} \quad (1.25)$$

Here,  $\beta=1/(k_B T)$  represents the reciprocal temperature,  $n_0$  and  $n_1$  represents the sampled configuration number, and the term  $f()$  represents the Fermi function and  $f(U_1-U_0+C)=1/(1+e^{(\beta)(U_1-U_0+C)})$ . The  $C$  value that satisfies this equation is solved numerically. Thus, the free energy change is calculated with  $C=\Delta A+1/\beta \ln n_1/n_0$ . The important thing here is that the logarithmic term is equal to zero when the configuration numbers in the equation are sampled equally. This is possible if the energy difference calculated in the configurations in the lower window is equal to the energy difference calculated in the configurations in the upper window. Therefore, this equality creates a self-consistent field [75], [79].

Another alchemical one is Jarzynski's method [76] and in contrast to previously mentioned methods which are applicable in equilibrium state of system, it is generally used to define the systems in non-equilibrium state. Also, the amount of work done will not be reversible and, as a result, won't be equal to the free energy if a physical or alchemical process occurs in a finite length of time [80]. The nonequilibrium work  $W$  for this transition, not the equilibrium free energy, can subsequently be determined from equation (1.26)  $\Delta G$ . Jarzynski discovered that the average of the nonequilibrium trajectory starting from an equilibrium ensemble may be used to represent the free energy of the transformation:

$$\Delta G = \beta^{-1} \ln \langle e^{-\beta W} \rangle_0 \quad (1.26)$$

Equation (1.26) and EXP become interchangeable if the switching occurs instantly since the immediate work is just a potential energy shift. Numerous studies have compared single-step perturbations to nonequilibrium routes. However, it seems that in most cases, equilibrium simulations are about the same or slightly more efficient than free energies estimated from ensembles of nonequilibrium simulations, both in theory and in practice [81]. It is also uncertain if future ligand-binding predictions employing free energy calculations based on Jarzynski's relationship will play a

significant part [82]. Recent years have seen a lot of research on this subject, in part because simulations and nonequilibrium experiments may both be treated effectively using this formalism [83]. The schematic representation of all these methods is given in Figure 1.2.

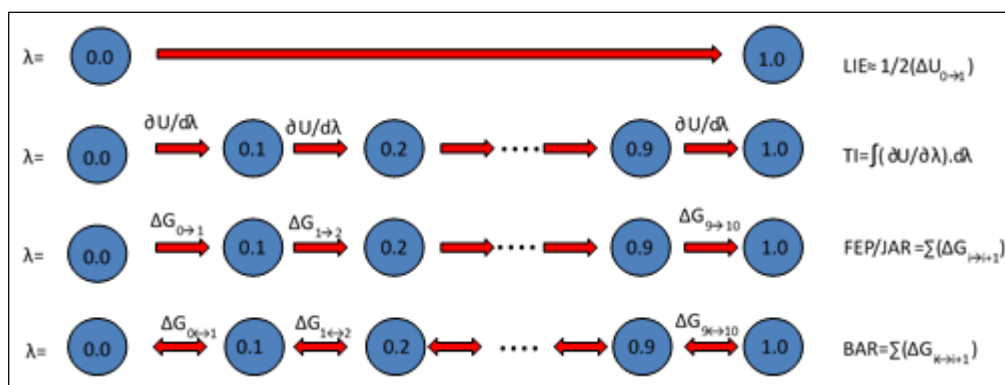


Figure 1.2: Mathematical definition of different MD methods.

In the absence of covalent interaction, these interactions are usually expressed with 2 different molecular mechanical energies, electrostatic (coulombic) and van der Waals, and  $\lambda_i$  is reduced from 0 to each 2 interactions (first electrostatic and then van der Waals) in small steps separately. It is set to 1. To obtain a reversible system, it is desirable to have as many  $\lambda$  windows as possible [84]. Therefore, an optimum choice is made between the number of windows and the computational cost. In the ideal case, at least 10 for electrostatic separation ( $\lambda_{\text{coul}}=0.0, 0.1...1.0$ ) and at least 10 for van der Waals separation ( $\lambda_{\text{vdW}}=0.0, 0.1...1.0$ ) total double decoupling. With a 20  $\lambda$  window, the dissolution free energy of a ligand can be obtained [85]. For the binding energy of the protein ligand system, 2 times as much MD simulation is needed as the separation of the ligand from the protein and the separation of the ligand from the solvent. Moreover, all calculations are made assuming a constant force field and partial load at the MM level. For this reason, increasing the accuracy and speed of these calculations still constitutes the center of scientific studies as an important problem to be overcome [57], [86].

There are many different approaches in the literature for the development of these computational methods. Almost all of these approaches have focused on increasing the accuracy and reducing the cost of the computer, and there is not yet an ideal method on which a consensus has been established.

## 1.4. Quantum Mechanics/Molecular Mechanics Methods (QM/MM)

As they incorporate the advantages and the strength of both the QM and MM methods, the Quantum Mechanics/Molecular Mechanics (QM/MM) approach is a well-known and significant system for simulating drug-like molecules-solvent interaction, proteins-solvent, or protein-ligand interactions. In a paper they published in 1976, Warshel and Levitt introduced the QM/MM approach [87]. The fundamental concept behind this methodology is the division of the system under discussion into two regions: a small portion, or the QM responding region of the system, and the remainder of the protein or drug patch, which is subjected to an MM force field (MM region) [88]; this idea is depicted in Figure 1.3. The interaction between the QM and MM zones is a key problem. The three groups of relations between titles (i) in the QM region, (ii) in the MM region, and (iii) in the QM and MM regions are generally reckoned. The more complex relationships between the two systems are accounted for by using a variety of methods, as opposed to the relations inside the QM and MM regions, which are dealt fluently at the QM and MM positions, independently [89]. The two fundamental strategies utilized in QM/MM computations are the subtractive coupling technique and the additive coupling strategy. The system's total energy (QM/MM energy) is calculated using a subtractive approach (the ONION method in Gaussian). After the system's total energy (containing the MM and QM regions) has been evaluated using the MM methodology, the isolated QM energy is then added [90], [91]. The final phase, after the MM energy of the QM subsystem has been computed and subtracted, adds the interaction inside the QM subsystem twice in order to correct the results. Eq. (1.27) and Fig. 1.3 both offer explanations [92].

$$V_{\text{QM/MM}} = V_{\text{MM(MM + QM)}} + V_{\text{QM(QM)}} - V_{\text{MM(QM)}} \quad (1.27)$$

The procedures for quantum chemistry and molecular mechanics are separate despite this coupling method, therefore the QM subsystem requires a force field with sufficient flexibility. According to Eq. 1.28, the cumulative technique is used to calculate the implicit energy of the entire system by summing the implicit energies of the MM and QM subsystems as well as the terms defining the coupling between the QM and MM systems (Eq. 1.28) [93].

$$V_{QM/MM} = V_{MM(MM)} + V_{QM(QM)} + V_{QM-MM(QM+MM)} \quad (1.28)$$

As a result, it should be highlighted that at the force field position  $V_{MM(MM)}$ , only relationships inside the MM region are defined; the relationships between the two subsystems are dealt with directly. Contrary to the subtractive coupling methodology,  $V_{QMMM(QM+MM)}$  can be defined utilizing complex methods [94], [95].

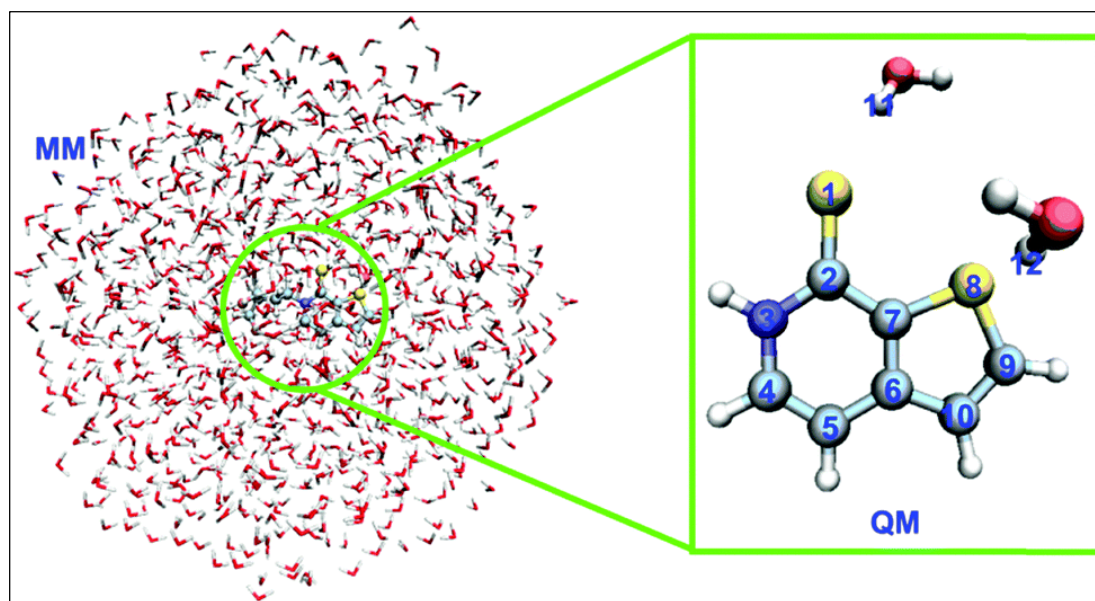


Figure 1.3: The view of QM/MM region in a solvated ligand system.

## 1.5. Machine Learning-Based (ML) Methods

Machine learning (ML), a phrase that has grown in popularity recently and is categorized as a branch of artificial intelligence (AI), analyses algorithms and enhances their performance using actual data already in existence that has been obtained via reliable research, or "learning from experience" like the neural system in the human brain [96], [97]. The systematic discovery and application of regularity (nonrandomness) in data for analysis or prediction is one of its key concerns. Robotics, brain-computer interfaces, recommender systems, chemistry, and most notably cheminformatics, are just a few of the fields where it has been successfully used. Support vector machines and artificial neural networks are two widely used and well-known techniques [98].

Machine learning (ML) methods are now more in vogue in chemistry thanks to the success of related fields including quantum mechanical property predicting, robotic chemical manufacturing, medicine, and material estimation [99]. Understanding the energetics and mechanics of large molecules in biological and chemical systems is a key component in the study of the natural sciences, especially Chemistry and Biology [100]. On the other hand, because to the impossibly high processing costs, theoretical analyses of such massive and complicated systems are typically constrained to the use of approximative methods, sacrificing precision for a speedup in the computational calculations. One of the most difficult issues in contemporary theoretical and computational chemistry is designing and implementing approximations that accelerate ab initio techniques and procedures without losing precision. The majority of common ways involve fractionalizing the physical system into small grids [101], linear scaling [102], semi-empirical (SE) methodologies [103], or the creation of empirical potentials that have been described for the replication of experimental or precise ab initio data [104]. The ultimate objective of this field of research is to provide quantum mechanical (QM) reference computation estimations (prognostications) that are as accurate while maintaining the computing cost close to that of classical force fields. With rigor far exceeding conventional physics-based methods, ML-based property predictors have been used to make accurate predictions of molecular atomization powers, forces, implicit energy shells, tiny partial charges, dipoles, and quadrupoles [99], [105]–[108]. Similar investigations that have been conducted have demonstrated that ML-based models can be simultaneously trained on a number of functionalities. ML property predictors have the potential to enhance the computational design of drugs and materials by bridging the speed versus accuracy gap between quantum physics and classical techniques [109].

Many of the computationally valuable integrals are substituted with empirically established parameters in SE approaches. As a result, there is a noticeable speed increase. However, due to the evaluated approximations, the delicacy is also significantly diminished when compared to high position originally styles. Additionally, the computational cost of SE methods remains significantly higher than that of classical force fields (FFs), which might restrict the scale of the system that can be examined [110].

By casting factors of the clicked, angular, dihedral, and non-bonded benefits fitted to an easy logical form, classical force fields or empirical interatomic capabilities

(EPs) further simplify the outline of interatomic connections. With much lower processing costs, EPs will be used in large-scale atomistic simulations. In order to improve the statistical slice and delicateness of molecular dynamics (MD) and Monte Carlo (MC) simulations, more precise EPs have long been desired. However, EPs are typically only reliable close to equilibrium. Condensed matter and polymer investigation, as well as the formulation of medications, make substantial use of these typically nonreactive empirical capabilities [111], [112]. As each FF and EP operation differs in delicateness depending on the system, each one must be carefully considered. In fact, it is frequently required to perform marks to determine the best FF combination for the situation at hand. Unfortunately, there are no precise or accurate methods for calculating the transferability of EPs [113], [114].

In order to anticipate an atomistic potential energy surface for a range of procedures, such as geometry optimization or molecular dynamics simulations, numerous machine learning (ML) techniques have been created. There have been reported ML potentials for biological [114], [115] and material [116], [117] systems. Dedicated ML potentials and transferrable ML potentials are the two groups of techniques that have been suggested for learning the potential energy of organic compounds. The potential energy surface of a single system or a small class of systems can be described by specialized ML potentials with the least amount of QM reference data available. For molecules with a reasonable number of degrees of freedom, these models frequently enable largely realistic powers and forces. Prior to every operation, QM computations are required for specialized ML capabilities. As a result, the effective computational scaling of dedicated models is that of the underlying QM system, making it impossible to utilize them to operate on massive natural systems like proteins and enormous pharmaceutical molecules. The quantity of QM data required to completely cover configurational space rises dramatically as the number of degrees of freedom within a system increases. This approach might be impractical for large enough systems. Similar models cannot be used for top outturn investigations on small molecule database because to the increased time required for QM data generation [118].

A reliable method to build various types of transferable [119], [120] and non-transferable [121] atomistic potentials enhanced using regression methods is emerging as machine learning (ML). The prediction of reaction pathways [122], QM excited state energies, formation energies, atomic forces, nuclear resonance chemical shifts,

and aiding in the discovery of innovative materials all successfully apply machine learning (ML) techniques to Chemistry [123], [124]. With a speed of up to 5 orders of magnitude, ML potentials have showed promise in accurately forecasting molecule energies using QM. Finding a precise molecular representation that supports and enhances learning inside the selected ML approach is the key to transferrable methods. For such representations to ensure energy conservation and be useful for ML models, Behler [125] outlined three requirements: they must be rotationally and translationally invariant, the exchange of two identical atoms must produce the same result, and given a set of atomic positions and kinds, the representation must describe a molecule's conformation in a highly distinctive manner. Many of these representations have been developed, but it has not yet been possible to extend them with chemical precision to complicated chemical environments, including all degrees of freedom for arbitrary organic molecules [126], [127].

With the perfect of avoiding direct QM computations prior to an operation, transferable ML capabilities attempt to directly pretend a full class of molecules. Additionally, after the model has been trained, a large number of systems can be applied to it without requiring additional QM computations. In the majority of cases, this results in an instantaneous scaling system. To create a training data set for current transferrable styles for organic molecules, tiny molecules are used to create an extremely huge and diverse data set of molecular conformations. To ensure that training on small systems results in extensibility to bigger systems, model position is used. A nearsightedness concept is acknowledged in much of chemistry. As a result, Behler and Parinello's invention of computing infinitesimal powers as a function of the atom local environment allows for both direct system size scaling and extendable capabilities [126]. The idea behind transferable models is to give up enough data diversity during the literacy process to force the model to be trained only once and to be told the initial infinitesimal drugs at play, as opposed to only describing the implicit energy face in a very small subset of infinitesimal configuration space. Generalizing features discovered from many systems is possible with transferable models. The amount of information required for each system while training is reduced since interaction types that are learned from some molecules within the training set do not have to be tested for all molecules. Despite being far more generic than specialized ML skills, transmittable ML capabilities tend to be slightly less precise, or better in near-ambient conditions, on the scale of chemical accuracy (1 kcal/mol error).

However, the position of delicacy attained by transferrable ML skills is more than enough to offer quantitative outcomes for a variety of practical operations [128]–[130].

In summary, for reduction in the costly QM calculations, due to their effective algorithms used to answer scientific questions about, but not limited to, chemical reactions, potential energy surfaces, forces, atomization energies, solvation free energies, and protein ligand interaction, machine learning (ML) techniques have attracted a lot of interest over the past ten years. One of the most promising aspects of the ML techniques is that the trained models can be applied to new systems (transferrable). One of the latest studies showed that ML methods such as “Accurate Neural network engine for Molecular Energies” (ANAKIN-ME or shortly ANI) can be used to predict QM energy at a given non equilibrium geometry for small organic compounds with millions times faster than conventional QM methods [118]. Additionally, several models using active learning such as ANI-1 [114], ANI-1x and ANI-1cxx [131] have been trained to calculate QM energies of small organic compounds containing C, H, O, N, atoms and the most recent version of these potentials, and ANI-2x [118] extended these to include F, Cl and S atoms at the wb97x/6-31G(d) level. Although they are very successful in predicting the single point energies (SPE) of the small compounds in vacuum, ANI’s success for calculation in the presence of explicit solvent (water) molecules need to be investigated.

## **1.6. Human Norovirus (HuNoV)**

Viruses can be classified based on their genetic content (RNA (ribonucleic acid) and DNA (deoxyribonucleic acid)), and containing envelope (bilayer membrane) or only capsid. All obligate viruses are intracellular microbes that dependent on the cellular machineries of the host cell for all stages of life cycle. Infection of cells is a multi-step process that usually is split into i) virus attachment to receptors on the outer of cell membrane via their surface protein, e.g human norovirus linking via P domain of its viral protein to some special unit on the cell surface, ii) fusion with the plasma membrane( some enclosed viruses) and membrane penetration from extracellular compartments to inside the cell, iii) uncoating of virus capsid/envelope, iv) replication, iv) generation of new viral particles, and v) exit outside of the target cell [132].

Human norovirus (HuNoV) remains the leading reason behind foodborne outbreaks and responsible for 9.4 million foodborne sicknesses, one hundred fifty

deaths and over \$ two billion in health-care costs annually in the US. The excess of these foodborne illnesses are caused by HuNoV [133]. HuNoV impose a significant international burden despite the lack of authorized antivirals or vaccines against it, with direct (healthcare) and indirect (loss of productivity) expenditures estimated at \$64 billion [134].

Human norovirus is transmitted via the fecal-oral route, together with direct person to person, consumption of spoiled food, contaminated water or contact with vomites. Infections with HuNoV cause canal symptoms (symptom, puking, abdomen pain and nausea) that are usually self-limiting in healthy grown-ups however are going to be severe for infants, elderly, and immune-compromised individuals . Noroviruses are also classified into a minimum of 7 genogroups (GI – GVII) and subdivided into various genotypes [135], [136]. Infections with HuNoV are caused by each GI and GII strains. Among them, the GII.4 HuNoVs as a dominant strains has a superiority kind causing the majority (60–90) of stomach flu outbreaks [137]. The virus is 27-40 nm in size, non-enveloped, single stranded ribonucleic acid (RNA) virus and belongs to the Caliciviridae family [138].

In humans, HuNoV binds to HBGAs that are necessary for a productive norovirus infection. Human HBGAs are complicated glycans expressed on the surface of red blood cells and tissue layer shells of secretory organs or is present as free antigens among the natural fluids of secretory organs such as slaver, milk and viscus contents [139]. The great number of GI HuNoVs adheres to H- and A-type HBGA and Lewis antigens, whereas GII HuNoVs parade an additional totally different HBGA binding pattern together with each the below and B-kind HBGAs [140]. HBGA molecules on the cell surface obtainable for virus fusion are glycoconjugates . The glycan portion usually extends farthest into the outer space of cell and virus binding cannot occur solely by interaction with membrane distal sugar; however conjointly to the membrane proximal macromolecule core involves in virus host cell interaction [141]. Binding to glycans is in varied cases mediated by electrostatic forces between viral proteins and charged sialic acid-containing oligosaccharides or parts of glycosaminoglycans. In some cases, the identical sugar motif could be present in glycolipids and glycoproteins like oligopeptidosaccharids, histo-group blood antigens (HBGAs) and so on [142], [143].

These attachments are essentials for virus infection in the most of the time. Thus, developing therapeutics that targets and deteriorate the interaction between virus and

host cell is one the main approach for developing antiviral drug against viral diseases [141].

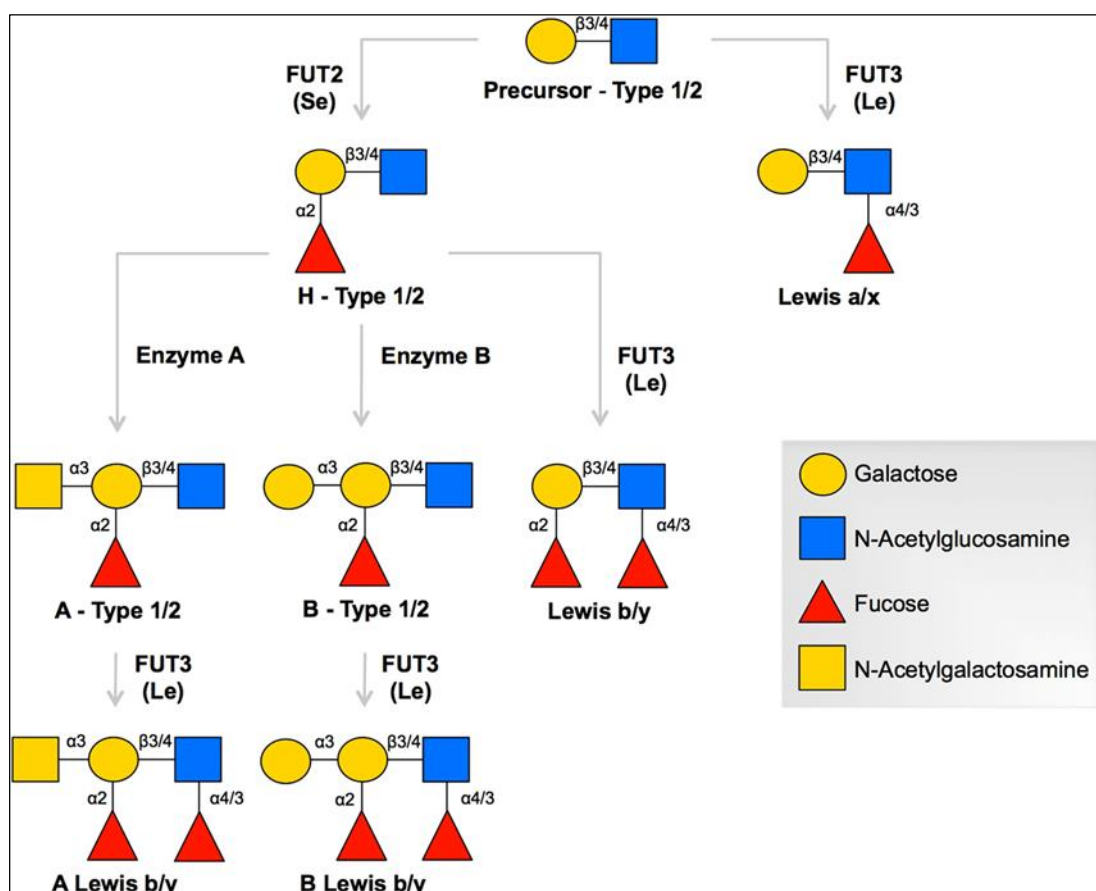


Figure 1.4: Different Types of HBGA based on the type of sugar unit and connection.

Virus binding to the cell surface shown to be essential for the virus life cycles in general and specifically for norovirus infection. Therefore, this interaction gains great attention for drug development against viral diseases. Viruses can bind to the host cell with completely different modes such as protein-based receptors, carbohydrates, and lipids. Some of these attachment factors concentrate virus on the cell surface, whereas others assist virus entry into cells [144].

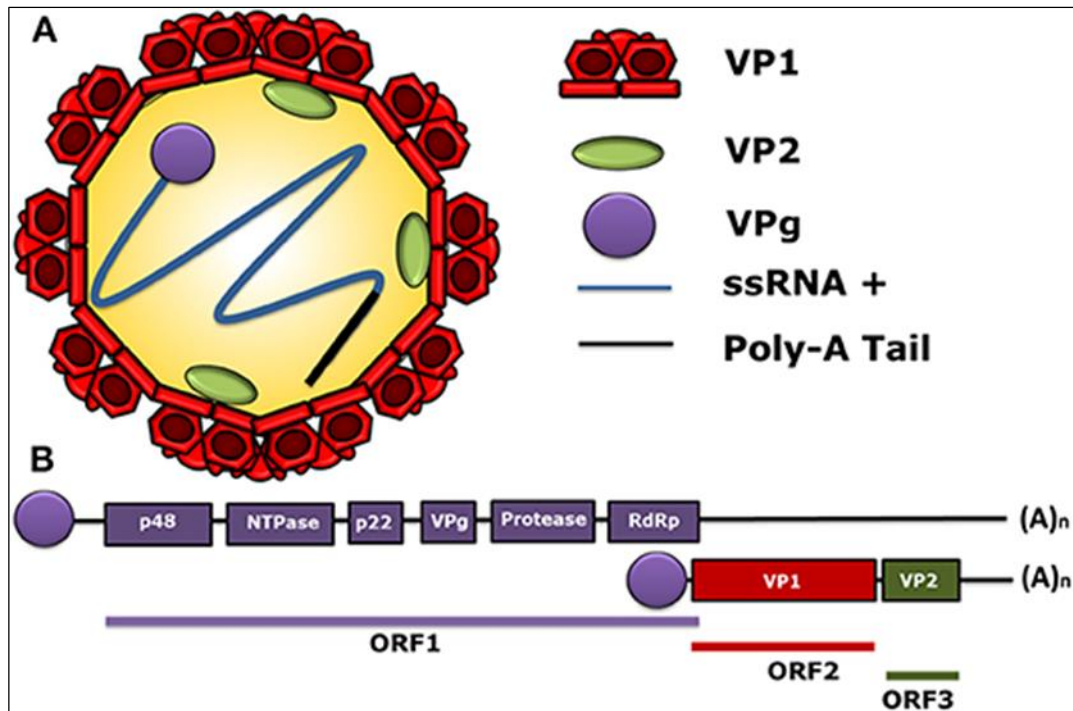


Figure 1.5: A general genome view of VP1.

The HuNoV genetic material comprised of 7–8 KB RNA and organized into 3 open reading frames (ORF). The ORF1 encodes an oversized polyprotein that is adhered into seven non-structural proteins. VP1 a vital structural molecules providing interaction with HBGAs is designated by ORF2 frame and lastly ORF3 encodes the minor structural supermolecule VP2. Attachment of the virus to human HBGAs mediated by the VP1 capsid supermolecule [145], [146]. Specifically, the VP1 split into 2 portions, the shell (S) and so the surface interacting sphere (P). The P portion is divided further into P1 (residues 222 – 274 and 418 – 539), that forms a stalk to project the P sphere down from the shell and P2 (residues 275 – 417) that is the foremost face exposed part of the virus. The P2 subdomain contains the HBGA- binding patches (HBS) and some negating protein epitopes [140], [147].

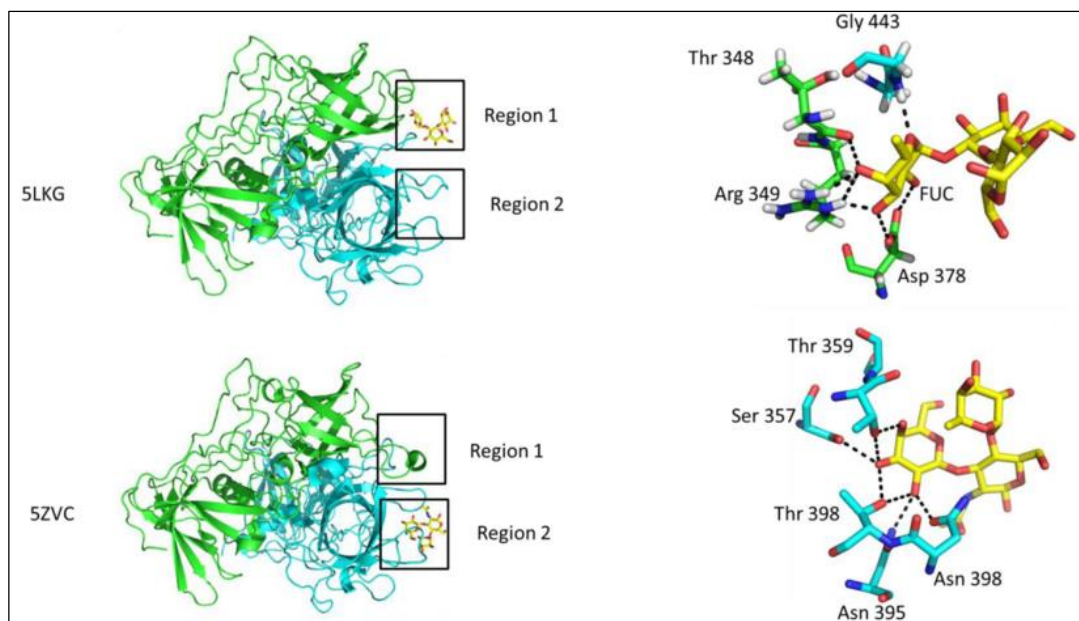


Figure 1.6: HBGA interaction mechanism and residues.

A mixture of two to eight neutral monosaccharides with various kinds of glycosidic linkages results in HBGAs. The HBGAs bind to the VP1 on the surface of the cell membrane, either as free glycans or as glycolipids, before the cell is infected. It is still unclear how the system works and why some glycans are selected over others. Therefore, an essential first step in creating new illness treatments is to comprehend how the sugar moieties and VP1 protein interact [148]. Fucose, glucose, and galactose moieties are frequently present in natural antigens. However, other sugar units are also engaged in the interaction between the HBGAs and VP1, which is mostly through the fucose moiety [149].

A shell (S) domain, which makes up the icosahedral shell's core in each VP1, and a protruding (P) domain, which creates arches that protrude from the shell, are the two main domains of each VP1. A 10-residue hinge connects these two domains. The P domain is thought to be in charge of host contacts and immune responses since it is situated on the external surface of the viral particle and has the most changeable sequence. The P domain also engages in dimeric interactions in the formed viral particle, which are thought to stabilize the capsid. Additionally, earlier research hypothesized that VP1 might be proteolytically converted into the soluble P polypeptide, a 30-kDa protein that may be involved in norovirus proliferation, host immune response, and disease. Subdomains P1 and P2, which represent the leg and head of the arch-like P dimer, respectively, can be created from the P domain [150].

The majority of the crystal structures revealed that the amino acids 343–347 and 371–376 of the HBGAs are what allow them to bind to the VP1 of GII (region 1) [151]. On the other hand, there have been reports that, after a few number of changes during the course of the virus's evolution, the virus can totally generate a separate binding site in the VP1 protein. Over 25 Å separate this location (region 2) from region 1, which is located between 352 and 357. Each region's position in VP1 is shown in Figure 1.6. According to certain research, only Lewis<sup>a</sup> antigen binds in region 2, while every other antigen adheres to region 1 [133], [151].

## 2. EXPERIMENT

### 2.1. Theory

#### 2.1.1. Average interaction energy (“AVG”)

A coupling parameter,  $\lambda$  [152], can be used to specify the incorporation of the ligand into a solvent (solvation free energy) or receptor (binding free energy) environment. The system's potential energy would be the following at each  $\lambda$ -state:

$$U_\lambda = (1 - \lambda)U_0 + \lambda \cdot U_1 \quad > U_\lambda = U_0 + \lambda \cdot \Delta U \quad (2.1)$$

where  $\Delta U$  represents the energy of interaction between the ligand and surroundings. The reference potential energy, or  $U_0$ , covers all energy terms other than the interaction between ligands and their environments. The partitioning feature is

$$Z(\lambda) = e^{-F(\lambda)/kT} = \int e^{-\lambda \Delta U/kT} e^{-U_0/kT} \quad (2.2)$$

where  $F$ , which depends on the  $\lambda$ -state, is the free energy. Rearranging the previous equation,

$$F(\lambda) = \frac{-1}{\beta} \ln[Z(\lambda)] \text{ where } \beta = \frac{1}{kT} \quad (2.3)$$

Thus,

$$\frac{\partial F}{\partial \lambda} = \langle \Delta U \rangle_\lambda \quad (2.4)$$

Only when the ligand-environment interaction is proportional to the coupling term  $\lambda$  is the aforementioned equation accurate. The outcome is [153]:

$$\Delta F = \alpha \cdot \langle \Delta U \rangle_0 + \beta \quad (2.5)$$

Here,  $\langle \Delta U \rangle_0$ , denote the ensemble average of in the ligand-surrounding fully interacting MD simulation ( $\lambda=0$ , fully interacting ligand-surroundings;  $\lambda=1$ : no

interaction). This formula has been extended to separate electrostatic and van der Waals interactions as [154]:

$$\Delta F = \alpha \cdot \langle \Delta U_{vdw} \rangle_0 + \beta \langle \Delta U_{ele} \rangle_0 + \gamma \quad (2.6)$$

where the empirical coefficients  $\alpha$ ,  $\beta$  and  $\gamma$  are generated by fitting to experimental values.

### 2.1.2. Exponential average with numerical correction (“nc-EXP”)

High fluctuations in FEP energies may bring considerable source of error to the free energy values, and thus may require a correction. Berg [155] has proposed a general solution for this problem by deriving the following equation;

$$\begin{aligned} \Delta F_{0 \rightarrow 1} &= \frac{-1}{\beta} \ln \langle e^{-\beta \Delta U_{0 \rightarrow 1}} \rangle_0 = -\frac{1}{\beta} \ln \left( \frac{1}{N} \sum_{i=1}^N e^{-\beta \Delta U_i} \right) \\ &= -\frac{1}{\beta} \ln \left( \frac{1}{N} \sum_{i=1}^N e^{-\beta (\overline{\Delta U} + \delta U_i)} \right) \end{aligned} \quad (2.7)$$

$$\Delta F = \overline{\Delta U} - \frac{1}{\beta} \ln \left( \frac{1}{N} \sum_{i=1}^N e^{-\beta \delta U_i} \right) \quad (2.8)$$

Here,  $\delta U_i = \Delta U_i - \overline{\Delta U}$  and  $\overline{\Delta U} = \frac{1}{N} \sum_{i=1}^N \Delta U_i$  are the error and average energy differences, respectively.

### 2.1.3. Exponential average method (“EXP”)

When calculating the free energy differences between two states, the free energy perturbation (FEP) method, which is based on Zwanzig's equation, is frequently used [156][157]. According to this technique, the free energy difference between states 0 and 1 is provided by;

$$\Delta F_{0 \rightarrow 1} = \frac{-1}{\beta} \ln \langle e^{-\beta \Delta U_{0 \rightarrow 1}} \rangle_0 \quad (2.9)$$

Here, the method known as in script as EXP in oppose to exponential method.

#### 2.1.4. Cumulant expansion (“cu-EXP”)

Alternatively, Konig et al [158] considered the random occurrence of low energy values in the trajectory and suggested the use of more efficient cumulant expansion rather than FEP directly by the following formula;

$$\Delta F = \sum_{k=1}^{\infty} C_k \frac{(-\beta)^{k-1}}{k!} = \overline{\Delta U} - \frac{\beta \sigma_{\Delta U_i}^2}{2} + \sum_{k=3}^{\infty} C_k \frac{(-\beta)^{k-1}}{k!} \quad (2.10)$$

If the distribution of  $\Delta U_i$  is Gaussian, the terms greater than the second order (i.e.  $k \geq 3$ ) will not contribute to the free energy. So the equation becomes;

$$\Delta F = \overline{\Delta U} - \frac{\beta \sigma_{\Delta U_i}^2}{2} \quad (2.11)$$

For equations explained above, the MM energies will take the place of the ANI potential, which reflects the QM energies. In order to do that, three separate energies at each MD frame were estimated, with ELS standing for the energy of the ligand complexed with the environment (solvent), L standing for the energy of the bare ligand that is extracted from the MD frame, and S standing for the energy of the environment that is also extracted from the same MD frame. And combining these three energy terms by  $\Delta U = U_{LS} - (U_L + U_S)$ . Finally, the averaged ensemble for the free energy was calculated using only one-end state simulations. Free energy is predicted by these two methods using ANI potential energies on the MD simulation trajectories. We have combined these calculations into a custom-built script, freely available at <https://github.com/otayfuroglu/deepQM>.

## 2.2. Computational Methods

### 2.2.1. Test Set for MD Simulations

To validate, prepare and check our .mdp protocol input files, we selected a database known as FreeSolv Database which were prepared by Mobley et al [159]. In this database, there are small organic molecules and also their topology and gro files with .mdp protocol files to run on Gromacs software. For validation and control test, a reproducing step was performed. This test was carried out by using topoly and gro files of small organic molecules in database. To run MD simulation, the .mdp protocol that belongs to database was also utilized with small changes owing to difference between version of software that is named as Gromacs. Energy minimization, NVT and NPT (two steps) ensembles were done, respectively. Before running simulation on cluster, an automated script was prepared and used to perform reproducing test for once. The simulations were run on a cluster with 56 cores for 4 weeks. After obtaining results for simulations, BAR was performed by following protocol. Then, experimental values for small organic molecules in database were plotted against the values that were calculated by Mobley group, that calculated by us and finally the values published by group were plotted against our result..

### 2.2.2. Classical MD and Generating Input Files

We have selected compounds (28 small molecules) for the purpose of testing our own .mdp protocols for EM, NVT and NPT ensembles after reproducing test set and then it was applied to the molecules for benchmarking. For the benchmarking studies, the topology, and gro files of chosen molecules were obtained from FreeSolv database. The database includes the structure of ligand-solvent complex systems and parameters of all-atom model of Amber ff99SB-ILDN force field [160] and Generalized Amber Force Field (GAFF) [161]. The classical MD simulations were done with Gromacs2020 software package. To initiate running protocol, automated script was used to prepare molecules to MD and start simulation.

Energy minimization was done in our.mdp choices up to a force of  $100 \text{ kJ.mol}^{-1}.\text{nm}^{-1}$  using the Verlet cutoff approach. The Particle Mesh Ewald method (PME) (6th order interpolation) was employed with a cutoff length of 10 Å for long-range electrostatic and Van der Waals interactions (65). The frequency of the neighbor list

update was set at  $1 \text{ ps}^{-1}$ . Each system was converged to its minimum energy within 5000 cycle of Steepest Descent minimization algorithm.  $\lambda$  defines the gradual elimination of the coulombic and van der Waals interactions for BAR. The higher the number of steps  $\lambda$  (lambda), which is used in BAR calculations and for which we define van der Waals + coulombic interactions, the higher the accuracy of the calculation, but since it will add additional cost to the calculations, the optimum  $\lambda$  (coul) and the  $\lambda$  (vdw) numbers have been determined. Starting from  $\lambda=0$  “bound state” with 100% interactions, transitioning to  $\lambda=1$  “unbound state” where interactions are 0%, first electrostatic terms in 5 equal steps ( $\lambda_{\text{coul}} = 0, 0.2, \dots, 1$ ) and then van der Waals terms were decoupled in 10 steps ( $\lambda_{\text{vdw}} = 0, 0.1, \dots, 1$ ) in a total of 15 windows. All neighboring  $\lambda$  windows are used in the BAR calculation (Bennet Acceptance Ratio, BAR). Shortly, the step number is 15 for .mdp protocols, which for coulombic interactions 5 steps were adjusted and for short range interactions (Van der Waals), 10 steps were used.

After minimization, Langevin Dynamics was used to equilibrate each system in three phases. A 5 ns NVT ensemble made up the first step. The Particle Mesh Ewald technique (PME) [162] was employed with a cutoff length of  $10 \text{ \AA}$  for both long-range electrostatic and Van der Waals interactions (6th order interpolation). To replicate the experimental values, simulations were run at 298.15 Kelvin. The following steps used NPT ensembles in which the systems were Parrinello-Rahman isotropic pressure coupled for 1 ns to a reference pressure of 1 atm after being equilibrated to the 1 atm pressure by Berendsen for 200 ps. An MD simulation lasting 5 ns was performed once the systems had attained equilibrium. Trajectories were saved every 1000 steps with a total of 5000 frames for each trajectory when the time step was set to 2 fs. Then, the database's experimental values were plotted against the BAR results for the NVT ensemble and the NPT ensemble.

In addition to aforementioned information, the next step is to perform MD simulations with minimized MD protocol and self-parametrized small organic molecules. By performing self mdp protocol minimization, we already carried out the initial step of this part successfully. For the second part that was defined as running self mdp protocols with self-parametrized molecules. For the purpose of obtaining them, sdf files of molecules in database were saved firstly and then molecules were prepared to MD by using our automated bash script in Ubuntu 20.04. In script, after converting file type there are two options which process can be continued optimizing

molecules and generating .esp files or generating .esp files without optimizing files. At first glance, sdf files of molecules were converted to xyz format type to generate g16 input type. With respect to g16 generation input, small molecules were optimized at the B3LYP/6-31G\* and ESP [163], [164] charges of molecules were produced at the HF/6-31G\* level using Gaussian 16 software. RESP charges and GAFF force field [165] atom types were generated using Antechamber, AmberTools v20. Finally, amber parameters were converted to Gromacs input files using amb2gro\_top\_gro.py script. Using these gromacs input files, molecules were kept continue to prepare for MD simulation like generating a box and solvating ligand. MD simulation was run by starting with energy minimization. In energy minimization, the aforementioned process was followed to perform. Taking mdp protocol test result into consideration, we decided to perform only NVT ensemble to apply molecules in database. The same protocol that was defined above was applied for NVT step; however, trajectories were saved in every 10000 steps with a total of 250 frames for each molecule trajectory. The running was accomplished on a cluster that possess 56 cores for 2 weeks.

### **2.2.3. Different Implicit Water Methods in G16**

It is hard to make calculation of solvation free energy of small molecules in an explicit water environment at the level of QM because of high degree of freedom in QM calculations. Thus, implicit solvent models were developed to calculate solvation free energy of small molecules rapidly and modeling the solvent applying as a continuum, in oppose to individual molecules, enable ab initio computation feasible for SFE calculation. Based on the goal, there are several implicit solvent models like CPCM [28], [166], IFPCM [29], [167], and SMD [168] that are implemented into G16. Instead of using molecules in explicit water medium to calculate SFE, the other implemented method into G16 known as implicit water methods can be used to get SFE of small molecules. With regard to this part of study, to start with, conversion of sdf files of molecules to xyz type were performed by the mean of Obabel and g16 input files were arranged by using g16\_generate.sh script developed to prepare multiple file as input for G16. Then, the addition of required running parameters for g16 followed file convert step and the final files were saved as g16 input type (com) and run on gas phase at wb97x/6-31G\* with scf=(xqc,maxconventionalcycle= 200) to calculate their single point energy. With calculation SPE's of ligands on gas phase, SPE's of

molecules in implicit water medium were calculated at the same function and basis set with the addition of solvent type, setting as water in default setting, in g16 by defining as `scrf=SMD`. The running process was repeated for other implicit water methods-implemented in G16 by adjusting `scrf` to IEFPCM and CPCM (`scrf=IEFPCM` and `scrf=CPCM`). This protocol which was performed for making single point energy calculations of ligands with g16 on gas, SMD, IEFPCM and CPCM phase, respectively was run nearly for 14 hours on a cluster that consists of 56 cores. At the end of protocol, before the outcomes that were obtained were plotted against experimental values of ligands in database, the unit of single point energy which was calculated via g16 was converted from Hartree to kcal/mol to compare experimental values and produced results.

#### **2.2.4. ANI as Utilizable ML Potentials and Post-MD Processor**

In recent years, it can be stressed that machine learning term has been introduced rapidly and actively in diverse areas and its scope of application field has expanded day by day. Probably, one of the tops of these fields is Computational Chemistry and it has been used for making various (broad range of) calculations like single point energy of molecules in QM level, performing classical MD, post-MD analysis and calculations and so forth. In our present study, the reason why we introduced ML potentials is that calculation of solvation free energy of small molecules in QM level result in high cost and taking long time despite the fact that results of QM has relatively high accuracy and classical MD simulation as an explicit water method has several difficulties to perform, especially taking alchemical methods like FEP and BAR into account. Basically, ML methods which are used in different sort of calculations in Computational Chemistry has a superiority over QM and ab initio methods in classical MD in terms of cost and time. To come conclusion, during our search, the purpose to use ANI is to demonstrate that ANI can be utilized for solvation free energy calculation of small molecules relatively or accurately and rapidly and be an alternative method of SFE calculations since ANI as ML based method may have a potential to carry out SFE calculation in a considerable extent.

In the present study, our main goal is to illustrate that ANI as a ML based method which was trained with 20 thousand molecules to calculate single point energies at quantum mechanical (QM) DFT [48], [169] level can be used to produce solvation free

energies of small organic molecules in FreeSolv Database after generating poses from 0th step of NVT (nvt\_0.xtc) ensemble and using them for carrying out post-MD calculations.

To begin with, our first step is to witness that ANI trained by transferable learning method can accurately calculate single point energy of small molecules at QM level and its trained data encloses the selected molecules in FreeSolv Database. To test benchmarking for accuracy and effect of ANI in single point energy calculations, performing an experiment which single point energies of molecule in selected database were calculated at g16 with wb97x/6-31g\* scf=(xqc,MaxConventionalcycle=160), 24 GB RAM and 24 cores on local computer was required to compare single point energies of small molecules which was derived from g16 and ANI. Regarding running SPEs on g16, the calculation of SPEs of molecules selected from FreeSolv Database was made by the means of ANI. In a clear way, molecules without any solvent were run to print SPEs by defining only coordinate files via ML based ANI. Like G16, ANI calculation was also performed on local computer. In conclusion, results of both methods were plotted against each other to evaluate  $R^2$  value.

We validated the results of SPEs of small molecules in database obtained by ANI calculations with that obtained by g16. After validation test, we demonstrated that ANI is a successful and useful tool to calculate SPEs of small molecules and molecules in FreeSolv Database are in the range of compounds that were used for training and testing. Taking all results into consideration, the final step is to test our ultimate goal which ANI can be used as a promising tool to calculate relatively or directly solvation free energies of small molecules via snapshots taking from MD simulation. Depending on this purpose, snapshots from MD simulations were indispensable for us to run ANI and we already performed MD simulation to get information about SFEs by only running energy minimization and NVT ensemble to make BAR calculations. In general, ANI as ML based method is a tool to enable making calculations by using out of MD simulation, which generally called as post-MD method. To produce snapshots, the 0th step (md\_0.xtc) of MD was used by generating poses via gmx trjconv. The same protocol that was defined above was applied for NVT step; however, trajectories were saved in every 10000 steps with a total of 250 frames for each molecule trajectory. Energy values of each molecule containing complex (water + ligand), only ligand and only water molecules were calculated for each frame one by one with ANI. Then, the energy calculation of each frame of molecules was converted and written in

4 different implemented methods that were defined in theory part via python code (calcFromAniOutpust.py). Finally, the energy values obtained from 4 methods were plotted against experimental values of small molecules in database.

### **2.2.5. Conformer Generation and SPE of HBGAs**

There are two types of HBGAs known as secretor and nonsecretor type. For this purpose, The secretor type of HBGAs as A type 1 and 2, B type 1 and 2, H type 1 and 2 and precursor type 1 and 2 were selected to generate conformers. Firstly, each type was optimized by using b3lyp/6-311++g(d,p) with G16 and then each output file of G16 was converted to mol2 file type. For each HBGA type, mol2 file was open on Spartan to generate conformers; however, the maximum number conformers which can be generated by Spartan is 200. For the conformers, the main torsion points were used to generate conformers and torsion number was set to 12 for fucose unit. All conformers of precursor type 1 and 2 were generated by Spartan. For other types, 144 conformers were generated for each unit in Spartan and then they were combined with a python script. SPE of all conformers were calculated with ANI and conformers for each type were sorted and conformers which have energy until 100 kcal/mol were selected to continue. These selected conformers were aligned to GII.4 VP1 protein (as shown in Figure 2.7) with gmx confrms command and the SPE of system was calculated for each aligned conformer to protein. Finally, psi ( $\Phi$ ) and phi ( $\Psi$ ) angles were plotted against complex energy.

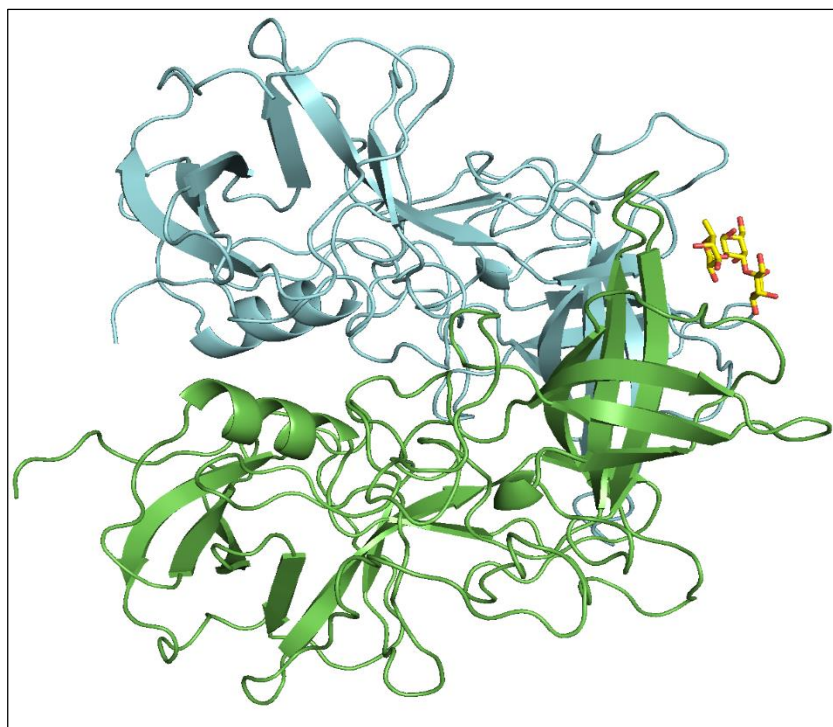


Figure 2.7: The view of Norovirus G.II-4 with its ligand (PDB ID:4X07).

Spartan WaveFunction, PyMol by Schrodinger LLC. and Igor Pro v.6.37 by WaveMetrics Inc. have been extensively used for visual and graphical representation.

### 3. RESULT and DISCUSSION

#### 3.1. Testing Database and Defining Input Parameters

A database known as FreeSolv Database which were prepared by Mobley et al [159] was required to control and test our purpose. Database consists of small organic molecules and for calculation in database to obtain SFEs vi an alchemical method called as BAR; small changes were done in running parameter files due to difference between software versions. Reproducing test was performed successfully by using topology and .mdp parameter files. Results from simulations were collected and brought together by the help of different gmx modules. and then, BAR was performed to find SFEs of molecules. There were some errors in calculations, but still retesting results were consistent with that calculated by Mobley group. In Figure 3.8, some points can be observed as out of range due to probably differences between version or errors on cluster. In short, as in shown figure, when calculated values on NPT level were plotted against that in database,  $R^2$  was calculated as 0.9998 and this test demonstrated that our reproducing test was carried out in a good way.

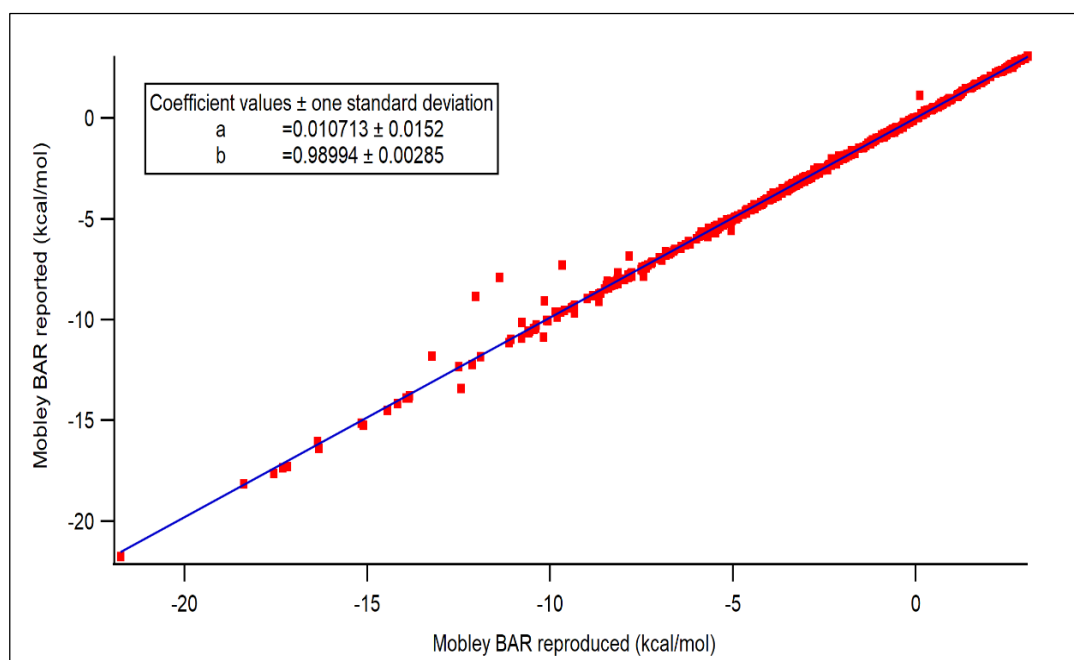


Figure 2.8: Testing data selected based on FreeSolv Database.

We have selected compounds (28 small molecules) for the purpose of testing our own .mdp protocols for EM, NVT and NPT ensembles after reproducing test set and

then it was applied to the molecules for benchmarking. For the benchmarking studies, the topology, and gro files of choosen molecules were obtained from FreeSolv database. As shown in Figure 3.9, there is no a significant diffirence between BAR results which were created from NVT ensemble outputs and NPT ensemble outputs. In general, based on the figure result and  $R^2$  value if time and  $\lambda$  number are sufficient, NVT ensemble might also be used to produce BAR result for calculation of solvation free energy.

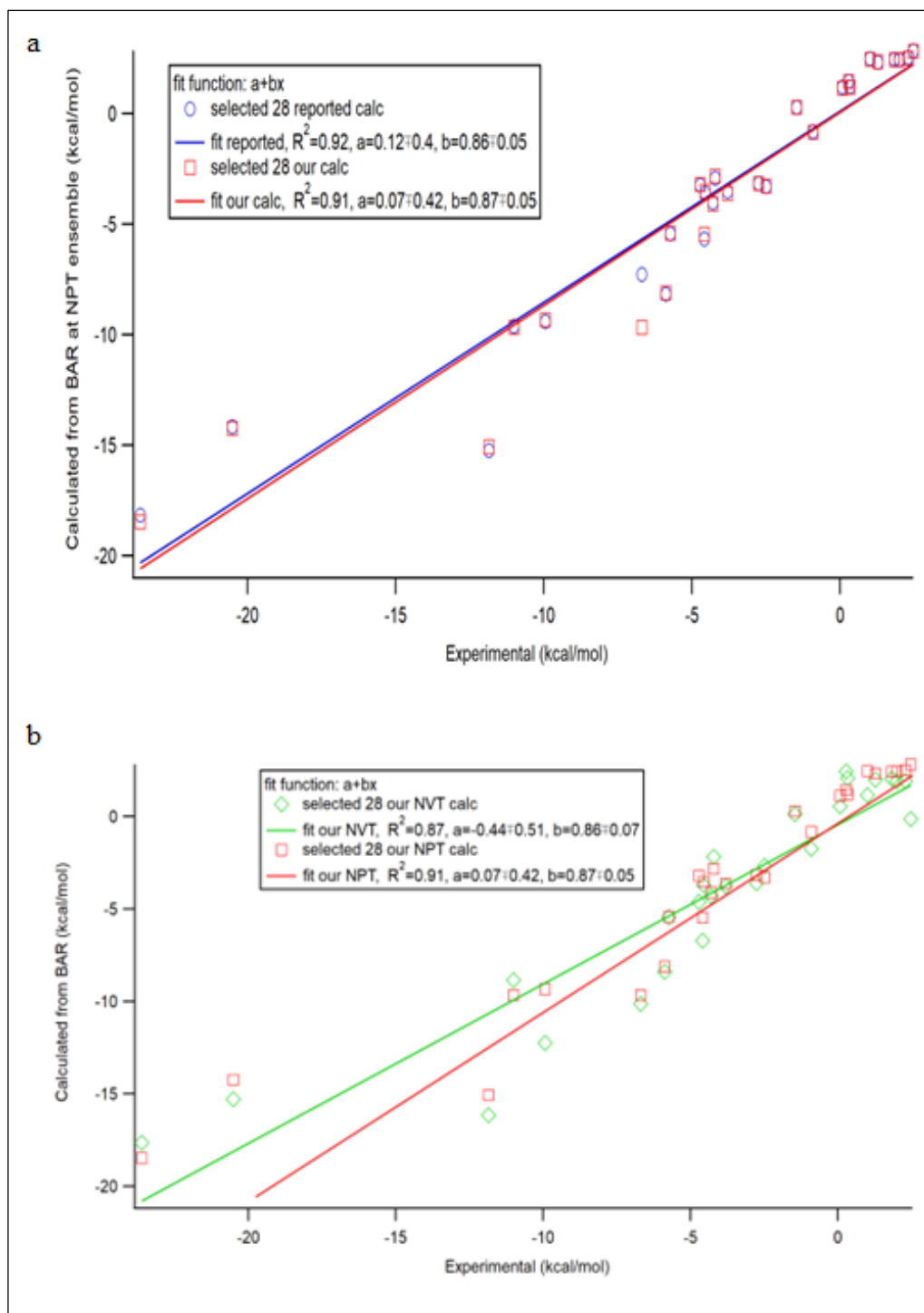


Figure 3.9: The BAR results for selected 28 molecules at NVT and NPT ensemble for 5 ns of simulation.

Our custom script (deepQM) to do free energy calculations from ANI applied as post MD simulations, is written in Python 3.9 and uses ASE libraries to calculate ANI single point energies of the given AB complex structure (e.g, A=ligand, L; B=solvent, S). As an endtoend process with the DFT [48], [169], [170] accuracy, solvation free energy is predicted from LS complex, free ligand and solvent (water) structures, all of

which are extracted from the trajectories of classical MD simulations of aqueous ligand system.

The workflow for the prediction of solvation free energy from multiple frames of an MD simulation is as follows: First, the atomic indices of ligand and solvents are read from the index groups of the index file and extracted into three separate xyz coordinate files (L, S and LS) (Figure 3.10). Next, ANI calculator is imported and set from ASE platform, and then, each group is read by the calculator to predict total single point energy (SPE). Then, the interaction energy between ligand and surroundings (water) are written to data files for each frame in both methods. In the final step, solvation free energy is estimated using the equations in theory part.

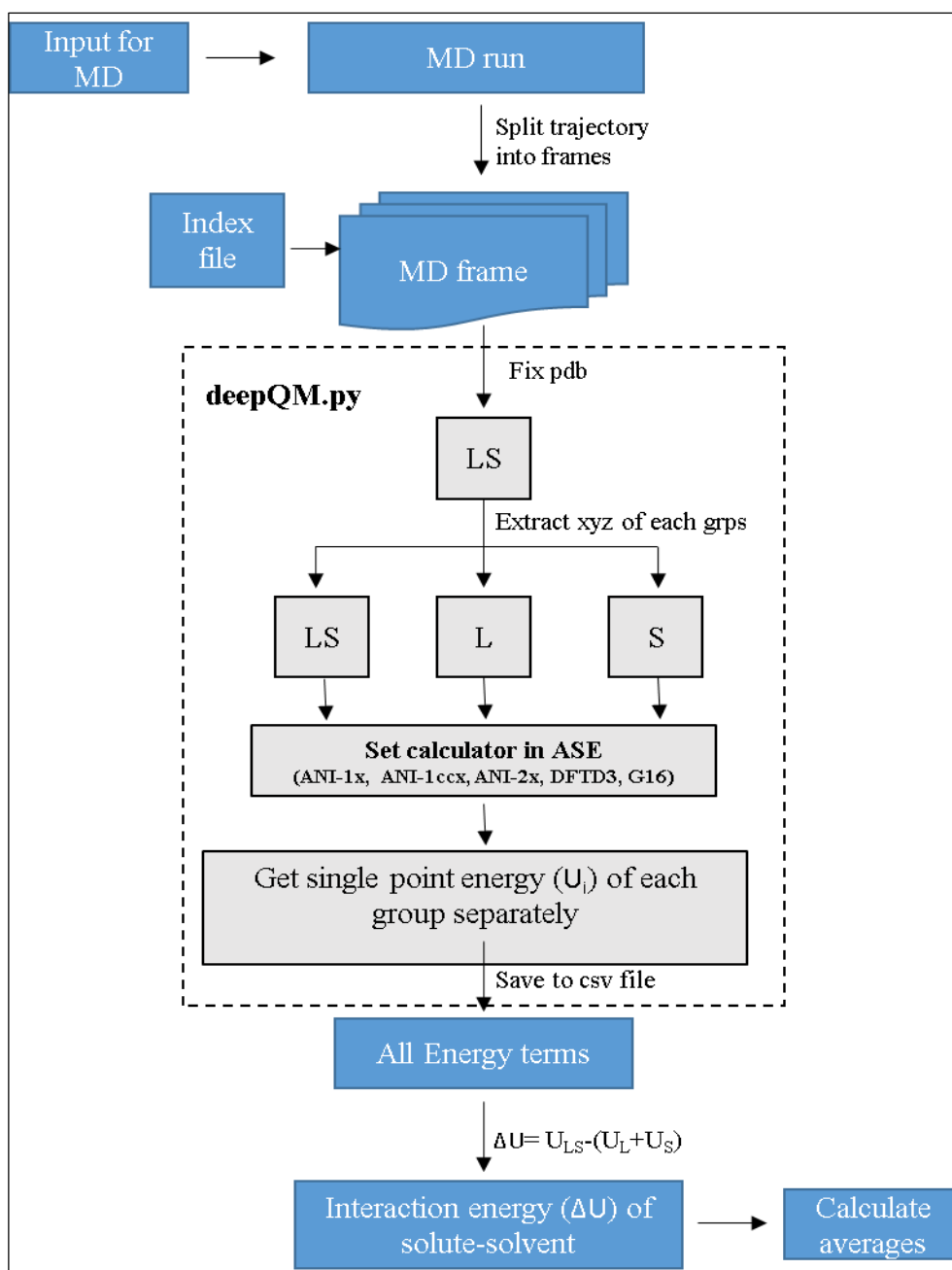


Figure 3.10: Workflow for ANI prediction of SFE by applying on single step MD trajectory.

### 3.2. ANI's performance on reproducing single point DFT energies

ANI-2x has been developed for covalently bound small organic molecules. To begin with, our first step is to assess whether ANI can accurately calculate single point energy (SPE) at the wb97x/6-31G(d) level of the isolated (gas phase) small molecules studied herein and available in FreeSolv database. Figure 3.11 shows the ANI vs G16 calculated SPE of nearly 500 molecules. It should be noted that ANI has been trained

for only C, H, O, N, S, F, Cl atoms, and thus the ligands with other heavy atoms than these were excluded from the calculations.

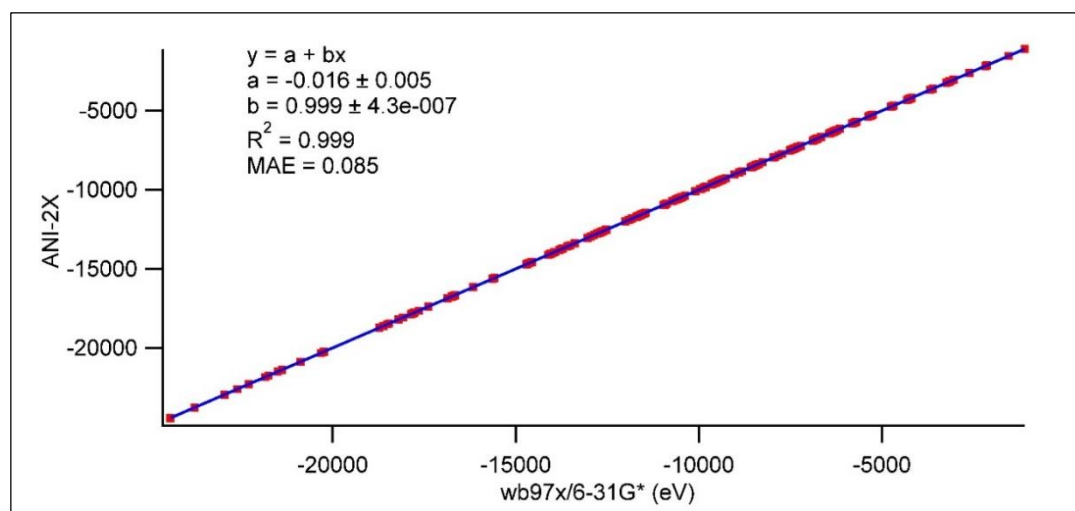


Figure 3.11: SPE of isolated molecules in FreeSolv predicted by ANI 2x vs DFT.

ANI-2x predicted SPE energies are in excellent agreement with the DFT energies. The mean absolute error is only 0.085 eV (1.96 kcal/mol). This deviation gets even smaller when the structure is at equilibrium geometry [171]. ANI is trained and uses atomic environments to calculate the energy of the given system. Therefore, in principle it should only care interatomic distances and predict the DFT energies no matter what the nature of the interaction. However, its reproduction of the DFT energies when non-covalent interactions (i.e. solute-solvent) exist still needs to be explored. For that purpose, we have started with C<sub>6</sub>H<sub>6</sub> (benzene) molecule and manually placed different number of H<sub>2</sub>O molecule so as that the water molecules around the ligand is held in the interaction distance. Table 3.1 shows the SPEs predicted by ANI and calculated by DFT in the existence of different numbers of H<sub>2</sub>O molecule around the ligand. ANI's prediction of the single point energies (not optimized) deviates from the DFT calculations in the presence of systems with non-covalent interactions. There could be mainly three reasons for this deviation: 1) the structure could be too far from the ANI's trained potential energy surface and thus beyond well represented regions of the neural network 2) non-covalent interactions may not be well covered in the training the ANI potentials 3) since the ANI was trained with small molecules, it may have a bias when the size of the system is increased. The first scenario would give a limitation to ANI's success of reproducing DFT energies.

However, the latter two can be fixed by introducing non-covalent interactions or scaling the ANI predictions in order to reproduce DFT energies.

Table 3.1 The interaction energy of benzene with 1 to 10 H<sub>2</sub>O molecules.

| n  | ANI (eV)     | DFT          | ANI-DFT (eV) |
|----|--------------|--------------|--------------|
| 1  | -8396.71206  | -8396.618172 | -0.093887262 |
| 2  | -10468.51749 | -10468.31767 | -0.199819184 |
| 3  | -12546.21956 | -12545.94694 | -0.272625984 |
| 4  | -14623.15944 | -14622.68645 | -0.472992388 |
| 5  | -16699.61713 | -16699.2506  | -0.36652547  |
| 6  | -18774.38569 | -18773.28446 | -1.101224306 |
| 7  | -20850.12259 | -20849.06604 | -1.05654602  |
| 8  | -22927.33709 | -22926.10341 | -1.233677591 |
| 9  | -25003.59647 | -25002.76063 | -0.835832431 |
| 10 | -27080.95508 | -27079.82898 | -1.126105807 |

To investigate further in detail about the nature of this deviation, we have replaced different amount of solvent in a periodic box and run NPT MD simulations for several ligands to extract frames and repeat the single point energy calculation of the ligand-nH<sub>2</sub>O complex system using both ANI and DFT. When the number of solvent is too few (i.e. unsaturated), the solvent molecules during the MD simulations cannot come to interaction distance around the solute. Therefore, the interaction between solute and solvent is basically zero in these cases, and the energy of the complex system is the sum of ligand and solvent's individual energies. In these non-interacting solvent-solute systems, ANI can perfectly reproduce single point energies of the ligand-nH<sub>2</sub>O complex system at the DFT level without significant error (Figure 3.12 a-d, for all other ligands refer to Figure 3.13). When the ligand is surrounded partially or completely by the solvent, the error in estimation of DFT energies of the system gets larger. However, there is still a linear correlation between ANI estimated and DFT calculated energies of the systems (Figure 3.12 e-i).

### 3.3. ANI's performance on reproducing QM interaction energies

Upon insertion of multiple numbers of water, the ANI predicted energy of the complex system deviates from DFT results. On the other hand, we are interested in the interaction energies rather than the absolute energies of the solute-solvent complex

systems. Therefore, we have also calculated the interaction energies between solute and solvent from  $\Delta U = U_{\text{complex}} - (U_{\text{lig}} + U_{\text{water}})$ .

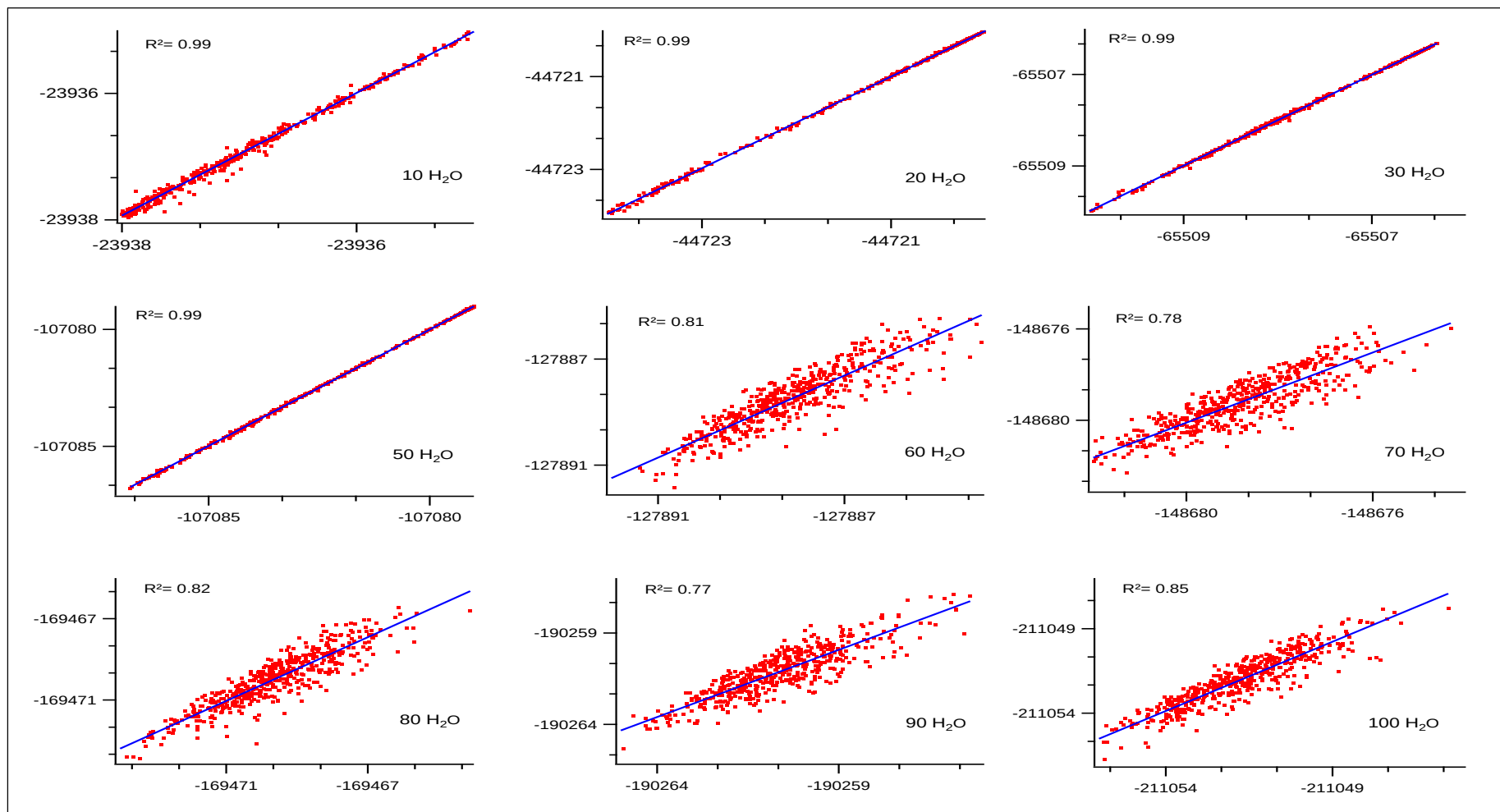


Figure 3.12: The SPE of the ligand(mobley\_1636752)-nH<sub>2</sub>O (n=10,20..100) systems calculated by ANI and DFT obtained MD trajectories.

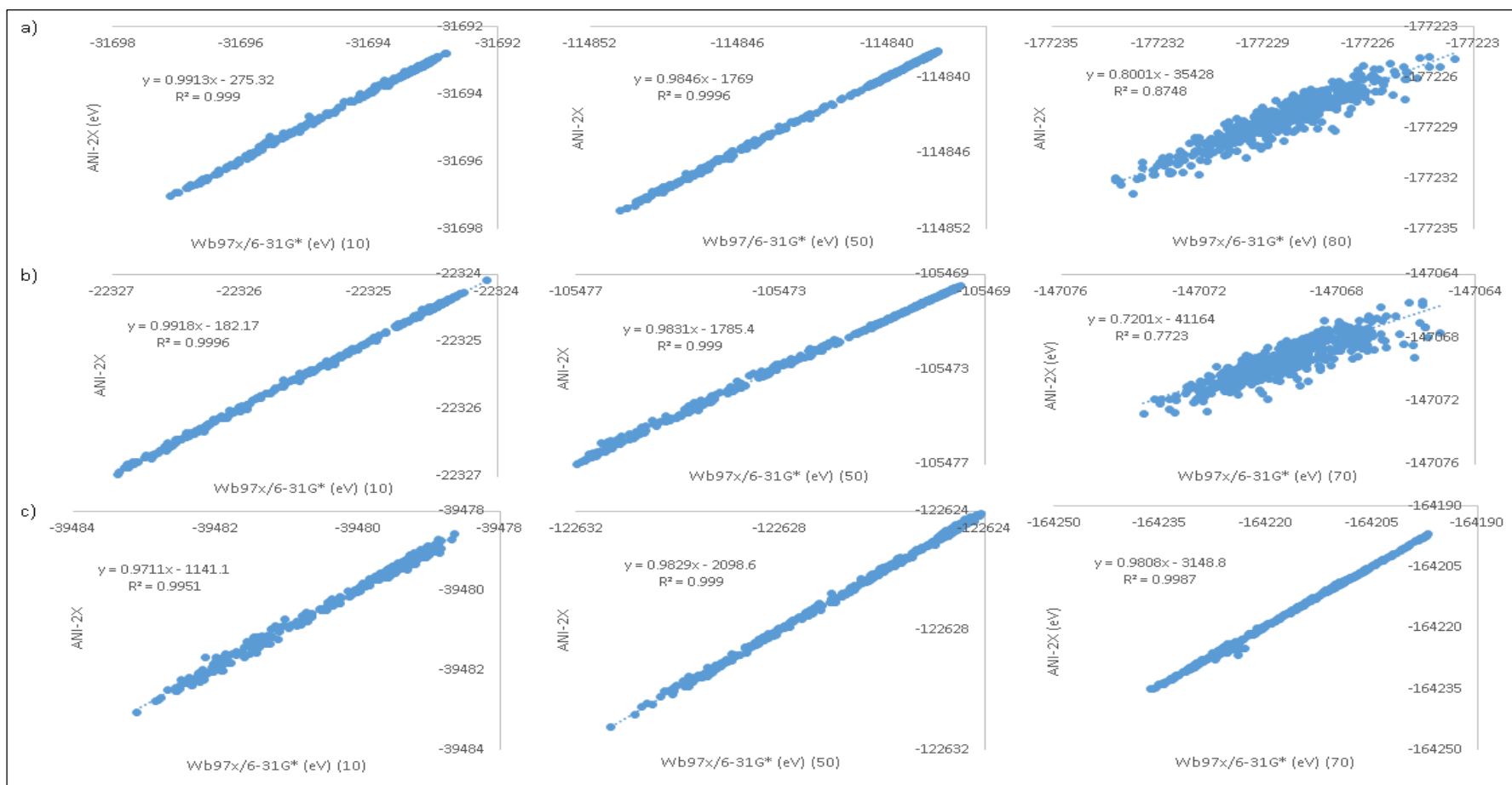


Figure 3.13: The SPE of the different ligands with different nH<sub>2</sub>O systems calculated by ANI and DFT obtained MD trajectories.

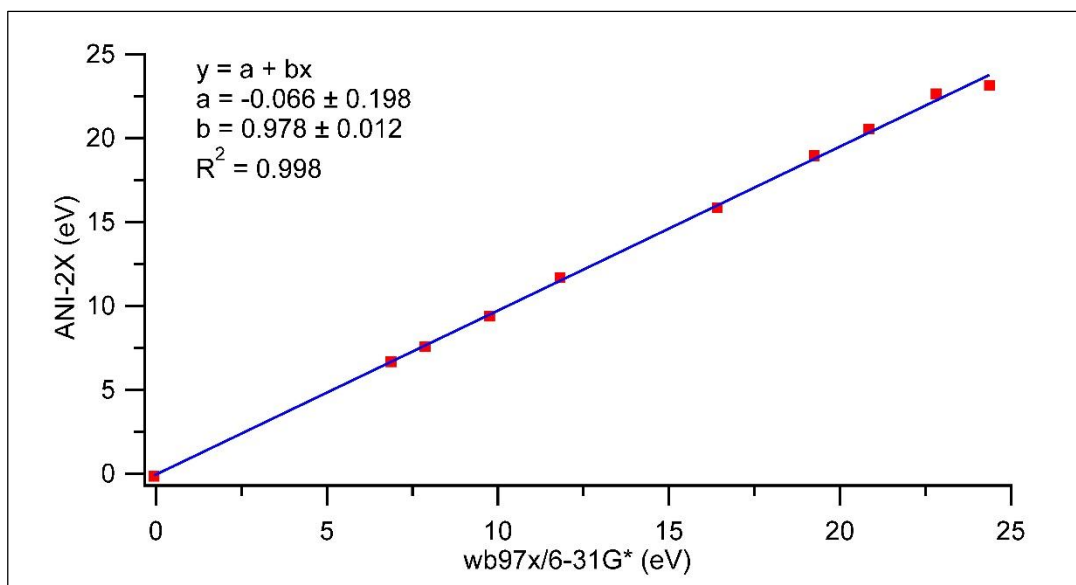


Figure 3.14: Interaction energies calculated by ANI and G16 for  $C_6H_6-nH_2O$  ( $n=1-10$ ),  $\Delta U = U_{complex} - (U_{lig} + U_{water})$ .

Figure 3.14 shows the interaction energies between ligand and solvent calculated with two different methods known as ANI and DFT for benzene- $nH_2O$ . Here, deviations of estimated interaction energies produced by ANI from DFT energies are much smaller, clearly indicating ANI can be replaced for DFT accurate calculation of interaction energies. The magnitude of the interaction energies calculated by both ANI and DFT are too large. This is because these energies are calculated from manually inserted water molecules around benzene ligand; i.e., energies at non equilibrium geometries, which might deviate from optimum geometry.

In addition, we have also calculated the interaction energies using MD simulation with different amount of solvents for selected ligands by stripping the solute and solvent from the trajectory of the complex system as outlined in Figure 3.10. When there are very limited number of solvents in the simulation box, the interaction energy is zero since the solvents cannot create adequate solvation shell around the ligand throughout the simulation time. When the ligand is surrounded partially or completely by the solvent during MD simulation, both ANI and DFT produce non zero interaction energies (Figure 3.15).

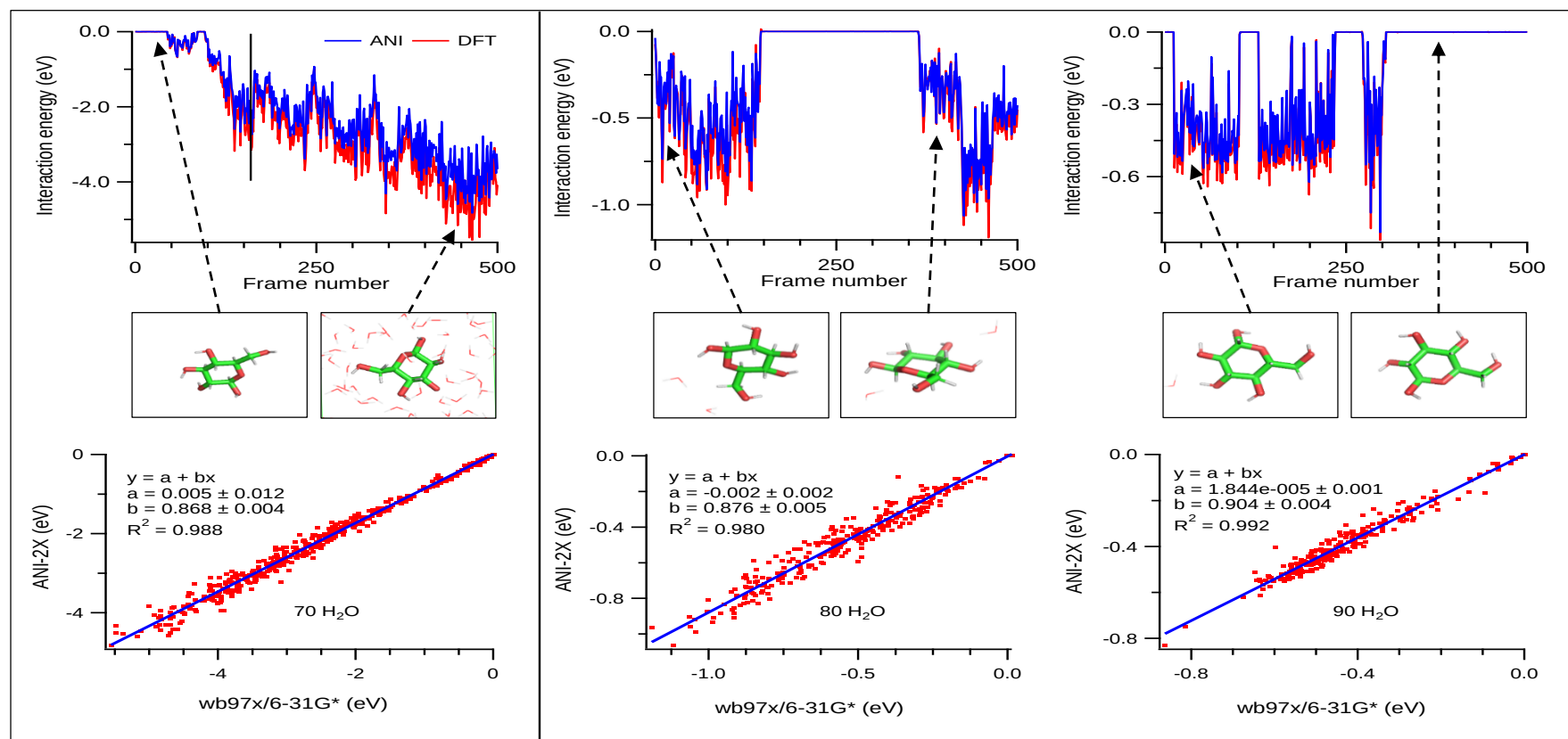


Figure 3.15: Comparison of interaction energies predicted by ANI and G16 throughout MD trajectories in the presence of 70-90 (a, b, and c, respectively) water for a selected ligand. The top plots and figures show the ligand unsolvated ( $y=0$  values) or partially solvated ( $y \neq 0$  values) frames along MD simulations. Bottom plots show the correlation between ANI and DFT predicted values for these non-interacting or partially interacting MD frames.

When the simulations were started with sufficient number of solvents, the solvation shell around the ligand becomes complete (saturated, fully interacting) during simulation. The number of solvents for a complete shell varies from 60 to 600 according to size of the ligand and simulation time. For methanol in 1 ns simulation time, this number is as small as 60 while for glucose it is 300. Figure 3.16 indicates that both ANI and DFT follow very similar pattern in terms of interaction energies throughout the 100 frames of 40 ns MD trajectory for the ligand in 268 water and there is some deviation of the ANI interaction energies from DFT values. The same pattern is also observed along 500 frames extracted from 10 ns MD simulation for the ligand in 100 water solvent (Figure 3.17). Additionally, in case of fully solvation ligand in 300 water molecules, ANI calculations of different molecules show a similar trend to DFT results and when ANI is scaled with a and b ( $a \cdot x + b$ , a and b are constant and x is represented by DFT and  $a=1.17$  and  $b=0.5$  (\* small deviations was observed  $\pm 0.05$ )), it is seen that ANI can highly mimic DFT results in terms of interaction energy as illustrated in Figure 3.18. However, this deviation is constant in the averaged energies and can easily be fixed by introducing a correction factor (i.e. a linear scaling) to the ANI predicted values since the average interaction energy changes no more by the excess number of solvents existed in the simulation box (Figure 3.19) and in Figures 3.20, the trend between ANI and DFT output continues after fully contacted and solvated. After saturation, any changes in energy were not observed. In . For these saturated systems, ANI predicted interaction energies can be given with a simple linear scaling from DFT energies as mentioned above.

In conclusion, ANI, although it deviates from DFT predicted average interaction energies, can be utilized to predict the DFT accurate average interaction energies with a correction which eventually can be translated to the solvation free energies of molecules in explicit solvents. Instead of performing costly DFT calculations (that can take days) on the trajectory of an end state MD simulation in order to predict average interaction energies, ANI can produce similar results with a remarkable speed up (within minutes).

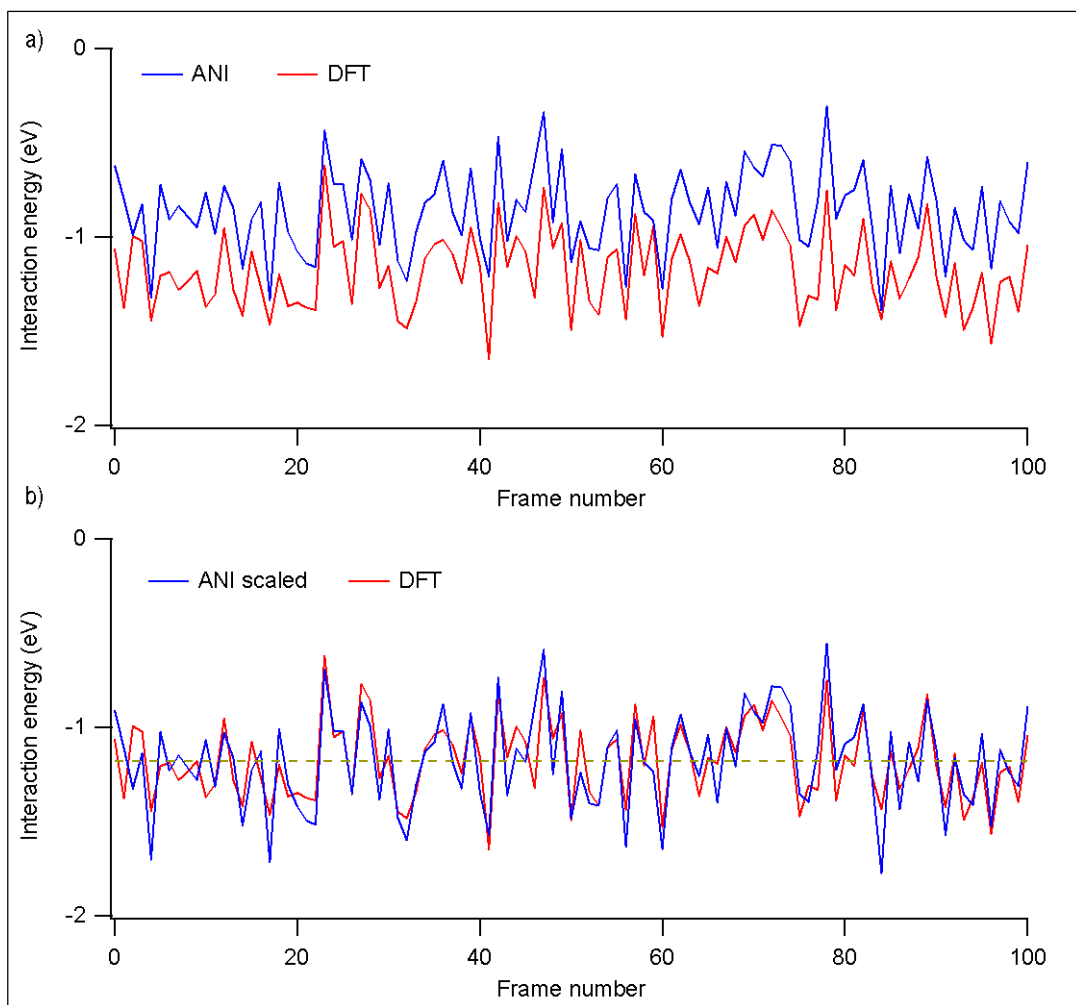


Figure 3.16: DFT and ANI calculated interaction energies for methanol (moley\_16366752)-268H<sub>2</sub>O system along 100 frames of a 40 ns MD simulation. a) ANI's deviation from DFT, b) ANI linearly scaled to reproduce DFT interaction energies ( $\Delta E_{\text{ANI}} = 0.18217 + 0.88499 \Delta E_{\text{DFT}}$ ). The green dashed line shows the average interaction energies of both ANI and DFT to be the same as -1.1772 eV.

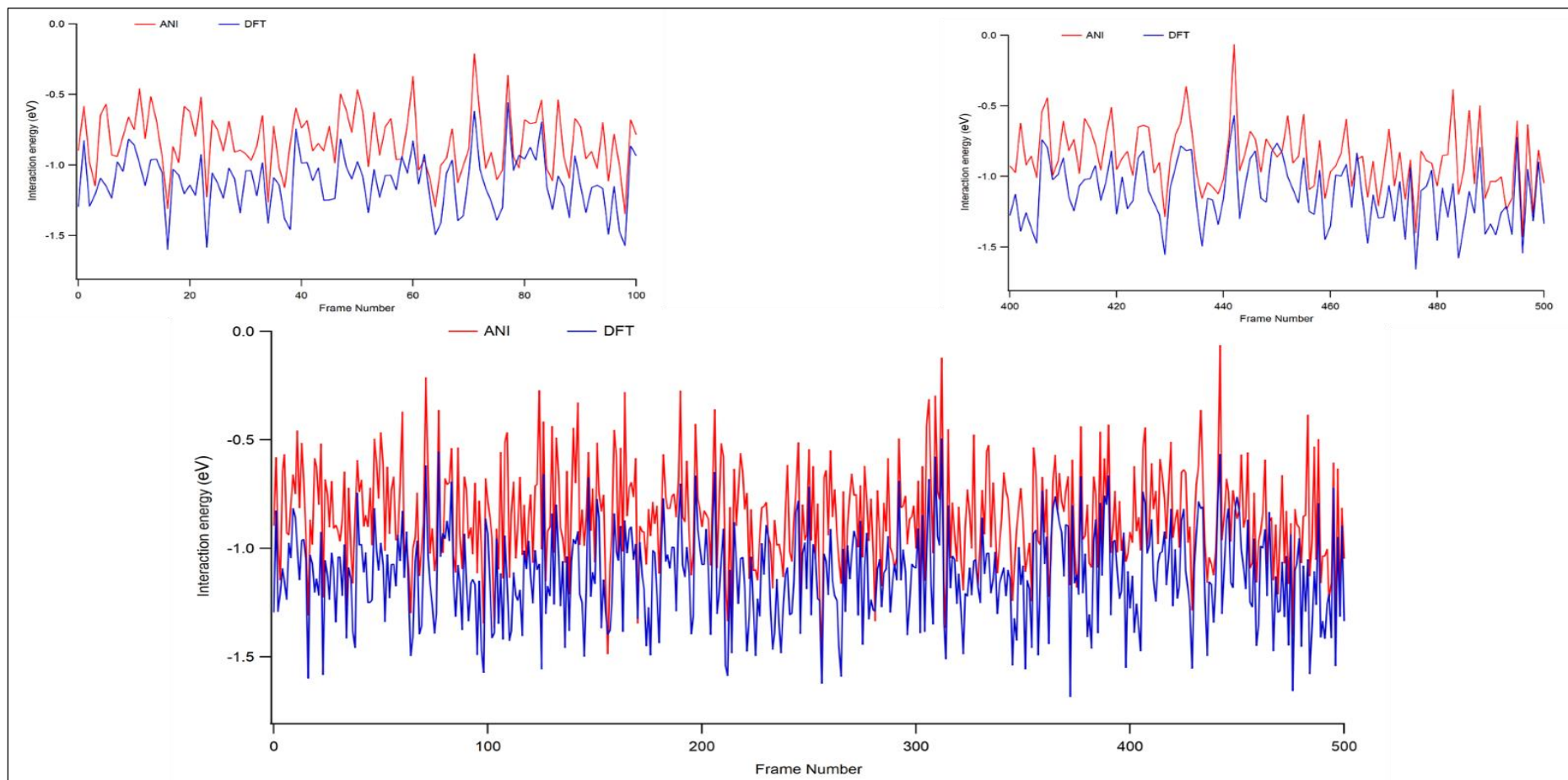


Figure 3.17: DFT and ANI calculated interaction energies for methanol (mobley\_1636752)-100H<sub>2</sub>O system along 500 frames of a 5 ns MD simulation. a) Top two figures show a specific range of bottom main figure by enlarging view.

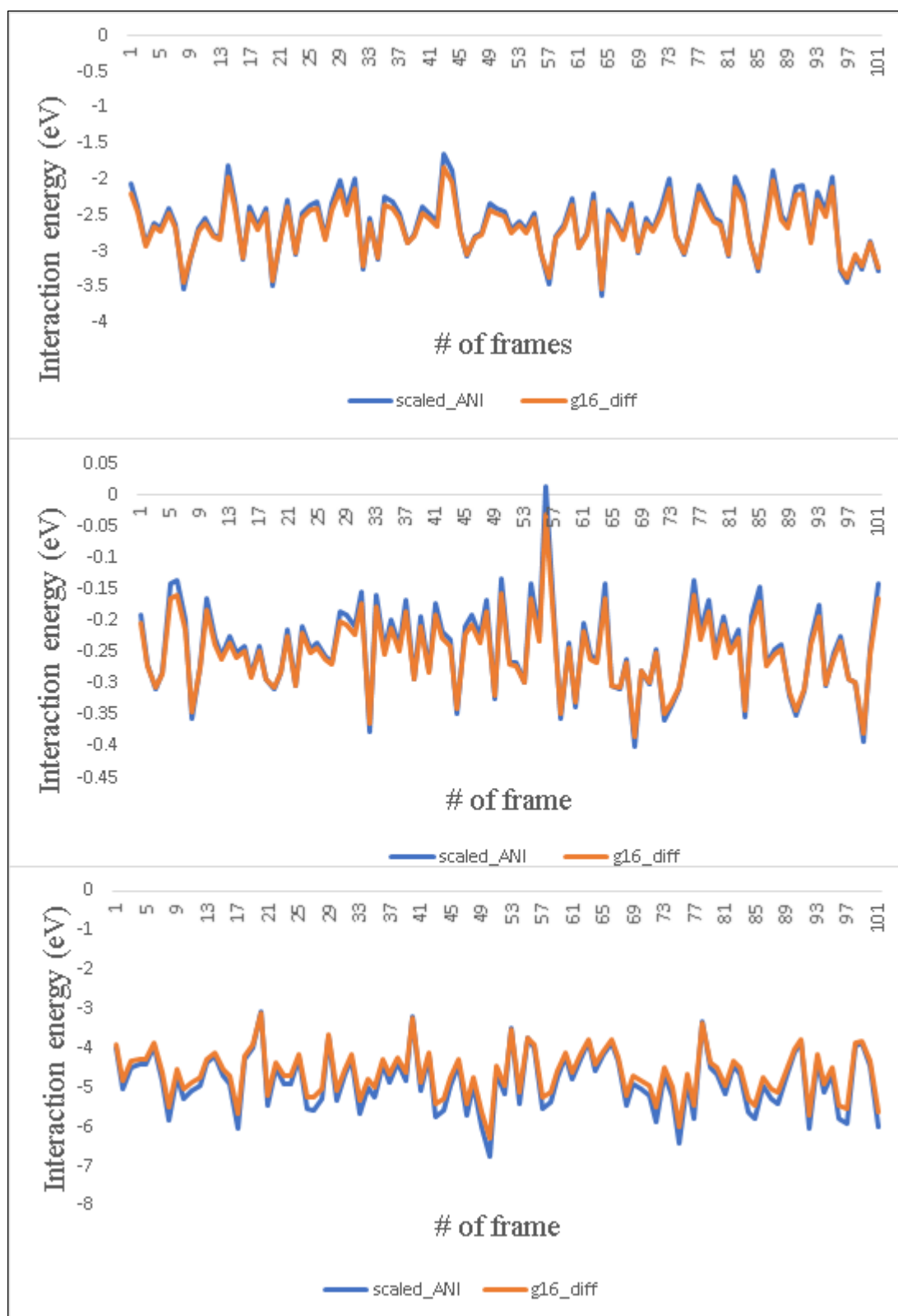


Figure 3.18: Comparison of scaled ANI and DFT interaction energy calculation for different molecules. They are mobley\_3259411, mobley\_9055303 and mobley\_9534740 (respectively) with 300 H<sub>2</sub>O for 101 frames from trajectories.

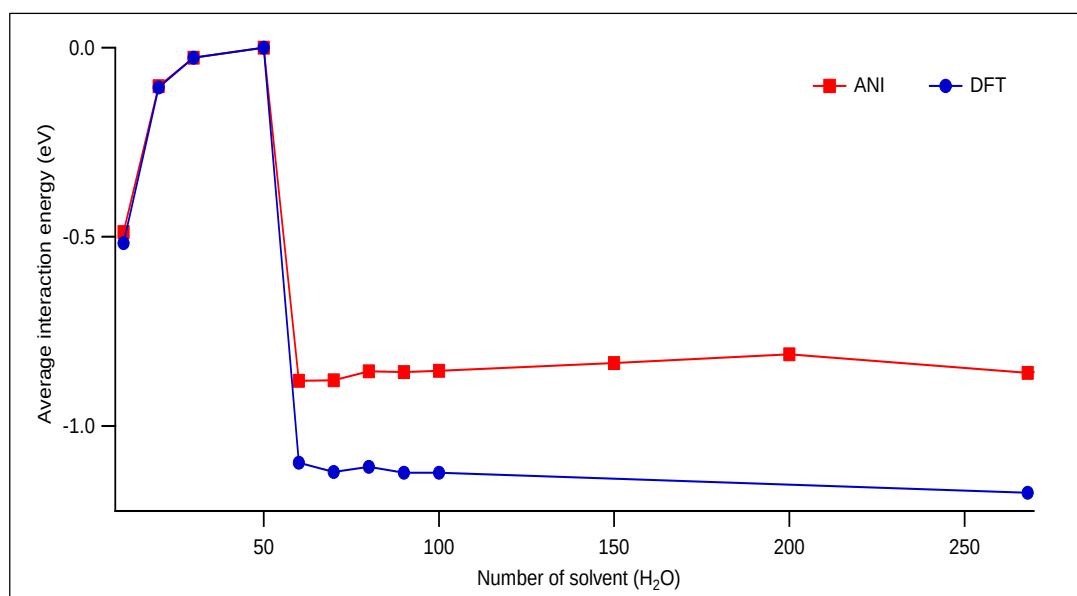


Figure 3.19: Average interaction energy of all 500 frames from MD simulations. Simulations are in the presence of different numbers of solvent molecules via ANI and DFT methods. The last points (n=268) use average of 100 frames from a 40ns MD simulation.

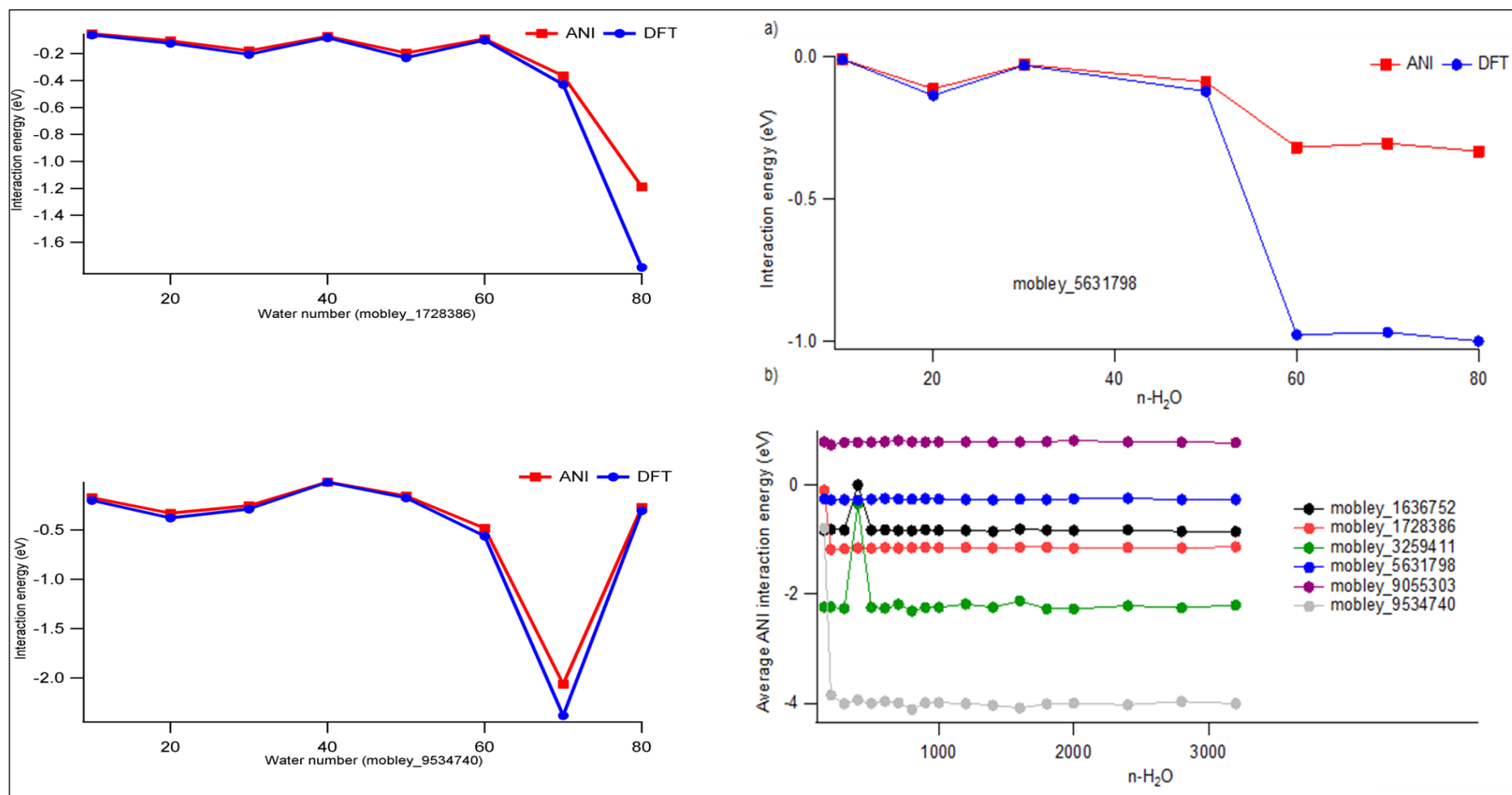


Figure 3.20: The view of interaction energy calculated by ANI and G16 for different molecules in different solvent volumes.

### 3.4. Prediction of solvation free energies

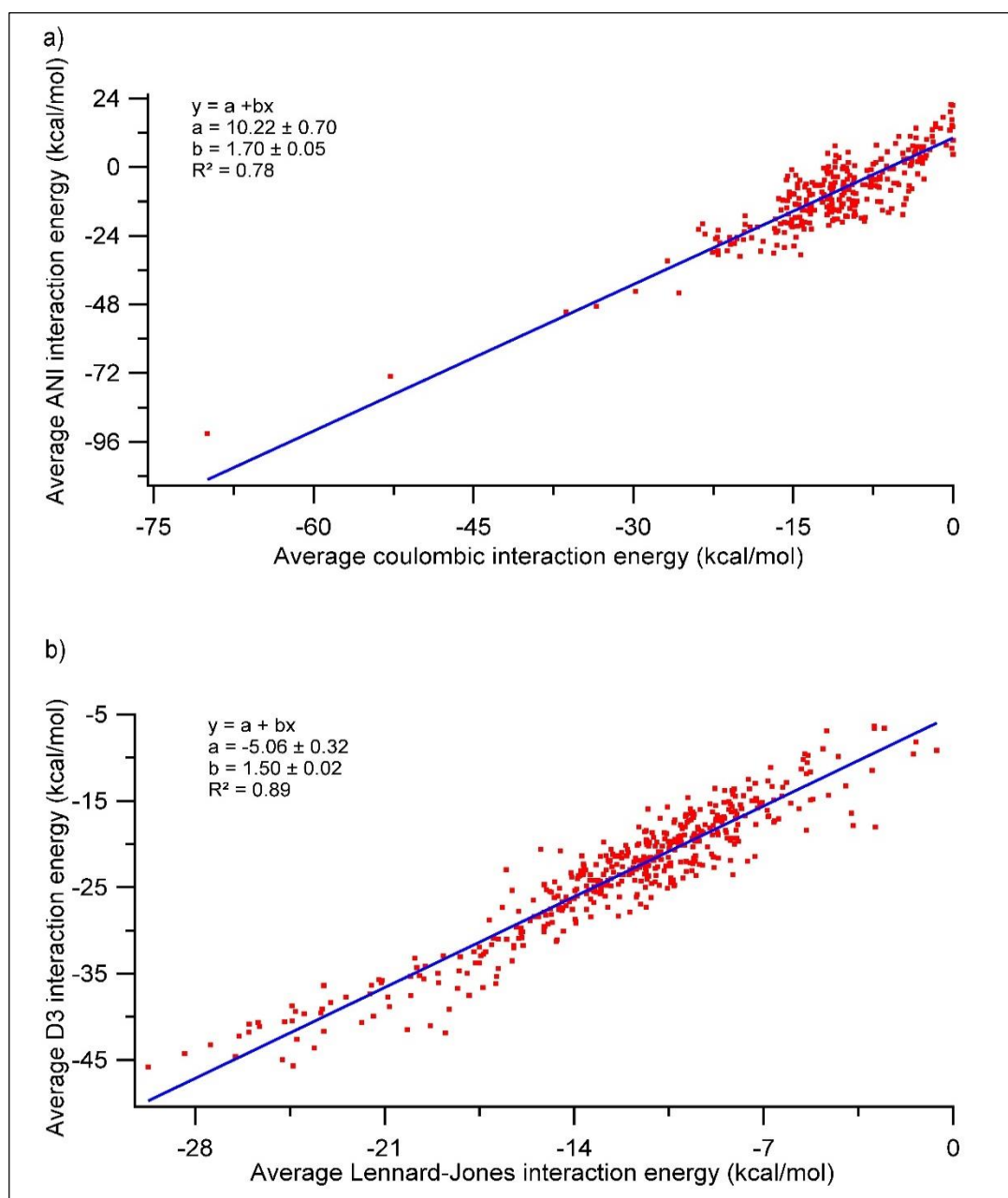


Figure 3.21: Average ANI and D3 interaction energies. They show correlation to coulombic and Lennard-Jones interaction terms, respectively, between the ligand and solvent for the fully solvated MD trajectories.

Our results for interaction energies predicted by ANI and DFT follow similar trends with a scalable error. These ML based DFT accurate interaction energies by ANI, applied as post-MD simulation of ligand-water complex systems, were further explored by means of applicability to predict solvation free energies.

As outlined in Figure 3.10, after obtaining SPE of solute-solvent, solvent and solute for each frame extracted from MD trajectories, equations 2.1-2.11 can be readily applied to estimate the solvation free energies [172]. In order to achieve this, we first compared the ANI predicted average interaction energies with average electrostatic MM energies and average D3 interaction energies with average van der Waals interaction energies between the solute and solvent during a simulation (Figure 3.21a). Average ANI interaction energies are correlated to average electrostatic interaction energies. Having correlation coefficient low may indicate the superiority of DFT accurate ML potentials. Similarly, DFT-D3 [51] interaction energies show high correlation with van der Waals term of the MM energy since both methods rely on use of Lennard-Jones 12-6 potential (Figure 3.21b).

The approaches of classical MD which are appeared by involving correlation between the MM electrostatic interaction energies and the solvation free energies are linear interaction energy (LIE) [68] and linear response approximation (LRA) [173]. The correlation coefficient for the electrostatic term was first found to be 0.5 while later studies updated this value to 0.43 and 0.38 [174]. In these molecules, we find this value as  $0.389 \pm 0.012$  [174]. Similarly, the coefficient of the van der Waals term was found to be  $0.136 \pm 0.020$ , in a good agreement with the reported value of 0.12 (Figure 3.22a). We have also calculated the solvation free energies using classical MM based BAR and DFT using different implicit solvent definitions. BAR [175] is also capable of reproducing experimental results with high accuracy ( $R^2=0.87$ ) (Figure 3.22b). Among the continuum methods of SMD, CPCM and IEFPCM, DFT using SMD polarizable continuum model showed best results with  $R^2=0.9$  (Figure 3.22c). For the other continuum model, the results can be seen in Figure 3.23a and Figure 3.23b (For IEFPCM,  $R^2=0.75$  and for CPCM,  $R^2=0.75$ ).

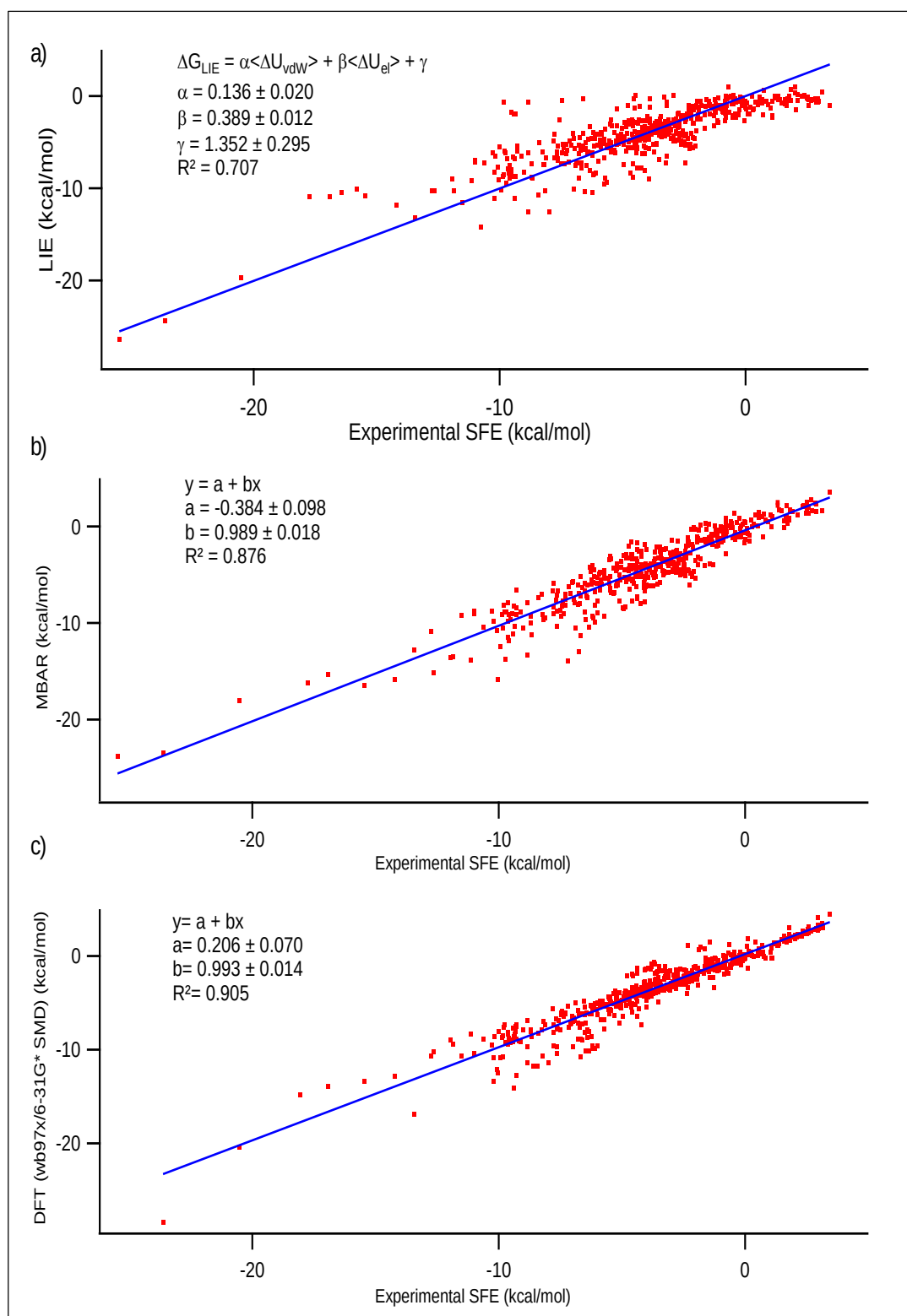


Figure 3.22: Three different approaches to predict solvation free energies. a) poorest but fastest results with LIE, b) accurate but costly BAR and c) the most accurate but continuum model DFT-SMD. For other continuum models.

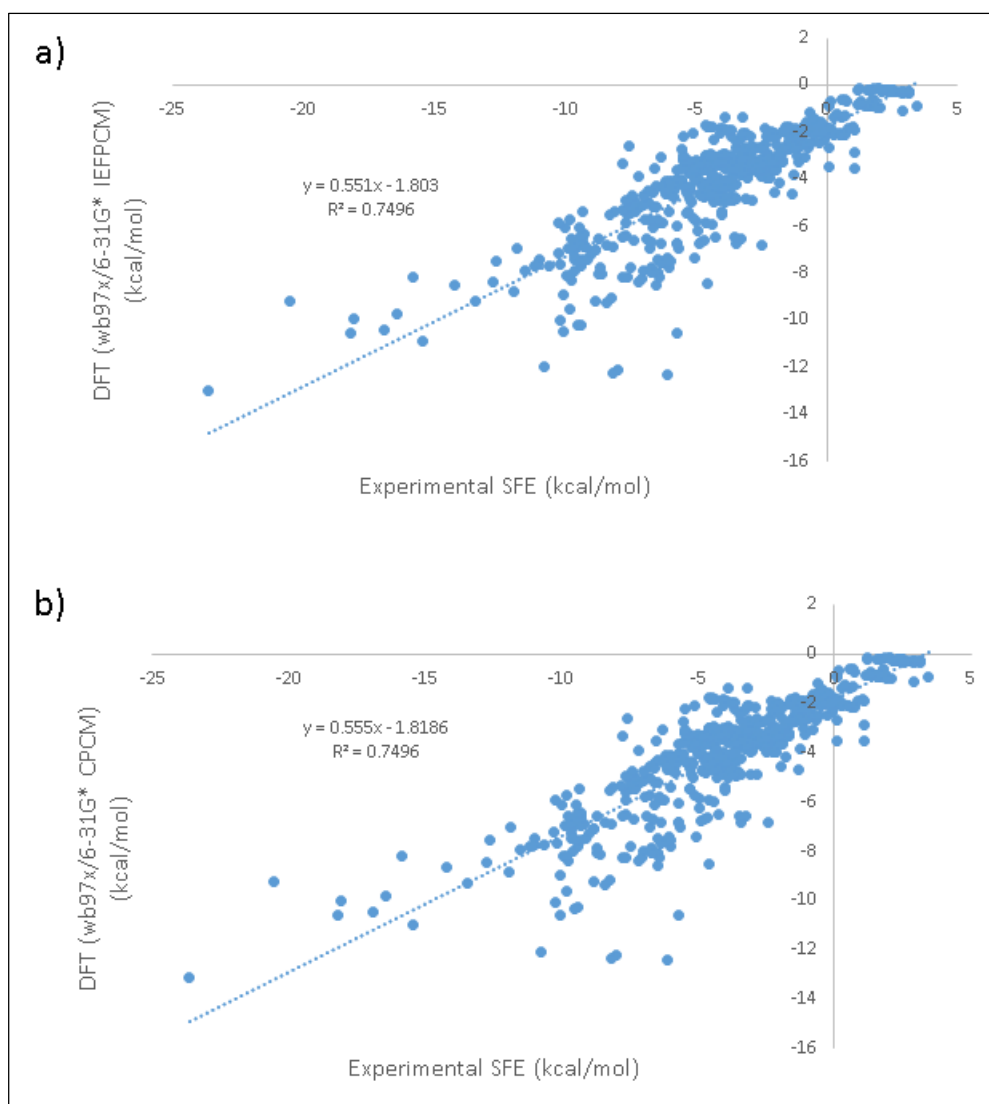


Figure 3.23: The other continuum model for solvation free energy calculation. The methods in implicit solvent medium besides SMD a) IEFPCM b) CPCM.

On the other hand, using ANI potentials and DFT-D3 energies as a replacement of columbic and van der Waals interactions respectively in equations (2.1) and (2.8), the coefficient values must be different. Figure 3.24 shows average ANI interaction energies fitted to the experimental solvation free energies by either excluding or including D3 correction terms. When compared to classical LIE approach, ANI replacement with MM energies increases the correlation coefficient from  $R^2=0.70$  to 0.80. In addition, the inclusion of D3 correction terms improves the results 1%. It should be noted that ANI calculation brings almost no additional computational cost to the simulations. Thus, ANI can be a promising alternative approach in the prediction of solvation (hydration) free energies.

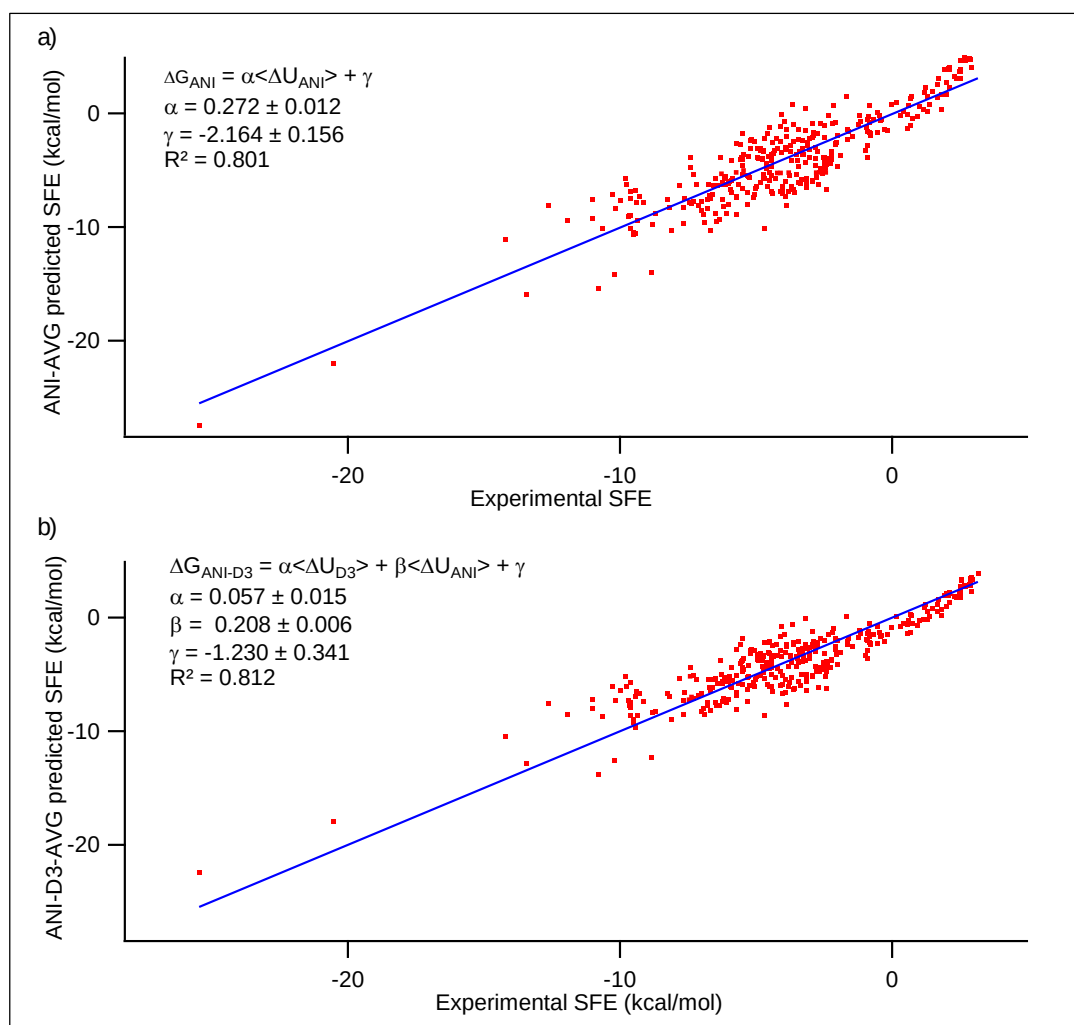


Figure 3.24: ANI predicted solvation free energies shows more accuracy than LIE and less computational cost than BAR.

In addition to the ANI modification of the LIE approach, we have also assessed the success of replacement of MM energies with ANI interaction energies in single step free energy calculations in equations (2.7) – (2.11). Figure 3.25 shows the trend among the 340 organic compounds calculated by ANI modified equations. Among these methods, exponential average predicts almost identical results with ANI modified LIE method, outperforming the MM based LIE.

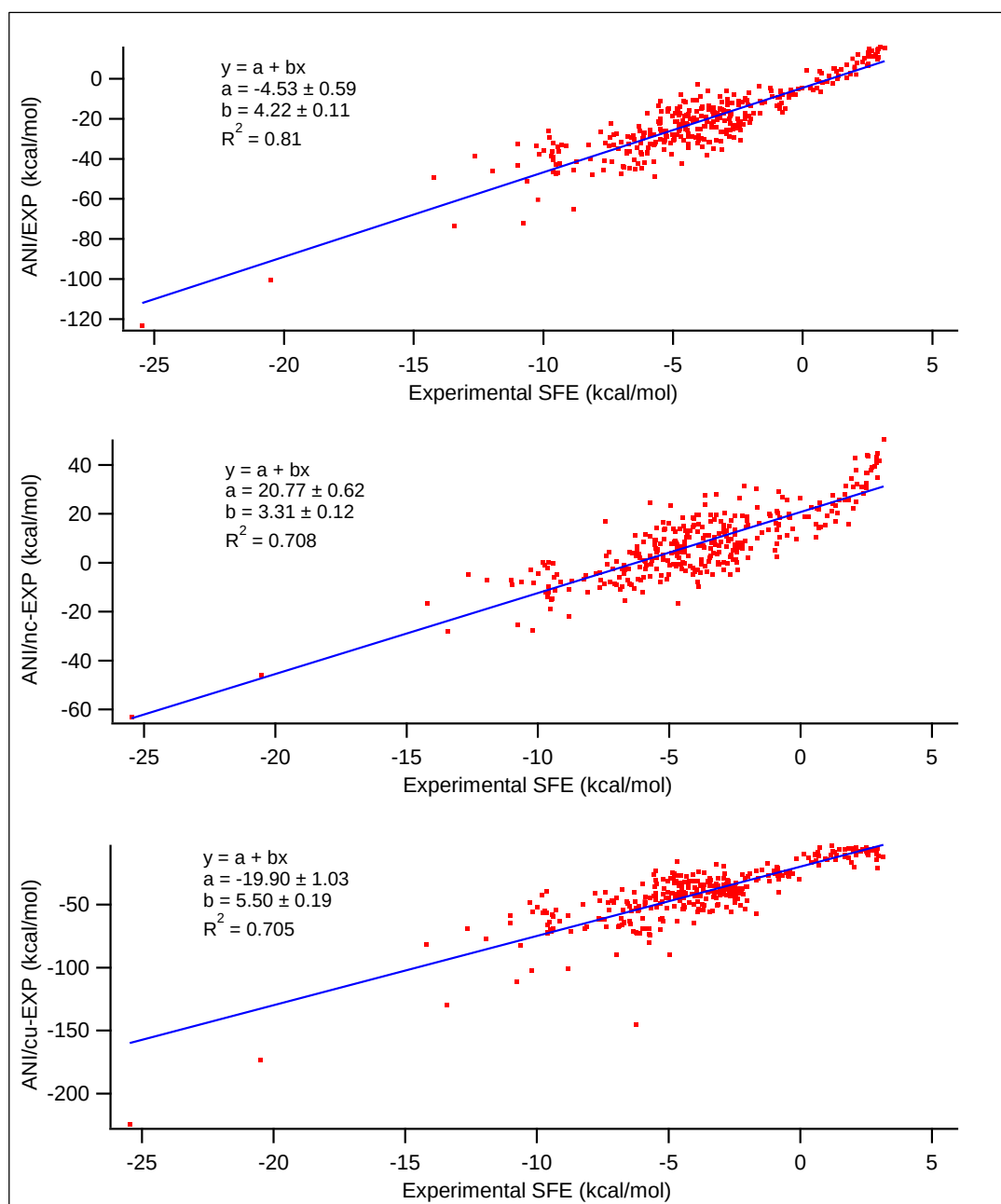


Figure 3.25: ANI's success of reproducing experimental SFEs using equations (2.7)-(2.11). Here, ANI energies were replaced with MM energies.

### 3.5. Conformer Search of HBGAs

After calculating solvation free energy calculation with ML based method known ANI, we proposed that ANI can also be utilized to estimate the interaction

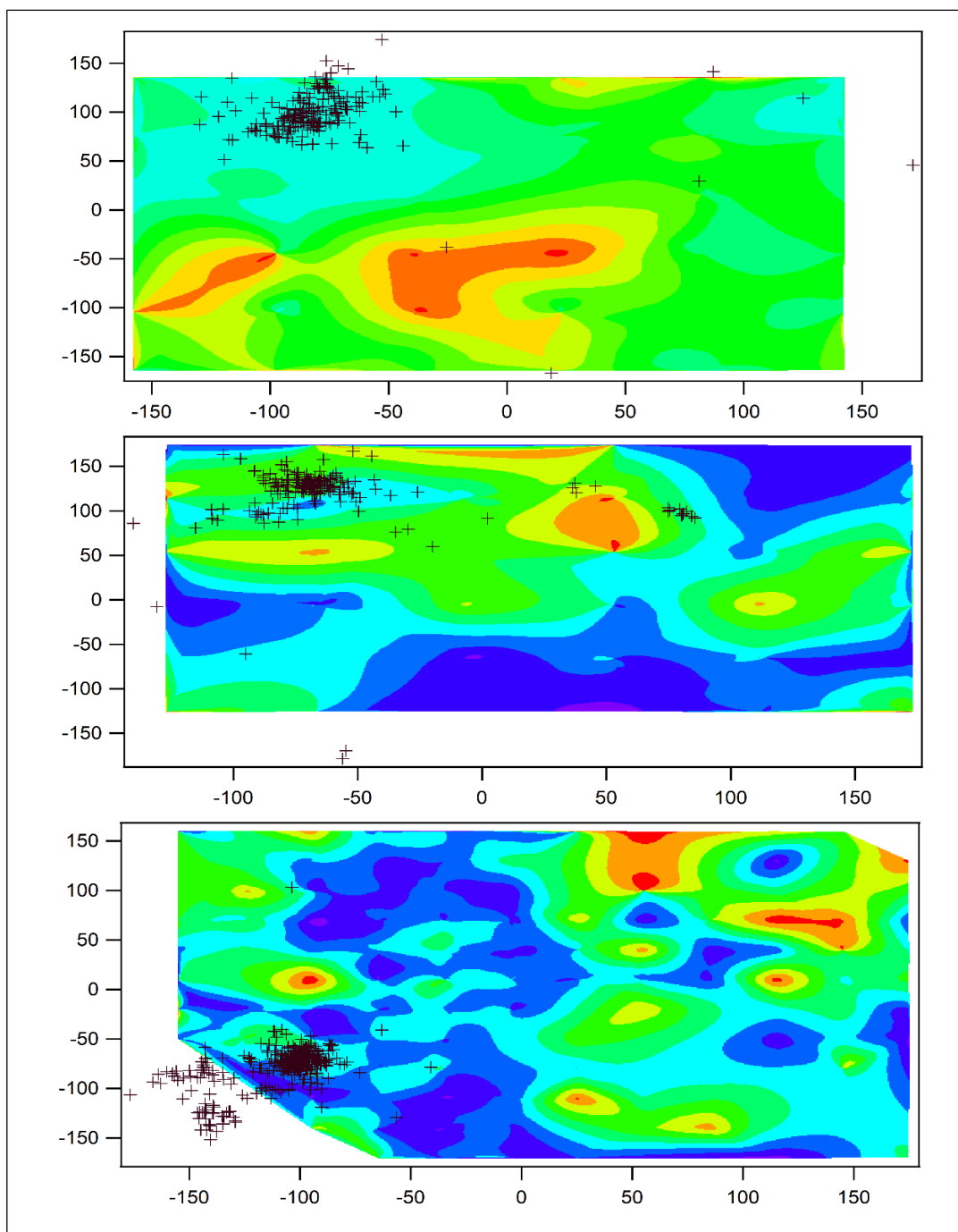


Figure 3.26: HBGA A Type 1 for only HBGA's SPE. The compare of psi and phi angles of 3 main torsion points of the generated conformers with crystal structure's psi and phi angles plotted with the relative energy of conformers based on conformer having the lowest SPE energy. Psi represents by horizontal axis (x) and Phi represents by horizontal axis (y). Last one shows the prediction for fucose unit.

energy between protein and different conformers of ligand. For this aim, we searched for human blood group antigens interaction with HuNov. To start with, the SPE of generated conformers of secretor kinds of HBGAs was calculated with ANI and they were sorted with a certain cut-off (100 kcal/mol) and then sorted ones are aligned. The

SPE of aligned structure was measured with ANI. Torsion points were selected as conjunction of two sugars in HBGA.

As shown Figure 3.26 (HBGA A Type 1 in case of only ligand), there was a nearly good estimation between calculated and experimental values for torsion 1 in HBGA A type 1 but this could not be said for torsion 2 and 3. Particularly, the estimation for fucose unit was out of range of experimental. In short, an acceptable result was not obtained for this type of HBGA in case of only ligand. The reason why the predictions were not in the range of interest is that the conformers may be generated incorrectly due to python combine script or program (Spartan WaveFunction) used to produce first torsions. Another reason is that there is no enough similar sample in terms of crystal structure for HBGA A Type 1.

As shown Figure 3.27 (HBGA A Type 1 in case of ligand-protein complex), there was a nearly good estimation between calculated and experimental values for torsion 1 and 2. In addition, the estimation for fucose unit was very close to the experimental values. In short, for this type of HBGA in case of ligand-protein complex, a good prediction values calculated by ANI were observed. The similar reasons, mentioned before, for non-satisfying results may be involved.

Figure 3.28 (HBGA A Type 2 in case of only ligand) indicates that ANI was probably failed to calculate energies at specific psi and phi angles of torsions and the predicted results against experimental values are very poor. In addition to aforementioned reasons, ANI algorithm may not be sufficient to calculate SPE of this complex molecules.

Nearly good estimation was observed between calculated and experimental values for torsion 1 in Figure 3.29 (HBGA A Type 2 in case of ligand-protein complex); however, it is difficult to say the same thing for torsion 2 and torsion 3 in fucose unit. In short, for this type of HBGA in case of ligand-protein complex, a prediction values calculated by ANI in a good range were not obtained. Sorted ligands were aligned to a protein of HuNoV G.II-4 with gromacs gmx confms command with specified index number (fucose unit). One of the problem for failed results is that probably there is a alignment error, which plays a role in that.

In conclusion for both HBGA A Type 1 and 2 in the state of both ligand and ligand-complex, the predictions were generally not exactly correlated with experimental values whereas some calculated values were in the range of experimental ones. Maybe, ANI is required more implementations to calculate better.

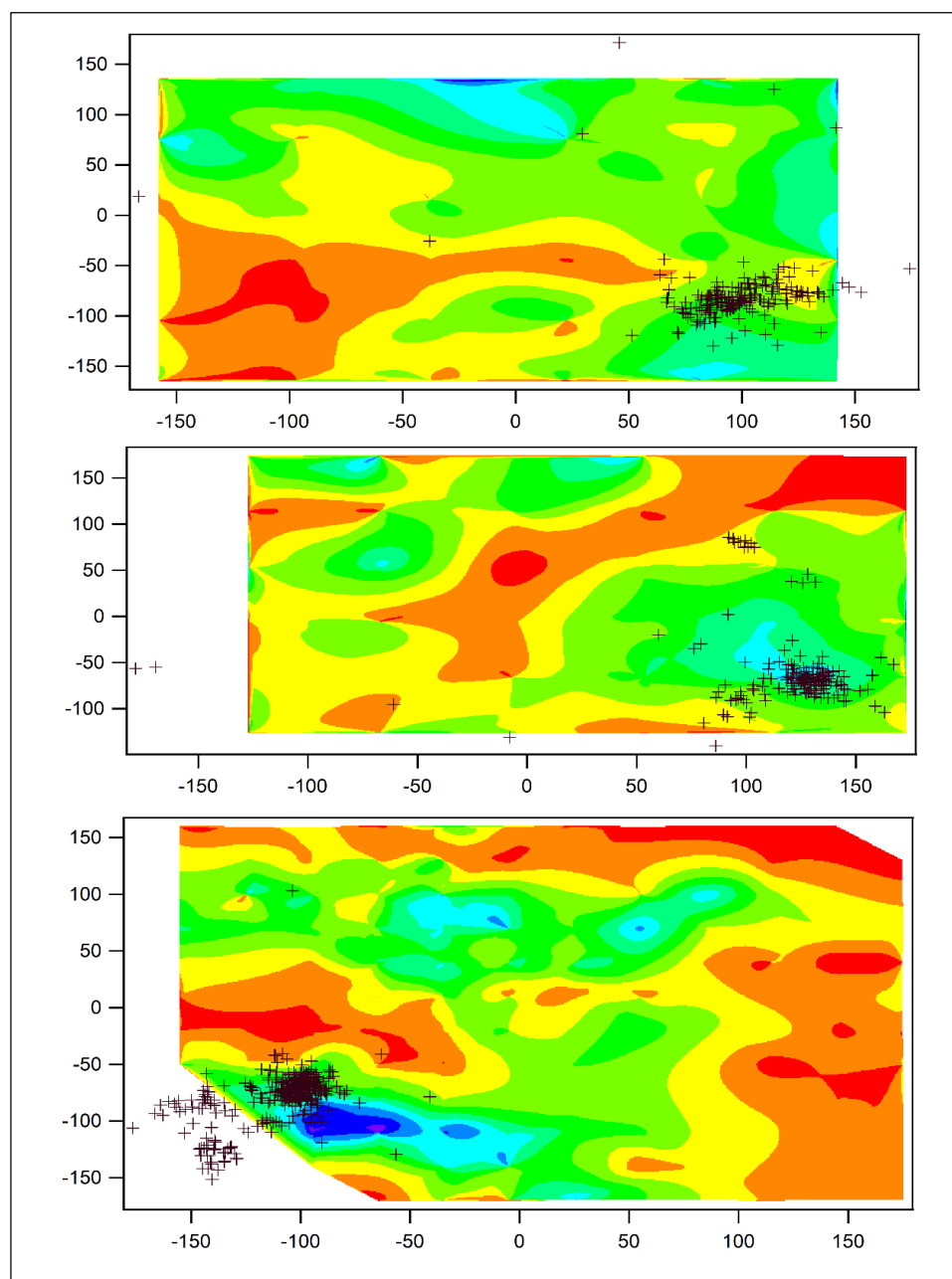


Figure 3.27: HBGA A Type 1 for HBGA-protein complex. The compare of psi and phi angles of 3 main torsion points of the generated conformers aligned to protein with crystal structure's psi and phi angles plotted with the relative energy of ligand-protein complex based on complex having the lowest SPE energy. Psi represents by horizontal axis (x) and Phi represents by horizontal axis (y). Last one shows the prediction for fucose unit.

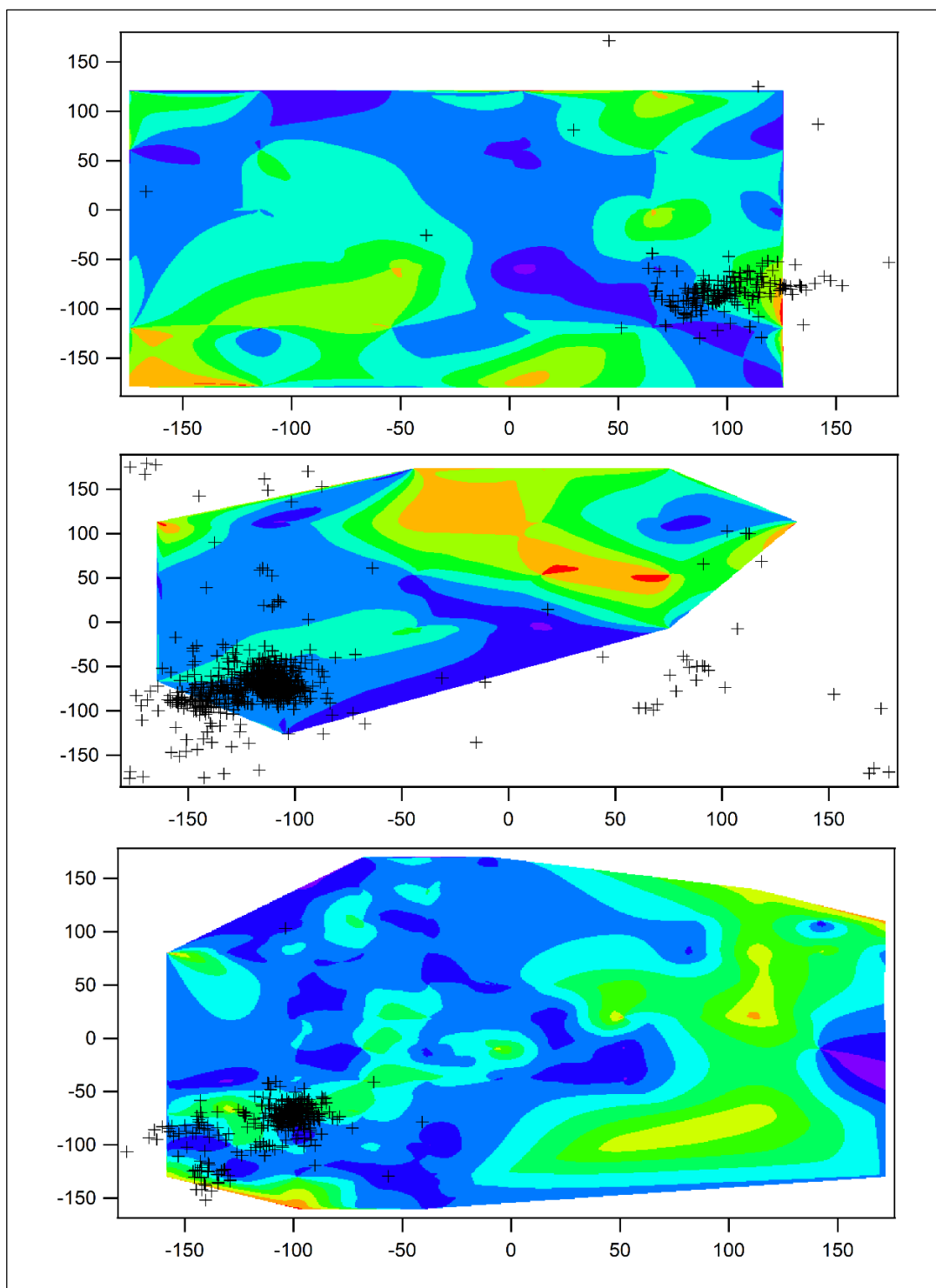


Figure 3.28: HBGA A Type 2 for only HBGA's SPE. The compare of psi and phi angles of 3 main torsion points of the generated conformers with crystal structure's psi and phi angles plotted with the relative energy of conformers based on conformer having the lowest SPE energy. Psi represents by horizontal axis (x) and Phi represents by horizontal axis (y). Last one shows the prediction for fucose unit.

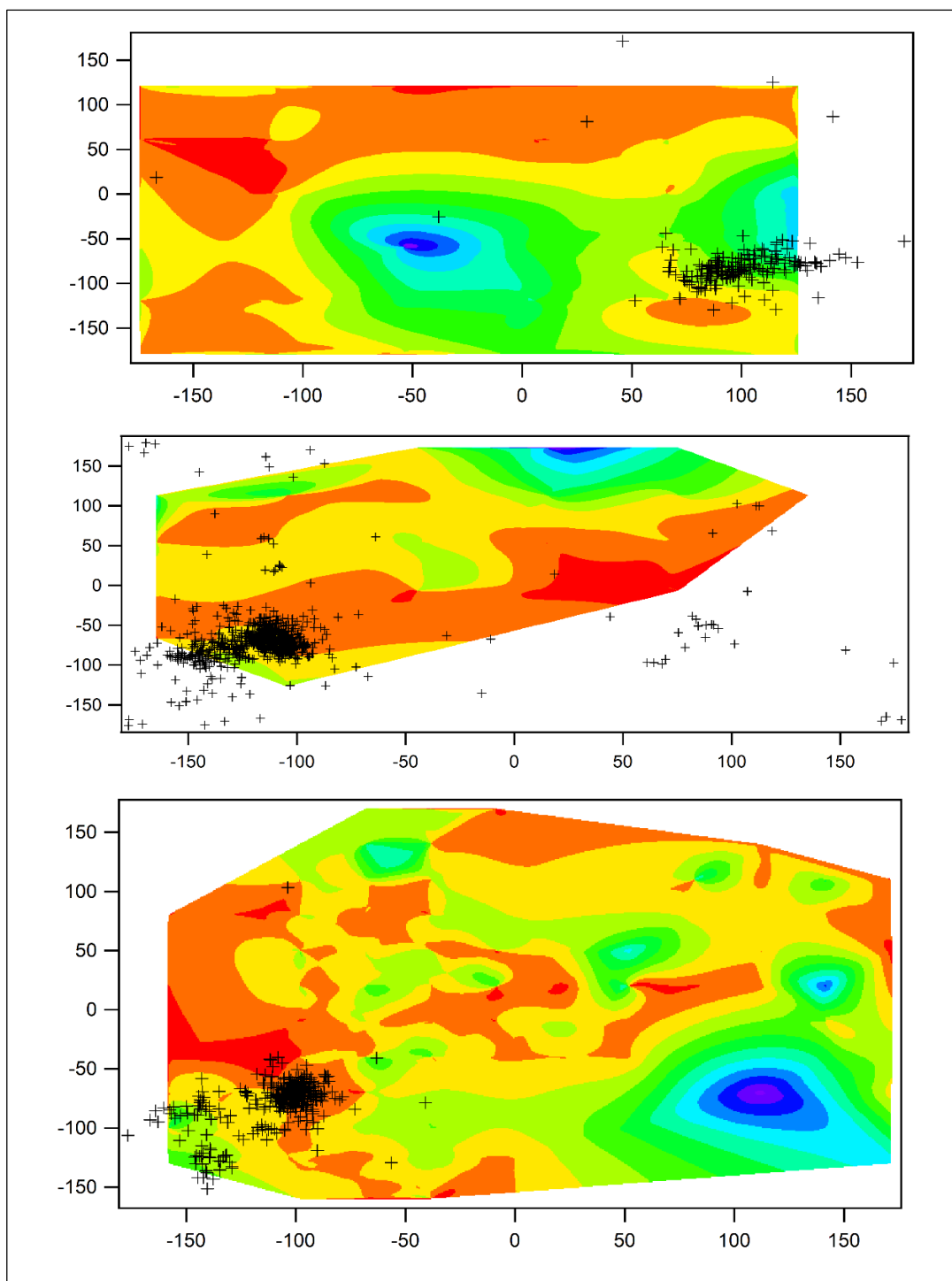


Figure 3.29: HBGA A Type 2 for HBGA-protein complex. The compare of psi and phi angles of 3 main torsion points of the generated conformers aligned to protein with crystal structure's psi and phi angles plotted with the relative energy of ligand-protein complex based on complex having the lowest SPE energy. Psi represents by horizontal axis (x) and Phi represents by horizontal axis (y). Last one shows the prediction for fucose unit.

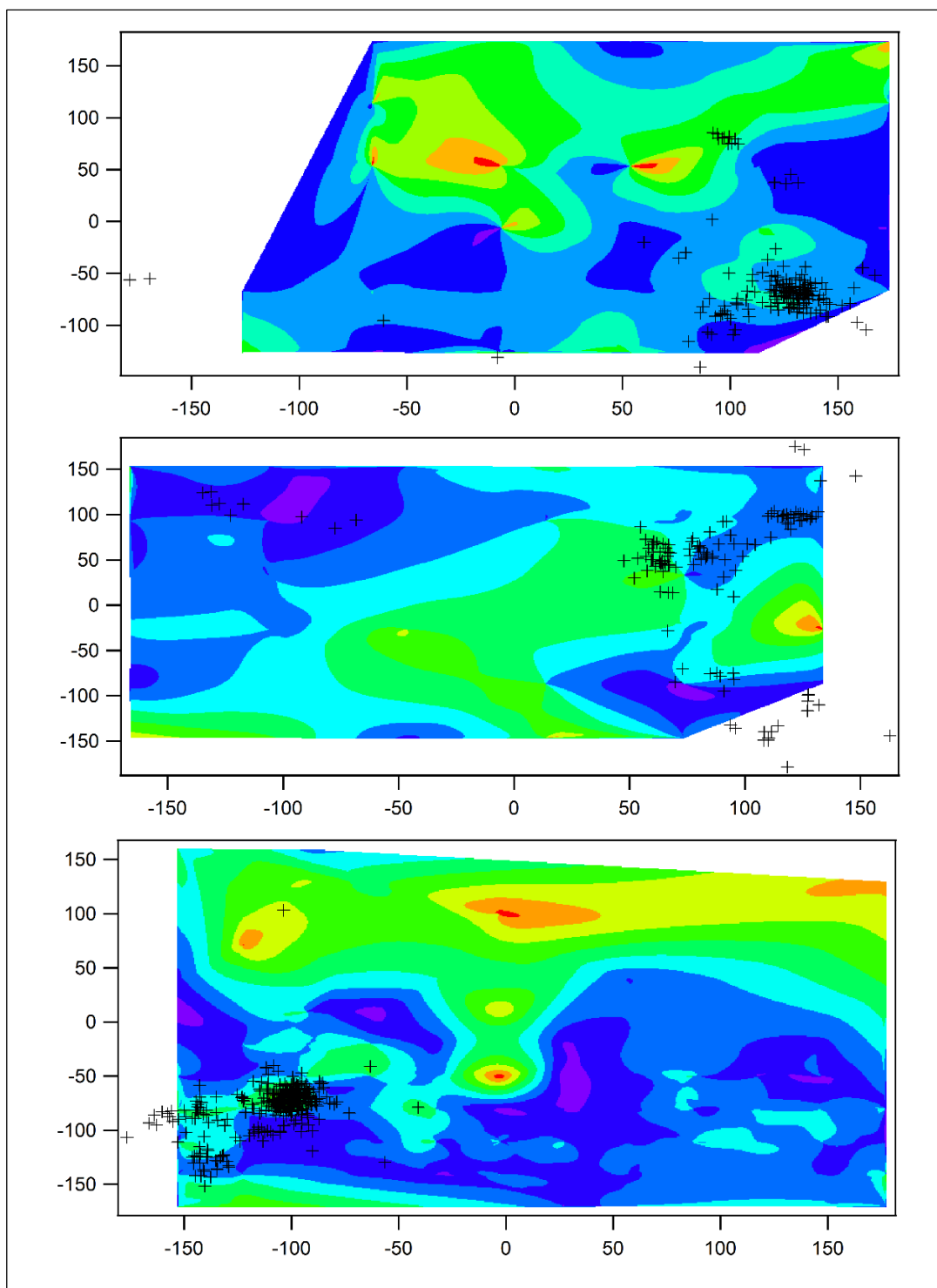


Figure 3.30: HBGA B Type 1 for only HBGA's SPE. The compare of psi and phi angles of 3 main torsion points of the generated conformers with crystal structure's psi and phi angles plotted with the relative energy of conformers based on conformer having the lowest SPE energy. Psi represents by horizontal axis (x) and Phi represents by horizontal axis (y). Last one shows the prediction for fucose unit.

In Figure 3.30 for HBGA B Type 1 in case of only ligand, there is no an acceptable output for all torsion points 1, 2 and 3. Particularly, the estimation for fucose unit was out of range of experimental. Except the reasons mentioned earlier, another one is that they differ from each other since while A type has 2 N-Acetylgalactosamine groups with one fucose unit in torsion constituents, B type has one N-Acetylgalactosamine group and galactosamine with that within torsion units. In short, a convenient result was not obtained for this type of HBGA in case of only ligand.

As shown Figure 3.31 (HBGA B Type 1 in case of ligand-protein complex), it is pointed out that similar to the case of HBGA B Type 1 for only ligand, outputs for complex have no a good sign for a correlation between calculated and experimental values. In short, for this type of HBGA in case of ligand-protein complex, prediction values calculated by ANI were non-significant to evaluate.

Figure 3.32 (HBGA B Type 2 in case of only ligand) illustrate that the failure of ANI results in calculating energies of conformers of HBGA at specific psi and phi angles of torsions and the predicted results against experimental values are very poor. In conclusion, there is also no good result for this one.

One of the best prediction results was witnessed between calculated and experimental values for torsion 1 in Figure 3.33 (HBGA B Type 2 in case of ligand-protein complex). In the contrary, it is hard to say the same thing for torsion 2 and torsion 3 in fucose unit. In short, the results calculated by ANI were obtained in acceptable range for torsion 1 but not for 2 and 3 points for this type of HBGA in case of ligand-protein complex

In summary, for both HBGA B Type 1 and 2 in the state of both ligand and ligand-complex, the results were not well correlated with experimental values whereas some calculated values were in the range of experimental ones, especially torsion 1 values in B type 2 ligand-protein complex and. Moreover, the result for torsion 1 of B Type 2 complex, ANI may still have a potential to use for calculation of conformers of complex structures.

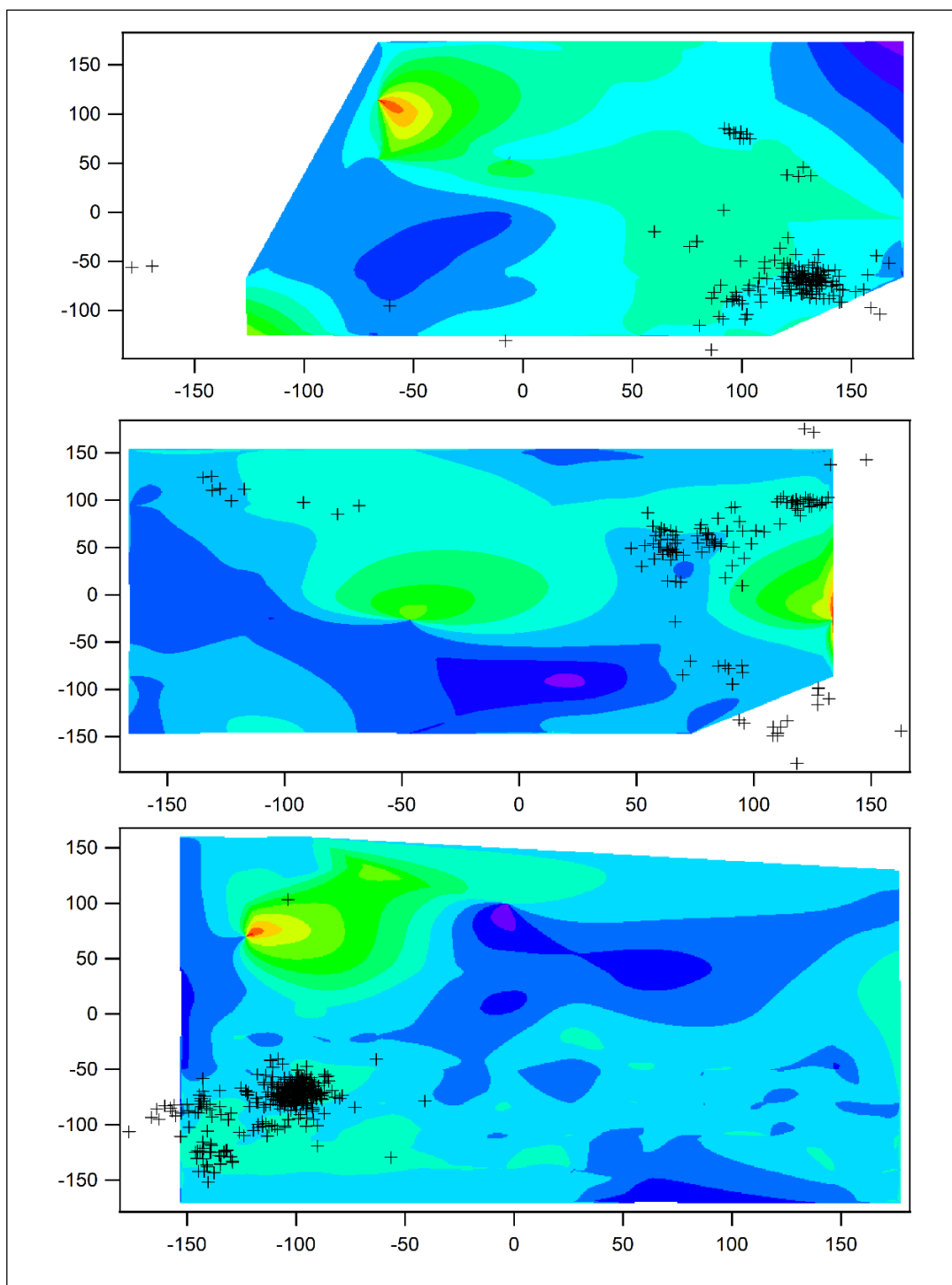


Figure 3.31: HBGA B Type 1 for HBGA-protein complex. The compare of psi and phi angles of 3 main torsion points of the generated conformers aligned to protein with crystal structure's psi and phi angles plotted with the relative energy of ligand-protein complex based on complex having the lowest SPE energy. Psi represents by horizontal axis (x) and Phi represents by horizontal axis (y). Last one shows the prediction for fucose unit.

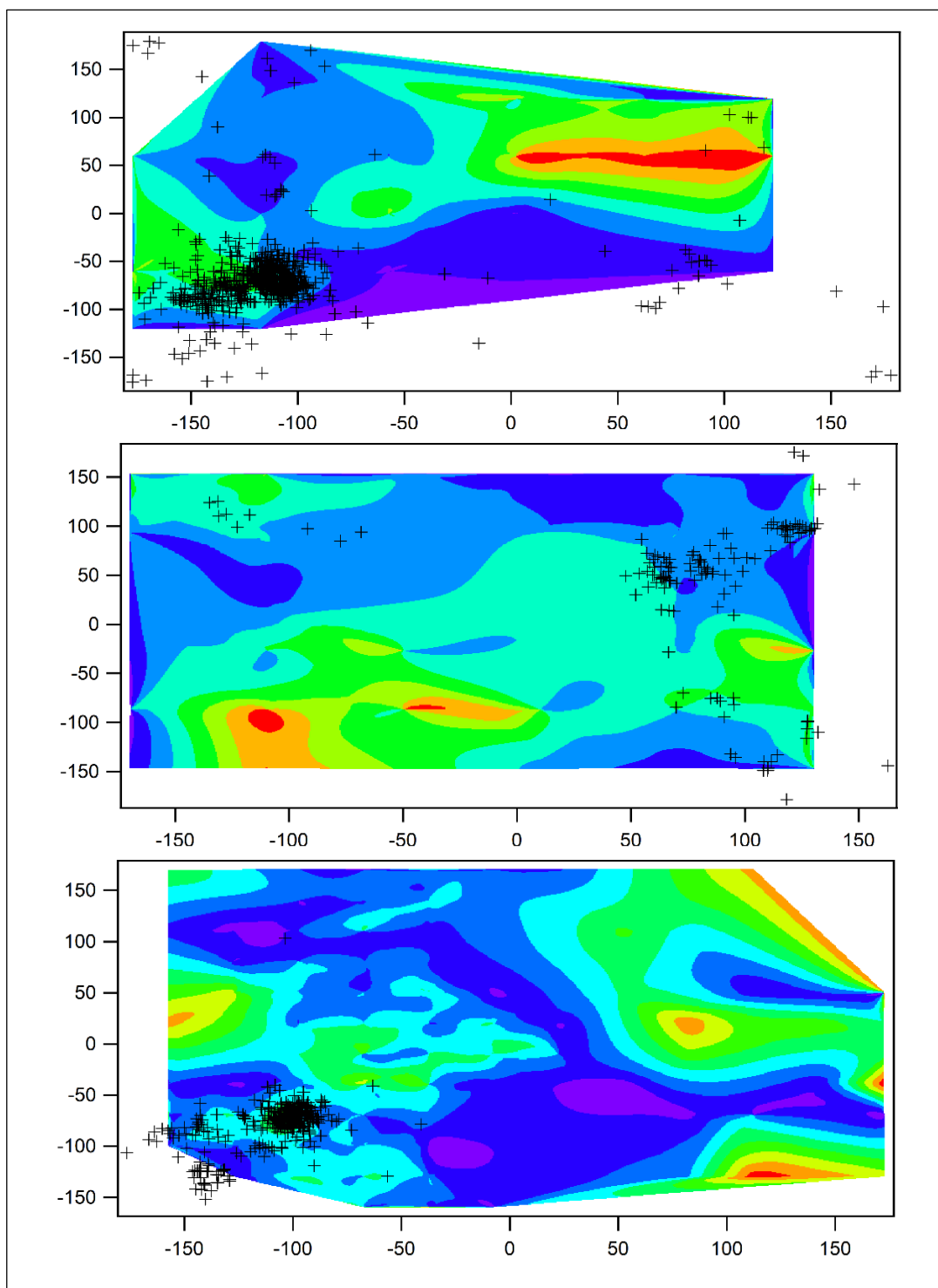


Figure 3.32: HBGA B Type 2 for only HBGA's SPE. The compare of psi and phi angles of 3 main torsion points of the generated conformers with crystal structure's psi and phi angles plotted with the relative energy of conformers based on conformer having the lowest SPE energy. Psi represents by horizontal axis (x) and Phi represents by horizontal axis (y). Last one shows the prediction for fucose unit.

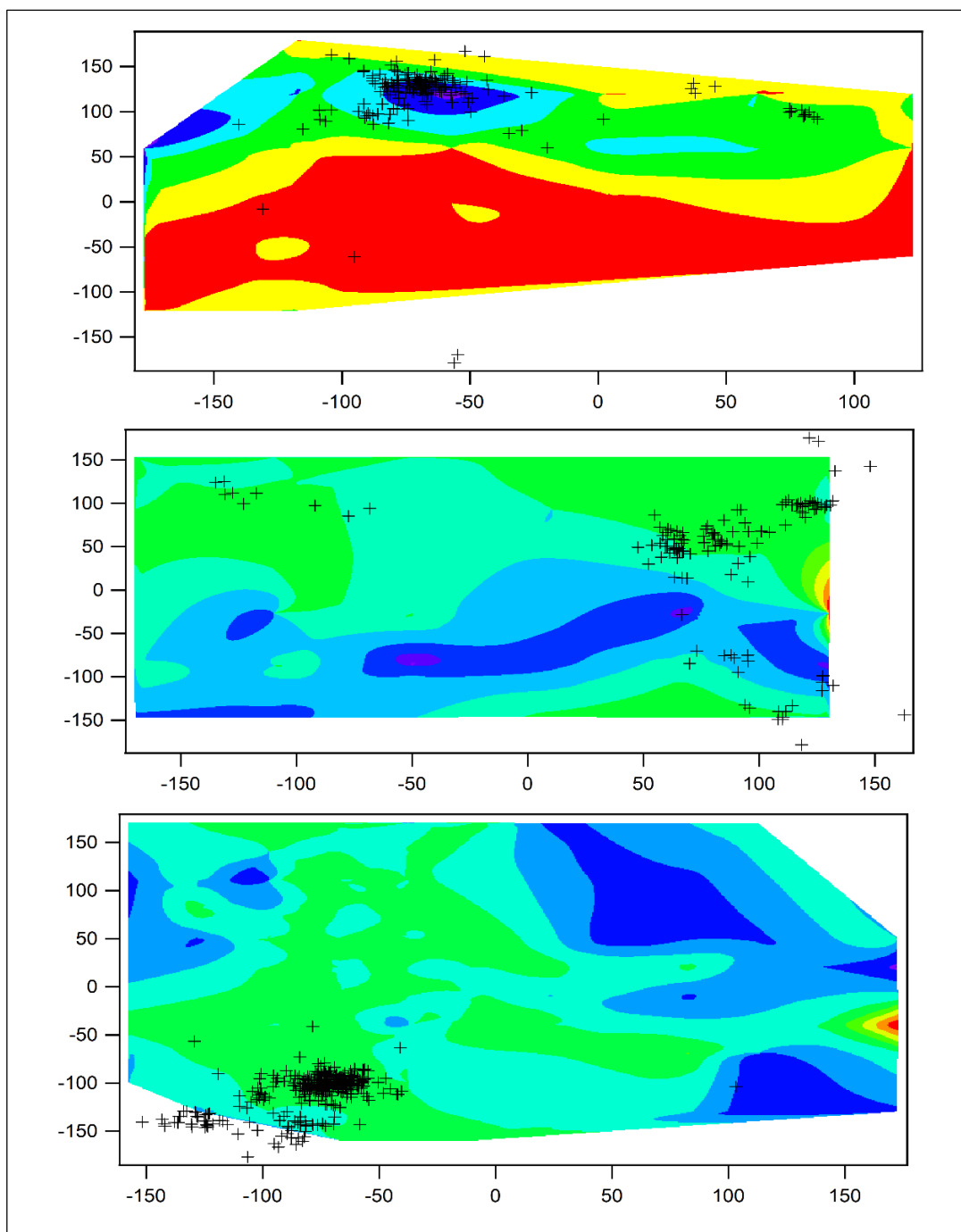


Figure 3.33: HBGA B Type 2 for HBGA-protein complex. The compare of psi and phi angles of 3 main torsion points of the generated conformers aligned to protein with crystal structure's psi and phi angles plotted with the relative energy of ligand-protein complex based on complex having the lowest SPE energy. Psi represents by horizontal axis (x) and Phi represents by horizontal axis (y). Last one shows the prediction for fucose unit.

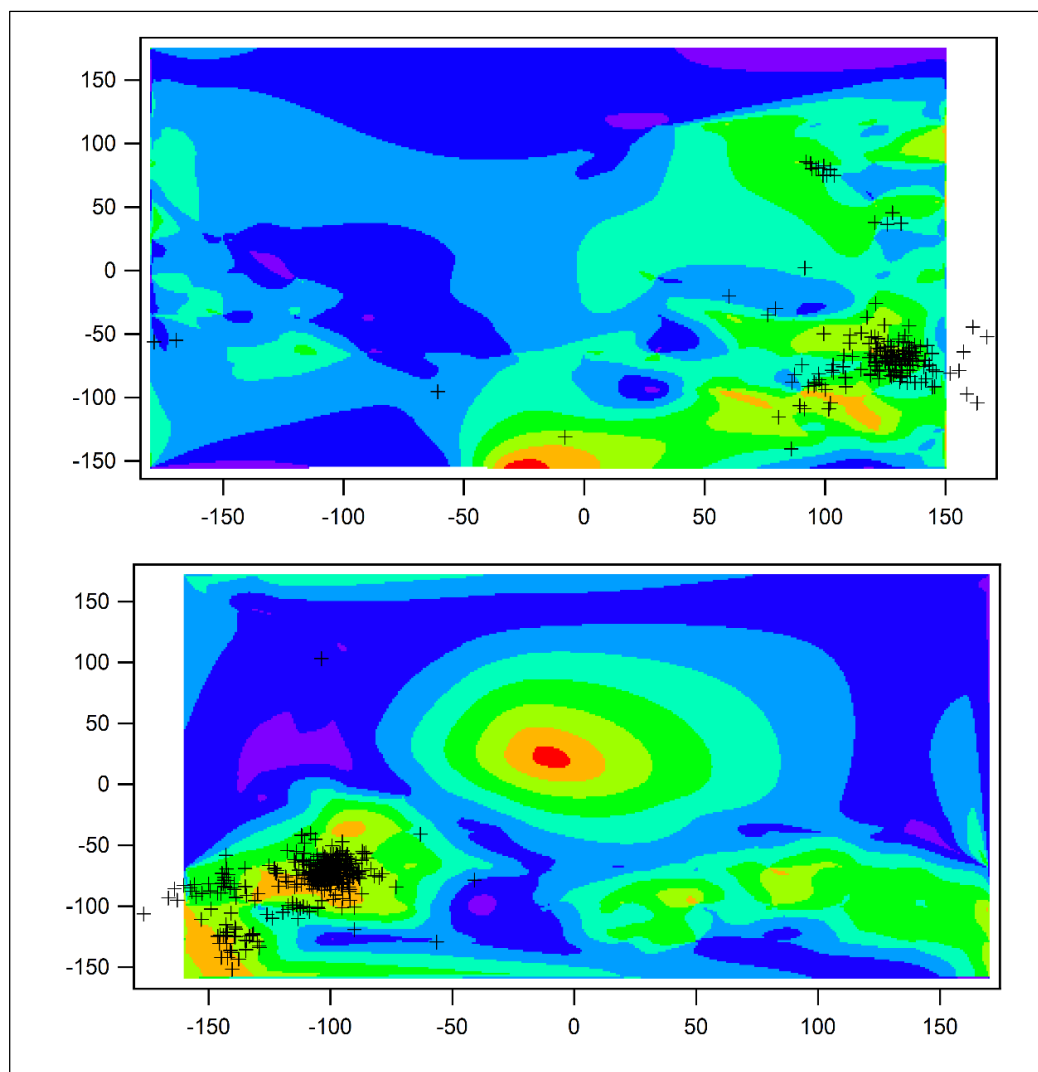


Figure 3.34: HBGA H Type 1 for only HBGA's SPE. The compare of psi and phi angles of 2 main torsion points of the generated conformers with crystal structure's psi and phi angles plotted with the relative energy of conformers based on conformer having the lowest SPE energy. Psi represents by horizontal axis (x) and Phi represents by horizontal axis (y). Last one shows the prediction for fucose unit.

There are two torsion points in HBGA H type in oppose to A type and B type and one of them is fucose unit and other is N-acetylgalactosamine. Figure 3.34 (HBGA H Type 1 in case of only ligand) shows that results for calculating energies of conformers of HBGA at specific psi and phi angles of torsions have been failed and the predicted results against experimental values are very poor. Nonetheless, as shown Figure 3.35 (HBGA H Type 1 in case of ligand-protein complex), the predicted results calculated by ANI show a close correlation with experimental values for torsion 1 and 2.

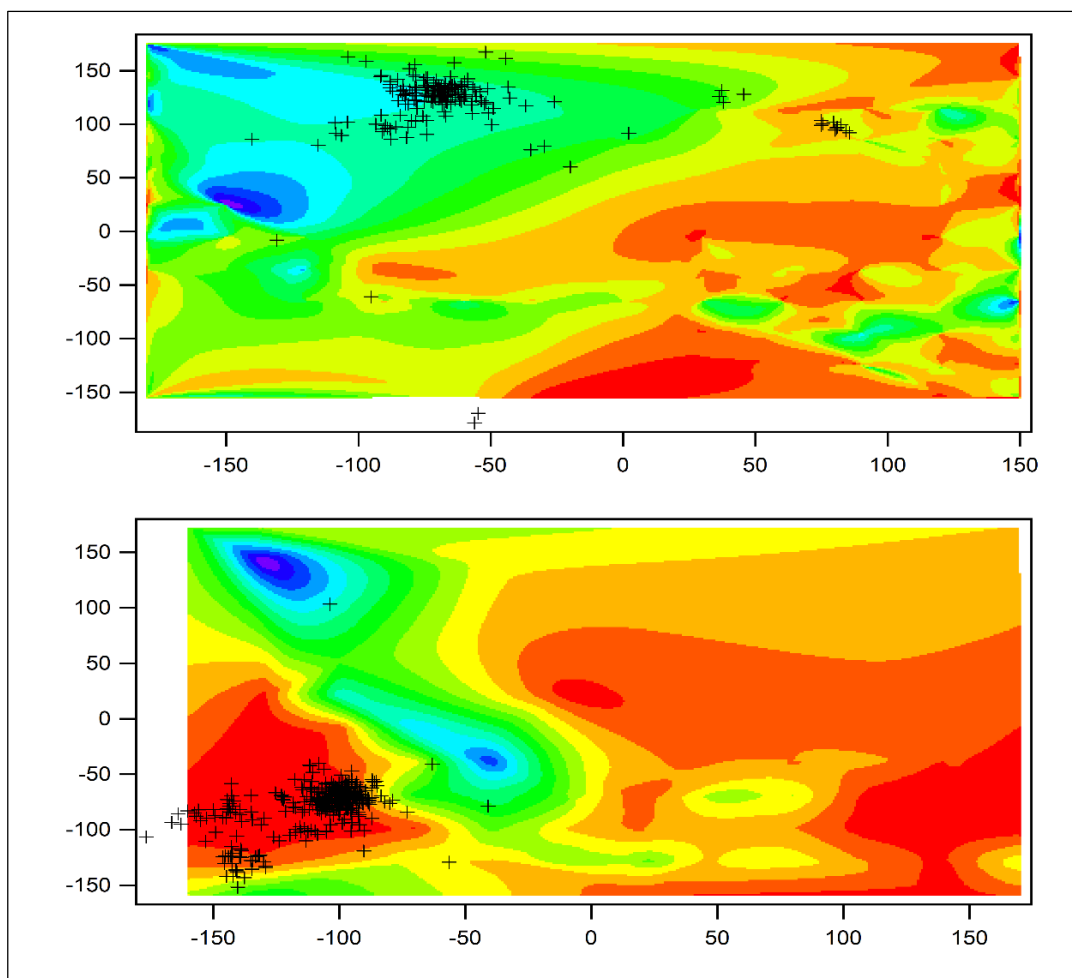


Figure 3.35: HBGA H Type 1 for HBGA-protein complex. The compare of psi and phi angles of 2 main torsion points of the generated conformers aligned to protein with crystal structure's psi and phi angles plotted with the relative energy of ligand-protein complex based on complex having the lowest SPE energy. Psi represents by horizontal axis (x) and Phi represents by horizontal axis (y). Last one shows the prediction for fucose unit.

In addition, the estimation for fucose unit is in a close range to the experimental values. In short, for this type of HBGA in case of ligand-protein complex, a good prediction values calculated by ANI were observed but not for the presence of only HBGA in environment.

Figure 3.36 (HBGA H Type 2 in case of only ligand) reveals that results for calculating energies of conformers of HBGA at specific psi and phi angles of both torsions (1 and 2) have been failed and the predicted results are very poor to compare experimental ones. Furthermore, similar results are observed for both torsion points in Figure 3.37 for HBGA H Type 2 in case of ligand-protein complex. In addition, there are some calculated points matched with experimental values but the almost all points

are out of the desired zone for torsion 1 and they are clustered in opposite angles. The prediction for fucose unit which has role in interaction is not in an acceptable range when compared to the experimental values. In conclusion, although there are some good outputs, failed ones are in excess for both torsions and ANI is not solely sufficient method to handle this by using as a direct method.

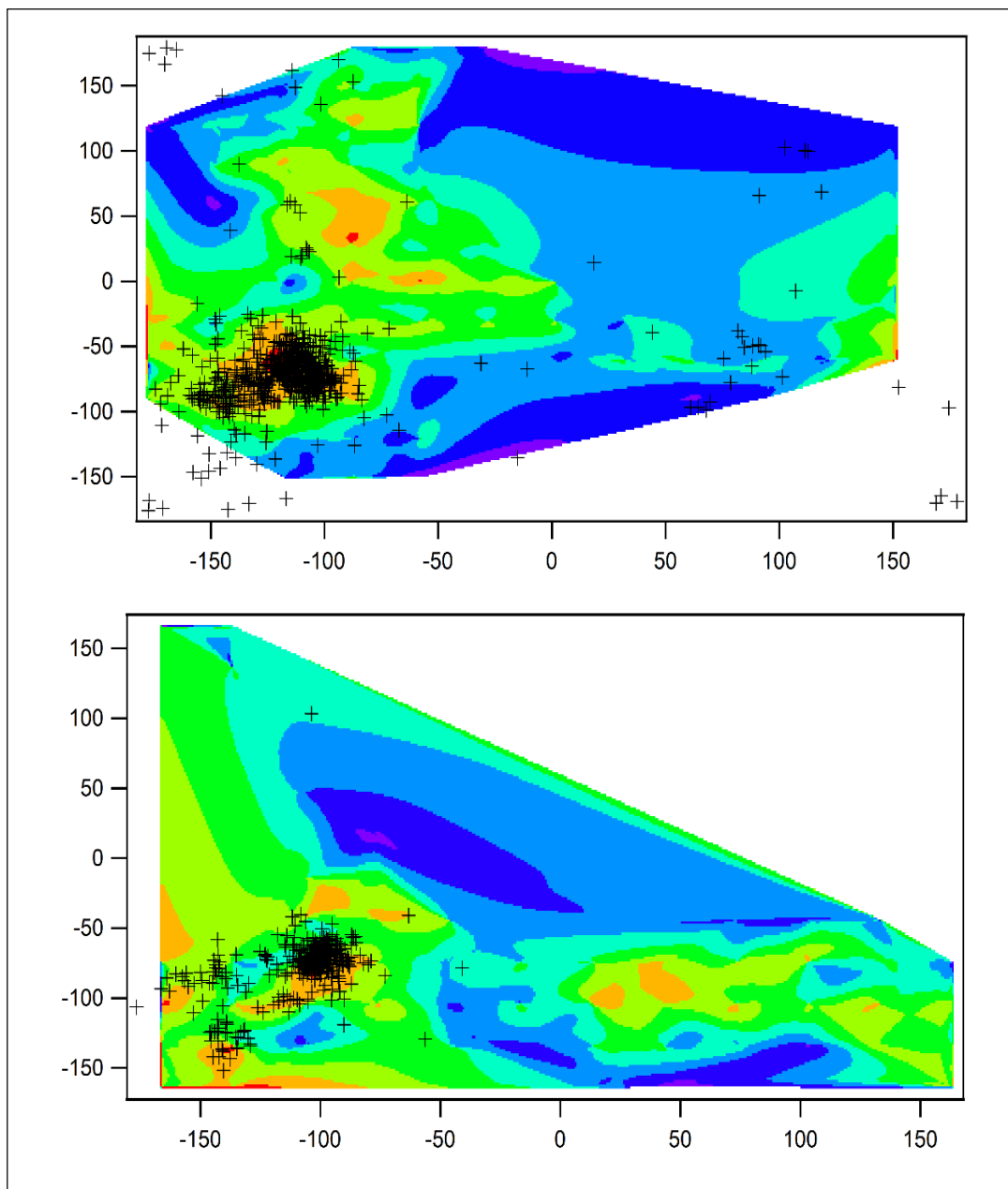


Figure 3.36: HBGA H Type 2 for only HBGA's SPE. The compare of psi and phi angles of 2 main torsion points of the generated conformers with crystal structure's psi and phi angles plotted with the relative energy of conformers based on conformer having the lowest SPE energy. Psi represents by horizontal axis (x) and Phi represents by horizontal axis (y). Last one shows the prediction for fucose unit.

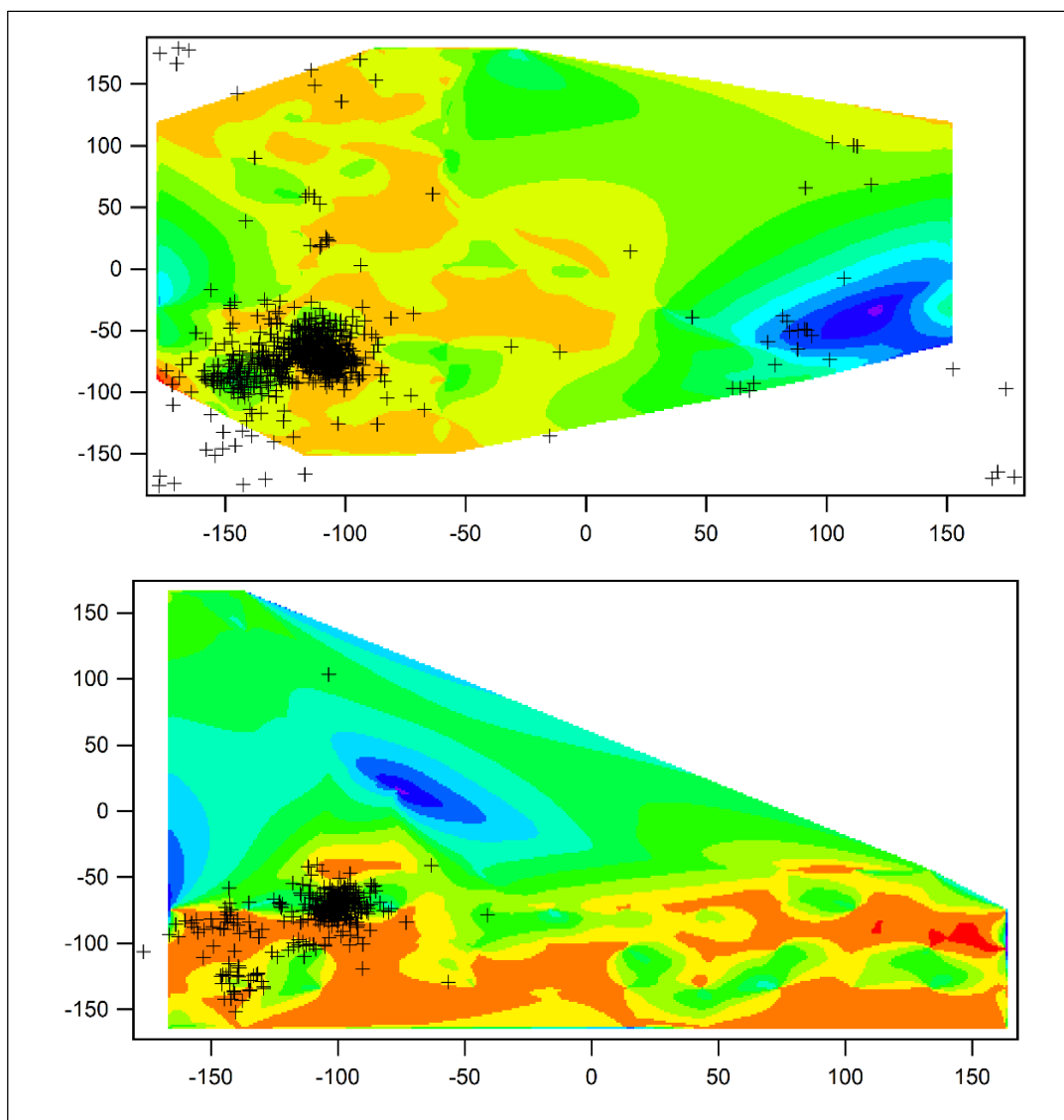


Figure 3.37: HBGA H Type 2 for HBGA-protein complex. The compare of psi and phi angles of 2 main torsion points of the generated conformers aligned to protein with crystal structure's psi and phi angles plotted with the relative energy of ligand-protein complex based on complex having the lowest SPE energy. Psi represents by horizontal axis (x) and Phi represents by horizontal axis (y). Last one shows the prediction for fucose unit.

In contrast to all HBGAs type, precursor type 1 and 2 have one torsion point and there is no fucose unit on them because they are the basic substructure of others. Moreover, they have galactosamine and N-acetylgalactosamine as sugar units and interacting units. In these kinds, the interaction with VP1 is carried out by N-acetylgalactosamine in oppose to others binding via fucose unit. In Figure 3.38,

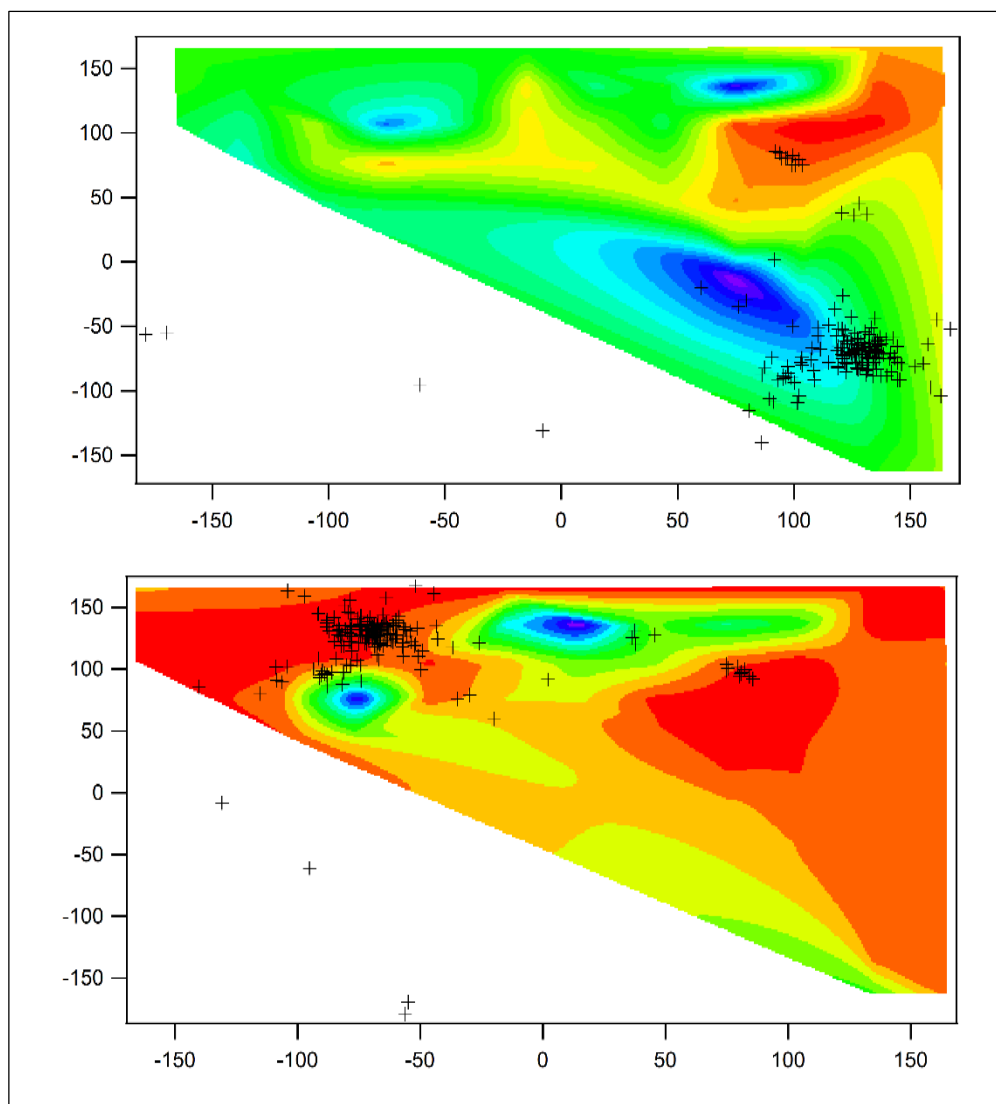


Figure 3.38: HBGA Precursor Type 1 for only HBGA SPE and ligand-protein complex, respectively. The compare of psi and phi angles of 1 main torsion point of the generated conformers with crystal structure's psi and phi angles plotted with the relative energy of conformers based on conformer having the lowest SPE energy. Second one shows the compare of psi and phi angles of 1 main torsion points of the generated conformers aligned to protein with crystal structure's psi and phi angles plotted with the relative energy of ligand-protein complex based on complex having the lowest SPE energy. Psi represents by horizontal axis (x) and Phi represents by horizontal axis (y).

the calculated values are not exactly but correlated with experimental values for both the presence of sole ligand and protein-ligand complex. In particular, the prediction for only ligand is better than protein-ligand complex.

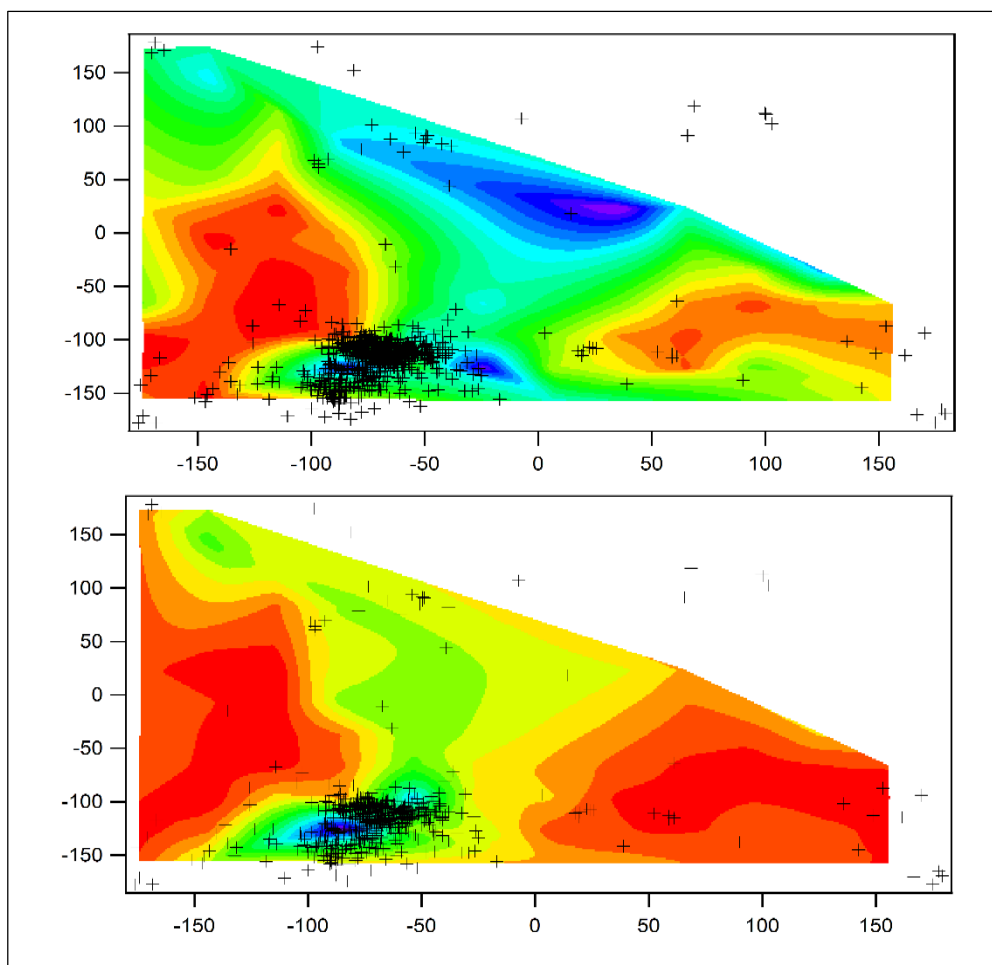


Figure 3.39: HBGA Precursor Type 2 for only HBGA SPE and ligand-protein complex, respectively. The compare of psi and phi angles of 1 main torsion point of the generated conformers with crystal structure's psi and phi angles plotted with the relative energy of conformers based on conformer having the lowest SPE energy. Second one shows the compare of psi and phi angles of 1 main torsion points of the generated conformers aligned to protein with crystal structure's psi and phi angles plotted with the relative energy of ligand-protein complex based on complex having the lowest SPE energy. Psi represents by horizontal axis (x) and Phi represents by horizontal axis (y).

As a result, it can be say that there is possible to use ANI for scanning dihedral angles and also conformers energies since as shown Figure 3.39, the results which were analyzed for precursor type 2 in case of both only ligand and protein-ligand complex supported that claim. The predicted results were highly compatible in both states: only ligand and protein-ligand complex. Also, in can be seen that many allowed areas (blue ones) were observed and one of the predicted outputs is in the range of experimental data. This indicates that the conformers and psi and phi angles belong to these in other zones can be allowed in protein and have an impact on interaction behaviour and mechanism.

To sum up, the conformers generation and scanning study show that ANI can probably be a potential candidate for enlightening the interaction energy and behaviour of protein-ligand complex with conformers. On the other hand, ANI may be coupled with other techniques to get better result like using as end-state method mentioned for solvation free energy calculation done above.

## 4. CONCLUSION

In this study, we have tested ANI's performance for systems involving noncovalent interactions and showed that these ML potentials which were developed can be utilized to calculate interaction energies at the Density Functional Theory level by simple linear scaling. We have also tested the performance of ANI for the estimation of hydration free energies of drug-like organic compounds, including particular element, using several end state methods applied on the trajectory of the MD simulations in which the ligand and solvent are fully interacting. As benchmarking studies, ANI predicted SFEs of small molecules were compared with other conventional methods such as LIE, MBAR and several continuum models defined in the DFT calculations. The results showed that ANI outperforms the MM based LIE. Although, the accuracy is less than MBAR and DFT-SMD model, the computational cost can be reduced tremendously. ML-based ANI as a next generation calculation method can take its place among other solvation free energy calculation methods in terms of accuracy and applicability as well as low cost and high speed.

Another applied area of ANI in this study, ML-based method, is to estimate crystal structure of human blood group antigens conformers in protein by help of calculating SPE energies of each conformer and conformers aligned protein complexes. Human Norovirus G.II-4 performs specifically interaction with one of HBGA's subunits known as fucose in general. Herein, we tried to shed light on the effect of main torsions in HBGAs along their interactions with VP1 by generating and sorting conformers and aligning to crystallized protein. Even though some acceptable outputs were obtained, previous graphs belonging to HBGAs and their torsion' impacts demonstrate that the results are not generally wanted in the range of interest. In general, the results are not well correlated with experimental crystal psi and phi values whereas some calculated values were in the range of experimental ones. Although not very encouraging results are obtained for overall secretor HBGA types, ANI may still have a potential to be utilized for calculation of conformers of complex structures by integrating with existed software. Only ANI is not sufficient to find crystal structures.

## REFERENCES

- [1] Salthammer T., Grimme S., Stahn M., Hohm U., Palm W.U., (2022), “Quantum Chemical Calculation And Evaluation Of Partition Coefficients For Classical And Emerging Environmentally Relevant Organic Compounds”, *Environmental Science and Technology*, 56 (1), 379–391.
- [2] Tomasi J., Mennucci B., Cammi R., (2005), “Quantum Mechanical Continuum Solvation Models”, *Chemical Reviews*, 105 (8), 2999–3093.
- [3] Skyner R.E., McDonagh J.L., Groom C.R., Van Mourik T., Mitchell J.B.O., (2015), “A Review Of Methods For The Calculation Of Solution Free Energies And The Modelling Of Systems In Solution”, *Physical Chemistry Chemical Physics*, 17 (9), 6174–6191.
- [4] Lewars E.G., (2016), “An Outline Of What Computational Chemistry Is All About BT - Computational Chemistry: Introduction To The Theory And Applications Of Molecular And Quantum Mechanics”, 2<sup>nd</sup> Ed., Springer International Publishing.
- [5] Wang X., Li Y., Gao Y., Yang Z., Lu C., Zhu T., (2018), “A Quantum Mechanical Computational Method For Modeling Electrostatic And Solvation Effects Of Protein”, *Scientific Reports*, 8 (1), 1–10.
- [6] Zou X., Yaxiong Kuntz I.D., (1999), “Inclusion Of Solvation In Ligand Binding Free Energy Calculations Using The Generalized-Born Model”, *Journal of the American Chemical Society*, 121 (35), 8033–8043.
- [7] Lipinski C.A., Lombardo F., Dominy B.W., Feeney P.J., (1997), “Experimental And Computational Approaches To Estimate Solubility And Permeability In Drug Discovery And Development Settings”, *Advanced Drug Delivery Reviews*, 23 (1), 3–25.
- [8] Corbett P.T., Leclaire J., Vial L., West K.R., Wietor J.L., Sanders J.K.M., Otto S., (2006), “Dynamic Combinatorial Chemistry”, *Chemical Reviews*, 106 (9), 3652–3711.
- [9] Jorgensen W.L., Duffy E.M., (2002), “Prediction Of Drug Solubility From Structure”, *Advanced Drug Delivery Reviews*, 54 (3), 355–366.
- [10] Onsager L., (1936), “Electric Moments Of Molecules In Liquids”, *Journal of the American Chemical Society*, 58 (8), 1486–1493.
- [11] Mehler E.L., Solmajer T., (1991), “Electrostatic Effects In Proteins: Comparison Of Dielectric And Charge Models”, *Protein Engineering, Design and Selection*, 4 (8), 903–910.
- [12] Gilson M.K., Sharp K.A., Honig B.H., (1988), “Calculating The Electrostatic Potential Of Molecules In Solution: Method And Error Assessment”, *Journal of Computational Chemistry*, 9 (4), 327–335.

- [13] Genheden S., Mikulskis P., Hu L., Kongsted J., Söderhjelm P., Ryde U., (2011), "Accurate Predictions Of Nonpolar Solvation Free Energies Require Explicit Consideration Of Binding-Site Hydration", *Journal of the American Chemical Society*, 133 (33), 13081–13092.
- [14] Anisimov V.M., Cavasotto C.N., (2011), "Hydration Free Energies Using Semiempirical Quantum Mechanical Hamiltonians And A Continuum Solvent Model With Multiple Atomic-Type Parameters", *The Journal of Physical Chemistry B*, 115 (24), 7896–7905.
- [15] Marenich A.V., Cramer C.J., Truhlar D.G., (2008), "Perspective On Foundations Of Solvation Modeling: The Electrostatic Contribution To The Free Energy Of Solvation", *Journal of Chemical Theory and Computation*, 4 (6), 877–887.
- [16] Corbeil C.R., Sulea T., Purisima E.O., (2010), "Rapid Prediction Of Solvation Free Energy. 2. The First-Shell Hydration (FiSH) Continuum Model", *Journal of Chemical Theory and Computation*, 6 (5), 1622–1637.
- [17] Eisenberg D., McLachlan A.D., (1986), "Solvation Energy In Protein Folding And Binding", *Nature*, 319 (6050), 199–203.
- [18] Boyer R.D., Bryan R.L., (2012), "Fast Estimation Of Solvation Free Energies For Diverse Chemical Species", *The Journal of Physical Chemistry B*, 116 (12), 3772–3779.
- [19] Sergiievskiy V.P., Fedorov M.V., (2012), "3DRISM Multigrid Algorithm For Fast Solvation Free Energy Calculations", *Journal of Chemical Theory and Computation*, 8 (6), 2062–2070.
- [20] Setny P., Zacharias M., (2010), "Hydration In Discrete Water. A Mean Field, Cellular Automata Based Approach To Calculating Hydration Free Energies", *The Journal of Physical Chemistry B*, 114 (26), 8667–8675.
- [21] Bernazzani L., Duce C., Micheli A., Mollica V., Tiné M.R., (2010), "Quantitative Structure–Property Relationship (QSPR) Prediction Of Solvation Gibbs Energy Of Bifunctional Compounds By Recursive Neural Networks", *Journal of Chemical & Engineering Data*, 55 (12), 5425–5428.
- [22] 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.
- [23] Marenich A.V., Cramer C.J., Truhlar D.G., (2009), "Performance Of SM6, SM8, And SMD On The SAMPL1 Test Set For The Prediction Of Small-Molecule Solvation Free Energies", *The Journal of Physical Chemistry B*, 113 (14), 4538–4543.
- [24] Choi H., Kang H., Park H., (2013), "New Solvation Free Energy Function Comprising Intermolecular Solvation And Intramolecular Self-Solvation Terms", *Journal of Cheminformatics*, 5 (1), 8.

- [25] Ratkova E.L., Palmer D.S., Fedorov M.V., (2015), “Solvation Thermodynamics Of Organic Molecules By The Molecular Integral Equation Theory: Approaching Chemical Accuracy”, *Chemical Reviews*, 115 (13), 6312–6356.
- [26] Kamerlin S.C.L., Haranczyk M., Warshel A., (2009), “Are Mixed Explicit/Implicit Solvation Models Reliable For Studying Phosphate Hydrolysis? A Comparative Study Of Continuum, Explicit And Mixed Solvation Models”, *ChemPhysChem*, 10 (7), 1125–1134.
- [27] Cancès E., Mennucci B., Tomasi J., (1997), “A New Integral Equation Formalism For The Polarizable Continuum Model: Theoretical Background And Applications To Isotropic And Anisotropic Dielectrics”, *The Journal of Chemical Physics*, 107 (8), 3032–3041.
- [28] Cossi M., Rega N., Scalmani G., Barone V., (2003), “Energies, Structures, And Electronic Properties Of Molecules In Solution With The C-PCM Solvation Model”, *Journal of Computational Chemistry*, 24 (6), 669–681.
- [29] Miertuš S., Scrocco E., Tomasi J., (1981), “Electrostatic Interaction Of A Solute With A Continuum. A Direct Utilizaion Of AB Initio Molecular Potentials For The Prevision Of Solvent Effects”, *Chemical Physics*, 55 (1), 117–129.
- [30] Marenich A.V., Olson R.M., Kelly C.P., Cramer C.J., Truhlar D.G., (2007), “Self-Consistent Reaction Field Model For Aqueous And Nonaqueous Solutions Based On Accurate Polarized Partial Charges”, *Journal of Chemical Theory and Computation*, 3 (6), 2011–2033.
- [31] Marenich A.V., Cramer C.J., Truhlar D.G., (2013), “Generalized Born Solvation Model SM12”, *Journal of Chemical Theory and Computation*, 9 (1), 609–620.
- [32] Cramer C.J., Truhlar D.G., (1991), “General Parameterized SCF Model For Free Energies Of Solvation In Aqueous Solution”, *Journal of the American Chemical Society*, 113 (22), 8305–8311.
- [33] Engelage E., Schulz N., Heinen F., Huber S.M., Truhlar D.G., Cramer C.J., (2018), “Refined SMD Parameters For Bromine And Iodine Accurately Model Halogen-Bonding Interactions In Solution”, *Chemistry – A European Journal*, 24 (60), 15983–15987.
- [34] Mu W.H., Chasse G.A., Fang D.C., (2008), “Test And Modification Of The Van Der Waals’ Radii Employed In The Default PCM Model”, *International Journal of Quantum Chemistry*, 108 (9), 1422–1434.
- [35] Marenich A.V., Cramer C.J., Truhlar D.G., (2009), “Universal Solvation Model Based On Solute Electron Density And On A Continuum Model Of The Solvent Defined By The Bulk Dielectric Constant And Atomic Surface Tensions”, *Journal of Physical Chemistry B*, 113 (18), 6378–6396.
- [36] Mirzaei S., Ivanov M.V., Timerghazin Q.K., (2019), “Improving Performance Of The SMD Solvation Model: Bondi Radii Improve Predicted Aqueous Solvation Free Energies Of Ions And PKa Values Of Thiols”, *The Journal of*

Physical Chemistry A, 123 (44), 9498–9504.

- [37] Thapa B., Schlegel H.B., (2016), “Density Functional Theory Calculation Of PKa’s Of Thiols In Aqueous Solution Using Explicit Water Molecules And The Polarizable Continuum Model”, *The Journal of Physical Chemistry A*, 120 (28), 5726–5735.
- [38] Orozco M., Luque F.J., (2000), “Theoretical Methods For The Description Of The Solvent Effect In Biomolecular Systems”, *Chemical Reviews*, 100 (11), 4187–4226.
- [39] Guillot B., (2002), “A Reappraisal Of What We Have Learnt During Three Decades Of Computer Simulations On Water”, *Journal of Molecular Liquids*, 101 (1), 219–260.
- [40] Mahoney M.W., Jorgensen W.L., (2000), “Diffusion Constant Of The TIP5P Model Of Liquid Water”, *The Journal of Chemical Physics*, 114 (1), 363–366.
- [41] Berendsen H.J.C., Grigera J.R., Straatsma T.P., (1987), “The Missing Term In Effective Pair Potentials”, *The Journal of Physical Chemistry*, 91 (24), 6269–6271.
- [42] Florová P., Sklenovský P., Banáš P., Otyepka M., (2010), “Explicit Water Models Affect The Specific Solvation And Dynamics Of Unfolded Peptides While The Conformational Behavior And Flexibility Of Folded Peptides Remain Intact”, *Journal of Chemical Theory and Computation*, 6 (11), 3569–3579.
- [43] Shepler B., (2006), “Introduction To Computational Quantum Chemistry Definition Of Computational Chemistry”,.
- [44] Chen S., Raugei S., Rousseau R., Dupuis M., Bullock R.M., (2010), “Homogeneous Ni Catalysts For H<sub>2</sub> Oxidation And Production: An Assessment Of Theoretical Methods, From Density Functional Theory To Post Hartree–Fock Correlated Wave-Function Theory”, *The Journal of Physical Chemistry A*, 114 (48), 12716–12724.
- [45] Raghavachari K., Trucks G.W., Pople J.A., Head-Gordon M., (1989), “A Fifth-Order Perturbation Comparison Of Electron Correlation Theories”, *Chemical Physics Letters*, 157 (6), 479–483.
- [46] Sherrill C.D., (2010), “Frontiers In Electronic Structure Theory”, *The Journal of Chemical Physics*, 132 (11), 110902.
- [47] Sousa S.F., Fernandes P.A., Ramos M.J., (2007), “General Performance Of Density Functionals”, *The Journal of Physical Chemistry A*, 111 (42), 10439–10452.
- [48] Kohn W., Sham L.J., (1965), “Self-Consistent Equations Including Exchange And Correlation Effects”, *Physical Review*, 140 (4A), A1133–A1138.
- [49] Becke A.D., (2014), “Perspective: Fifty Years Of Density-Functional Theory In

Chemical Physics”, The Journal of Chemical Physics, 140 (18), 18A301.

- [50] Grimme S., (2011), “Density Functional Theory With London Dispersion Corrections”, WIREs Computational Molecular Science, 1 (2), 211–228.
- [51] Grimme S., Antony J., Ehrlich S., Krieg H., (2010), “A Consistent And Accurate Ab Initio Parametrization Of Density Functional Dispersion Correction (DFT-D) For The 94 Elements H-Pu”, The Journal of Chemical Physics, 132 (15), 154104.
- [52] Medvedev M.G., Bushmarinov I.S., Sun J., Perdew J.P., Lyssenko A.K., (2017), “Density Functional Theory Is Straying From The Path Toward The Exact Functional”, Science, 355 (6320), 49–52.
- [53] Cárdenas R., Martínez-Seoane J., Amero C., (2020), “Combining Experimental Data And Computational Methods For The Non-Computer Specialist”, Molecules , 25 (20). 2020.
- [54] Salmaso V., Moro S., (2018), “Bridging Molecular Docking To Molecular Dynamics In Exploring Ligand-Protein Recognition Process: An Overview ”, Frontiers in Pharmacology, 9, 923.
- [55] Maximova T., Moffatt R., Ma B., Nussinov R., Shehu A., (2016), “Principles And Overview Of Sampling Methods For Modeling Macromolecular Structure And Dynamics”, PLOS Computational Biology, 12 (4), e1004619.
- [56] Traoré S., Allouche D., André I., de Givry S., Katsirelos G., Schiex T., Barbe S., (2013), “A New Framework For Computational Protein Design Through Cost Function Network Optimization”, Bioinformatics, 29 (17), 2129–2136.
- [57] Hehre W.J., (2003), “A Guide To Molecular Mechanics And Quantum Chemical Calculations”, 2003.
- [58] Leach A.R., (2001), “Molecular Modelling : Principles and Applications”, 2<sup>nd</sup> ed., Pearson Education Limited.
- [59] 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.
- [60] Monticelli L., Tieleman D.P., (2013), “Force Fields For Classical Molecular Dynamics BT - Biomolecular Simulations: Methods And Protocols”, vol 924, Humana Press.
- [61] Lyubartsev A.P., Rabinovich A.L., (2016), “Force Field Development For Lipid Membrane Simulations”, Biochimica et Biophysica Acta (BBA) - Biomembranes, 1858 (10), 2483–2497.
- [62] Alder B.J., Wainwright T.E., (1959), “Studies In Molecular Dynamics. I. General Method”, The Journal of Chemical Physics, 31 (2), 459–466.

- [63] Hess B., Kutzner C., van der Spoel D., Lindahl E., (2008), “GROMACS 4: Algorithms For Highly Efficient, Load-Balanced, And Scalable Molecular Simulation”, *Journal of Chemical Theory and Computation*, 4 (3), 435–447.
- [64] Salomon-Ferrer R., Case D.A., Walker R.C., (2013), “An Overview Of The Amber Biomolecular Simulation Package”, *WIREs Computational Molecular Science*, 3 (2), 198–210.
- [65] Phillips J.C., Hardy D.J., Maia J.D.C., Stone J.E., Ribeiro J.V., Bernardi R.C., Buch R., Fiorin G., Hénin J., Jiang W., McGreevy R., Melo M.C.R., Radak B.K. et al., (2020), “Scalable Molecular Dynamics On CPU And GPU Architectures With NAMD”, *The Journal of Chemical Physics*, 153 (4), 44130.
- [66] De Vivo M., Masetti M., Bottegoni G., Cavalli A., (2016), “Role Of Molecular Dynamics And Related Methods In Drug Discovery”, *Journal of Medicinal Chemistry*, 59 (9), 4035–4061.
- [67] Limongelli V., (2020), “Ligand Binding Free Energy And Kinetics Calculation In 2020”, *WIREs Computational Molecular Science*, 10 (4), e1455.
- [68] Aqvist J., Medina C., Samuelsson J.E., (1994), “A New Method For Predicting Binding Affinity In Computer-Aided Drug Design”, *Protein engineering*, 7 (3), 385–391.
- [69] Åqvist J., Luzhkov V.B., Brandsdal B.O., (2002), “Ligand Binding Affinities From MD Simulations”, *Acc. Chem. Res.*, 35, 358.
- [70] Su Y., Gallicchio E., Das K., Arnold E., Levy R.M., (2007), “Linear Interaction Energy (LIE) Models For Ligand Binding In Implicit Solvent: Theory And Application To The Binding Of NNRTIs To HIV-1 Reverse Transcriptase”, *Journal of Chemical Theory and Computation*, 3 (1), 256–277.
- [71] Cruz J., Wickstrom L., Yang D., Gallicchio E., Deng N., (2020), “Combining Alchemical Transformation With A Physical Pathway To Accelerate Absolute Binding Free Energy Calculations Of Charged Ligands To Enclosed Binding Sites”, *ACS Applied Materials and Interfaces*, 16 (4), 2803–2813.
- [72] Sitkoff D., Sharp K.A., Honig B., (1994), “Accurate Calculation Of Hydration Free Energies Using Macroscopic Solvent Models”, *Journal of Physical Chemistry*, 98 (7), 1978–1988.
- [73] Kirkwood J.G., (1935), “Statistical Mechanics Of Fluid Mixtures”, *The Journal of Chemical Physics*, 3 (5), 300–313.
- [74] Zwanzig R.W., (1954), “High-Temperature Equation Of State By A Perturbation Method. I. Nonpolar Gases”, *The Journal of Chemical Physics*, 22 (8), 1420–1426.
- [75] Bennett J.M., Catovsky D., Daniel M.T., Flandrin G., Galton D.A.G., Gralnick H.R., Sultan C., (1976), “Proposals For The Classification Of The Acute Leukaemias French-American-British (FAB) Co-Operative Group”, *British*

Journal of Haematology, 33 (4), 451–458.

- [76] Jarzynski C., (1997), “Nonequilibrium Equality For Free Energy Differences”, *Physical Review Letters*, 78 (14), 2690–2693.
- [77] Kim J.T., Hamilton A.D., Bailey C.M., Domoal R.A., Wang L., Anderson K.S., Jorgensen W.L., (2006), “FEP-Guided Selection Of Bicyclic Heterocycles In Lead Optimization For Non-Nucleoside Inhibitors Of HIV-1 Reverse Transcriptase”, *Journal of the American Chemical Society*, 128 (48), 15372–15373.
- [78] Widom B., (1963), “Some Topics In The Theory Of Fluids”, *The Journal of Chemical Physics*, 39 (11), 2808–2812.
- [79] 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”, *The Journal of Chemical Physics*, 122 (14), 144107.
- [80] Dellago C., Hummer G., (2014), “Computing Equilibrium Free Energies Using Non-Equilibrium Molecular Dynamics”, *Entropy*, 16 (1), 41–61.
- [81] Jarzynski C., (2006), “Rare Events And The Convergence Of Exponentially Averaged Work Values”, *Physical Review E*, 73 (4), 46105.
- [82] Chris O.F., van G.W., (2005), “Free Energies Of Ligand Binding For Structurally Diverse Compounds”, *Proceedings of the National Academy of Sciences*, 102 (19), 6750–6754.
- [83] Shirts M.R., Mobley D.L., Brown S.P., (2010), “Free-Energy Calculations In Structure-Based Drug Design”, 1<sup>st</sup> ed., Cambridge University Press.
- [84] Adamovic I., Gordon M.S., (2005), “Dynamic Polarizability, Dispersion Coefficient C6 And Dispersion Energy In The Effective Fragment Potential Method”, *Molecular Physics*, 103 (2–3), 379–387.
- [85] Oostenbrink C., Van Gunsteren W.F., (2003), “Single-Step Perturbations To Calculate Free Energy Differences From Unphysical Reference States: Limits On Size, Flexibility, And Character”, *Journal of Computational Chemistry*, 24 (14), 1730–1739.
- [86] Steinbrecher T., Case D.A., Labahn A., (2006), “A Multistep Approach To Structure-Based Drug Design: Studying Ligand Binding At The Human Neutrophil Elastase”, *Journal of Medicinal Chemistry*, 49 (6), 1837–1844.
- [87] 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.
- [88] van der Kamp M.W., Mulholland A.J., (2013), “Combined Quantum Mechanics/Molecular Mechanics (QM/MM) Methods In Computational Enzymology”, *Biochemistry*, 52 (16), 2708–2728.

- [89] Lonsdale R., Harvey J.N., Mulholland A.J., (2012), “A Practical Guide To Modelling Enzyme-Catalysed Reactions”, *Chemical Society Reviews*, 41 (8), 3025–3038.
- [90] Friesner R.A., Guallar V., (2004), “Ab Initio Quantum Chemical and Mixed Quantum Mechanics/Molecular Mechanics (QM/MM) Methods For Studying Enzymatic Catalysis”, *Annual Review of Physical Chemistry*, 56 (1), 389–427.
- [91] Siegbahn P.E.M., Himo F., (2011), “The Quantum Chemical Cluster Approach For Modeling Enzyme Reactions”, *WIREs Computational Molecular Science*, 1 (3), 323–336.
- [92] Woods C.J., Manby F.R., Mulholland A.J., (2008), “An Efficient Method For The Calculation Of Quantum Mechanics/Molecular Mechanics Free Energies”, *The Journal of Chemical Physics*, 128 (1), 14109.
- [93] Mireia G.V., Jiali G., Martin K.G., T.D., (2004), “How Enzymes Work: Analysis By Modern Rate Theory And Computer Simulations”, *Science*, 303 (5655), 186–195.
- [94] Papamichael E.M., Stamatis H., Stergiou P.Y., Foukis A., Gkini O.A., (2019), “Advances in Enzyme Technology”, 3<sup>rd</sup> ed., Elsevier B.V.
- [95] Guo W.W., Zhang T.S., Fang W.H., Cui G., (2018), “QM/MM Studies On The Excited-State Relaxation Mechanism Of A Semisynthetic DTPT3 Base”, *Physical Chemistry Chemical Physics*, 20 (7), 5067–5073.
- [96] Goldman B.B., Walters W.P., (2006), “Chemoinformatics for Drug Discovery y”, 1<sup>st</sup> ed., Elsevier.
- [97] Westermayr J., Gastegger M., Schütt K.T., Maurer R.J., (2021), “Perspective On Integrating Machine Learning Into Computational Chemistry And Materials Science”, *The Journal of Chemical Physics*, 154 (23), 230903.
- [98] Rupp M., (2015), “Machine Learning For Quantum Mechanics In A Nutshell”, *International Journal of Quantum Chemistry*, 115 (16), 1058–1073.
- [99] 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.
- [100] Behler J., (2014), “Representing Potential Energy Surfaces By High-Dimensional Neural Network Potentials”, *Journal of Physics: Condensed Matter* 26 (18), 183001.
- [101] Fedorov D.G., Kitaura K., (2004), “The Importance Of Three-Body Terms In The Fragment Molecular Orbital Method”, *The Journal of Chemical Physics*, 120 (15), 6832–6840.
- [102] Cavasotto C.N., Aucar M.G., Adler N.S., (2019), “Computational Chemistry In Drug Lead Discovery And Design”, *International Journal of Quantum*

Chemistry, 119 (2), e25678.

- [103] Elstner M., (2006), “The SCC-DFTB Method And Its Application To Biological Systems”, *Theoretical Chemistry Accounts*, 116 (1), 316–325.
- [104] Stewart J.J.P., (2009), “Application Of The PM6 Method To Modeling Proteins”, *Journal of Molecular Modeling*, 15 (7), 765–805.
- [105] Glielmo A., Sollich P., De Vita A., (2017), “Accurate Interatomic Force Fields Via Machine Learning With Covariant Kernels”, *Physical Review B*, 95 (21), 214302.
- [106] Botu V., Batra R., Chapman J., Ramprasad R., (2017), “Machine Learning Force Fields: Construction, Validation, And Outlook”, *The Journal of Physical Chemistry C*, 121 (1), 511–522.
- [107] Lubbers N., Smith J.S., Barros K., (2018), “Hierarchical Modeling Of Molecular Energies Using A Deep Neural Network”, *The Journal of Chemical Physics*, 148 (24), 241715.
- [108] Gassner H., Probst M., Lauenstein A., Hermansson K., (1998), “Representation Of Intermolecular Potential Functions By Neural Networks”, *The Journal of Physical Chemistry A*, 102 (24), 4596–4605.
- [109] Roman Z.S., Smith S.J., Leszczynski J., Isayev O., (2022), “Accurate And Transferable Multitask Prediction Of Chemical Properties With An Atoms-In-Molecules Neural Network”, *Science Advances*, 5 (8), eaav6490.
- [110] Gastegger M., Schwiedrzik L., Bittermann M., Berzsényi F., Marquetand P., (2018), “WACSF—Weighted Atom-Centered Symmetry Functions As Descriptors In Machine Learning Potentials”, *The Journal of Chemical Physics*, 148 (24), 241709.
- [111] Halgren T.A., (1996), “Merck Molecular Force Field. I. Basis Form Scope, Parameterization, And Performance Of MMFF94”, *Journal of Computational Chemistry*, 17 (5–6), 490–519.
- [112] Maier J.A., Martinez C., Kasavajhala K., Wickstrom L., Hauser K.E., Simmerling C., (2015), “Ff14SB: Improving The Accuracy Of Protein Side Chain And Backbone Parameters From Ff99SB”, *Journal of Chemical Theory and Computation*, 11 (8), 3696–3713.
- [113] van Duin A.C.T., Dasgupta S., Lorant F., Goddard W.A., (2001), “ReaxFF: A Reactive Force Field For Hydrocarbons”, *The Journal of Physical Chemistry A*, 105 (41), 9396–9409.
- [114] 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.
- [115] Chmiela S., Tkatchenko A., Sauceda H.E., Poltavsky I., Schütt K.T., Müller K.R., (2022), “Machine Learning Of Accurate Energy-Conserving Molecular

Force Fields”, *Science Advances*, 3 (5), e1603015.

- [116] Schütt K.T., Sauceda H.E., Kindermans P.J., Tkatchenko A., Müller K.R., (2018), “SchNet – A Deep Learning Architecture For Molecules And Materials”, *The Journal of Chemical Physics*, 148 (24), 241722.
- [117] Thompson A.P., Swiler L.P., Trott C.R., Foiles S.M., Tucker G.J., (2015), “Spectral Neighbor Analysis Method For Automated Generation Of Quantum-Accurate Interatomic Potentials”, *Journal of Computational Physics*, 285, 316–330.
- [118] Devereux C., Smith J.S., Huddleston 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.
- [119] Manzhos S., Carrington T., (2006), “A Random-Sampling High Dimensional Model Representation Neural Network For Building Potential Energy Surfaces”, *The Journal of Chemical Physics*, 125 (8), 84109.
- [120] Schütt K.T., Arbabzadah F., Chmiela S., Müller K.R., Tkatchenko A., (2017), “Quantum-Chemical Insights From Deep Tensor Neural Networks”, *Nature Communications*, 8 (1), 13890.
- [121] Kolb B., Zhao B., Li J., Jiang B., Guo H., (2016), “Permutation Invariant Potential Energy Surfaces For Polyatomic Reactions Using Atomistic Neural Networks”, *The Journal of Chemical Physics*, 144 (22), 224103.
- [122] 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.
- [123] Häse F., Valteau S., Pyzer-Knapp E., Aspuru-Guzik A., (2016), “Machine Learning Exciton Dynamics”, *Chemical Science*, 7 (8), 5139–5147.
- [124] Rupp M., Ramakrishnan R., von Lilienfeld O.A., (2015), “Machine Learning For Quantum Mechanical Properties Of Atoms In Molecules”, *The Journal of Physical Chemistry Letters*, 6 (16), 3309–3313.
- [125] Behler J., (2015), “Constructing High-Dimensional Neural Network Potentials: A Tutorial Review”, *International Journal of Quantum Chemistry*, 115 (16), 1032–1050.
- [126] Behler J., Parrinello M., (2007), “Generalized Neural-Network Representation Of High-Dimensional Potential-Energy Surfaces”, *Physical Review Letters*, 98 (14), 146401.
- [127] 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.

- [128] Smith J.S., Nebgen B.T., Zubatyuk R., Lubbers N., Devereux C., Barros K., Tretiak S., Isayev O., Roitberg A.E., (2019), “Approaching Coupled Cluster Accuracy With A General-Purpose Neural Network Potential Through Transfer Learning”, *Nature Communications*, 10 (1), 2903.
- [129] Grisafi A., Fabrizio A., Meyer B., Wilkins D.M., Corminboeuf C., Ceriotti M., (2019), “Transferable Machine-Learning Model Of The Electron Density”, *ACS Central Science*, 5 (1), 57–64.
- [130] Chmiela S., Sauceda H.E., Müller K.R., Tkatchenko A., (2018), “Towards Exact Molecular Dynamics Simulations With Machine-Learned Force Fields”, *Nature Communications*, 9 (1), 3887.
- [131] Smith J.S., Zubatyuk R., Nebgen B., Lubbers N., Barros K., Roitberg A.E., Isayev O., Tretiak S., (2020), “The ANI-1ccx And ANI-1x Data Sets, Coupled-Cluster And Density Functional Theory Properties For Molecules”, *Scientific Data*, 7 (1), 134.
- [132] Woolhouse M., Scott F., Hudson Z., Howey R., Chase-Topping M., (2012), “Human Viruses: Discovery And Emergence”, *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367 (1604), 2864–2871.
- [133] 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.
- [134] Yildiz M., Kocak A., (2018), “Molecular Dynamics Studies Of Histo-Blood Group Antigen Blocking Human Immunoglobulin A Antibody And Escape Mechanism In Noroviruses Upon Mutation”, *Journal of Computational Biology*, 26 (9), 962–974.
- [135] Saito M., Tsukagoshi H., Ishigaki H., Aso J., Ishii H., Okayama K., Ryo A., Ishioka T., Kuroda M., Saruki N., Katayama K., Kimura H., Topin J., et al., (2018), “Structural Biology Of Noroviruses”, *Journal of Virology*, 9 (9), 1–11.
- [136] Saito M., Tsukagoshi H., Ishigaki H., Aso J., Ishii H., Okayama K., Ryo A., Ishioka T., Kuroda M., Saruki N., Katayama K., Kimura H., (2020), “Molecular Evolution Of The Capsid (VP1) Region In Human Norovirus Genogroup II Genotype 3”, *Heliyon*, 6 (5), e03835.
- [137] Ozaki K., Matsushima Y., Nagasawa K., Motoya T., Ryo A., Kuroda M., Katayama K., Kimura H., (2018), “Molecular Evolutionary Analyses Of The RNA-Dependent RNA Polymerase Region In Norovirus Genogroup II”, *Frontiers in Microbiology*, 9 (December), 1–11.
- [138] Shanker S., Choi J.M., Sankaran B., Atmar R.L., Estes M.K., Prasad B.V.V., (2011), “Structural Analysis Of Histo-Blood Group Antigen Binding Specificity In A Norovirus GII.4 Epidemic Variant: Implications For Epochal Evolution”, *Journal of Virology*, 85 (17), 8635–8645.
- [139] Scallan E., Hoekstra R.M., Angulo F.J., Tauxe R.V., Widdowson M.A., Roy S.L., Jones J.L., Griffin P.M., (2011), “Foodborne Illness Acquired In The

United States—Major Pathogens”, *Emerging Infectious Disease journal*, 17 (1), 7.

- [140] Esseili M.A., Gao X., Boley P., Hou Y., Saif L.J., Brewer-Jensen P., Lindesmith L.C., Baric R.S., Atmar R.L., Wang Q., (2019), “Human Norovirus Histo-Blood Group Antigen (HBGA) Binding Sites Mediate The Virus Specific Interactions With Lettuce Carbohydrates”, *Viruses*, 11 (9).
- [141] Petronella N., Ronholm J., Suresh M., Harlow J., Mykytczuk O., Corneau N., Bidawid S., Nasheri N., (2018), “Genetic Characterization Of Norovirus GII.4 Variants Circulating In Canada Using A Metagenomic Technique”, *BMC Infectious Diseases*, 18 (1), 1–11.
- [142] Taube S., Jiang M., Wobus C.E., (2010), “Glycosphingolipids As Receptors For Non-Enveloped Viruses”, *Viruses*, 2 (4), 1011–1049.
- [143] Barbé L., Le Moullac-Vaidye B., Echasserieau K., Bernardeau K., Carton T., Bovin N., Nordgren J., Svensson L., Ruvoën-Clouet N., Le Pendu J., (2018), “Histo-Blood Group Antigen-Binding Specificities Of Human Rotaviruses Are Associated With Gastroenteritis But Not With In Vitro Infection”, *Scientific Reports*, 8 (1), 1–14.
- [144] Reeck A., Kavanagh O., Estes M.K., Opekun A.R., Gilger M.A., Graham D.Y., Atmar R.L., (2010), “Serological Correlate Of Protection Against Norovirus-Induced Gastroenteritis”, *The Journal of Infectious Diseases*, 202 (8), 1212–1218.
- [145] Campillay-Véliz C.P., Carvajal J.J., Avellaneda A.M., Escobar D., Covián C., Kalergis A.M., Lay M.K., (2020), “Human Norovirus Proteins: Implications In The Replicative Cycle, Pathogenesis, And The Host Immune Response”, *Frontiers in Immunology*, 11.
- [146] Weichert S., Koromyslova A., Singh B.K., Hansman S., Jennewein S., Schrotten H., Hansman G.S., (2016), “Structural Basis For Norovirus Inhibition By Human Milk Oligosaccharides”, *Journal of Virology*, 90 (9), 4843–4848.
- [147] Mallory M.L., Lindesmith L.C., Graham R.L., Baric R.S., (2019), “GII.4 Human Norovirus: Surveying The Antigenic Landscape”, *Viruses*, 11 (2), 2019.
- [148] Gandhi K.M., Mandrell R.E., Tian P., (2010), “Binding Of Virus-Like Particles Of Norwalk Virus To Romaine Lettuce Veins=”, *Applied and Environmental Microbiology*, 76 (24), 7997–8003.
- [149] Morozov V., Hanisch F.G., Wegner K.M., Schrotten H., (2018), “Pandemic GII.4 Sydney And Epidemic GII.17 Kawasaki308 Noroviruses Display Distinct Specificities For Histo-Blood Group Antigens Leading To Different Transmission Vector Dynamics In Pacific Oysters”, *Frontiers in Microbiology*, 9 (NOV), 1–9.
- [150] Bu W., Mamedova A., Tan M., Xia M., Jiang X., Hegde R.S., (2008), “Structural Basis For The Receptor Binding Specificity Of Norwalk Virus”,

Journal of virology, 82 (11), 5340–5347.

- [151] Woods J.R., (2018), “Predicting The Structures Of Glycans, Glycoproteins, And Their Complexes”, *Chemical Reviews*, 118 (17), 8005–8024.
- [152] Su Y., Gallicchio E., Das K., Arnold E., Levy R.M., (2007), “Linear Interaction Energy (LIE) Models For Ligand Binding In Implicit Solvent: Theory And Application To The Binding Of NNRTIs To HIV-1 Reverse Transcriptase”, *Journal of Chemical Theory and Computation*, 3 (1), 256–277.
- [153] Aqvist J., Medina C., Samuelsson J.E., (1994), “A New Method For Predicting Binding Affinity In Computer-Aided Drug Design”, *Protein Engineering*, 7 (3), 385–391.
- [154] Hummer G., (2001), “Fast-Growth Thermodynamic Integration: Error And Efficiency Analysis”, *The Journal of Chemical Physics*, 114 (17), 7330–7337.
- [155] Kofke D.A., (2006), “On The Sampling Requirements For Exponential-Work Free-Energy Calculations”, *Molecular Physics*, 104 (22–24), 3701–3708.
- [156] Zwanzig R., (1983), “From Classical Dynamics To Continuous Time Random Walks”, *Journal of Statistical Physics*, 30 (2), 255–262.
- [157] Hummer G., Szabo A., (1996), “Calculation Of Free-energy Differences From Computer Simulations Of Initial And Final States”, *The Journal of Chemical Physics*, 105 (5), 2004–2010.
- [158] 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.
- [159] Mobley D.L., Guthrie J.P., (2014), “FreeSolv: A Database Of Experimental And Calculated Hydration Free Energies, With Input Files”, *Journal of Computer-Aided Molecular Design*, 28 (7), 711–720.
- [160] 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.
- [161] Shivakumar D., Deng Y., Roux B., (2009), “Computations Of Absolute Solvation Free Energies Of Small Molecules Using Explicit And Implicit Solvent Model”, *Journal of Chemical Theory and Computation*, 5 (4), 919–930.
- [162] Darden T., York D., Pedersen L., (1993), “Particle Mesh Ewald: An N·log(N) Method For Ewald Sums In Large Systems”, *The Journal of Chemical Physics*, 98 (12), 10089–10092.
- [163] Besler B.H., Merz Jr. K.M., Kollman P.A., (1990), “Atomic Charges Derived From Semiempirical Methods”, *Journal of Computational Chemistry*, 11 (4),

- [164] Singh U.C., Kollman P.A., (1984), “An Approach To Computing Electrostatic Charges For Molecules”, *Journal of Computational Chemistry*, 5 (2), 129–145.
- [165] 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.
- [166] Barone V., Cossi M., (1998), “Quantum Calculation Of Molecular Energies And Energy Gradients In Solution By A Conductor Solvent Model”, *The Journal of Physical Chemistry A*, 102 (11), 1995–2001.
- [167] Miertuš S., Tomasi J., (1982), “Approximate Evaluations Of The Electrostatic Free Energy And Internal Energy Changes In Solution Processes”, *Chemical Physics*, 65 (2), 239–245.
- [168] Marenich A.V., Cramer C.J., Truhlar D.G., (2009), “Universal Solvation Model Based On Solute Electron Density And On A Continuum Model Of The Solvent Defined By The Bulk Dielectric Constant And Atomic Surface Tensions”, *The Journal of Physical Chemistry B*, 113 (18), 6378–6396.
- [169] Pople J.A., Gill P.M.W., Johnson B.G., (1992), “Kohn—Sham Density-Functional Theory Within A Finite Basis Set”, *Chemical Physics Letters*, 199 (6), 557–560.
- [170] Hohenberg P., Kohn W., (1964), “Inhomogeneous Electron Gas”, *Physical Review*, 136 (3B), B864–B871.
- [171] 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.
- [172] Boresch S., Woodcock H.L., (2017), “Convergence Of Single-Step Free Energy Perturbation”, *Molecular Physics*, 115 (9–12), 1200–1213.
- [173] Carlson H.A., Jorgensen W.L., (1995), “An Extended Linear Response Method For Determining Free Energies Of Hydration”, *The Journal of Physical Chemistry*, 99 (26), 10667–10673.
- [174] Rifai E.A., van Dijk M., Geerke D.P., (2020), “Recent Developments In Linear Interaction Energy Based Binding Free Energy Calculations”, *Frontiers in Molecular Biosciences*, 7, 114.
- [175] Shirts M.R., Chodera J.D., (2008), “Statistically Optimal Analysis Of Samples From Multiple Equilibrium States”, *The Journal of Chemical Physics*, 129 (12), 124105.

## **BIOGRAPHY**

Mutesir TEMEL finished Marmara Primary School in Istanbul and after that he graduated from Turk Kizilayi Kartal Anatolian High School at 2012 as a first. He continued my education life at at the department of Molecular Biology and Genetics at Bogazici University and graduated at 2018. With following times, Mutesir applied for Master Degree to Gebze Technical University for Chemistry Department and is still here. During his master degree, he had a chance to experience different areas of Chemistry. To sum up, in his opinion, Mutesir TEMEL got a lot of knowledge about different process, techniques and methods to apply and combine for producing more beneficial results.

## **APPENDICES**

### **Appendix A: Thesis Supporting Project**

This thesis and its calculations were supported by Tubitak ULAKBIM and numerical calculations reported in this study were partially performed at TUBITAK ULAKBIM, 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).