

T.R.
GEBZE TECHNICAL UNIVERSITY
GRADUATE SCHOOL

TARGETING FDA DRUGS AGAINST SARS-COV-2 MAIN PROTEASE
USING FREE ENERGY METHODS

İREM NUR ZENGİN

A THESIS OF MASTER OF SCIENCE
DEPARTMENT OF CHEMISTRY

ADVISOR: ASSOC. PROF. DR. ABDULKADİR KOÇAK

APRIL 2025

T.R.
GEBZE TECHNICAL UNIVERSITY
GRADUATE SCHOOL

**TARGETING FDA DRUGS AGAINST SARS-COV-2
MAINPROTEASE USING FREE ENERGY METHODS**

İREM NUR ZENGİN

A THESIS OF MASTER OF SCIENCE
DEPARTMENT OF CHEMISTRY

ADVISOR: ASSOC. PROF. DR. ABDULKADİR KOÇAK

APRIL 2025

T.C.
GEBZE TEKNİK ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

FDA İLAÇLARININ SERBEST ENERJİ YÖNTEMLERİ
KULLANILARAK SARS-COV-2 ANA PROTEAZA
KARŞI HEDEFLENMESİ

İREM NUR ZENGİN

YÜKSEK LİSANS TEZİ
KİMYA ANABİLİM DALI

DANIŞMAN: DOÇ. DR. ABDULKADİR KOÇAK

NİSAN 2025

MASTER of SCIENCE JURY APPROVAL FORM

A thesis submitted by İrem Nur ZENGİN, defended on 25/04/2025 before the jury formed with the 21/04/2025 date and 2025/23 numbered decision of the GTU Graduate Administration Board, has been accepted as a MASTER of SCIENCE thesis in the Department of Chemistry

JURY

MEMBER

(THESIS ADVISOR): Assoc. Prof. Dr. Abdulkadir KOÇAK

MEMBER : Prof. Dr. İbrahim Fazıl ŞENGÜL

MEMBER : Assist. Prof. Dr. İsmail Erol

APPROVAL

Gebze Technical University Graduate Administration Board

...../...../..... date and/..... numbered decision.

SIGNATURE/SEAL

To my Family

ABSTRACT

The COVID-19 pandemic, caused by SARS-CoV-2, has impacted global health systems and economies. In response, researchers are exploring drug repurposing as a strategy to quickly find potential treatments, given the long time and high costs involved in developing new drugs from scratch. Drug repurposing involves re-targeting existing FDA-approved drugs, which already have extensive safety and efficacy data. This approach is particularly useful for rapidly spreading diseases like COVID-19.

In this study, nearly 2,500 FDA-approved drugs will be screened through virtual screening techniques, including molecular docking, molecular dynamics simulations, and free energy calculations, to identify promising candidates for COVID-19 treatment.

We first evaluate ANI, a machine learning potential with quantum mechanics accuracy, to rescore poses generated by GOLD docking software for known inhibitors of the SARS-CoV-2 main protease (Mpro). Next, we apply high-throughput virtual screening of FDA-approved drugs against Mpro and its common variants (beta, delta, omicron) using ANI and GOLD. The results from various binding energy calculations will guide experimental efforts to find effective COVID-19 therapeutics. After all these results, we think that the initial structure of the molecules can also change the docking and simulation results, so that we introduce a low energy conformer generation algorithm using ANI-ML potentials at DFT accuracy and benchmark in reproducing bioactive conformations. And we show that the method is efficient when the initial starting structure is far from equilibrium, when the ML potentials are stuck in non-smooth regions, and when the quality of the conformers in a less conformer size is demanded.

Keywords: Drug Repurposing, Machine Learning, Docking, Molecular Dynamics, Free Energy.

ÖZET

COVID-19 pandemisi, SARS-CoV-2 virüsünün yol açtığı küresel bir sağlık krizi olarak toplumları, ekonomileri ve sağlık sistemlerini derinden etkilemiştir. Pandemiye karşı etkili terapötik stratejilerin hızla geliştirilmesi gerekliliği, araştırmacıları zaman ve maliyet açısından daha avantajlı olan ilaç yeniden konumlandırma (repurposing) yöntemlerine yönlendirmiştir.

Bu çalışmada, yaklaşık 2.500 FDA onaylı ilacın, SARS-CoV-2'ye karşı potansiyel etkinliğini araştırmak amacıyla sanal tarama teknikleri kullanılmıştır. Moleküler yerleştirme (docking), moleküler dinamik (MD) simülasyonları ve serbest enerji hesaplamaları gibi hesaplamalı yöntemler bir arada kullanılarak, etkili inhibitör adayları belirlenmiştir. Öncelikle, SARS-CoV-2 ana proteazı (Mpro) için GOLD yazılımıyla elde edilen bağlanma pozisyonları, ANI (Atomistik Nöral Ağ Potansiyeli) ile yeniden puanlanarak doğrulanmıştır. Daha sonra, Mpro için yüksek verimli sanal taramalar gerçekleştirilmiştir.

En umut verici ilaç adayları, kapsamlı MD simülasyonları ve bağlanma serbest enerjisi analizleri (MM/GBSA, LIE gibi yöntemlerle) ile ileri düzeyde değerlendirilmiştir. Bu hesaplamalardan elde edilen sonuçlar, COVID-19 tedavisinde etkili olabilecek yeniden konumlandırılmış ilaçların belirlenmesine katkı sağlamış ve deneysel çalışmalara yön verici nitelikte olmuştur.

Ayrıca, elde edilen sonuçlar doğrultusunda moleküllerin başlangıç yapılarının, moleküler yerleştirme ve simülasyon sonuçlarını önemli ölçüde etkileyebileceği gözlemlenmiştir. Bu nedenle, bu çalışmada düşük enerjili konformerlerin üretimini sağlayan, DFT düzeyinde doğruluğa sahip ANI-ML potansiyellerini kullanan yeni bir algoritma olan DeepConf tanıtılmıştır. DeepConf, özellikle başlangıç yapısının denge konumundan uzak olduğu durumlarda, makine öğrenmesi potansiyellerinin düzgün olmayan bölgelerde takılı kalma riskine karşı etkili çözümler sunmaktadır. Ayrıca, daha az sayıda konformerler yüksek kaliteli sonuçlar elde etme ihtiyacında olan sistemler için de verimli bir yaklaşım sağlamaktadır.

Anahtar Kelimeler: İlaç Yeniden Konumlandırma, Yapay Zeka, Moleküler Yerleştirme, Moleküler Dinamik, Serbest Enerji Hesaplamaları.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my thesis advisor, to Assoc. Prof. Dr. Abdulkadir Koak, for their continuous support, guidance, and patience throughout my research. Their insightful feedback and encouragement have been invaluable in shaping this work. I am also sincerely thankful to my co-advisor mer Tayfuroėlu, whose expertise and mentorship have greatly contributed to my academic development and to the completion of this thesis.

My heartfelt thanks go to all the researchers, modellab members, and colleagues I have had the pleasure to work with during this period. Their collaboration and friendship made this journey both productive and enjoyable.

I am especially grateful to my family and friends for their endless love, encouragement, and belief in me. Without their unwavering support, this achievement would not have been possible.

I also thank to the Tubitak ULAKBİM because all calculations that are reported in this thesis were 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 Numbers: **120Z732** and **223Z125**).

TABLE OF CONTENTS

	<u>Page</u>
ABSTRACT	vi
ÖZET	vii
ACKNOWLEDGEMENTS	viii
TABLE OF CONTENTS	ix
LIST OF SYMBOLS AND ABBREVIATIONS	xi
LIST OF FIGURES	xii
LIST OF TABLES	xiv
1. INTRODUCTION	1
1.1. Computational Methods in Drug Discovery	1
1.2. Target Identification	2
1.3. Evolutionary Methods	4
1.4. Geometrical Methods	5
1.5. Energy-Based Methods	5
1.6. Structure-Based Methodologies	6
1.6.1. Molecular Docking	6
1.6.1.1. Search Algorithms	7
1.6.1.2. Scoring Algorithms	8
1.6.2. Molecular Docking Limitations	11
1.7. Sampling, Scoring, and Optimization	12
1.8. Machine Learning in De Novo Drug Design	13
1.9. Limitations in De Novo Drug Design	14
1.10. Approaches for Drug Target Identification	16
1.11. The Role of Informatics in Chemistry	17
1.11.1. Importance of Free Computational Resources in Drug Discovery	18
1.12. Compound Databases	18
2. METHODS	21
2.1. Benchmarking ANI potentials as a rescoring function and screening FDA drugs for SARS-CoV-2 Main Protease	21
2.1.1. CASF-2016 Dataset	21
2.1.2. GOLD docking protocol	22
2.1.3. ANI-2x as rescoring	22
2.1.4. MD simulations	23
2.1.5. Free energy calculations	23
2.2. Leveraging ANI-ML Potentials for Exploring Local Minima with Application to Bioactive Conformations	24
3. RESULTS	25
3.1. Benchmarking ANI potentials as a rescoring function and screening FDA drugs for SARS-CoV-2 Main Protease	25
3.1.1. Introduction	25
3.1.2. Results and discussion	30
3.1.2.1. Benchmarking with other docking software on CASF-2016 dataset	30
3.1.2.2. Assessment on SARS-CoV-2 with GOLD	32

3.1.2.3. Screening FDA drugs for SARS-CoV-2 Mpro	37
3.1.3. Summary	44
3.2. DeepConf: Leveraging ANI-ML Potentials for Exploring Local Minima with Application to Bioactive Conformations	45
3.2.1. Introduction	45
3.2.2. Results and Discussion	50
3.2.2.1. Workflow	50
3.2.2.2. ANI-ML potentials for PES	53
3.2.2.3. Success of DeepConf vs. Auto3D on global/local minima match to G16.	59
3.2.2.3.1. Starting with a near equilibrium structure	59
3.2.2.3.2. Starting with off-equilibrium structure	62
3.2.2.4. Finding bioactive conformations with DeepConf and Auto3D	65
3.2.3. Summary	83
4. CONCLUSIONS	84
REFERENCES	85
BIOGRAPHY	104
PUBLICATIONS AND PRESENTATIONS FROM THE THESIS	105
APPENDICES	106

LIST OF SYMBOLS AND ABBREVIATIONS

3D	: Three-Dimensional
Cryo-EM	: Cryogenic Electron Microscopy
ASE	: Atomic Simulation Environment
BAR	: Bennett Acceptance Ratio
BFE	: Binding Free Energy
CADD	: Computer-Aided Drug Design
CSD	: Cambridge Structure Database
DL	: Deep Learning
DNDD	: De Novo Drug Design
DRL	: Deep Reinforcement Learning
ESP	: Electrostatic Potential
FF	: Force Field
HF	: Hydrogen Fluoride
HTS	: High Throughput Screening
HTVS	: High Throughput Virtual Screening
LBVS	: Ligand-Based Virtual Screening
LIE	: Linear Interaction Energy
MC	: Monte Carlo
MD	: Molecular Dynamics
MK	: Merz-Kollman
ML	: Machine Learning
MLR	: Multivariate Linear Regression
MM	: Molecular Mechanics
MUD	: Mean Unsigned Difference
NMR	: Nuclear Magnetic Resonance
NP	: Natural Product
PDB	: Protein Data Bank
PES	: Potential Energy Surface
PL	: Protein-Ligand
PLS	: Protein-Ligand-Solvent
PME	: Particle-Mesh Ewald
QM	: Quantum Mechanics
RESP	: Restricted Electrostatic Potential
RL	: Reinforcement Learning
RMSD	: Root-Mean-Square Deviation
SAR	: Structure Activity Relationship
SBDD	: Structure-Based Drug Design
SPE	: Single Point Energy
WHO	: World Health Organization

LIST OF FIGURES

	<u>Page</u>
Figure 1.1: Approaches for target identification .	16
Figure 3.1: The success of different scoring functions on predicting the first poses as the closest structure to the crystal with RMSD less than 2.0 Å. ANI is ranked 8th among 34 functions.	30
Figure 3.2: The success of different scoring functions on predicting the first poses as the exact crystal structure. ANI is ranked 5th among 34 functions.	31
Figure 3.3: The Spearman correlation coefficients between RMSD and scores for each scoring function being tested. The x-axis indicates the RMSD bins (≤ 2 , ≤ 3 , etc.) Heat map ranging between 0 and 1 indicates low/high correlations.	32
Figure 3.4: The structure of the SARS-CoV-2 main protease (PDB ID:7L14) used in GOLD docking. The nearby residues around ligand by 10 Å are highlighted. ANI scores were calculated using $\Delta G_{ANI} = E_{complex} - (E_{protein} + E_{ligand})$	33
Figure 3.5: GOLD-PLP and ANI scoring on the poses generated by Gold docking on 14 selected inhibitors of SARS-CoV-2 main protease. ANI shows superiority to GOLD in the top one solution.	35
Figure 3.6: Docking scores generated by a GOLD-PLP and ANI with respect to experimental BFEs. ANI scores were produced by scaling with 0.127 as outlined in.	36
Figure 3.7: Our unique screening protocol involving consensus scoring of ANI and GOLD classical MD simulations and consensus BFEs by end-state methods.	38
Figure 3.8: Application of elimination (red)/selection (blue) criteria to MD trajectories $\Delta(RMSD) \leq 3.0$ Å for the ligand, and $\Delta(\#H\text{-bonds}) \geq -0.5$ between ligand and protein.	39
Figure 3.9: An example RMSD of protein backbone atoms for selected (blue) and eliminated (red).	40
Figure 3.10: DeepConf flowchart.	51
Figure 3.11: PES scan of the H-F distance computed using G16 and ANI-2x .	54
Figure 3.12: PES of H-O-H angle and O-H bond distances by ANI-2x and G16.	55
Figure 3.13: DeepConf approach to overcoming optimization challenges in non-smooth regions of the potential energy surface.	56
Figure 3.14: Initial test compounds and their G16 predicted global and local minima.	57

Figure 3.15. Schematic representation of performance of conformer generation algorithms for capturing bioactive conformation.	67
Figure 3.16: Distribution of minimum RMSD values of all conformers generated by different algorithms.	72
Figure 3.17: Distribution of relative energy differences of conformers generated by different algorithms.	79

LIST OF TABLES

	<u>Page</u>
Table 1.1: Examples of binding site identification software.	4
Table 1.2: Summary of the sampling algorithms that are used commonly.	10
Table 1.3: Scoring function examples in molecular docking calculations.	10
Table 1.4: Examples and classification of de novo software.	12
Table 1.5: Common chemical databases freely available to access.	20
Table 3.1: RMSD values of amino acids residing in the docking grid.	33
Table 3.2: Long MD simulations and their calculated free energy values for selected 110 compounds.	41
Table 3.3: The final drug list we recommended in the FDA drug repurposing study to target SARS-CoV-2 Mpro.	44
Table 3.4: ANI predicted the number of minima and its success in terms of reproducing G16 predicted ones from manually created intermediate conformations for each torsion point.	58
Table 3.5: Success of conformer generation on selected compounds with different parameter settings (parset) when the initial structure is near equilibrium.	60
Table 3.6: Parameter settings used in conformer generation for the initial dataset of selected ligands. Both programs are adjusted to similar settings in order to assess their performance.	61
Table 3.7: Effect of maximum step / convergence threshold on conformer generation performances.	62
Table 3.8: Optimization performance of DeepConf when initial structures are beyond equilibrium geometry. The success shows the number of instances where optimization is converged to a geometry with $\text{RMSD} \leq 0.5 \text{ \AA}$.	63
Table 3.9: Optimization of 100 frames extracted from MD simulations for the selected compounds. Matched refers to the number of structures with $\text{RMSD} \leq 0.5 \text{ \AA}$ to any of the minima on Figure S1. Global is the global minimum and min1,min2,etc refer to the local minima. Corresponding values under the columns show the number of structures yielded (upon optimization) to that specific minimum. Unmatched refers the structure is converged to a different minima (likely saddle point or non-smooth region). For G16 calculations.	65
Table 3.10: Optimization performance of DeepConf when initial structures are beyond equilibrium geometry. $\text{\AA}$ to any of the minima on Figure S1. Global is the global minimum and min1,min2,etc refer to the local minima. Corresponding values under the columns show the number of structures yielded (upon optimization) to that specific minimum.	

Unmatched refers the structure is converged to a different minima (likely saddle point or non-smooth region). For DeepConf aloritym.	65
Table 3.11: Overall performance of DeepConf and Auto3D in reproduction of model/ideal structures from ideal structures.	71
Table 3.12: Overall performance of Confgen and Auto3D in reproduction of model/ideal structures	73
Table 3.13: The performance of conformer generation algorithms on reproducing QM optimized energies.	77
Table 3.14: The performance of conformer generation algorithms on reproducing relative energies between model and ideal structures.	78
Table 3.15: The performance of conformer generation algorithms on reproducing the true global and local minima	80
Table 3.16: Effect of maximum step / convergence threshold	82
Table 3.17: Computational performance of Auto3D and DeepConf	82

1. INTRODUCTION

1.1. Computational Methods in Drug Discovery

Computational methods in drug discovery have been used in many years for the pharmaceutical industry. With the pandemic, the use of these tools has increased even more and become a fundamental methodology. Computational methods have two main advantages in drug discovery. These are higher speed and lower cost. Researchers around the world are working to develop treatments and vaccines to struggle the diseases, and several drugs have been approved for emergency use by regulatory agencies. Since synthesizing a new drug molecule will be a very costly and time-consuming application, drug repurposing, which is a method of retargeting existing drugs, can be applied in such rapidly spreading diseases. Free Energy methods, which is commonly used in repurposing, are used to select reasonable number of molecules to try in vitro experiments for treatment of pandemics.

In computational aided drug design, structure-based drug design (SBDD-target-based drug discovery) and Ligand-based virtual screening (LBVS) and have been used [1],[2] SBDD is based on 3D structural information of the protein based on examining ligand protein interaction to clarify experimental results [3].

The development in structural resolution techniques such as nuclear magnetic resonance (NMR), X-ray crystallography, and cryogenic electron microscopy (cryo-EM) in recent years has also increased SBDD studies, since the regularity of 3D structures if both protein and ligand [4]. As of 2023, approximately 200,000 structures are loaded in The Protein Data Bank (PDB), international database. PDB is updated every week, adding new structures to the database [5]. The development of structural resolution techniques and databases has also enabled the development of software packages that enable the calculations of binding sites predictions of inhibitors and insight into further developments [6].

The SBDD studies begin with target identification which is supported by genetics, molecular biology, and bioinformatics methodologies. After the target identification, the protein extraction and purification performed to structural determination of the

target generally using X-ray crystallography, NMR and Cry-EM [7]. The biological structures performed by different computational methodologies such as molecular dynamics, molecular docking, and de novo drug design the three core methodologies of structure-based drug design (SBDD).

The drug candidates generally selected from databases to be evaluated in structure-based virtual screening (VS) methodologies. At the end, the top molecules are synthesized to in-vitro tests [8].

1.2. Target Identification

Public and proprietary datasets are used to identify potential oncology drug targets. Significant resources, finances, and time are invested in discovering candidate drugs for clinical testing, including biologics and small molecules. Prior validation of potential targets is crucial to ensure selection of those with the highest potential to positively impact disease and benefit patients.

When a new drug develops a specific disease, the first step is to understand the physiological and pathological mechanisms behind it. This helps identify potential molecular targets. Common targets include proteins like enzymes, metabotropic and ionotropic receptors, and antibodies, as well as nucleic acids such as nuclear and mitochondrial DNA and messenger RNA. Among these, proteins are the most common targets because they play a key role in regulating metabolic processes [9],[10].

Drugs are mostly act like inhibitors on enzymatic targets. When targeting receptors, they can function as either agonists, mimicking the effects of endogenous substances, or antagonists, blocking the signals triggered by these substances. The development of Structure-Based Drug Design (SBDD) requires detailed knowledge of the topological arrangement of target proteins. This involves acquiring high-resolution 3D data of macromolecules, which is typically obtained through techniques like X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy (cryo-EM). These structures are deposited in publicly accessible databases, once determined. Protein Data Bank (PDB), InterPro, ExPASy, and Relibase Among are the most popular and widely used databases [11].

If the target structure is unknown, SBDD can be performed by using homology modelling. Homology modelling relies on silico approximations to create a 3D structure based on a homologous protein with a known structure. There are three main methods for predicting structures, these are comparative modelling, threading, and *ab initio* modeling, all of which are reliable and can be validated. The homology modelling process begins with sequence alignment or registration, allowing for insertions and deletions to align the target sequence with the experimental structure by using dynamic programming algorithms. In the next step, amino acid residues in the experimental structure are mutated to match those of the target protein. If multiple experimental structures are available, a hybrid model can be created, combining elements from these structures to more closely approximate the target protein. Finally, the model undergoes refinement and validation to ensure its structural accuracy and reasonableness [12].

These similarities on the target structure, allows the designing new drugs with a higher likelihood of effectively targeting the studied pathology. Tools such as the Ramachandran plot are invaluable during the validation stage. This plot evaluates protein model quality by identifying residues in energetically favorable and unfavorable regions, based on dihedral angles Ψ versus ϕ for each amino acid, except glycine. For a model to be considered reliable, at least 90% of residues should reside in favorable regions [13].

Firstly, to identify the binding site, the target protein is diagnosed, and the biological function of protein confirmed, the next challenge in structure-based drug design (SBDD) involves locating and validating the specific regions within the protein structure that are most likely to bind with ligands. Binding sites, also known as cavities, can be divided into two distinct categories. If a ligand interacts with the catalytic site of a target protein, this site is referred to as the orthostatic site. However, if the interaction occurs at a different region that induces conformational changes in the protein, ultimately affecting the catalytic site, it is known as an allosteric site [14]. Understanding the structure of binding cavities and their interaction energies with ligands is essential for any structure-based drug design (SBDD) process. To achieve this, various strategies have been developed to identify binding sites within target proteins. These strategies include evolutionary approaches that analyze residue

similarities between functionally related proteins, geometrical methods that focus on the protein's structural features, and energy-based techniques that assess interaction energies using probe molecules [15],[16].

1.3. Evolutionary Methods

The evolutionary method relies on the principle that homologous proteins, which share a common ancestry, often retain conserved residues across different species. These conserved sequences are likely to be functionally significant, especially in regions responsible for biological or enzymatic activities, to ensure that essential physiological processes are maintained without malfunctions [17]. Based on this assumption, evolutionary methods use information related to protein cavities to identify potential binding sites in homologous or functionally related proteins by comparing their primary residue sequences.

Table 1 highlights several algorithms that are based on evolutionary approaches. However, a major limitation of these methods is the requirement for prior knowledge about existing protein sequences. Given the vast diversity of proteins across various organisms, obtaining comprehensive sequence information can be challenging. Additionally, if the target protein and the reference protein have low sequence similarity, the accuracy of cavity predictions can be compromised, negatively impacting the success of structure-based drug design (SBDD) studies.

Table 1.1: Examples of binding side identification software according to type of algorithm.

Evolutionary methods	Geometrical methods	Energy-based methods
ConSurf	LIGSITE	GRID
Rate4Site	CAST	DrugSite
GarLig	SURFNET	QsiteFinder
	PocketPicker	MCSS
	VOIDOO	
	KVFinder	

1.4. Geometrical Methods

Geometrical methods focus on identifying protein cavities based on their structural features, such as size, shape, and surface properties. Cavities are generally visualized as concave, hole-like regions on the protein surface [18]. These methods analyze the three-dimensional (3D) structure of proteins to detect binding sites by using parameters like ionizability, intermolecular interaction tendencies, and covalent bonding sites [19]. In geometry-based techniques, probe objects, often spheres, are used to scan the protein surface and fit into cavities based on size [16],[19]. Other methods involve tessellation or creating a 3D grid over the protein structure to identify voids of significant size on its surface. These techniques are commonly categorized into three approaches: probe sphere filling, grid system scanning, and alpha-shape modeling [20].

Geometrical methods are widely implemented in various cavity detection software due to their simplicity and efficiency. One of their biggest advantages is that they require only the 3D structure of the protein, without needing additional prior knowledge. Since these algorithms typically use crystal structures from databases such as the Protein Data Bank (PDB), the protein is treated as a static entity, ignoring its inherent flexibility and dynamic conformational changes [21].

The primary disadvantages of these methods include: (i) the arbitrary nature of setting tessellation and grid parameters, which can affect the accuracy of surface probing, and (ii) inaccuracies caused by protein orientation impacts and potential changes of conformations [22].

1.5. Energy-Based Methods

Energy-based methods aim to identify binding sites by evaluating the interaction energies between the target protein and small molecules or fragments. These methods are particularly useful in predicting whether a ligand will bind to a specific site and thus serve as a potential drug candidate. The core idea is to assess the binding probability by calculating interaction energies between different parts of the receptor and probes.

In these algorithms, hydrogen bond energy, the van der Waals (vdW) interaction energy, or Lennard-Jones potentials hydrogen bond energy, between the target protein and probe molecules (e.g., water or methyl groups) are analyzed to locate binding sites [21],[23] .

Although energy-based methods provide valuable insights into binding affinities, they are more computationally intensive compared to other techniques. This poses a significant drawback, as reducing computation time is critical in SBDD research [21]. Despite their higher computational cost, energy-based methods offer precise identification of energetically favorable binding sites, which can greatly enhance the accuracy of ligand-binding predictions.

1.6. Structure-Based Methodologies

1.6.1. Molecular Docking

Molecular docking is one of the widely used target-based drug discovery methods. In recent years, more than 52,000 research on target-based drug discovery have been carried out, with over half of them appearing between 2017 and 2021 (Web of Science, Copyright Clarivate 202). The primary goal of molecular docking is to predict the binding score and the binding mode of a ligand that best fits a target protein. Docking methods aim to accurately determine how a ligand interacts with a receptor binding site and estimate the binding strength [24].

In protein-ligand docking, the process involves generating multiple conformations and orientations (poses) of the ligand within the protein's binding site [25]. To identify the most probable binding conformation, docking involves two key steps: (1) exploring a large conformational space representing possible binding modes, and (2) predicting the interaction energy accurately associated with each pose. Docking programs perform these steps through iterative cycles where the ligand's conformation is evaluated using scoring functions until a minimum energy solution is achieved [5].

There are three main types of molecular docking techniques, these are categorized by the flexibility of the molecules:

While the rigid docking, the ligand and the protein are treated as fixed structures. Only three translational and three rotational degrees of freedom are considered during sampling [25].

In semi-flexible docking method, the ligand is flexible while the protein remains rigid. This approach allows the ligand's conformational degrees of freedom to be sampled in addition to translational and rotational movements [25].

Flexible Docking: Both the ligand and the protein are considered flexible. The protein is not treated as a passive rigid structure during the binding process, allowing for more accurate predictions [25].

A molecular docking program consists of two crucial components: (1) sampling the possible conformations of the ligand within the binding site, and (2) ranking these conformations using a scoring function [26],[27]. Sampling involves generating potential ligand-binding poses near the binding site and can be divided into ligand sampling and protein flexibility. Scoring functions evaluate the binding strength for each pose using physical or empirical energy functions [28].

The computational cost of docking calculations increases with the number of degrees of freedom. Therefore, both sampling and scoring functions must be optimized to balance accuracy and computational efficiency [25].

1.6.1.1. Search Algorithms

Sampling algorithms evaluate the capability of docking programs to predict ligand-binding poses [28]. Sufficient sampling of both ligand and protein states is essential for accurate docking results, as well as for properly estimating the binding energies of potential protein-ligand complexes. A key challenge in sampling is ensuring that the docking program sufficiently explores possible binding states and how increasing the sampling space impacts scoring accuracy and docking outcomes. This includes sampling the internal degrees of freedom within the ligand and exploring the poses between the ligand and the protein receptor [29].

Sampling algorithms are typically classified into three main categories [28].

Shape Matching: These methods consider the geometric overlap between the ligand and the protein. While they can predict docking conformations quickly, their accuracy tends to be lower [30].

Systematic Search: These methods systematically explore all possible degrees of freedom of the ligand during the docking process [31].

Stochastic Search: These algorithms randomly modify the degrees of freedom of the ligand (translational, rotational, and conformational) at each step, generating diverse solutions. However, multiple independent runs of the algorithm are often required to maximize the chances of finding the global energy minimum [31].

Genetic algorithm: Genetic algorithms (GAs) represent a class of stochastic optimization methods inspired by the principles of evolutionary biology, including selection, crossover, and mutation. Within the context of molecular docking, GAs are employed to iteratively optimize ligand poses by evolving a population of candidate solutions through successive generations. Each candidate, or individual, encodes a specific conformation and orientation of the ligand relative to the receptor. The Lamarckian genetic algorithm, an extension of the classical Darwinian approach, incorporates a local search step after each mutation or crossover event, allowing beneficial modifications discovered through local energy minimization to be encoded back into the genotype. This hybrid strategy effectively balances global exploration with local exploitation, thereby increasing the efficiency of conformational sampling and improving the likelihood of locating low-energy, biologically relevant binding poses [32],[33].

1.6.1.2. Scoring Algorithms

In molecular docking, scoring algorithms are responsible for differentiating correct ligand-binding poses from incorrect ones within a reasonable computation time. These functions estimate the binding affinity between the protein and the ligand by applying various assumptions and simplifications [26].

Two key theoretical aspects influence the performance of scoring functions:

The ability of the scoring function to identify the global minimum in the ligand-binding landscape at the correct location. The accuracy of the binding energy estimate at the identified minimum [34].

Scoring algorithms serve three primary purposes in the docking process:

Pose Prediction: The scoring function must distinguish the experimentally validated binding mode from other modes generated during the docking process [35].

Virtual Screening: The ability to differentiate between binders (active compounds) and non-binders (inactive compounds) [31].

Binding Affinity Estimation: The prediction of binding constants and the ability to rank multiple compounds based on their binding strength [31].

Scoring functions are generally classified into three traditional categories, with a fourth category emerging recently:

Force Field-Based Scoring: These functions use classical molecular mechanics energy terms to calculate the binding free energy of protein-ligand complexes. They consider van der Waals forces, electrostatic interactions, and solvent effects [36].

Knowledge-Based Scoring: These scoring functions are based on the inverse Boltzmann statistical principle, which assumes that the frequency of different atom pairs at various distances is related to their interaction. The observed frequencies are converted into a distance-dependent potential of the mean force. Knowledge-based scoring functions offer advantages in terms of reduced computational cost and predictive accuracy compared to other types of scoring functions [37].

Empirical Scoring: These functions assess protein-ligand binding by summing contributions from various individual terms, each representing a significant energetic factor in the binding process. Techniques such as multivariate linear regression (MLR) or partial least-squares (PLS) analysis are commonly used to identify the factors influencing the final binding score [38],[39].

Machine Learning-Based Scoring: These dynamic models use machine learning techniques to predict binding poses and affinities. Machine learning-based scoring functions have gained attention as potential rescoring tools for structure-based virtual screening (VS). They can implicitly learn from training data and use adjustable

parameters to improve performance, offering flexibility and convenience [40]. Common machine learning algorithms used for scoring include support vector machines, random forests, neural networks, and deep learning [37].

Table 1.2: Summary of the sampling algorithms that are used commonly [41],[45].

Algorithm	Description	Examples
Shape matching	Considers the geometrical overlap between two molecules, predicting docking conformations at a fast speed but with lower accuracy.	ZDOCK, SYSDOCK
Exhaustive search	Explores values of each degree of freedom in a combinatorial manner, rotating all dihedral angles of the ligand.	Glide, eHiTS
Fragmentation	Ligand is separated into smaller fragments, followed by selection and docking of a base fragment.	FlexX
Conformation ensemble	Rigidly docks a set of previously generated ligand conformations into the binding site.	DOCK 4.0, FLOG
Monte-Carlo	Applies random Cartesian moves to the system and accepts or rejects based on Boltzmann probability.	LigandFit, rDock
Genetic algorithm	Evaluates evolution of a population of possible solutions via genetic operators, optimizing a fitness function.	rDock, AutoDock, GOLD
Tabu Search methods	An iterative procedure that moves solutions to optimization problems using a Meta-Heuristic methodology.	PSI-DOCK
Swarm optimization methods	Particles use knowledge of other particles to speculate about the promising region to explore.	PLANTS, PSO-VINA

Table 1.3: Scoring function examples in molecular docking calculations [38].

Force field-based	Empirical	Knowledge-based	Machine learning-based
AutoDock	ICM-Score	ICM-PMF	RF-Score
DOCK/FF	SYBYL/F-Score	SYBYL/PMF-Score3	RF-IChem
SYBYL/G-Score	SYBYL/ChemScore	DrugScore	SVM-IChem
SYBYL/D-Score	LUDI X-Score		

1.6.2. Molecular Docking Limitations

Designing new molecular structures with therapeutic potential is a significant challenge in the development of drugs for commercial use. To address this, various strategies have been introduced to identify and create new drug candidates. Given the vast amount of information available on the chemical space related to different diseases, it is possible to leverage this data to design new molecules. One such method is de novo drug design (DNDD), where small molecular fragments are assembled computationally to optimize their interactions with the target's active site using growth algorithms [46].

De novo design involves creating molecules that meet specific structural and functional requirements necessary to achieve desired biological or therapeutic effects [47]. There are two primary approaches to initiating a de novo design process:

Structure-Based Approach: In this method, the molecular design starts with knowledge of the active site's 3D structure and potential interaction points. The information about the receptor's structure guides the development of the new molecule.

Ligand-Based Approach: This method begins with a known active ligand, but without detailed knowledge of the target's active site. The molecular design is based on the chemical properties of the existing ligand [48].

After selecting the appropriate approach, the next step involves building and evaluating molecules that meet the required constraints and characteristics [47],[48]. Various software tools are available to support de novo design, and they are categorized based on the methodology they use, as shown in

Table 1.4.

Table 1.4: Examples and classification of de novo software by methodology used [48],[49].

Software	Ligand-based	Structure-based
LUDI		X
SPROUT		X
Chemical genesis	X	X
PRO_LIGAND	X	X
TOPAS	X	
ADAPT		X
Diamond Lattice		X
LEGEND	X	X
MCDNLG		X
DLD		X
RASSE		X

1.7. Sampling, Scoring, and Optimization

When generating a list of potential drug candidates in de novo drug design, two primary methodologies are used. The atom-based approach modifies individual atoms within a molecule to explore a wide range of possible molecular structures. This intuitive method can produce novel compounds, but it carries the risk of creating unstable structures or synthetically inaccessible molecules [48],[50]. On the other hand, the fragment-based approach designs molecules by mutating predefined molecular fragments. This method significantly reduces the chemical space to be explored, making it faster and more likely to yield chemically feasible molecules [50],[51].

During molecular construction, candidate molecules are evaluated based on various molecular descriptors using scoring algorithms to assess their potential druggability. In structure-based approaches, the scoring is based on docking simulations that measure interactions between molecular fragments and the target's active site. These fragments are then used as seeds to build the rest of the molecule. In ligand-based approaches, molecules are compared to a reference ligand based on their molecular similarities.

Optimization of candidate molecules often relies on evolutionary computation, a process inspired by natural selection that involves mutation, crossover, and selection of the best-performing candidates. Evolutionary computation methods include genetic algorithms, genetic programming, evolution strategies, and evolutionary programming [46],[52]. These algorithms typically start with an initial population of molecules and iteratively refine them by selecting the most promising candidates to form the next generation, thus improving the quality of the proposed drug candidates over time.

The main advantages of these optimization algorithms are their simplicity, adaptability, and efficiency in exploring the chemical space for potential drugs [52].

1.8. Machine Learning in De Novo Drug Design

With the increasing availability of extensive chemical and biological data from databases such as ChEMBL, COCONUT [53], and ZINC, machine learning has become a valuable tool in de novo drug design. These methods allow researchers to analyze and predict the behavior of bioactive compounds, significantly reducing the time needed to explore the chemical space in drug discovery [46].

Machine learning advancements have enabled computers to identify promising molecules by analyzing chemical and biological data. These models can recognize structural patterns (fingerprints) that enhance molecular activity, improving the efficiency of de novo design processes. Techniques such as Deep Learning (DL) [54],[55], Reinforcement Learning (RL) [56], Deep Reinforcement Learning (DRL) [57], and Reinforcement Learning for Structural Evolution (ReLeaSE) [58], have been developed to assist in the de novo design of molecules [46],[58].

These machine learning models typically use molecular representations, such as SMILES strings, and corresponding biological activity data as input. The main differences between these methods lie in the model architecture and the specific molecular properties being analyzed.

1.9. Limitations in De Novo Drug Design

Despite its advantages, de novo drug design has several limitations that need to be addressed to improve its effectiveness. The primary limitation is the accuracy of scoring functions. Current de novo tools often fail to account for conformational changes in the target protein's structure, particularly in the active site, and overlook the entropic effects caused by interactions between the ligand and the solvent. These omissions reduce the reliability of ligand-pocket interaction energy predictions, which are crucial for designing effective molecules [59].

Additionally, while de novo methodologies can predict strong target selectivity and interaction energies, they often fail to account for important physicochemical properties required for successful drug development. These include solubility, cell permeability, and affinity to transport proteins, which are critical for pharmacokinetic studies. Addressing these limitations is essential for enhancing the success rate of de novo approaches in drug discovery [46],[49].

Drug discovery is a key area in medicinal sciences, focusing on a series of integrated steps to find treatments for diseases considered incurable. This process involves complex procedures aimed at the chemical and pharmaceutical optimization of new compounds to treat human illnesses [60]. The discovery process begins with identifying the biological cause of the disease and determining potential drug targets. Modern drug discovery is credited to Paul Ehrlich, who introduced the concept of selective drugs to combat infectious diseases. He also laid the groundwork for discovering synthetic drugs to treat bacterial infections [44].

The initial and essential step in drug discovery is target identification, which involves determining the molecular targets, such as proteins or nucleic acids, that small molecules interact with. This can be achieved using computational tools, biochemical methods, or genetic approaches. An ideal target must meet specific criteria to ensure clinical efficacy and safety when modulated by small molecules.

Historically, drug discovery followed a “one target, one drug” approach. However, there has been a shift toward a “one drug, multiple targets” paradigm [61]. Both approaches involve validating targets based on their ability to regulate biological functions by interacting with small molecules. Once the target validation is complete,

the next step is hit identification, where initial molecules, known as hits, are found and used as a basis to develop more potent derivatives [62],[63].

Between 1999 and 2008, the FDA approved 259 therapeutic agents, of which 75 were first-in-class drugs with new molecular mechanisms of action. Notably, phenotypic screening contributed more significantly to the development of these first-in-class small molecule agents than target-based approaches. Specifically, 28 drugs were approved through phenotypic screening, compared to 17 drugs from target-based screening [64].

Current drug therapies are developed based on approximately 500 molecular targets, with G-protein coupled receptors accounting for 45%, enzymes for 28%, hormones for 11%, ion channels for 5%, and nuclear receptors for 2% (Drews, 2000). In cancer treatment, the FDA has approved 154 drugs, which are categorized into three groups: five drugs targeting non-protein or unknown targets, 64 cytotoxic drugs, and 85 drugs targeting specific proteins [65].

The identification of disease-associated targets is critical in the screening of small molecules in target-based approaches. This typically involves genome-wide data from various sources [66]. Additionally, target deconvolution, the process of determining the cellular mechanisms of hits identified through phenotypic screening, is essential for understanding the targets of small molecules [67].

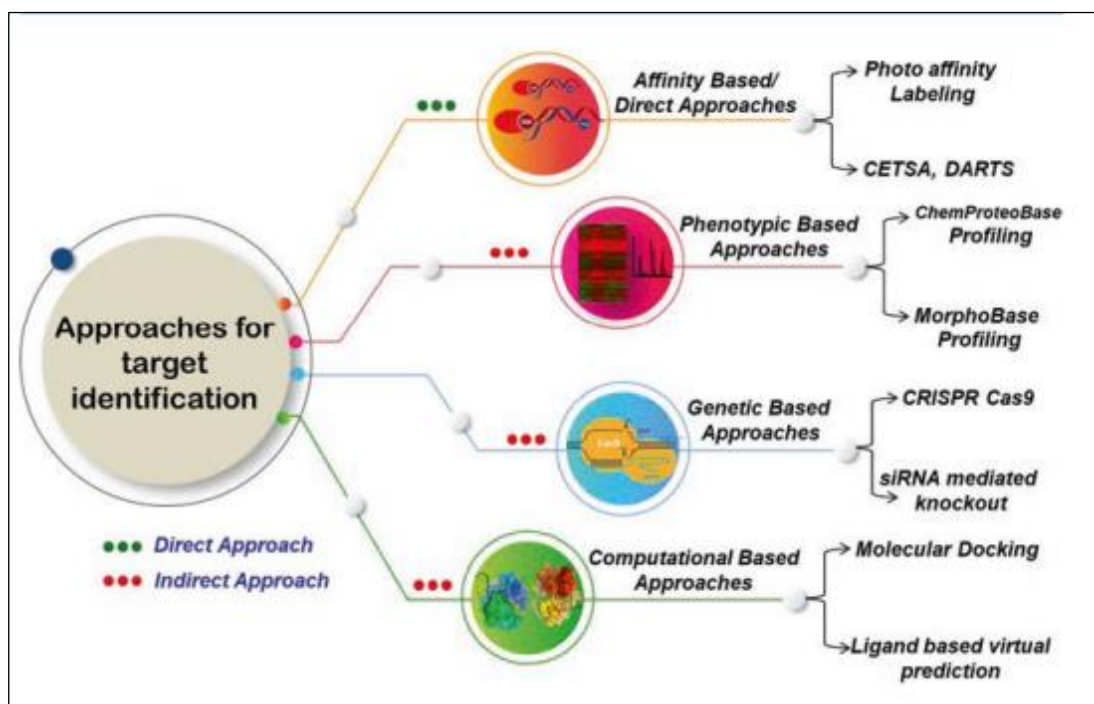


Figure 1.1: Approaches for target identification.

1.10. Approaches for Drug Target Identification

Several methods have been proposed to identify drug targets for bioactive molecules, which can be categorized into four main groups:

- **Affinity-Based Approaches:** These methods focus on the direct interaction between small molecules and target proteins [68].
- **Phenotype-Based Approaches:** This category involves comparing the biological profiles of small molecules with those of known drugs to identify potential targets [61].
- **Genetic-Based Approaches:** These approaches aim to pinpoint genes associated with drug resistance or sensitivity to small molecules, helping to identify the underlying genetic factors [69].
- **Computational Approaches:** These methods predict drug targets by analyzing the chemical complementarity between proteins and potential drug molecules [70],[71].

1.11. The Role of Informatics in Chemistry

Informatics plays a crucial role in various chemistry-related applications, contributing to the establishment of well-known fields like bioinformatics and cheminformatics. Additionally, it has driven the evolution of newer subfields, such as food informatics, epi-informatics, and, more recently, natural products informatics. The widespread use of informatics across multiple disciplines has been accelerated by the growing availability of open, well-documented resources, many of which have been developed as web applications, making them more accessible to the scientific community.

This chapter highlights recent advancements in the development of public chemoinformatic tools designed for various tasks, with a particular focus on drug discovery applications. Given the ongoing COVID-19 pandemic, special attention is given to resources that have been created and released in recent months to support global drug discovery efforts.

Cheminformatics has significantly impacted various fields, including drug discovery, organic synthesis (planning and analysis), analytical chemistry, and more recently, food chemistry and natural product chemistry. Its influence, particularly in drug discovery, has been widely discussed in the literature [72]–[76].

Over the past decade, advances in conceptual and methodological approaches in cheminformatics have been accompanied by an increasing number of open-source programs. These tools, available as public web servers, stand-alone software, or customizable code, are well-documented and stored in public repositories that ensure free access and easy maintenance. Additionally, specialized journals in cheminformatics and computational sciences now require proper documentation and accessibility of software to encourage data sharing, method reproducibility, and transparency [77],[78].

There are numerous reviews of open cheminformatics resources for various applications. For instance, Singh et al. reviewed online web servers used for virtual screening and docking of small molecules [78], while Wu et al. explored databases and tools to predict ADME/Tox properties (Absorption, Distribution, Metabolism, Excretion, and Toxicity). Additionally, Gonzalez-Medina et al. conducted a comprehensive review of open resources for analyzing structure–activity relationships

(SAR), including web servers and workflows that can be implemented using public software [66].

1.11.1. Importance of Free Computational Resources in Drug Discovery

The availability of free computational tools is beneficial for drug discovery across various diseases, but it is especially critical for tackling complex, rare, or neglected diseases. These resources are also crucial during major health crises, such as the COVID-19 pandemic, which has affected populations globally. Such health emergencies require the deployment of tools that can assist in every phase of the drug discovery process [76].

This chapter highlights recent advancements in free and open-source cheminformatics tools that support key activities in drug discovery, with a particular focus on resources for COVID-19 drug research. Additionally, the chapter addresses free tools designed to assess NP-likeness, an emerging concept in the development of natural product-based drugs.

The content is divided into three primary sections. The first section provides an overview of free compound databases for virtual screening, with a focus on natural product (NP) databases and compound collections aimed at COVID-19 drug discovery. The second section covers tools for virtual screening and target fishing, which help identify potential drug targets. The third section discusses advancements in cheminformatics tools for NP-based drug discovery, with an emphasis on free resources to evaluate NP-likeness.

1.12. Compound Databases

Compound databases form the foundation of cheminformatics by systematically organizing the vast chemical space. This space continues to grow, with new chemical databases being continuously developed. Recent advancements have led to the creation of numerous on-demand and virtual libraries, which have been extensively discussed in the literature [73],[79]. Several reviews have been published on chemical libraries

specifically designed for drug discovery [80],[81]. Some of these libraries are target-specific or indication-focused, such as epigenetic target compound collections.

This section highlights recent developments in focused chemical libraries, particularly those created to support COVID-19 drug discovery efforts.

Coronaviruses (CoVs) are a family of enveloped RNA viruses characterized by their crown-like appearance, caused by the glycoproteins on their surface. In December 2019, a novel coronavirus, SARS-CoV-2, was identified in cases of respiratory infections that came to be known as COVID-19. The virus spread rapidly, prompting the World Health Organization (WHO) to declare it a pandemic in March 2020 [82].

Extensive research efforts, including structural studies, microbiology, proteomics, and chemical library screening, have provided valuable insights for computational approaches such as bioinformatics, cheminformatics, molecular modeling, and drug discovery. By March 2021, more than 110,000 peer-reviewed articles on COVID-19 had been published on PubMed, with an additional 12,500 preprints uploaded to MedRxiv and BioRxiv.

In response to the global health crisis, publicly accessible chemical and structural databases have been developed to support COVID-19 drug discovery. Table 1 summarizes some of these key free resources.

These databases provide information on viral proteins, active compounds, and approved drugs. For instance, the NCBI SARS-CoV-2 database is the largest, with 54,572 protein sequences. In contrast, the CAS COVID-19 Antiviral Candidate Compound Dataset contains nearly 50,000 molecules with potential activity against SARS-CoV-2.

Most databases offer browsing and filtering capabilities, allowing users to search for relevant information, such as protein identifiers, virus names, drug names, and clinical approval status. Users can typically download search results in various formats like CSV, PDF, or Excel, depending on the database.

Some databases also include analytical tools to enhance data interpretation. These tools may include 3D structure viewers, chemical structure comparison tools, sequence logo generators, and phylogenetic tree builders. Additionally, certain databases, such as

DBCOVP and NCBI SARS-CoV-2, allow users to submit new protein sequences for analysis by completing a form and entering the sequence in FASTA format.

Table 1.5: Common chemical databases freely available to access.

Database	Contents	URL
PubChem	Contains more than 109 million compounds, 270 million substances, and 1.2 million BioAssays.	https://pubchem.ncbi.nlm.nih.gov
Natural products database (NatProdDB)	Contains naturally occurring diverse compounds used in virtual screening.	http://natprodb.uefs.br
COCONUT	Aggregated dataset of natural products from different sources.	https://coconut.naturalproducts.net
ChEMBL	Manually curated database of bioactive molecules with 1.9 million compounds and 16 million activities.	https://www.ebi.ac.uk/chembl
ZINC	Compound database connecting biological activities with commercial availability.	http://zinc15.docking.org
Protein Data Bank (PDB)	Database providing 3D structures for 170,000 large biological molecules.	https://www.rcsb.org
Nuclear Receptor Activity (NURA)	Curated database of small molecules modulating nuclear receptors for pharmacological applications.	https://zenodo.org/record/3999420#.YCC-Di2xBQI

2. METHODS

2.1. Benchmarking ANI potentials as a rescoring function and screening FDA drugs for SARS-CoV-2 Main Protease

2.1.1. CASF-2016 Dataset

In order to test ANI's performance as a rescoring function, we have used CASF-2016 dataset, which has been created for benchmarking purposes in docking algorithms. The dataset is composed of 285 protein-ligand complexes and each complex has a native binding pose as well as maximum 100 decoy ligand binding poses, selected from a normal distribution of 1000 poses generated by three different docking software with $\text{RMSD} \leq 10$ Å from the crystal structures [68]–[70]. In the study, 34 different docking algorithms/software were tested in terms of “ranking power”, “scoring power”, “docking power” and “screening power”. Docking power is defined as whether a scoring function can properly differentiate the native binding pose from all the decoys within the top 1, 2, or 3 scores. In addition, when the different decoys create a potential energy surface in the binding region, the native pose should correspond to the minimum energy structure. This has been defined as “binding funnel analysis” and evaluated based on Spearman's rank correlation between the 10 bins of different RMSD windows and average scores for these bins [83]. We used the docking poses provided in this CASF-2016 dataset directly in assessment of the performance of ANI-2x. We have calculated the binding affinities by ligand interacting with only residues in docking region (grid box/sphere) by saturating the discontinuing atoms with hydrogens using Pdbfixer [84]. Since the ANI-2x has been trained for molecules containing only H, C, O, N, S, F, and Cl atoms, the total the dataset reduced to 254 from the original set of 285 proteins. In addition, we also tested the ANI's performance on SARS-CoV-2 main protease inhibitors by applying our own docking protocol using GOLD software.

2.1.2. GOLD docking protocol

The inhibitors of SARS-CoV-2 main protease that are reported in Protein Data Bank (PDB) were retrieved (complex PDB IDs: 7N44, 7L10, 7L11, 7L12, 7L14, 7M8M, 7M8P, 7M8O, 7M8N, 7M8X, 7M8Y, 7M90, 7M8Z, 7M91) and docked to the protein structure with PDB ID: 7L14. Docking was performed using the GOLD Suite v.5.3 software by Cambridge Structure Database (CSD) [85]. For GOLD, the ChemPLP scoring function was used since it is known to give better results in prediction of binding mode than other scoring functions implemented in the software. The genetic algorithm with a minimum of 1,000,000 and maximum 1,250,000 iterations was used. A grid sphere of 10 Å radius was defined. The center of the sphere was defined by the reference ligand in the structure of 7L14. All rotatable bonds in protein were frozen while the ligand was defined as flexible. The bond orders, hydrogens, atom types, and partial charges were produced by the Hermes software's edit utilities. GOLD software was constructed to produce 25 poses for each ligand without early termination.

2.1.3. ANI-2x as rescoring

ANI scoring relies on the interaction-energy between the receptor and the ligand similar to force field-based scoring functions. The interaction energy is calculated by ANI-ML potentials rather than non-covalent coulombic and van der Waals interactions of conventional force field-based scoring. The details of the calculation can be found on <https://github.com/otayfuroglu/deepQM> and the related ANI_LIE work by Akkus et al.[86], as will be discussed in the next section. The main difference is to use docking pose instead of calculating frames from an MD simulation trajectory. Apart from the input pdb file with protein and ligand complex, the user is required to give an index file containing the index groups to be calculated. The index file is the one produced by Gromacs (i.e., listing comma separated residue numbers followed by the group name in the square bracket). The ligand and the residues in the docking region (or entire protein) are defined as separate groups in the index file. Thus, ANI scores can be readily calculated as the difference in energy between the index groups corresponding to complex and free

components (protein and ligand) so that $\Delta E_{ANI} = E_{complex} - (E_{protein} + E_{ligand})$. The interaction energy ΔE_{ANI} is then translated to free energy using a scaling factor, as discussed in the next section. Using this approach, each pose reported on CASF-2016 dataset or generated by our docking with GOLD was recalculated by ANI scoring function.

2.1.4. MD simulations

A similar protocol to our previous works has been used in MD simulations [86]–[94]. Briefly, ligands were first optimized at B3LYP/6-31++G(d,p), and Merz-Kollman (MK) electrostatic potential (ESP) charges were computed at HF/6-31G* level using Gaussian 16 software. GAFF2 parameters and restricted electrostatic potential charges (RESP) were generated using the antechamber utility in AmberTools 22. The amber99sb-ildn force field was used for the protein's topology. Solvation of protein-ligand complexes employed the TIP3P water model in a dodecahedron box with 10 Å dimensions for each axis. System neutralization occurred at a salt concentration of 0.15 M Na⁺ and Cl[−] ions. Energy minimization, utilizing the steepest descent algorithm with a Verlet cutoff scheme, was performed to a maximum force of 100 kJ mol^{−1} nm^{−1}. Electrostatic interactions were computed using the particle-mesh Ewald (PME) method, while bonds involving hydrogen atoms were constrained with a harmonic potential. Prior to the final simulations, a 5 ns NVT-MD equilibration at 310 K was conducted using a Langevin thermostat followed by 200 ps and 1 ns NPT-MD equilibrations at 1 atm using Berendsen and Parrinello-Rahman barostats, respectively.

2.1.5. Free energy calculations

The single-trajectory approach MMGBSA calculations were calculated using the gmx_MMPBSA fault parameters with the internal dielectric constant, $\epsilon_{int}=2$ for polar solvation terms. The SASA-only model was applied with parameters of $\gamma=0.0072$ kcal/mol·Å² and $b=0$ for nonpolar terms.

In addition to MM-PBSA calculations, we also used ANI_LIE, a recent end-state binding free energy method introduced by Akkus et al. [86]. The details of the method have been discussed elsewhere [86]. Briefly, the method uses the linear interaction energy (LIE) approach utilizing the potential energies predicted by ANI-ML potentials as a replacement to molecular mechanics (MM) energy terms in LIE formalism. The free energy is calculated where $\langle \Delta E_{ANI}^{L-P} \rangle_{PLS}$ is the average interaction energy between the protein and ligand produced by extracting the MD frames for energy groups of protein (P), ligand (L) and protein-ligand (PL).

For ANI_LIE calculations were performed using the simplest form neglecting the D3 term and solvation effects (i.e., equation 4 in Ref. [86]) with default β and γ empirical parameters of 0.127 and -5.11, respectively. As reported by Akkus et al. [86], these parameters were optimized SARS-CoV-2 M^{pro} and have been successfully used in other studies [87],[89],[95],[96].

2.2. Leveraging ANI-ML Potentials for Exploring Local Minima with Application to Bioactive Conformations

In the QM calculations, we used Gaussian 16 software with the WB97X/6-31G* level since ANI-2x is trained at this DFT level. Custom Python scripts along with Openbabel and RDKit were used extensively in RMSD calculation, file conversion, and conformer generation processes. Corresponding GitHub repositories have been used in Auto3D calculations. The atomic simulation environment (ASE) was used in our custom script for the energy minimizations.

3. RESULTS

3.1. Benchmarking ANI potentials as a rescoring function and screening FDA drugs for SARS-CoV-2 Main Protease

Adopted with permission from Zengin, I.N., Koca, M.S., Tayfuroglu, O. et al. Benchmarking ANI potentials as a rescoring function and screening FDA drugs for SARS-CoV-2 Mpro. *J Comput Aided Mol Des* 38, 15 (2024) and DeepConf: Leveraging ANI-ML Potentials for Exploring Local Minima with Application to Bioactive Conformations

Here, we introduce the use of ANI-ML potentials as a rescoring function in the host–guest interaction in molecular docking. Our results show that the “docking power” of ANI potentials can compete with the current scoring functions at the same level of computational cost. Benchmarking studies on CASF-2016 dataset showed that ANI is ranked in the top 5 scoring functions among the other 34 tested. In particular, the ANI predicted interaction energies when used in conjunction with GOLD-PLP scoring function can boost the top ranked solution to be the closest to the x-ray structure. Rapid and accurate calculation of interaction energies between ligand and protein also enables screening of millions of drug candidates/docking poses. Using a unique protocol in which docking by GOLD-PLP, rescoring by ANI-ML potentials and extensive MD simulations along with end state free energy methods are combined, we have screened FDA approved drugs against the SARS-CoV-2 main protease (Mpro). The top six drug molecules suggested by the consensus of these free energy methods have already been in clinical trials or proposed as potential drug molecules in previous theoretical and experimental studies, approving the validity and the power of accuracy in our screening method.

3.1.1. Introduction

Designing a therapeutic molecule to treat a disease can be quite expensive and time consuming. Preclinical processes such as identification of drug targets, High Throughput Screening (HTS) experiments on the target, generating a library of hit compounds, optimizing potency of bioactive compounds and preceding clinical trials require a

tremendous amount of resources. Typically, a drug can cost more than 2 billion US dollars and 10-15 years to be approved and enter the market [97],[99]. Since repurposed by an FDA approved drug is a much faster and less expensive strategy it has become one of the most popular approaches in recent drug development endeavors. Even testing all FDA approved drugs experimentally may not be affordable in terms of economic burden and time. To help ease this problem, computational methods have been utilized in rational drug design.

In searching for drug candidates as potential inhibitors, binding mode and binding affinity are two critical questions that ought to be addressed. Docking methods are capable of screening ultra large libraries of compounds by aforementioned approaches (billions of molecules) with much reduced cost and time. Docking methods rely on a search algorithm and scoring function that are physics based, empirical, knowledge-based and Machine Learning based which compromise on accuracy to increase the speed [9]. Despite the great success of the docking methods, there is still a need for improving scoring functions so as to reproduce experimental results.

Machine Learning (ML) methods have just started to gain attention to advance the docking methodology recently similar to other computational techniques. The main objective of the ML techniques is to find an accurate and a fast solution to an existing problem by learning from previous experimental data. This is achieved by supervised, unsupervised and reinforcement methods [100]. Supervised methods are based on training an algorithm on a set of inputs to find an output while unsupervised learning algorithms find a pattern in data set predict results from this arrangement. On the other hand, reinforcement learning, the input progress in an environment and uses the data learned from the experience [97]. Recent studies have shown the outperformance of ML based scoring functions over classical ones, which directly use a known mathematical function [101],[103]. ML based scoring functions are usually used by means of rescoring [104] due to the dependence on the training dataset [105],[106].

ANI-1, ANI-1x and ANI-1cxx have been trained to calculate DFT and CCSD(T) energies of small organic compounds containing C, H, O and N atoms in non-equilibrium conformations [107]–[109]. The extended version, ANI-2x, has been shown to predict

DFT energies of equilibrium or non-equilibrium conformations of molecules containing C, H, O, N, S, F and Cl atoms. It has been shown to reproduce the energies at the accuracy of the wb97x/6-31G* level but millions of times faster than the actual QM calculations [110]. The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the cause of the coronavirus disease pandemic in 2019 [111]. Since the first case of COVID-19 in 2019, more than 6.8 million deaths have been reported worldwide (World Health Organization. Weekly Epidemiological Update on COVID-19., n.d.). The coronavirus disease 2019 (COVID-19) pandemic, caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has destroyed health systems, societies, and economies. After the first identification of the SARS-CoV-2 strain coming from Wuhan, several variants of concern (VOCs) have been identified. Clinical reports and epidemiological features of the infection indicate a relatively mild disease flow and increased human-to-human virus spreading [113],[115]. Improved COVID-19 treatment and prevention techniques are urgently required given the increasing SARS-CoV-2 infection rate and lack of efficient treatment options. Researchers around the world are working to develop treatments and vaccines to struggle the disease, and several drugs have been approved for emergency use by regulatory agencies. For now, only two SARS-CoV-2 oral drugs (Paxlovid and Molnupiravir) have been clinically approved [116],[117]. Nevertheless, the antiviral potency of these compounds is not adequate to deal with the pandemic. Paxlovid (Nirmatrelvir with Ritonavir) has been reported to permit the SARS-CoV-2 replication upon completion of 5 day long oral take [118]. In addition to limited number of vaccines and drugs, the emergence of drug-resistant variants of SARS-CoV-2 brings additional concerns [119],[120] about the use of those drugs [121]. Therefore, developing anti-viral oral therapeutics for COVID-19 is still highly demanded.

Similar to other viral infections, once entering the host cell cytoplasm, the SARS-CoV-2 viral genome is translated into approximately 30 proteins. At first, 16 of these proteins are translated as two polyproteins, and in order to continue infection, these polyproteins must be split apart into the two proteins that virally encoded proteases; the major viral protease known as 3CLpro (or Mpro), and the protease papain-like protease (PLpro), facilitate the cleavage of these polyproteins into structural and non-structural proteins (NSPs)

[122],[123], which play a crucial function in the transcription/replication during the infection. Since Mpro is a critical enzyme in the life cycle of the virus [124], it has been a validated high-profile antiviral drug target, and its inhibitors have been shown to have strong antiviral activity in cell cultures and animal models [125]–[134]. Other coronavirus enzymes, like the helicase and the RNA-dependent RNA polymerase, has potential to be target for the development of antiviral drugs, such efforts are currently limited because these enzymes don't have crystal structures [135],[136].

Main protease (Mpro) is a three-domain (domains I-III) cysteine protease consisting of 306 amino acids. Anti-parallel β -barrel structures takes place domain I and domain II (residues 8–101/ residues 102–184, respectively), and domain III (residues 201–303) is connected to domain II through a long loop region (residues 185–200) that contains 5 α -helices arranged in one substantially antiparallel globular cluster [137]. The glutamine residue in the P1 position of the substrate is cleaved by Mpro using the protease Cys145-His41 dyad, where the cysteine thiol serves as the nucleophile in the proteolytic reaction.

Designing a therapeutic molecule from the ground up to treat an illness can become extremely costly and a time-taking process. In particular, for pandemics with such high transmission rates such as SARS-CoV-2, it requires prompt actions for controlling the disease. Instead, drug repurposing strategy might offer inexpensive and faster solutions to the problem. Repurposed drugs that have existing clinical data on the effective dose, treatment duration, side effects, and toxicity could be rapidly translated into the treatment of patients.

High throughput virtual screening by means of drug repurposing have been used to identify safe-in-human drugs with potential anti-SARS-CoV-2 properties. Most of these studies are based on molecular docking. Although it is the gold standard method to find the binding mode, docking is a very coarse method and almost never predicts the correct experimental binding affinity trend among the inhibitors. One strategy to overcome this limitation is to use consensus scoring [122].

More sophisticated methods to calculate the potential binding free energy of inhibitor candidate to the protein ranges from post molecular dynamics simulations such as Molecular Mechanics Poisson-Boltzmann Surface Area (MM-PBSA) to perturbation methods such as Bennett acceptance ratio (BAR), the latter being much more accurate yet quite costly. Although there have been numerous attempts that combine docking and MD based free energy methods in virtual screening, the number of candidates is mostly reduced to tens of hit compounds prior to MD simulations in these studies due to computational cost. Therefore, the top pose from docking is usually used in MD simulations and free energy calculations. Although different docking algorithms are successful in prediction of correct binding mode in the top three poses [138], MD simulations towards screening are performed only on the first pose. This limits the success of the binding free energy calculations (BFE) when the ligand is totally mis-oriented (such as flipped) among the top poses in docking since one cannot expect the MD simulations' correcting these drastic changes no matter how long the simulations performed. Therefore, running separate MD simulations for all the top poses may become a necessity to correctly predict BFEs.

One-trajectory approach end-state BFE calculations are quite attractive since they require only one MD simulation for each protein–ligand (PL) complex system with less computational cost. Although relatively better than docking scores, the accuracy in end-state BFE methods is still low due to over-simplifications such as implicit solvent definition in the case of MM-P(G)BSA and molecular mechanics (MM) definition of the Hamiltonian of the system. On the other hand, several new implementations have been introduced to end-state methods to improve accuracy. We have recently implemented the use of ML based potentials as a post MD simulation to improve the accuracy of BFE calculations [86],[88]. Herein, we introduce a unique virtual screening protocol to overcome the drawbacks.

Here, we first investigate and unravel the capability of ANI potentials as a rescoring function in molecular docking. Our results show that the docking power of ANI potentials can compete with the current scoring functions at the same level of computational cost. We then screened a library of 2500 clinically used drugs, either approved for human use

or with extensive safety data in humans (phase 2 or 3 clinical trials), for their ability to bind SARS-CoV-2 Mpro using consensus docking scores of ANI/GOLD and free energies by MMPBSA/ANI_LIE.

3.1.2. Results and discussion

3.1.2.1. Benchmarking with other docking software on CASF-2016 dataset

In the first part, we tested the scoring produced by ANI predicted interaction energies on a standard dataset containing diverse protein-ligand complexes. Using ANI, we have rescored all the poses belonging to 254 protein-ligand complexes and compared to the results of the original publication of CASF-2016 dataset, which reports the performance of 34 different software/algorithm for the same poses. Figure 3.1 shows the success rate of predicting the poses with less than 2 Å in the top ranked solutions. The ANI based scoring has been applied when only residues in docking grid are considered as the host in the host-guest interaction, outperforming most of the other 30 scoring functions.

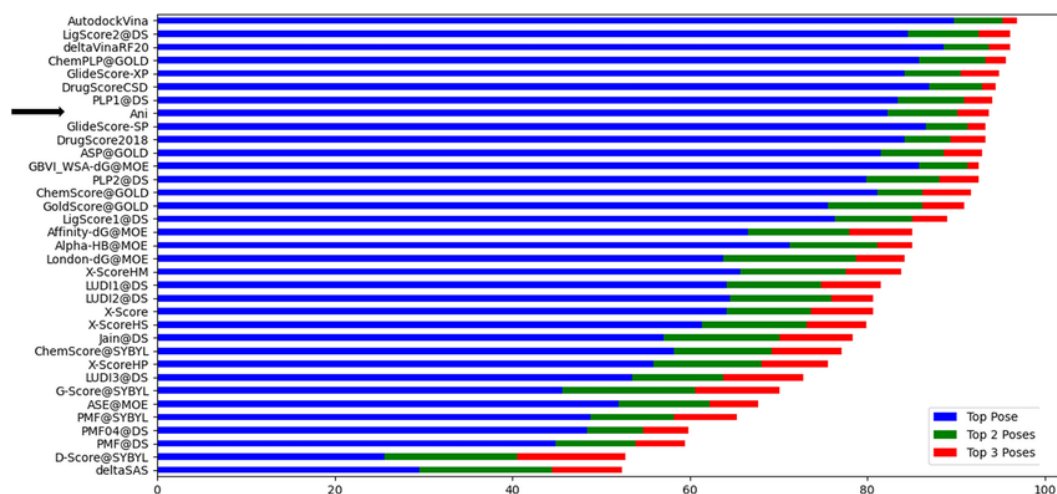


Figure 3.1: The success of different scoring functions on predicting the first poses as the closest structure to the crystal with RMSD less than 2.0 Å. ANI is ranked 8th among 34 functions.

We have also compared the success rate of finding the exact crystal structure in the first ranked solutions.

In addition to docking power, Yang et al. also defined as “binding funnel analysis” [83], which refers to the docking efficiency for a scoring function. The idea is that the poses that are closer to the crystal structure and those that are far from the crystal structure create a potential energy surface that will look like a funnel. The poses that are closer to the crystal structure will be in the lower energy region of this funnel while the poses that are far from x-ray structure will lie on the higher energy regions. We have performed the same analysis for ANI as well as borrowing the rest of the methods from literature [83] for our set of 254 proteins each with 5 ligands and 100 poses.

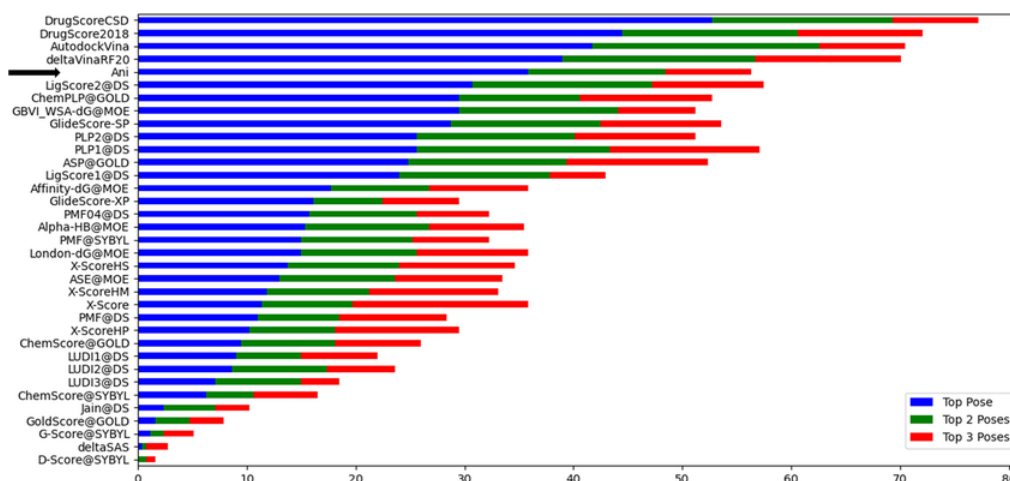


Figure 3.2: The success of different scoring functions on predicting the first poses as the exact crystal structure. ANI is ranked 5th among 34 functions.

We have created bins of RMSD windows with 0-2, 0-4, .. 0-10 Å and grouped poses according to these bins. For each RMSD window, Spearman’s correlation coefficient was calculated (Figure 3.3). The analysis show that the docking efficiency of ANI based scores are one of the most accurate methods, giving high Spearman’s correlation coefficients in the lower RMSD bins.

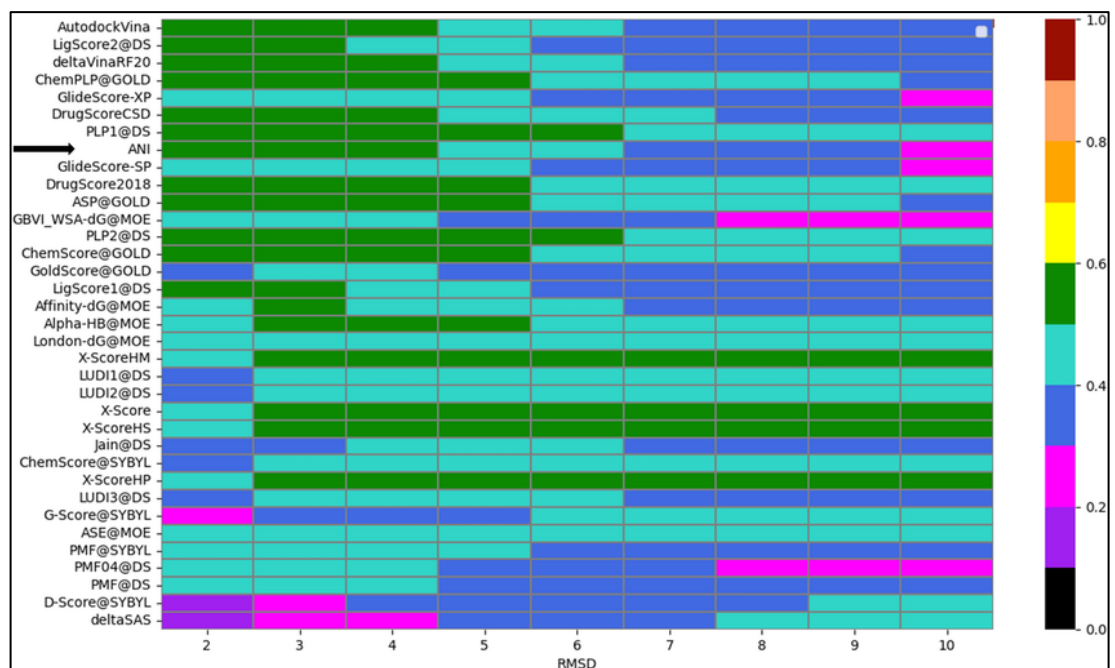


Figure 3.3: The Spearman correlation coefficients between RMSD and scores for each scoring function being tested. The x-axis indicates the RMSD bins (≤ 2 , ≤ 3 , etc.) Heat map ranging between 0 and 1 indicates low/high correlations.

3.1.2.2. Assessment on SARS-CoV-2 with GOLD

In the second part, we assessed the performance of ANI scorings on a specific protein family of SARS-CoV-2 Mpro rather than diverse set of proteins since our focused study is to screen the FDA drugs against this protein. A representative structure of SARS-CoV-2 main protease, used in the docking was shown in Figure 3.4.

In order to assess the ANI's performance using our own docking protocol in GOLD for SARS-CoV-2 main protease-inhibitor complexes, self-docking is a necessity. Nevertheless, some of the complexes have missing atoms/residues. To overcome this problem, docking all the ligands into a representative receptor structure that is fully resolved (i.e. 7L14) can be a better approach. This would also eliminate possible errors due to the use of different receptors in comparison with the ligand affinities.

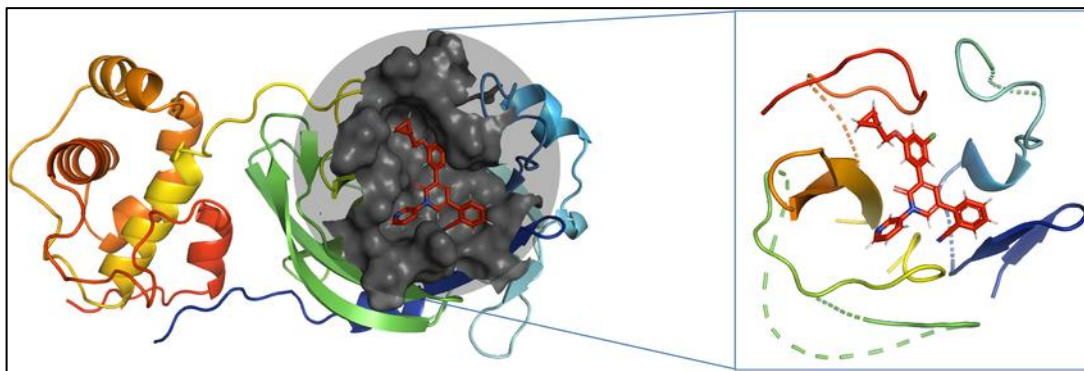


Figure 3.4: The structure of the SARS-CoV-2 main protease (PDB ID:7L14) used in GOLD docking. The nearby residues around ligand by 10 Å are highlighted. ANI scores were calculated using $\Delta G_{ANI} = E_{complex} - (E_{protein} + E_{ligand})$.

However, the complexes available in the PDB show structural differences and residues in the active site show side chain movements. Therefore, selecting only one receptor as a representation for all crystals has also complications to find correlation between docking scores and native binding pose. In order to avoid these error sources, we have selected the most common conformation as a representative structure (i.e. 7L14) for docking all the inhibitors. We have specifically selected 13 other protein-inhibitor complexes available in PDB that have similar scaffold in the binding region. The list of residues in the grid sphere and their RMSD values to the reference structure 7L14 is given in Table 3.1. In all of these crystal structures, the positions of the residues in the active site is protected. This allows us to compare the docking of the native ligands of these proteins to a reference protein (7L14).

Table 3.1: RMSD values of amino acids residing in the docking grid sphere to the reference structure of 7L14

Res. Num.	7N44	7L10	7L11	7L12	7M8M	7M8P	7M8O	7M8N	7M8X	7M8Y	7M90	7M8Z	7M91
21	0.5	0.5	0.1	0.4	0.3	0.5	0.1	0.4	0.5	0.5	0.4	0.4	0.4
22	0.6	0.5	0.1	0.4	1.0	0.9	0.2	0.3	0.4	0.5	0.4	0.4	0.5
23	0.7	0.6	0.1	0.5	0.9	0.8	0.4	0.5	0.5	0.8	0.4	0.7	0.7
24	1.0	0.5	0.1	0.5	1.6	0.8	1.3	1.3	0.3	1.5	0.3	0.5	0.6
25	0.5	0.5	0.2	0.4	0.4	0.4	0.2	0.5	0.5	0.5	0.4	0.3	0.4
26	1.3	0.5	0.1	0.5	0.3	0.3	0.3	0.5	0.5	1.3	0.5	0.3	0.4
27	0.3	0.3	0.2	0.3	0.2	0.1	0.2	0.2	0.2	0.3	0.3	0.2	0.2
28	0.2	0.2	0.1	0.2	0.1	0.1	0.2	0.2	0.1	0.1	0.2	0.1	0.2
29	0.2	0.1	0.1	0.1	0.1	0.1	0.2	0.1	0.2	0.1	0.1	0.2	0.2

Table 3.1: Continued.

38	0.3	0.1	0.1	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
39	0.4	0.1	0.1	0.3	0.1	0.1	0.2	0.2	0.2	0.2	0.2	0.2	0.1
40	0.6	0.3	0.1	0.6	0.2	0.1	0.3	0.2	0.2	0.5	0.2	0.3	0.3
41	1.3	0.2	0.1	1.2	0.2	0.2	1.2	0.3	1.2	0.4	1.2	0.3	0.3
42	0.5	0.3	0.1	0.5	0.2	0.2	0.2	0.3	0.3	0.4	0.2	0.2	0.2
43	0.7	0.4	0.2	0.9	0.2	0.2	0.5	0.3	0.4	0.7	0.4	0.3	0.3
44	1.1	0.5	0.3	1.1	0.1	0.2	0.4	0.6	0.5	1.0	0.5	0.5	0.6
49	2.2	1.1	0.3	2.0	0.6	0.8	0.9	0.8	1.5	1.5	1.3	1.4	1.0
50	3.0	1.4	0.2	3.3	0.2	0.5	0.7	0.9	2.4	2.6	2.5	1.4	1.2
51	2.0	0.9	0.3	2.1	0.2	0.4	0.5	0.6	1.2	1.7	1.5	1.5	0.6
52	1.1	0.4	0.2	1.0	0.2	0.5	0.4	0.7	0.2	0.8	0.2	0.4	0.3
54	1.1	0.3	0.2	1.0	0.3	0.2	0.3	0.4	0.3	0.8	0.3	0.4	0.4
57	1.1	0.6	0.2	1.1	0.3	0.3	0.3	0.4	0.8	0.9	0.6	0.7	-
85	0.6	0.2	0.1	0.5	0.2	0.1	0.2	0.1	0.2	0.4	0.3	0.2	0.3
114	0.1	0.2	0.0	0.1	0.1	0.2	0.2	0.2	0.2	0.2	0.1	0.2	0.2
f116	0.2	0.3	0.1	0.2	0.1	0.1	0.2	0.2	0.3	0.1	0.2	0.3	0.3
117	0.1	0.2	0.1	0.1	0.1	0.2	0.2	0.1	0.3	0.2	0.2	0.2	0.2
118	0.3	0.3	0.1	0.2	0.1	0.2	0.2	0.2	0.3	0.2	0.2	0.2	0.2
119	1.2	1.2	0.1	0.2	1.1	0.2	1.2	1.1	1.2	1.1	1.1	1.1	1.1
120	0.2	0.3	0.2	0.2	0.1	0.2	0.5	0.3	0.2	0.1	0.3	0.2	0.2
126	0.3	0.3	0.0	0.4	0.2	0.3	0.2	0.1	0.5	0.4	0.1	0.3	0.3
135	0.2	0.1	0.1	0.2	0.2	0.3	0.4	0.2	0.2	0.3	0.2	0.1	0.1
136	0.3	0.1	0.1	0.2	0.2	0.2	0.3	0.2	0.2	0.3	0.2	0.1	0.1
138	0.3	0.1	0.1	0.3	0.4	0.1	0.1	0.2	0.2	0.3	0.2	0.2	0.2
139	0.2	0.2	0.1	0.2	0.2	0.1	0.1	0.2	0.2	0.3	0.2	0.2	0.2
140	0.2	0.1	0.1	0.2	0.3	0.2	0.2	0.2	0.2	0.2	0.2	0.1	0.1
141	0.3	0.3	0.1	0.3	0.3	0.3	0.6	0.4	0.4	0.3	0.2	0.2	0.2
142	1.7	-	1.8	2.1	1.9	1.8	1.7	1.9	0.4	1.7	1.6	1.6	-
143	0.3	0.3	0.1	0.2	0.1	0.0	0.2	0.2	0.3	0.2	0.2	0.2	0.3
144	0.2	0.2	0.1	0.2	0.1	0.1	-	0.2	0.2	0.1	0.2	0.2	0.2
145	0.3	0.2	0.1	0.2	0.1	0.2	0.2	0.2	0.1	0.2	0.2	0.2	0.2
146	0.2	0.1	0.1	0.1	0.1	0.1	0.2	0.2	0.1	0.1	0.2	0.1	0.2
147	0.1	0.1	0.1	0.1	0.1	0.1	0.2	0.2	0.1	0.1	0.1	0.1	0.2
161	0.2	0.1	0.0	0.1	0.1	0.2	0.1	0.2	0.1	0.1	0.2	0.1	0.1
162	0.3	0.1	0.2	0.2	0.1	0.2	0.2	0.3	0.1	0.2	0.2	0.1	0.1
163	0.2	0.1	0.1	0.1	0.2	0.2	0.2	0.3	0.1	0.2	0.1	0.1	-
164	0.3	0.1	0.1	0.3	0.2	0.2	0.2	0.3	0.1	0.2	0.3	0.2	0.2
165	0.3	0.2	0.1	0.2	0.7	0.7	1.0	0.7	0.3	1.1	0.8	0.6	0.3
166	0.3	0.5	0.5	0.2	0.6	0.6	0.4	0.6	0.4	0.2	0.4	0.4	0.5
167	0.6	0.2	0.3	0.6	0.2	0.6	0.4	0.4	0.3	0.6	0.3	0.2	0.3
168	1.1	0.4	0.6	1.0	0.2	0.9	0.8	0.7	0.6	1.1	0.3	0.2	0.4
169	1.1	0.4	0.6	1.2	0.2	1.1	0.9	0.9	0.6	1.2	0.1	0.2	0.3
170	0.5	0.3	0.3	0.6	0.2	0.6	0.4	0.5	0.3	0.5	0.1	0.1	0.1
171	0.5	0.1	0.2	0.6	0.2	0.4	0.4	0.4	0.3	0.5	0.2	0.2	0.2
172	0.3	0.1	0.1	0.3	0.3	0.2	0.3	0.2	0.1	0.2	0.2	0.1	0.1
173	0.2	0.1	0.1	0.1	0.3	0.2	0.2	0.2	0.1	0.2	0.2	0.1	0.2
174	0.3	0.1	0.1	0.2	0.2	0.2	0.1	0.2	0.1	0.1	0.3	0.1	0.1
175	0.3	0.1	0.1	0.2	0.1	0.2	0.2	0.3	0.1	0.2	0.3	0.1	0.1
181	0.4	0.2	0.1	0.3	0.2	0.3	0.3	0.2	0.2	0.4	0.4	0.2	0.2
184	0.4	0.3	0.1	0.3	0.2	0.5	0.5	0.5	0.2	0.3	0.5	0.3	0.3
185	0.3	0.3	0.1	0.2	0.2	0.5	0.3	0.5	0.2	0.2	0.3	0.3	0.3

The top ranked solutions in GOLD and ANI mostly agree and are quite successful in predicting crystal structures. For half of the docked ligands, the top solutions for GOLD and ANI are the same. In addition, the top 3 ranked solutions for GOLD for each of the ligands were also within the first three solutions of ANI.

The success rate of finding the lowest RMSD structure in the top ranked solution was 35.7% for GOLD and 28.6% for ANI. In addition, the average RMSD values of the first poses in GOLD and ANI were 1.48 Å and 1.43 Å, respectively.

For $\text{RMSD} \leq 1.5$ Å, these values are 57.1% and 50.0% ≤ 1.5 Å, respectively; and 92.9% and 85.7%, respectively for $\text{RMSD} \leq 2.0$ Å. ANI's success in the top 1 pose is almost 7% better than GOLD's success. In general, the top 1 score is considered in the high throughput virtual screening (HTVS) studies aiming to find the putative binders for more advanced computational analysis such as molecular dynamics (MD). Given the fact that the ANI scores are very effective in finding the correct binding conformation at the first pose, it demonstrates the capacity of the ANI based rescoring in virtual screening studies.

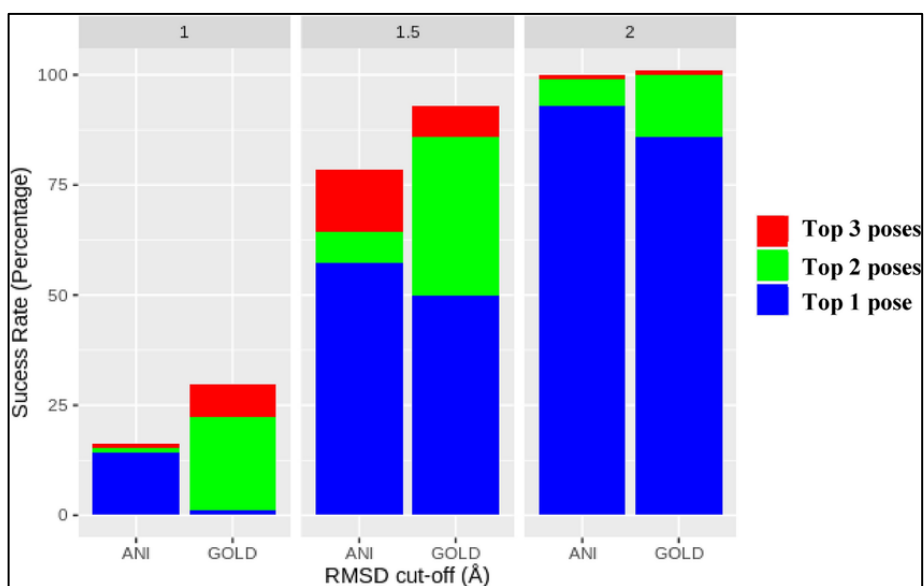


Figure 3.5: GOLD-PLP and ANI scoring on the poses generated by Gold docking on 14 selected inhibitors of SARS-CoV-2 main protease. ANI shows superiority to GOLD in the top one solution.

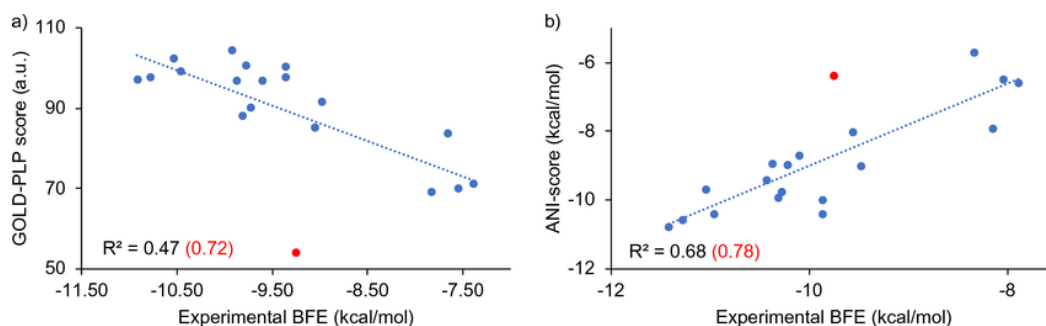


Figure 3.6: Docking scores generated by a GOLD-PLP and ANI with respect to experimental BFEs. ANI scores were produced by scaling with 0.127 as outlined in [41].

The trend between the docking scores of the best poses obtained with the GOLD software and the experimental binding energies shows that the Pearson correlation value is $R^2 = 0.47$ (Figure 3.6). GOLD scores are arbitrary and the highest scoring values correspond to the highest binding energy (negatively, the lowest binding energy), so the correlation should be inverse. It appears that the GOLD scores are actually in good agreement with the experimental values for the 19 compounds studied. In fact, when the D3F ligand is considered as an outlier, the success of GOLD scores reaches $R^2 = 0.72$. This ligand contains multiple nitro ($-\text{NO}_2$) groups and each $\text{N}=\text{O}$ bond is assumed double bonds (causing N atom to make 5 bonds) in GOLD due to the difficulty to define the partial atomic charges for these groups as discussed in the GOLD manual [85]. This might explain too low docking scores for this ligand. The ANI interaction energy between the amino acids in the binding site and the ligand showed a better trend with the experimental binding energies. Pearson correlation coefficient was found to be $R^2=0.68$, outperforming the correlation of GOLD scores. Similarly, when the D3F ligand (red dots in the plots) is considered as an outlier, this coefficient increases up to 0.78.

Instead of assessing the performance of ANI and GOLD scores individually, we have also compared when both methods are combined in consensus scoring. A much higher trend with $R^2=0.84$ was observed to the experimental values when the top three poses commonly recommended for both GOLD and ANI are considered, clearly supporting the idea of combining these two different methods (GOLD+ ANI) as consensus scoring.

3.1.2.3. Screening FDA drugs for SARS-CoV-2 Mpro

Our unique screening protocol is given in **Figure 3.7**. MD simulations were initiated using the top three poses obtained from docking. The docking of all FDA-approved drug (2500) molecules in the Zinc database was performed using GOLD. 25 poses for each drug molecule produced by GOLD and rescored with ANI for an additional filtering tool. Since ANI can only perform calculations for C, H, O, N, S, F, and Cl atoms, the number of drugs that are screened decreased to 1460. Our consensus scoring based on GOLD and ANI is as follows: if the top scored three poses generated by GOLD are also ranked in the top three scores with ANI, then these poses are used to MD simulation in the later stage. Otherwise, this complex is discarded. As a result of this filtering process, only 771 out of the best 4,380 GOLD docking poses for 1,460 drug molecules were also among the top 3 poses according to ANI. All 771 poses belonging to 669 different drug molecules were subjected to MD simulations for 10 ns in the next stage (a total of 7.71 μ s of simulations).

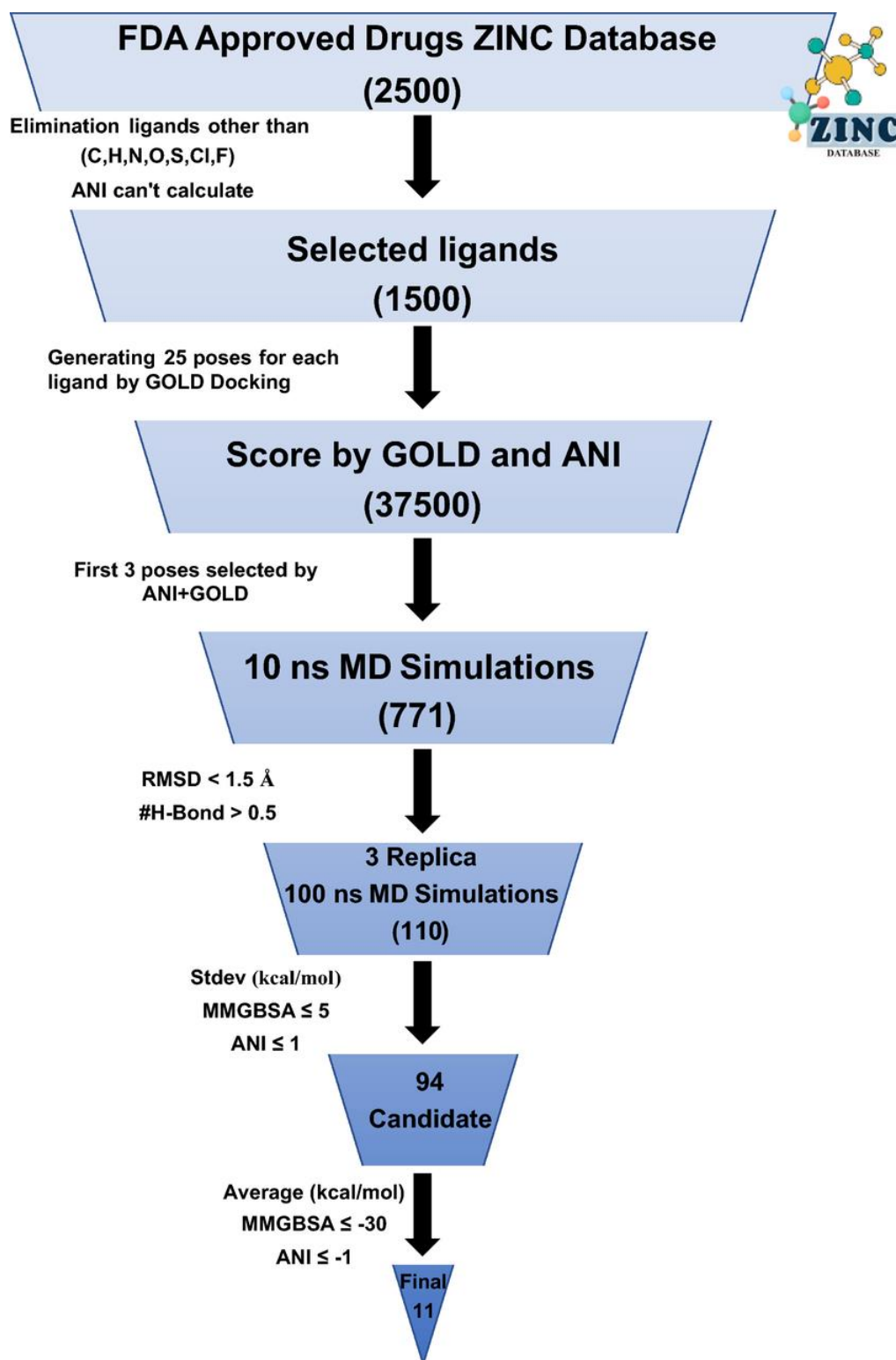


Figure 3.7: Our unique screening protocol involving consensus scoring of ANI and GOLD, classical MD simulations and consensus BFEs by end-state methods.

After the MD simulations, the resulting conformations were evaluated from several aspects to further filter possible SAR-CoV-2 main protease inhibitor candidates. Firstly, it was monitored how far the ligand moved away from the binding site during the MD simulation. If the ligand stably binds to the binding site, it will not move far away from the initial structure. We measured this by fitting the MD trajectory on proteins and calculating the RMSD values of the ligands. The cut-off for the RMSD for this stability test was 3.0 Å. Based on this evaluation, 467 drug molecules remained stable in the binding site. The second criterion for filtering out the ligands is protection of the number of H-bonds. By monitoring the number of H-bonds between the ligands and protein, we eliminated those structures that lose more than 0.5 H-bonds throughout the simulation from our hit list. Figure 3.8 shows the examples for these criteria. Thus, a total of 346 molecules were identified as potential candidates for binding to the protein.

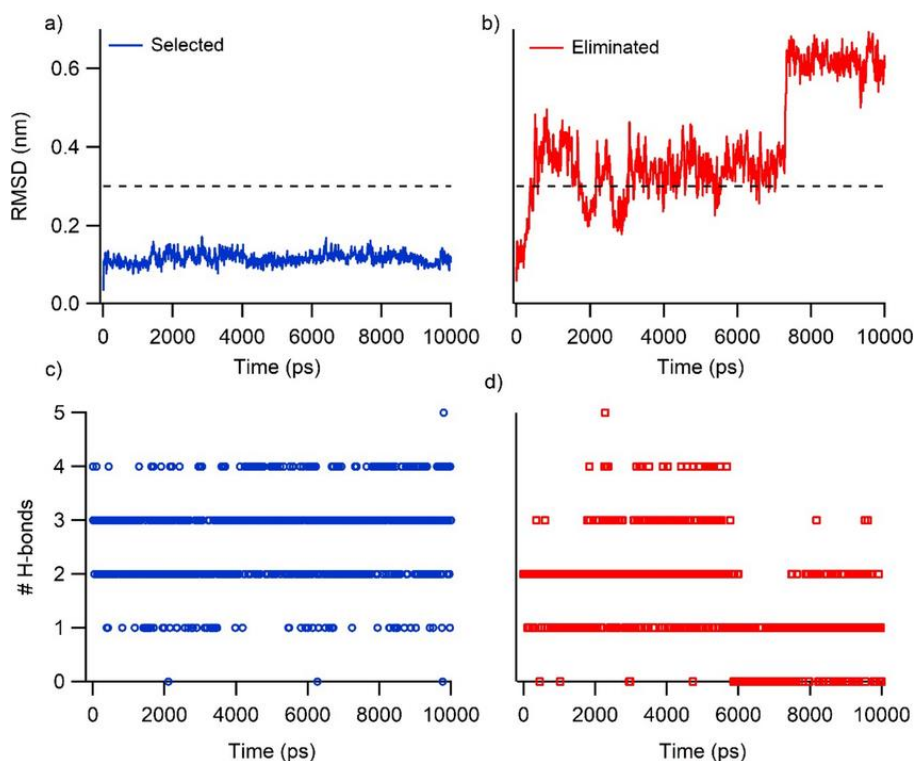


Figure 3.8: Application of elimination (red)/selection (blue) criteria to MD trajectories. $\Delta(\text{RMSD}) \leq 3.0 \text{ \AA}$ for the ligand, and $\Delta(\# \text{H-bonds}) \geq -0.5$ between ligand and protein.

As a final criterion in the classical MD simulations, the stability of the protein-ligand complex was assessed by means of RMSD of back bone atoms of the protein with a cutoff of 0.15 Å (Figure 3.9). This eliminated further drug poses, leaving over 110 complexes.

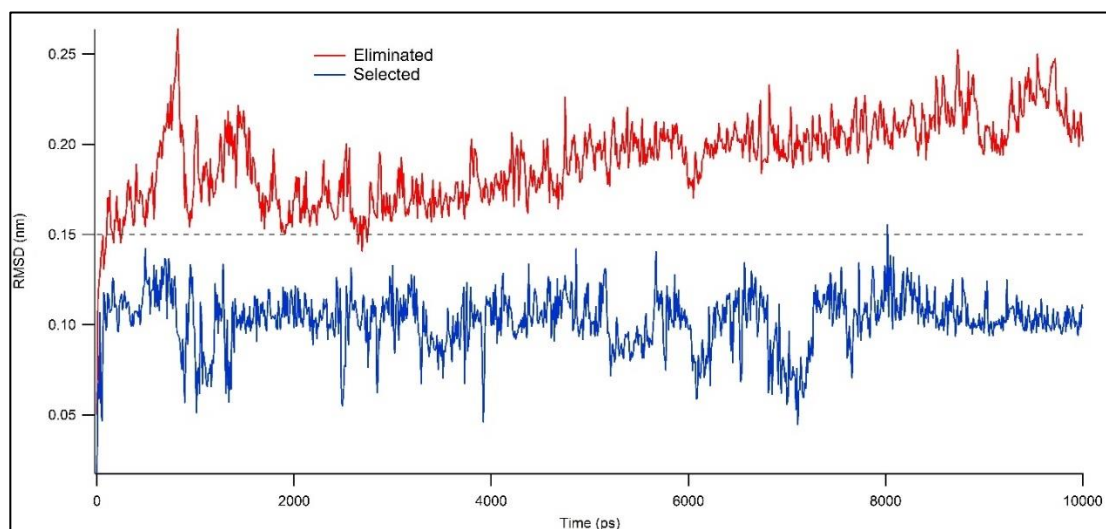


Figure 3.9: An example RMSD of protein backbone atoms for selected (blue) and eliminated (red).

In the next step, we performed end-state free energy calculations on remaining 110 compounds each followed by three replica 10 ns long MD simulations (a total of 2.2 μ s additional MD simulations). For these selected 110 compounds, free energies were calculated using MMGBSA and ANI_LIE methods (**Table 3.5**). Ideally, if a drug-pose interacts with M^{pro} strongly, the free energy values would be much lower. However, if any of the BFE values of the replica simulations is not in the same order as with the other two simulations, then the value predicted is not reliable and should not be trusted. Therefore, we have applied another criterion of standard deviation among the free energies calculated by the replica simulations (5 kcal/mol MMGBSA and 1 kcal/mol ANI_LIE).

Table 3.2: Long MD simulations and their calculated free energy values for selected 110 compounds.

	GOLD score	MMPBSA			ANI		
Name		1	2	3	1	2	3
Cabazitaxel	58.3	-44.7	-41.4	-46.7	-11.1	-10.2	-10.9
Rivaroxaban	61.0	-45.7	-39.5	-36.3	-11.1	-10.2	-9.8
Dapagliflozin	69.6	-31.3	-33.2	-35.9	-11.2	-9.6	-10.2
Miglitol	43.9	-21.3	-22.6		-10.6	-11.4	-8.9
Pentostatin	51.1	-27.4	-29.3	-20.8	-10.1	-10.5	-9.1
Azacitidine	44.3	-5.6	-18.3	-20.9	-8.1	-10.3	-10.8
Apixaban	64.2	-35.1	-22.6	-33.6	-9.8	-8.1	-9.7
Tetracycline	52.7	-30.2	-33.4	-21.5	-9.4	-9.6	-8.5
Acalabrutinib	68.3	-44.0	-38.2	-35.7	-10.1	-8.8	-8.6
vinorelbine	50.7	-46.0	-49.4	-32.8	-9.3	-9.8	-8.1
Sulfadoxine	52.1	-25.0	-23.9	-22.4	-9.1	-8.9	-8.3
Dorzolamide	43.5	-14.5	-24.1	-15.4	-8.2	-9.5	-8.6
Zaleplon	61.7	-36.7	-32.2	-34.6	-8.5	-8.6	-9.1
Sarecycline	58.3	-26.6	-28.4	-30.0	-8.2	-8.7	-9.3
Sotorasib	62.0	-31.9	-32.8	-29.8	-8.9	-8.6	-8.7
Voriconazole	51.6	-25.1	-32.3	-20.3	-9.0	-8.7	-8.3
protein_gold_soln_ob257	62.9	-40.4	-39.3	-35.3	-8.6	-8.9	-8.3
protein_gold_soln_ob852	47.4	-21.6	-22.4	-19.0	-9.0	-8.6	-8.1
protein_gold_soln_ob595	53.5	-24.4	-23.4	-21.1	-8.9	-8.5	-8.3
Prucalopride	54.0	-32.0	-32.4	-30.6	-8.7	-8.5	-8.4
(-)-O-Desmethylvenlafaxine	53.2	-23.4	-23.0	-24.5	-9.1	-8.5	-8.0
Fluticasone furoate	53.8	-34.5	-28.4	-20.3	-9.2	-7.8	-8.5
Minocycline	48.4	-26.5	-35.1	-26.8	-8.5	-8.8	-8.0
Rofecoxib	50.0	-26.6	-26.0	-26.3	-8.4	-8.3	-8.4
(S)-Praziquantel	54.0	-33.2	-27.8	-28.7	-8.8	-8.0	-8.3
Meclocycline	51.7	-28.7	-25.1	-23.7	-7.9	-8.7	-8.4
Desmethyl Tramadol	49.4	-21.7	-19.9	-26.9	-8.4	-8.1	-8.5
Lurbinectedin	57.3	-29.2	-32.1	-28.8	-9.0	-8.1	-7.8
(R)-dyphylline	50.8	-18.9	-14.2		-8.9	-7.6	-7.8
protein_gold_soln_ob602	67.1	-35.2	-33.5	-29.9	-8.6	-8.2	-8.0
Ascorbic Acid	42.7	-9.0	-16.0	-16.5	-7.4	-8.7	-8.6
mitomycin C	45.3	-25.9	-17.0		-7.9	-8.1	-8.6
Thiethylperazine	65.4	-34.1	-34.5	-34.7	-8.2	-8.2	-8.2
Nelarabine	49.0	-23.0	-20.0	-6.6	-9.0	-8.2	-7.3
MEBUTAMATE	36.6	-12.7	-20.1	-17.5	-7.4	-8.9	-8.2
(R)-donepezil	65.2	-41.3	-27.9	-37.3	-8.5	-7.4	-8.7
Sulfacetamide	44.7	-17.7	-22.8	-16.1	-7.9	-9.0	-7.7
Norfludiazepam	47.7	-22.2	-23.6	-21.0	-7.4	-8.8	-8.1
Sulfadiazine	46.0	-16.3	-21.9	-17.4	-8.1	-8.1	-8.0
N-Desmethyl ulipristal acetate	56.6	-38.7	-35.5	-37.6	-8.2	-7.9	-8.0
Clopidogrel	51.9	-26.5	-23.9	-30.5	-8.1	-7.6	-8.3
Podofilox	53.4	-28.7	-29.2	-31.9	-8.1	-8.0	-7.9
(s)-Trihexyphenidyl	67.4	-32.1	-27.5	-32.2	-8.0	-7.9	-7.9
Temozolomide	37.1	-6.9	-13.4	-14.2	-7.3	-8.3	-8.2
Desipramine	62.7	-27.8	-28.5	-20.2	-8.3	-8.3	-7.3
Benzhydrocodone	48.3	-25.5	-33.8	-26.2	-7.9	-7.9	-7.8

Table 3.2: Continued.

Larotrectinib	73.9	-26.4	-28.5	-24.3	-7.6	-8.0	-8.0
(2R,3S)-2-(2,4-difluorophenyl)-3-(4-methylenepiperidin-1-yl)-1-(1H-1,2,4-triazol-1-yl)butan-2-ol	58.5	-23.8	-20.0	-24.5	-7.7	-7.7	-8.1
Lorlatinib	53.5	-29.4	-27.4	-21.6	-7.8	-7.9	-7.6
(1S)-2-amino-1-(2,5-dimethoxyphenyl)ethan-1-ol	46.1	-15.1	-20.8	-17.6	-7.5	-8.1	-7.7
Nandrolone	45.8	-25.9	-29.9	-20.3	-7.8	-8.7	-6.8
Atovaquone	56.5	-29.8	-27.8	-31.9	-7.8	-7.2	-8.3
Sulfapyridine	46.4	-18.7	-19.1	-19.2	-7.7	-7.8	-7.8
Equilin	46.7	-26.8	-25.3	-25.2	-7.6	-7.9	-7.8
Hydroxyprogesterone caproate	50.1	-32.8	-27.3	-33.3	-7.6	-7.6	-8.1
SCHEMBL13481056	49.2	-20.1	-19.5	-22.6	-8.1	-7.1	-8.0
Sulfamethazine	47.4	-21.5	-21.8	-21.8	-7.6	-7.9	-7.7
imidazoline	44.9	-24.7	-23.3	-19.6	-8.0	-7.6	-7.3
O-Desmethyldiamadol	49.5	-19.3	-25.8	-18.4	-7.0	-8.9	-6.9
Hydralazine	38.0	-17.3	-16.8	-11.5	-7.9	-7.8	-7.2
Menadiol	45.1	-16.4	-15.7	-16.1	-7.5	-7.8	-7.5
Difluprednate	54.3	-27.6	-27.2	-22.3	-7.4	-8.0	-7.3
Benzphetamine	56.7	-26.6	-20.4	-22.0	-8.0	-7.3	-7.4
Estrone - folliculin (has more name)	46.5	-27.3	-28.7	-25.9	-7.7	-7.8	-7.2
(5S)-5-ethyl-3,5-dimethyl-1,3-oxazolidine-2,4-dione	30.2	-17.5	-17.0	-17.0	-7.4	-7.7	-7.6
R-epinastine	44.9	-17.3	-23.8	-18.2	-7.3	-7.9	-7.4
Ondansetron, (3R)-	62.6	-27.2	-20.7	-25.9	-7.7	-7.1	-7.8
Menadiol	43.3	-16.5	-15.4	-13.5	-7.6	-7.3	-7.6
Propofol	41.1	-18.5	-21.0	-19.1	-7.5	-7.6	-7.5
2-[[[(S)-1,2,3,4-Tetrahydronaphthalene]-1-yl]-2-imidazoline	53.7	-23.9	-15.1	-21.3	-7.9	-7.0	-7.6
Trioxsalen	43.3	-23.1	-21.5	-22.1	-7.6	-7.3	-7.5
Sulfanilamide	37.9	-11.6	-11.0	-12.4	-7.7	-7.4	-7.3
(R)-Mepivacaine	46.8	-21.9	-21.4	-19.8	-7.4	-7.5	-7.5
Sulfanilamide	34.1	-13.8	-12.9	-13.3	-7.5	-7.5	-7.3
Oxcarbazepine	44.6	-16.6	-17.8	-18.4	-7.3	-7.6	-7.5
Oxypurinol	29.9	-7.2	-8.3	-14.3	-7.2	-7.3	-7.7
Lamotrigine	46.0	-19.7	-19.0	-16.8	-7.3	-7.8	-7.0
Ethylestrenol	58.3	-26.3	-23.4	-28.1	-7.0	-7.4	-7.6
Sertraline	55.1	-17.0	-22.2	-25.5	-7.2	-7.4	-7.4
Ethylestrenol	48.7	-28.2	-28.7	-26.5	-7.2	-7.7	-7.1
Varenicline	40.0	-18.7	-20.5	-18.1	-7.0	-7.6	-7.5
(R)-Mirtazapine	45.8	-22.0	-22.6	-23.4	-7.4	-7.2	-7.3
Dapagliflozin	49.5	-22.0	-23.6	-25.7	-7.1	-7.4	-7.4
Mitotane, (S)-	46.6	-24.9	-26.8	-27.4	-7.3	-7.1	-7.5
Alprazolam	49.8	-21.8	-21.1	-26.8	-7.3	-7.0	-7.6
Methoxsalen	42.1	-18.5	-19.7	-21.8	-7.0	-7.3	-7.5
Cyclizine	49.5	-22.0	-23.0	-20.3	-7.3	-7.3	-7.2
Norethynodrel	47.3	-24.6	-20.3	-26.9	-7.0	-7.0	-7.6
Clonidine	41.8	-16.8	-22.7	-17.0	-7.3	-7.3	-7.0
(R)-Ketamine	41.3	-21.9	-18.2	-19.2	-7.6	-6.9	-7.1
Bictegravir	59.9	-25.7	-18.6	-19.8	-8.0	-6.7	-6.8

Table 3.2: Continued.

Anagrelide	44.8	-20.6	-17.1	-23.6	-7.1	-7.2	-7.3
(S)-Sibutramine	46.8	-23.7	-24.0	-22.3	-7.4	-7.1	-7.0
Ticlopidine	56.2	-20.6	-21.9	-25.8	-7.0	-7.3	-7.2
Nicotinamide	32.4	-10.9	-11.1	-10.4	-7.2	-7.2	-7.0
Nestorone	45.6	-21.1	-22.8	-23.0	-7.0	-7.1	-7.2
methyl (S)-phenyl[(R)-piperidin-2-yl]acetate	51.4	-16.2	-19.6	-15.7	-7.1	-7.3	-6.8
(1S)-1-(adamantan-1-yl)ethan-1-amine	36.6	-13.4	-17.4	-16.7	-6.9	-7.1	-7.1
Isoniazid	32.8	-11.1	-1.4	-14.2	-7.2	-5.8	-8.1
Pyrimethamine	45.1	-23.6	-21.9	-22.8	-6.9	-6.9	-7.0
protein_gold_soln_ob1484	28.7	-10.8	-6.9	-4.5	-7.3	-6.8	-6.6
Pirfenidone	47.8	-10.0	-14.8	-15.2	-6.5	-6.9	-6.9
Fomepizole	27.6	-10.9	-3.5	-7.0	-7.5	-6.0	-6.8
Methyltestosterone	48.0	-15.5	-25.2	-9.7	-6.6	-7.0	-6.4
trans-Phenmetrazine	39.1	-13.1	-8.7	-6.6	-7.1	-6.6	-6.3
4-Methoxyphenol	35.4	-10.3	-13.8	-1.5	-7.0	-7.0	-5.8
Menadione	41.3	-16.5	-16.7	-13.0	-6.5	-6.8	-6.4
Deferiprone	35.6	-13.8	-14.7	-10.2	-6.5	-6.6	-6.6
protein_gold_soln_ob866	48.8	-21.4	-21.2	-8.7	-6.6	-6.6	-6.2
Pyrazinamide	32.0	-7.5	-3.5	-3.8	-6.8	-6.3	-6.2

In the final step, ANI_LIE and MM-GBSA were used in combination in a way that drug molecules with the lowest binding energies relative to other drugs were suggested by both methods were selected. When the average BFE values of three replica MD simulations for a drug molecule were below -8 kcal/mol by ANI_LIE and -30 kcal/mol by MM-GBSA, the drug molecule was assumed plausible inhibitor for the SARS-CoV-2 M^{pro}. The rest of the candidates were eliminated with the assumption that they would not be sufficient for inhibition. Thus, a total of 11 drug molecules remained for which significant binding energy was suggested by 2 different free energy methods. These drug molecules are listed in Table 3.3.

Table 3.3: The final drug list we recommended in the FDA drug repurposing study to target SARS-CoV-2 Mpro.

	GOLD Score (a.u.)	ANI Score (kcal/mol)	MM-PBSA (kcal/mol)		ANI_LIE (kcal/mol)	
Name			Avg.	Stdev	Avg.	Stdev
Cabazitaxel	58.3	-9.0	-44.3	2.6	-10.8	0.5
Rivaroxaban*	61.0	-8.4	-40.5	4.8	-10.4	0.7
Dapagliflozin	69.6	-10.9	-33.5	2.3	-10.3	0.8
Acalabrutinib*	68.3	-7.9	-39.3	4.2	-9.1	0.8
Zaleplon	61.7	-8.0	-34.5	2.2	-8.8	0.3
Sotorasib	62.0	-8.3	-31.5	1.5	-8.7	0.2
Lopac-T-9025	62.9	-8.5	-38.3	2.7	-8.6	0.3
Prucalopride	54.0	-8.0	-31.7	0.9	-8.6	0.1
Lurbinectedin	57.3	-7.7	-30.0	1.8	-8.3	0.6
Thiethylperazine	65.4	-9.0	-34.4	0.3	-8.2	0.0
N-Desmethyl ulipristal acetate	56.6	-7.0	-37.3	1.6	-8.0	0.2
* Compounds that have been in clinical trials or suggested in literature						

In all three methods, structures with the lowest binding energies were selected as potential drug molecules by consensus. Among the top 5 drug molecules suggested by the common three methods with the highest binding energies, 3 had already been proposed as potential drug molecules in previous theoretical and experimental studies. This confirms the effectiveness of the screening method used here.

The drug molecules listed here are the most plausible SARS-CoV-2 Mpro inhibitors and it is worth performing further experimental analysis to reveal their potency. In our previous studies, we have shown that ANI_LIE values are much more accurate in predicting absolute binding free energies [86]. Therefore, Cabazitaxel, Rivaroxaban and Dapagliflozin might have sub-nanomolar inhibition concentrations for Mpro due to BFEs below -10 kcal/mol.

3.1.3. Summary

In this study, we have shown the capability of using ANI potentials as a rescoring function in molecular docking. Our benchmarking studies showed that this method can outperform most of the conventional scoring functions. In particular, the prediction of the top 1 solution showed one of the best performances in prediction of the true binding mode. The method can be adopted in any docking software to screen the drug like molecules as potent inhibitors to proteins. As a case study, we introduce our unique screening protocol which

incorporates consensus scoring of ANI and GOLD, classical MD simulations and consensus BFEs by end-state methods.

3.2. DeepConf: Leveraging ANI-ML Potentials for Exploring Local Minima with Application to Bioactive Conformations

Adopted with permission from Tayfuroglu, O, Zengin, I.N., Koca, M.S., Kocak, A. DeepConf: Leveraging ANI-ML Potentials for Exploring Local Minima with Application to Bioactive Conformations. *J. Chem. Inf. Model.* (2025) 65 (6), 2818-2833.

Here, we introduce a low energy conformer generation algorithm using ANI-ML potentials at the DFT accuracy and benchmark in reproducing bioactive conformations. We show that the method is efficient when the initial starting structure is far from equilibrium, when the ML potentials are stuck in non-smooth regions, and when the quality of the conformers in a less conformer size is demanded. We specifically focus on conformations due to rotations around the single bonds. For the first time, we assess the performance of ANI-ML potentials using our conformer generation algorithm, DeepConf, in addition to previously reported Auto3D (*J. Chem. Inf. Model.* **2022**, 62, 5373–5382) using the same potentials to reproduce bioactive conformations as well as providing a guideline for bioactive conformation evaluation processes. Our results show that the ANI-ML potentials can reproduce the bioactive conformations with mean value of the root-mean-square-deviation (RMSD) less than 0.5 Å, outperforming the limit of conventional methods. The code offers several features including but not limited to geometry optimization, fast conformer generations via single point energies (SPE), different minimization algorithms, different ML-potentials, or high-quality conformers in the smallest amount of ensemble sizes.

3.2.1. Introduction

Isomerism (as constitutional isomerism and stereoisomerism) play a key role in physical organic chemistry as well as drug design and discovery [139],[140]. Experimental and

computational methods can easily elucidate the structural (constitutional) isomers since their physical and chemical properties are different due to different connectivity [141]. Similarly, configurational stereoisomers (around a double bond) usually require remarkably high energy for interconversion, making them easy to model computationally. For organic chemists, conformational isomers may not be a major concern since 2D representation is mostly sufficient to depict synthesized compounds. Likewise, experimental procedures for synthesis and characterization are mostly performed at the macroscopic level, so all the accessible conformational isomers are already reflected on the measurements as an equilibrated mixture under experimental conditions (e.g., room or physiological temperature, solution phase, etc.). However, for computer-aided drug design studies (CADD), conformational stereoisomers (conformers) around single rotatable bonds (i.e., torsion points) can become cumbersome due to their vast number of energetically accessible rearrangements, making the true 3D representation of a molecule vital. Indeed, most drug-like organic compounds contain multiple rotatable bonds, creating a conformational space that must be included in computational studies. [142],[143].

The energy change in response to the dihedral angle change around the four adjacent atoms in a single (rotatable) bond can be too small to accurately predict by computational methods. Even if the dihedral angle is predicted accurately, it only represents one of the many possible conformers (local minima). Finding the true 3D structure with the lowest energy of all conformers (global minimum) requires the energies of all minima to be calculated accurately. As a further complication in computer aided drug design, all conformers need to be evaluated at their biological target, as the most stable complex between a biological target molecule such as proteins and a drug-like molecule (i.e. the bound state) may not necessarily correspond to the global minimum of the drug-like compound in the free (unbound) state [142]. Moreover, the structural strain on the bioactive compound, which is relatively small at the torsion points, can be easily outweighed by the possible strong interactions with the biotarget compound, leading to the bound state of drug-like compounds possessing a non-equilibrium structure. For this reason, most computational tools, such as molecular docking, allow free rotation around

these torsion points and use the ligand strain energy in the binding affinity calculations as a compromise to the energy gain between the bound and the unbound states.

Indeed, there is an entire field in computational chemistry dedicated to the accurate and fast prediction of binding free energy (BFE) changes upon ligand binding to its biological target compound [86]–[88],[93],[144]–[175]. Thus, besides finding low energy conformers of a drug-like compound, it is also necessary to find the relative energies among the conformers and potential energy surface (PES) connecting these conformations. Numerous strategies, beyond the screening of conformers using molecular docking, have been developed, including induced-fit docking, ensemble docking, molecular dynamics (MD) and Monte Carlo (MC) simulations, all of which are involved at some level in the conformational space of the ligand in BFE calculations.

Most of the concern in isomerism arises from conformational isomers due to torsional points around single bonds, as almost all the conformational isomers are energetically accessible and populated to some extent. Thus, the number of torsion points defines (and mostly limits) the number of conformational isomers to be sampled to obtain more accurate energetics of the molecule. The best scenario for examining low energy lying conformations is to systematically calculate all possible conformations and then find all local minima along with the global minimum structure. However, as the number of torsion points increases, the number of different conformations to assess grows exponentially. For instance, a simple flip (180° rotation) around a single bond creates two possible conformations to evaluate, and if there is six such torsion points in a small molecule, a total of $2^6=64$ calculations (minima on PES) must be completed to identify the lowest energy (global minimum) and other low-energy (local minima) structures. For this reason, most conformer generation algorithms use stochastic search algorithms rather than systematically evaluating all conformers.

Ideally, all minima on the PES should be produced by multidimensional (3N-6) scanning of all coordinates at the most accurate quantum mechanical (QM) levels. Otherwise, a systematic evaluation of all possible conformers at the most accurate level is required to gain insight into the global minimum and all local minima, along with their relative energies and populations. Since there are energy barriers between different minima on the

PES, minimization algorithms using energy gradients (i.e., minimizing energy with respect to coordinates) only find the nearest minimum where the energy gradient (force) reaches to the zero threshold. Thus, an extensive search with different initial conformations is still required to locate all minima.

Over the years, numerous commercial (e.g., iCon [176], Omega [177], ConfGen [178], MOE[179], CREST [180]) and open-source tools (e.g., FROG2 [181],[182], Confab [183], OpenBabel [184], and RDKit [185]) have been developed for fast and accurate conformer generation. Most recently, Seidel et al. [142] and earlier Friedrich et al. [186] reviewed the performance of some of these software tools. Due to the enormous number of conformer calculations, most algorithms use molecular mechanics (MM)-based minima searches rather than more accurate QM methods. However, searching for minima using MM methods introduces additional problems. First, they rely on the accuracy of the force field (FF) used and to cover all chemical library, they need to be generalized (such as UFF, GAFF, MMFF94), which results in a loss of accuracy. Second, they use empirical parameters for bonds, angles, and dihedral definitions for covalent interactions, as well as fixed partial atomic charges and Lennard-Jones parameters for non-covalent interactions. They require well-defined bonding information in the initial structure. If the initial structure is incorrectly drawn (by the user) or the pre-assigned bond order is incorrect, MM based minimization algorithms may totally fail to locate the correct nearest minimum and other minima. Furthermore, even if the bonding is correctly defined and the local minimum is in the correct conformation, the relative energies are still in MM level, which might deviate from the actual values. For these reasons, there is a need to perform minimizations and search for conformers at QM accuracy. Searching for global and local minima using QM methods such as DFT can become computationally prohibitive due to the enormous number of conformation evaluations.

One strategy to balance QM accuracy and computational cost is to use machine learning (ML) techniques to model the multidimensional PES of the system at QM accuracy. To date, several neural network potentials (NNPs) have been developed to represent the multidimensional PES of systems at QM accuracy. In principle, if the PES of the system is well-represented by NNPs, the energy minimization and conformer search can be

performed using ML potentials rather than costly QM calculations or less accurate MM calculations.

The most popular ML potentials are ANI [187]–[190] (with versions of ANI-1, ANI-2x, and ANI-2xt) and AIMNET, which have been successfully utilized for energy minimization and low-lying minimum search using several gradient based minimization algorithms [191],[192]. A general issue with ML potential is that the PES is not expressed in an explicit mathematical function form. As a result, its smoothness depends heavily on the quality of the training data, and it may not be as smooth or physically interpretable as the PES from QM or MM methods. Having a non-smooth PES raises problems in gradient-based minimizations as they can easily be stuck on saddle-points where the energy derivation with respect to coordinate is zero. While such challenges are inherent to both smooth and non-smooth potential energy surfaces (PES), non-smoothness worsens the problem by introducing abrupt discontinuities in PES, often arising from undertrained ML models with insufficient or low-quality data. This issue is particularly pronounced for geometries far from equilibrium, where sampling is sparse.

In addition, drawing errors in typical chemical software (e.g., ChemDraw, MarvinSketch, GaussView, or Spartan) can lead to incorrect bond types, misassigned connectivity (e.g., single bonds instead of double bonds or missing ring closures), or improper bond lengths. While these programs generally provide ‘cleaning’ utilities based on molecular mechanics (MM), errors persist if the connectivity is initially misdefined.

Herein, we attempt to improve the minimization and local minima search based on ANI-ML potentials by introducing a strategy that generates multiple conformations, clusters them according to RMSD, and refines the representative conformer in each cluster. This multi-conformer approach broadens the search space and reduces the likelihood of the optimization becoming trapped in local minima, particularly in non-smooth regions of the PES. We start with small systems as a case study to overcome non-smooth regions, demonstrating that our conformer generation algorithm allows for successfully reaching local minima. Then, we extend our approach to a larger dataset, where bioactive conformations are generated using ANI-ML potentials, and evaluate its performance on this broader scope.

3.2.2. Results and Discussion

3.2.2.1. Workflow

DeepConf is written in Python 3.8 and uses ASE libraries to calculate energy and force components in the conformer generations and energy minimization processes. The workflow for the prediction of local minima and conformer generation is given in Figure 3.10.

The program starts reading the common 3D files (xyz, mol2, sdf and pdb file formats) from a directory of the ligands and fixes the structure by adding required hydrogen atoms if needed, upon user request. Depending on the user's choice, pre-optimization of the ligand is performed using several optimization methods (optimization_method=BFGS, LBFGS, FIRE, GPMIn or Berny) and the energy/force calculator (calculator_type=ANI-2x or G16) until the convergence threshold is achieved (set by thr_fmax and maxiter, corresponding to maximum force and maximum iterations). Then, the program identifies rotatable bonds (torsion points) using the RDKit library, with or without the ETKDG approach.

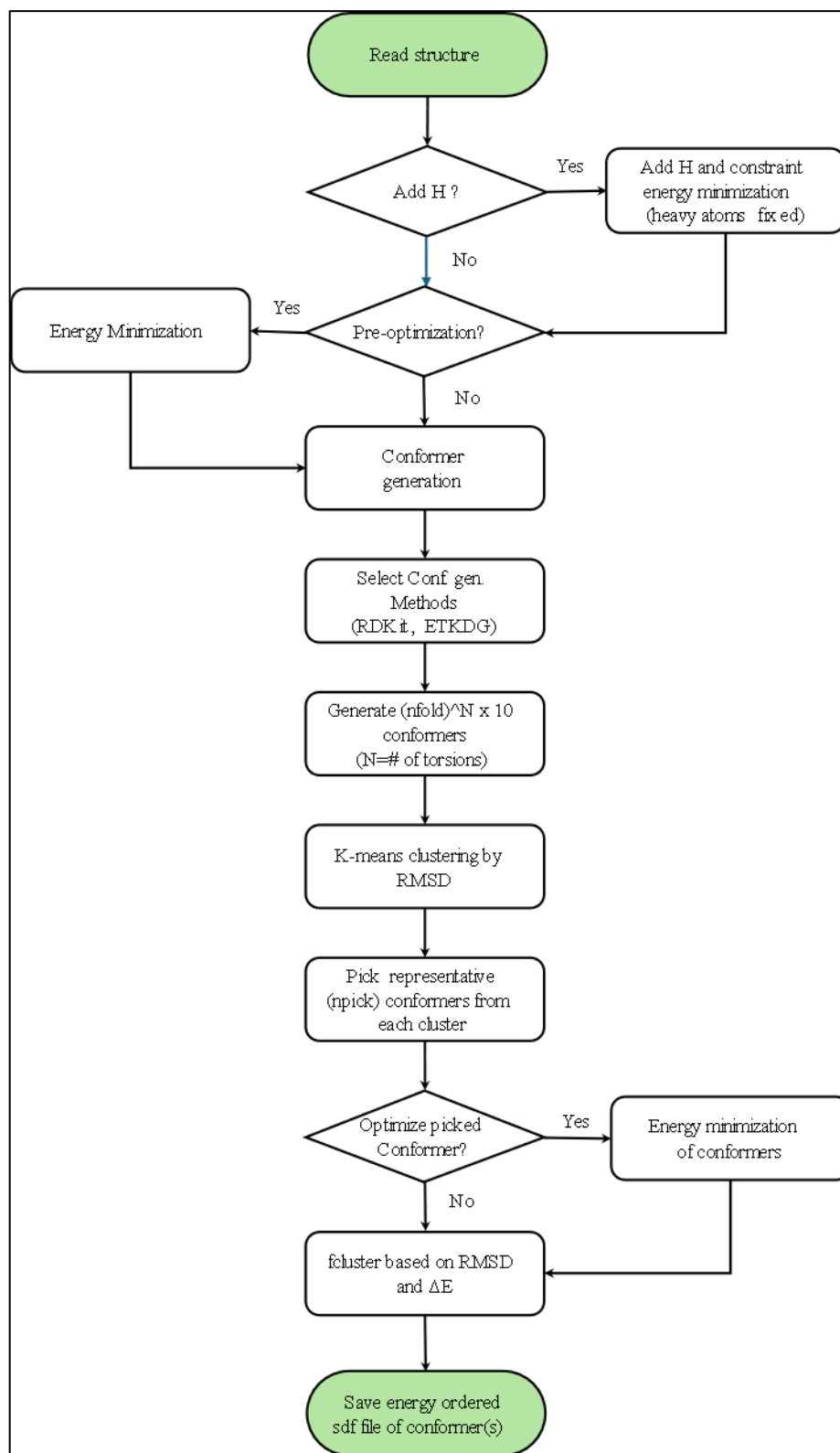


Figure 3.10: DeepConf flowchart.

After identifying the torsion points, the program requests the RDKit/ETKDG to sample diverse conformations. To reduce redundancy and select representative structures, the generated conformers are clustered based on root-mean-square deviation (RMSD) of atomic coordinates using K-means clustering implemented in the SciPy library. The lowest-energy structures in each cluster are chosen as the final representative conformations. This approach ensures that the final set of conformers is both diverse and energetically favorable. To enhance conformer diversity, the program uses the *nfold* parameter to request a specified number of conformers from ETKDG, with the default set to 2. This generates a total of $(\text{nfold})^N \times 10$ conformations, where *N* represents the number of torsions. It is important to note that chemical preferences for torsions (e.g., C(sp³)-C(sp³) favoring ~60°/120°) are handled by ETKDG, regardless of the *nfold* value. For instance, if there are 5 torsions and each one defined to flip 180° (i.e., *nfold*=2), a total of $2^5 \times 10 = 320$ structures are produced. Additionally, the maximum number of conformations can be set by the user to avoid generating too many conformations. Next, the program generates $(\text{nfold})^N$ different clusters, where *N* represents the number of torsions, using K-means clustering and distributes generated structures into these clusters to ensure that each fold is categorized into a distinct cluster. For example, for a molecule with *N* = 5 torsions and *nfold* = 2, the program generates $2^5 = 32$ clusters and distributes the 320 generated structures among them. Then, it calculates the single-point energies (SPE) of the structures and selects one representative structure (the one with the lowest energy) from each cluster. While the conformations are expected to be physically valid within the limits of RDKit/ETKDG, some unphysical structures, such as atom clashes, could arise. However, the SPE calculations, combined with clustering, effectively eliminate these unphysical structures, as they will have significantly higher energies compared to other conformers in the cluster. At this point, the user can set extra structures for consideration from each of the clusters (default is *npick*=0, corresponding to no additional structure pick from clusters and pick only the lowest energy one). In addition, if the user wants to optimize the picked structures from cluster, the program uses the minimization methods and calculators to minimize the structures. The last part is to eliminate the structures that are converging to the same equilibrium geometries. This is done by using F-clustering, a

hierarchical clustering method implemented in the SciPy library, based on RMSD (`opt_prune_rms_thresh`) and energy (`opt_prune_diffE_thresh`) thresholds. The final structures represent the global and local minima structures.

The program offers flexibility in searching for conformers along with geometry optimizations. One could just perform geometry optimizations by turning off the conformer generation process or rapidly generate conformers without geometry optimizations on picked conformers. Alternatively, one could search for all local minima along with the global minimum by turning on the optimization on the picked conformers. We can highlight the solutions the program provides to a few complications encountered in conventional conformer generation algorithms:

- If the initial starting structure is not drawn correctly (by the user) or bonding is incorrect (by the user or the MM predefinitions for bond orders), our script fixes that by performing pre-optimization process, which is not affected by bonding information.
- If the user wants to obtain the most plausible conformations among different torsions as quickly as possible, the program selects the best torsion angles corresponding to the lowest-energy structure purely based on single-point energies (SPE), even without further optimizations.
- If the user would like to generate most local minima extensively with the smallest ensemble size, the program attempts to find them all smartly using the flowchart given in **Figure 3.10**.
- If the optimization is stuck due to non-smoothness of the ML potential, the program attempts to find several other points that might lead to better solutions for the energy minimization problem.

3.2.2.2. ANI-ML potentials for PES

ANI-ML potentials can stop earlier than the real minimum reached because of the nature of the non-smooth PES. We start with hydrogen fluoride (HF), a diatomic molecule with one of the simplest representations of PES (only on H-F distance), as an illustration of

how ML-potentials can be non-smooth. Figure 3.11 shows the potential energy profile of H-F distance calculated by G16 and ANI-2x. If one were to perform a geometry optimization of HF starting with an initial structure of H-F distance outside the 0.6-1.2 Å range, the optimization with ANI-2x would fail due to saddle points near 0.5 and 1.9 Å whereas G16 still brings the structure to the equilibrium distance of 0.96 Å. Indeed, we started at several different initial distances and observed convergence of energy at the saddle point in the case of LBFGS algorithm and ANI-2x calculator. Although it is not the case in this study, it should be noted that ANI is trained on neutral compounds and is not capable of accurately modeling spin states or charges that may arise when atomic distances exceed the bond dissociation limit.

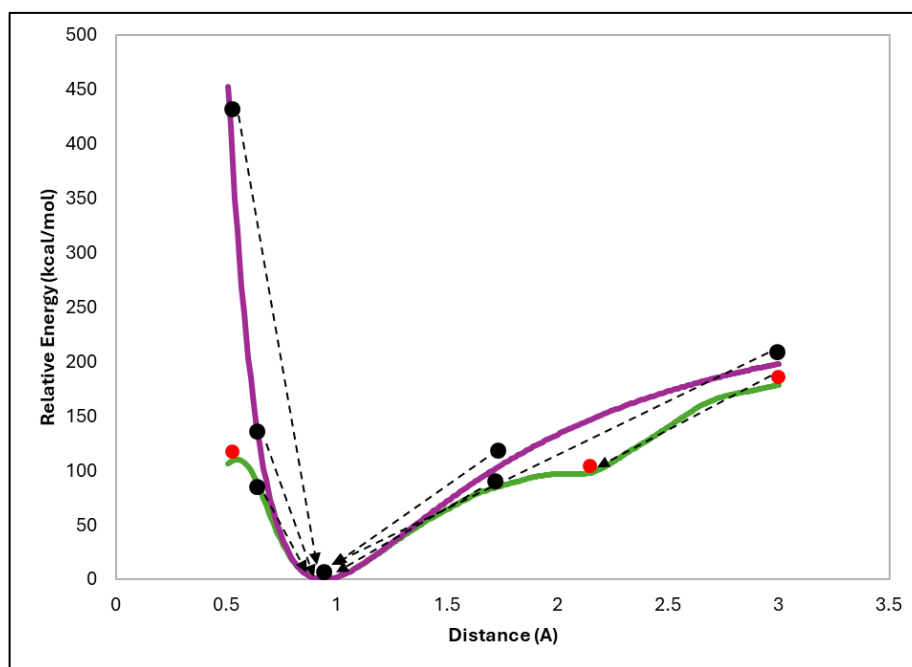


Figure 3.11: PES scan of the H-F distance computed using G16 (blue) and ANI-2x (green). Black and red spheres, respectively, represent the initial structures reaching the equilibrium geometry and failing to converge during gradient-based optimization.

To better illustrate when the system is larger, we also performed a two-dimensional scan (the symmetric stretching and bending vibrational modes) of the coordinates of H₂O. **Figure 3.12** shows the non-smooth regions on PES generated by ANI-2x.

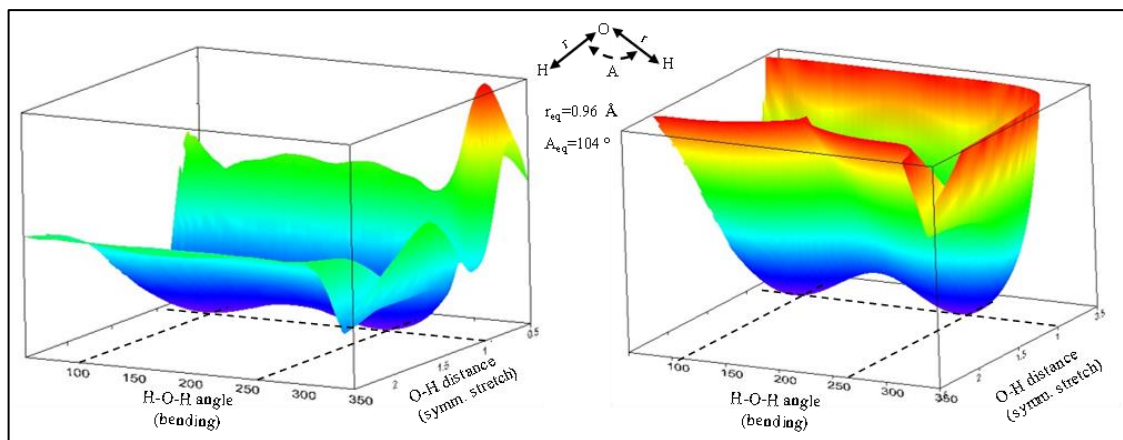


Figure 3.12: PES of H-O-H angle and O-H bond distances by ANI-2x (left) and G16 (right).

In larger compounds with several torsion points, the multidimensional PES often contains numerous non-smooth regions. Our approach begins with several diverse points on this representative PES, generated by RDKit/ETKDG, to more efficiently converge to equilibrium geometries.

When structures are near equilibrium, ANI-ML potentials perform well in locating the nearest minimum. However, when far from equilibrium, ANI-ML potentials can become stuck in non-smooth regions of the multidimensional PES. These non-smooth regions occur more frequently in dihedral angles because ANI-ML potentials are not trained to include interactions beyond 1-4. Previous studies have reported significant deviations of ANI-ML potentials from QM calculations at torsional energy barriers [193].

Our approach addresses this issue by creating diverse solutions at torsional points. This is achieved by boosting RDKit/ETKDG conformations, followed by selecting the lowest-energy structure from this diverse set. This strategy ensures that even in challenging regions of the PES, a structure has more opportunities to reach the nearest local minimum within the smallest ensemble size possible (Figure 3.13).

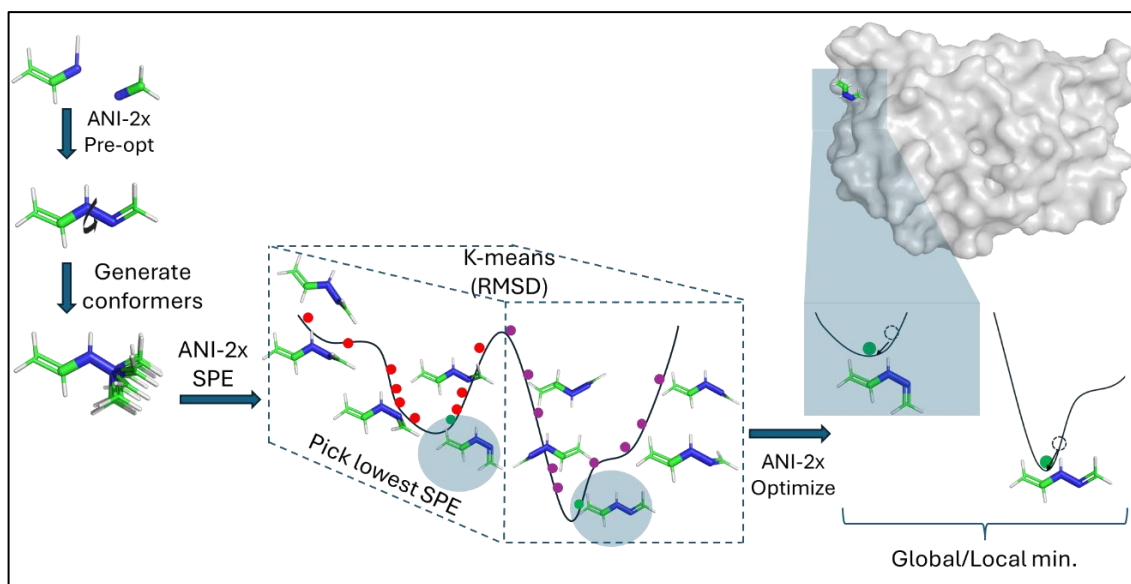


Figure 3.13: DeepConf approach to overcoming optimization challenges in non-smooth regions of the potential energy surface.

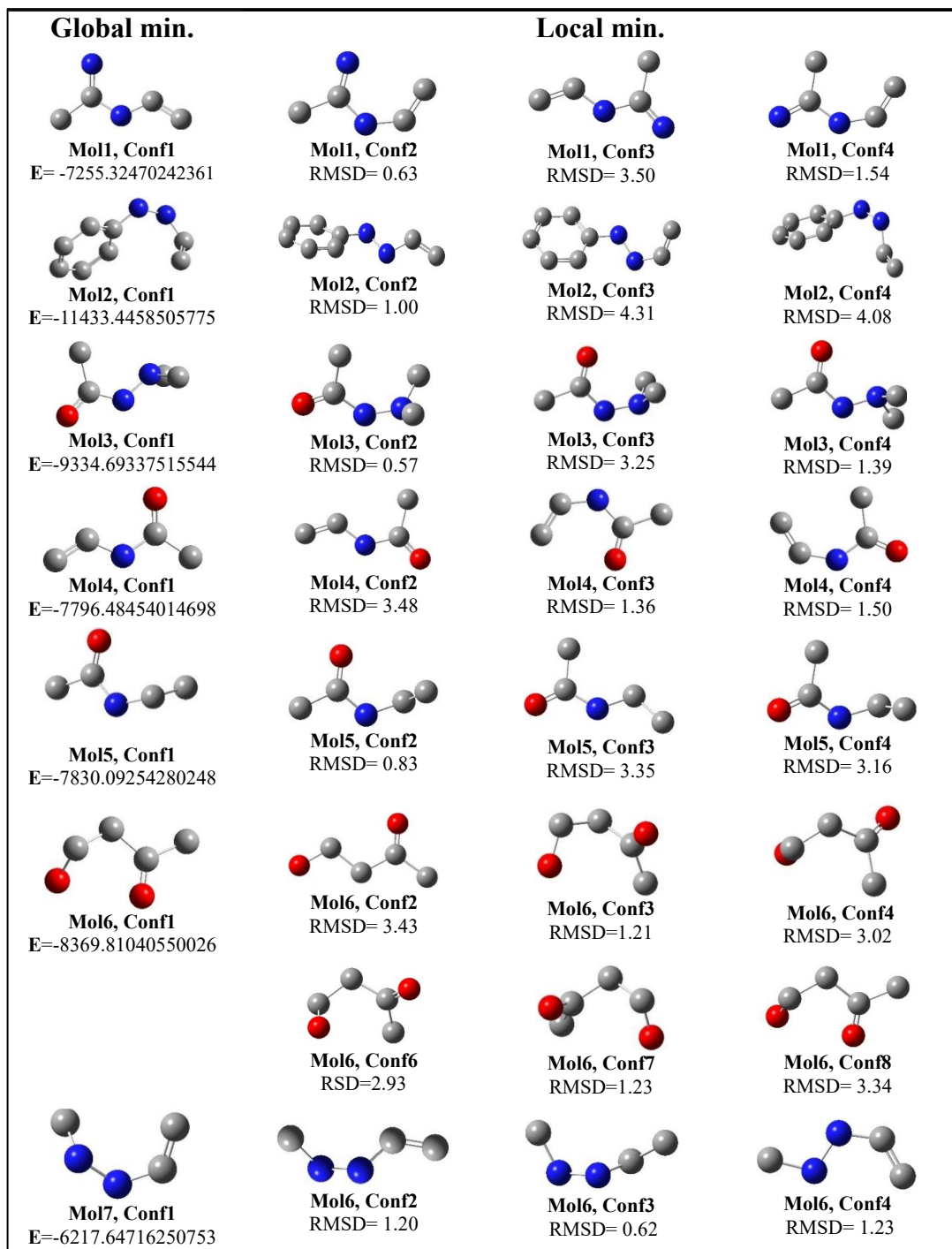


Figure 3.14: Initial test compounds and their G16 predicted global and local minima. Our first step with ANI is to assess it on several small compounds, in which there are only a few torsion points.

Using the compounds in Figure 3.4, we first manually generated the most plausible conformers including intermediate dihedral angles (up to 200 conformations for each compound) via rotation on each torsion. Next, we optimized each structure at the WB97X/6-31G* level to create our reference structures of global minimum and other possible local minima that the QM calculations predict by eliminating structures that converge to the same equilibrium conformers. These reference structures were used to assess the performance of DeepConf and Auto3D, both utilize ANI-ML potentials.

When we used the manually generated conformations by rotating around the torsion points (i.e., including several intermediates at each torsion point), ANI based optimizations yield the same conformers predicted by QM calculations (Table 3.4). This is as expected because we manually generated intermediate structures connecting each minimum and thus no conformer generation algorithm is required (i.e., DeepConf and Auto3D are the same at this point as ANI-2x). However, the extensive sampling for each torsion via several different dihedral angles eventually leads to ANI covering all the G16 minima. This also indicates that even if ANI can be stuck in a non-local geometry due to non-smooth PES, the optimization problem can be overcome by taking several samples from intermediate structures and optimizing them.

Table 3.4. ANI predicted the number of minima and its success in terms of reproducing G16 predicted ones from manually created intermediate conformations for each torsion point.

ID	G16 number of minima	ANI number of minima	Number of minima ANI to G16 RMSD$\leq$0.5 Å
Mol1	4	5	4
Mol2	4	8	4
Mol3	4	6	6
Mol4	4	4	4
Mol5	4	6	4
Mol6	8	8	8
Mol7	5	6	5
Mol8	4	5	4

3.2.2.3. Success of DeepConf vs. Auto3D on global/local minima match to G16.

3.2.2.3.1. Starting with a near equilibrium structure.

In the next step, we assessed the DeepConf and Auto3D predicted minima with respect to G16 predicted ones starting with G16 optimized global minima for these compounds. In this part, instead of manually creating intermediate structures, we started with G16 equilibrium and used the two conformer generation algorithms to find global and local minima. We defined the following success criteria for minima generation by the two conformer algorithms. If the conformer generation algorithms suggested minima match the G16 predicted ones by means that RMSD is less than or equal to 0.5 Å, then we considered this as successful. The rationale for choosing 0.5 Å as the success threshold is that the G16-predicted conformers for these small ligands are separated by at least this RMSD value.

Table 3.5 shows the success of conformer generation using either several parameter settings in DeepConf or Auto3D for the selected ligands when the G16 optimized global minimum is used as the initial structures in conformer generation process. Both DeepConf and Auto3D are adjusted to use the most similar parameter settings to assess their performance. In Auto3D, five out of eight structures have RMSD ≤ 0.5 Å of Auto3D assigned global minimum to the G16 assigned global minimum, yielding an overall success of 63%. On the other hand, when DeepConf is used, the overall success can increase to 75%. Since DeepConf has versatility in picking different numbers of structures for each torsion point by two different options as “nfold” and “npick”, one could obtain improved results. These features are not available in Auto3D, which totally relies on Rdkit and/or ETKDG to generate numbers of conformers. Although DeepConf also uses Rdkit and ETKDG, we can adjust the number of rotations for each torsion (nfold) and we further cluster these structures by K-means clustering and pick more structures from each cluster (npick) that allows the assessment of multiple points (corresponding to more than one dihedral angle for each cluster) from each of the torsion points. With DeepConf, we also observed slight improvement in finding local minima.

Table 3.5: Success of conformer generation on selected compounds with different parameter settings (parset) when the initial structure is near equilibrium. For the details of parsets, refer to Table 3.6.

DeepConf: parset1					DeepConf: parset2				DeepConf: parset3			
		global to G16 global		local to G16 local		global to G16 global		local to G16 local		global to G16 global		local to G16 local
ID	# conf.	RM SD (Å)	ΔE (kcal/ mol)	success	# conf.	RM SD (Å)	ΔE (kcal/ mol)	success	# conf.	RM SD (Å)	ΔE (kcal/ mol)	success
Mol1	2	0.01	0.1	33%	2	0.01	0.1	33%	2	0.01	0.1	33%
Mol2	11	0.71	-0.1	67%	3	0.17	-1.9	33%	8	0.17	-1.9	67%
Mol3	3	0.85	0.0	33%	2	0.86	0.8	33%	3	0.86	0.8	33%
Mol4	2	0.01	0.0	33%	2	0.02	-0.2	33%	2	0.02	-0.2	33%
Mol5	2	0.81	0.0	0%	2	0.80	0.0	0%	2	0.80	0.0	0%
Mol6	7	0.05	0.0	43%	2	0.09	0.2	14%	6	0.09	0.2	43%
Mol7	5	0.61	0.0	50%	3	0.35	-0.1	25%	5	0.66	-0.3	50%
Mol8	3	0.01	-0.1	33%	2	0.01	-1.3	33%	4	0.01	-1.3	67%
Overall I		50%	0.0	37%		75%	0.6	26%		63%	0.6	41%
DeepConf: parset4					DeepConf: parset5				Auto3D			
		global to G16 global		local to G16 local		global to G16 global		local to G16 local		Auto3D global to G16 global		Auto3D local to G16 local
ID	# conf.	RM SD (Å)	ΔE (kcal/ mol)	success	# conf.	RM SD (Å)	ΔE (kcal/ mol)	success	# conf.	RM SD (Å)	ΔE (kcal/ mol)	success
Mol1	2	0.01	0.2	33%	2	0.03	0.2	33%	1	0.01	0.0	0%
Mol2	29	0.20	-1.8	100%	11	0.25	-1.7	100%	8	0.17	-2.1	67%
Mol3	4	0.86	0.8	33%	3	0.86	0.7	33%	2	0.86	0.6	33%
Mol4	2	0.01	-0.2	33%	2	0.02	-0.2	33%	2	0.02	-0.1	33%
Mol5	2	0.79	0.1	0%	2	0.05	0.1	0%	3	0.05	-0.1	33%
Mol6	10	0.10	0.2	71%	7	0.55	0.2	43%	10	0.55	0.2	43%
Mol7	6	0.61	-1.0	75%	6	0.02	-1.0	50%	5	0.61	-1.1	50%
Mol8	3	0.01	-1.3	33%	3	0.01	-1.3	33%	3	0.05	-1.3	33%
Overall I		63%	0.7	47%		75%	0.7	41%		63%	0.7	37%

Table 3.5 also reports the ΔE values of DeepConf and Auto3D relative to G16-calculated energies for the global minima. The data suggests that the energies of global minima are accurately reproduced by both algorithms. Small negative or positive deviations fall within the precision margins of the ANI-ML potentials and may also result from slight structural shifts. To validate this, we recalculated energies using G16 at the equilibrium geometries produced by Auto3D and DeepConf. These deviations should not be misinterpreted as ML potentials identifying lower-energy structures. This does not impact the success of the conformer generation, which is based on RMSD rather than energies.

Table 3.6: Parameter settings used in conformer generation for the initial dataset of selected ligands. Both programs are adjusted to similar settings in order to assess their performance.

ConfGen	Parset1	Parset2	Parset3	Parset4	Parset5	Parset6	Parset7
optimization_method	LBFGS	LBFGS	LBFGS	FIRE	FIRE	FIRE	FIRE
pre_optimization_lig	no	no	no	no	no	no	no
genconformer	yes	yes	yes	yes	yes	yes	yes
ETKDG	yes	yes	yes	yes	yes	yes	yes
num_conformers	1	1	1	1	1	1	1
max_attempts	100000	100000	100000	100000	100000	100000	100000
prune_rms_thresh	0.5	0.5	0.5	0.5	0.5	0.5	0.5
opt_prune_rms_thresh	0.3	0.3	0.3	0.3	0.3	0.3	0.3
opt_prune_diffE_thresh	0.001	0.001	0.001	0.001	0.00001	0.001	0.001
calculator_type	ani2x	ani2x	ani2x	ani2x	ani2x	ani2x	ani2x
optimization_conf	yes	yes	yes	yes	yes	no	no
optimization_lig	no	no	no	no	no	no	no
thr_fmax	0.02	0.02	0.02	0.02	0.0002	0.02	0.02
maxiter	50000	50000	50000	50000	50000	50000	50000
nfold	4	2	2	4	2	2	2
npick	0	0	2	0	2	0	0
cluster	yes	yes	yes	yes	yes	yes	no

Among all the parsets we tested, we conclude that parset 2 is best for computational efficiency with slightly reduced accuracy, parset 4 is best for comprehensively covering both local and global minima at a higher computational cost, and parset 5 is the optimal choice for a balance between accuracy and computational cost (**Table 3.6** and **Table 3.7**).

One should note that when G16 is used in the optimization process, it considers four different criteria to converge (i.e., maximum force, RMS force, maximum distance, and RMS distance) whereas DeepConf and Auto3D considers only energy with respect to coordinate (i.e., maximum force) to converge.

Table 3.7: Effect of maximum step / convergence threshold on conformer generation performances.

	auto3d_f002			auto3d_f00001			confgen_f002			confgen_f00002		
	Auto3D global to G16 global			Auto3D global to G16 global								
ID	#auto3d_numconf	min top	min in all	#auto3d_numconf	min top	min in all	parset 5	mintop	min in all	parset 5	min top	min in all
38W	26	1.1	0.6	13	1.1	0.8	53	1.1	0.9	23	1.1	1.1
2WB	28	0.8	0.6	15	0.8	0.8						
2M3	117	2.2	0.9	68	2.8	0.7	48	2.2	1.3	34	2.5	1.1
2KW	35	2.6	1.0	6	2.6	1.1	48	2.2	1.0			
21G	22	0.5	0.5	20	0.4	0.4	34	0.5	0.3			
1GB	105	2.5	1.0	28	1.1	1.1	47	2.1	1.6	27	2.1	1.6
0YO	4	1.5	0.1	2	1.5	1.5	7	1.5	0.1	10	1.4	0.1
0VG	69	2.7	1.1	25	2.7	1.9	43	2.8	1.6			
0C7	99	0.6	0.6	31	0.5	0.2				36	0.5	0.2
05B	13	1.0	0.7	9	0.9	0.7	11	1.8	1.6	10	1.6	1.5
mean	51.8	1.6	0.7	21.7	1.4	0.9	36.4	1.8	1.0	23.3	1.5	0.9
%success<0.5		1	2		2	2		1	2		1	2
%success<1		3	9		4	6		1	3		1	2
%success<1.5		6	10		7	9		3	5		3	4

3.2.2.3.2. Starting with off-equilibrium structure

To evaluate the optimization processes for DeepConf, the G16-equilibrated test molecules in Table 2 were distorted using two different approaches (‘distortion 1’ and ‘distortion 2’) to generate 100 off-equilibrium geometries each. In the first approach (“distortion 1”), directly the x,y,z coordinates were randomly changed so as to produce structures that have RMSD to the initial (G16 equilibrated) structure between 0.5 Å (i.e., minimum RMSD) and 2.0 Å (i.e., maximum RMSD). In the second approach, the structures are placed in a box and the box is uniformly scaled between 0.96 Å and 1.10 Å. Thus, the molecule is contracted or expanded uniformly. Finally, the coordinates are randomly scaled between 0.5 and 1.5 to produce off-equilibrium geometries (“distortion 2”). QM optimizations confirmed that all distorted geometries converged to either global or local minima.

Table 3.8 compares how effectively DeepConf recovers low-lying structures from these distorted inputs. DeepConf directly optimizes from coordinates and successfully recovers the nearest local or global minimum in most cases (98% for ‘distortion 1’ and 60% for ‘distortion 2’). We should also note that Auto3D is not designed for direct geometry optimization rather it produces the lowest energy conformer no matter what starting structure is given even though the parameters are set to yield 1 conformer only. On the other hand, DeepConf is designed to overcome such issues by attempting to optimize directly from coordinates.

Table 3.8: Optimization performance of DeepConf when initial structures are beyond equilibrium geometry. The success shows the number of instances where optimization is converged to a geometry with $\text{RMSD} \leq 0.5 \text{ \AA}$.

ID	DeepConf	
	distortion 1	distortion 2
Mol1	99.0%	51.0%
Mol2	98.0%	53.0%
Mol3	99.0%	63.0%
Mol4	98.0%	56.0%
Mol5	96.0%	67.0%
Mol6	99.0%	71.0%
Mol7	97.0%	69.0%
Mol8	97.0%	50.0%
Overall	98%	60%

In addition to the previously described distortion methods (‘distortion 1’ and ‘distortion 2’), we introduced a third distortion approach (‘distortion 3’) to further evaluate the optimization processes for DeepConf. Distortion 3 was generated by performing 1 ns classical MD simulations in vacuum at an elevated temperature of 1000 K using GAFF parameters and RESP charges in an NVT ensemble. From each trajectory, 100 frames were extracted and subjected to optimizations using QM and DeepConf. All of these structures converged to either global or local minima during QM optimization. As expected, most MD-sampled conformations remained close to equilibrium, restrained by the harmonic nature of classical MD simulations. Consequently, most of them are also converged to one of the minima by DeepConf. However, there were still several instances

where DeepConf optimizations did not match any of the known minima within the RMSD threshold. These mismatches likely reflect the challenges posed by non-smooth regions of the PES. Detailed comparisons of these results, including the distributions of recovered minima, are provided in the supplementary materials (Table 3.9). We also tested the performance of different optimization algorithms implemented in our script for navigating non-smooth regions of the PES. As a case study, we optimized the H-F bond starting from an initial separation of 3.0 Å using multiple minimization methods. For such a simple system, the computational cost was negligible, and all optimizations were completed within seconds. However, we observed that LBGS and FIRE often became stuck in the non-smooth region, yielding invalid geometries upon optimization. The success of these methods depended strongly on the initial distance.

Conversely, BFGS and GPMIn successfully converged to the equilibrium geometry across a range of starting distances. When applied to the selected set of 8 compounds, all four algorithms produced similar results in terms of equilibrium structures. GPMIn was the most successful in navigating non-smooth regions but was significantly impacted by the number of optimization steps, with its performance slowing exponentially as the step count increased. BFGS also showed strong performance but failed to converge for a non-negligible number of initial geometries. While a detailed analysis of these observations is beyond the scope of this study, we believe FIRE and BFGS represent the best trade-offs between computational cost and convergence for most systems.

Table 3.9: Optimization of 100 frames extracted from MD simulations for the selected compounds. Matched refers to the number of structures with $\text{RMSD} \leq 0.5 \text{ \AA}$ to any of the minima on Figure S1. Global is the global minimum and min1,min2,etc refer to the local minima. Corresponding values under the columns show the number of structures yielded (upon optimization) to that specific minimum. Unmatched refers the structure is converged to a different minima (likely saddle point or non-smooth region). For G16 calculations.

G16										
ID	Matched/ Converged	Global	min1	min2	min3	min4	min5	min6	min7	unmatched
Mol1	100/100	26	2	66	6					0
Mol2	100/100	9	17	13	25	36				0
Mol3	100/100	57	5	10	28					0
Mol4	100/100	94	0	6	0					0
Mol5	98/98	29	35	17	10	7				0
Mol6	97/99	17	3	2	11	17	35	9	3	2
Mol7	100/100	7	28	36	3	10	16			0
Mol8	100/100	71	2	10	17					0

Table 3.10: Optimization performance of DeepConf when initial structures are beyond equilibrium geometry. The success shows the number of instances where optimization is converged to a geometry with $\text{RMSD} \leq 0.5 \text{ \AA}$ to any of the minima on Figure S1. Global is the global minimum and min1,min2,etc refer to the local minima. Corresponding values under the columns show the number of structures yielded (upon optimization) to that specific minimum. Unmatched refers the structure is converged to a different minima (likely saddle point or non-smooth region). For DeepConf algorithm.

DeepConf										
ID	Matched/ Converged	Global	min1	min2	min3	min4	min5	min6	min7	unmatched
Mol1	100/100	26	2	66	6					0
Mol2	81/100	4	33	19	25	0				19
Mol3	100/100	60	2	14	24					0
Mol4	100/100	94	0	6	0					0
Mol5	89/100	26	28	17	10	8				11
Mol6	95/100	16	7	2	11	19	30	7	3	5
Mol7	100/100	7	34	33	1	14	11			0
Mol8	100/100	71	3	10	16					0

3.2.2.4. Finding bioactive conformations with DeepConf and Auto3D

One of the ultimate goals in drug discovery and design is to computationally screen many drug candidates to the biological target molecules and is to find the most plausible binding mode as well as binding strength using different strategies such as molecular docking, MD

simulations and free energy calculations. For instance, in molecular docking both search algorithms and scoring functions are limited to the attempts the software makes during the docking process. Hence, the conformer generation prior to docking can improve both search and scoring efficiency by creating a good starting structure. For this reason, most conformer generation algorithms are assessed for their performance on finding the bioactive conformations.

In this step, we assess the performance of auto3d and DeepConf to reproduce the bioactive conformations in the absence of target biomolecule. We have followed a unique protocol which can be a guideline for benchmarking conformer generation algorithms' performance to capture the bioactive conformation.

The Protein Data Bank (PDB) reports 'model' structures as the ligand conformations found in the presence of biomacromolecules such as proteins (i.e., bioactive conformations). PDB also reports "ideal" structures to represent a chemically reasonable, idealized structure of the ligand that is optimized to ensure correct bond lengths, bond angles, and stereochemistry. The "ideal" structure is not explicitly meant to be the global minimum; its purpose is to serve as a high-quality, chemically valid model for applications such as docking, visualization, and further computational studies. Our protocol begins with the idea that the "model" and "ideal" structures correspond to two different energetically low-lying structures. Thus, after ensuring "ideal" and "model" structures for each ligand that represent different conformations, starting with an "ideal" structure (one of the minima) and finding "model" structure (another of the minima) with conformer generation process would indicate the success of the algorithm. To assess the true performance of a conformer generation algorithm, we have considered the following scenarios and eliminated most structures according to the criteria defined in each scenario (Figure 3.15).

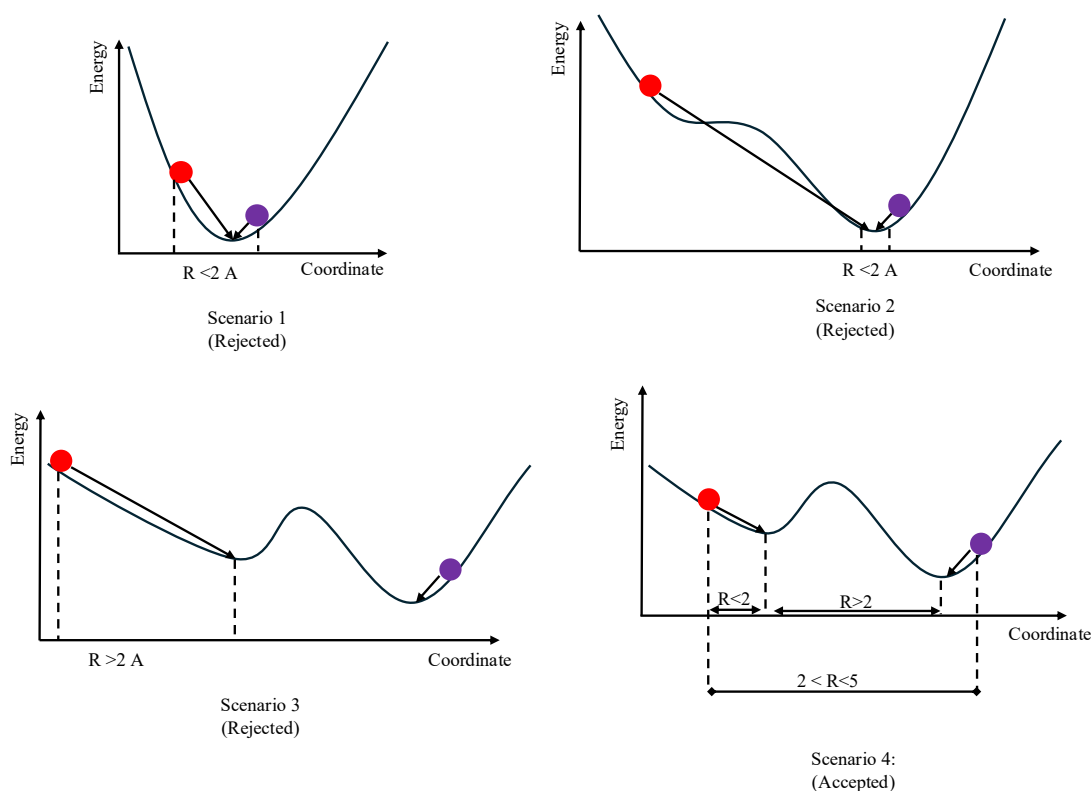


Figure 3.15. Schematic representation of performance of conformer generation algorithms for capturing bioactive conformation. The blue and red balls represent the ideal and model structures, respectively. Solid (one-sided) arrows show the structural change upon a QM optimization.

Scenario 1: The “ideal” and “model” structures of a ligand may correspond to the same conformation (minimum). If this is the case, QM optimization of “ideal” structure will already yield to the same structure with QM optimized “model” structure. Including these structures in conformer generation process would bring additional (un-realistic) success. Thus, we have excluded such cases by comparing the RMSD of heavy atoms of “ideal” and “model” structures. We rejected structures with $\text{RMSD} \leq 2 \text{ \AA}$ (ideal to model). By applying this criterion, we have also avoided the cumbersome QM optimization of thousands of ligands.

Scenario 2: Ideal and model structures may correspond to different conformers but the energy barrier connecting them might be too small. In this case, like scenario 1, QM optimization of “ideal” structure would still yield the same conformer as QM optimized

“model” structure. We have compared the optimized structures of “ideal” and “model” structures and rejected those with $\text{RMSD} \leq 2\text{\AA}$ (ideal to model).

Scenario 3: Although ideal and model structures may correspond to different conformers, the model structure might be too far from its local minimum. If this is the case, assumedly the ligand’s bioactive conformation corresponds to a structure that is mostly affected by nearby atoms (of the bio-macromolecule) and the ligand has a strain energy to a considerable extent. Since we optimized the model structure in the absence of the bio-macromolecule, the ligand relaxes to the closest un-strained conformation, which is too far from the bio-active conformation. Optimization of model structure would change the geometry a lot. Since it is not near a minimum, no optimization algorithm would find these structures. Therefore, we excluded these structures by only accepting $\text{RMSD} \leq 2\text{\AA}$ of QM optimized “model” against initial (un-optimized) “model” structures. This also avoids miscalculations of the performance by RMSD for the conformer generation since the un-optimized and QM optimized bioactive conformations differ drastically.

Scenario 4: This is the case where we have assessed the performance of conformer generation (accepted scenario). In this case, the model and ideal initial structures are at least $2\text{\AA}$ apart in terms of RMSD; optimization of both structures still converges to different minima; and the initial model structure already holds near minimum geometry.

We believe that this protocol outlined here should give the true performance of any conformer generation algorithm when assessed for bio-active conformation capturing when started from ideal structures of the ligands in the absence of the target bio-macromolecule.

We have downloaded 48485 structures, and among these only 43357 ligands had both model.sdf and ideal.sdf structures available in the database. To ease the computation, we have only included the structures that have the number of heavy atoms in the range of 15-30 and that have torsion points that are in the range of 2-5. We have also eliminated structures that have other atoms than C, H, N, O, F and Cl in their structure since ANI-ML potentials can only manage these atoms (i.e., NNP is trained only for these atoms). Only 34107 ligands have passed from these filters. Among these, 762 ligands had the $2\text{\AA} \leq \text{RMSD} \leq 5\text{\AA}$ values of initial “model” and “ideal” structures. Normally, there is no upper limit to RMSD values, and one could choose all that are above 2.0 Å (i.e., scenario 1 is applied). However, our choice of 5 Å maximum is to optimize structures at the DFT level in a feasible time. We have optimized these 762 ligands at the WB97X/6-31G* level and re-compared the RMSD values (between model and ideal structures) upon optimization. Only 439 structures with $\text{RMSD} \geq 2\text{\AA}$ between model and ideal (i.e., scenario 2 is applied) and that $\text{RMSD} \leq 2\text{\AA}$ between optimized model and un-optimized model structures (i.e., scenario 3 is applied) have remained after this step. These structures are assured to hold different “model” and “ideal” QM optimized conformations. Our aim for assessing the performance of the conformer generation is the ability to find the QM optimized “model” structure when QM optimized “ideal” structure is introduced to the program among these final 439 selected structures (Hata! Başvuru kaynağı bulunamadı.).

Table 3.11 shows the summary of the performance of both DeepConf and Auto3D along with conventional conformer generation by RDKit using two different classical force fields to reproduce the model structures starting from ideal structures. Auto3D has produced an average of 62 conformers per ligand. In the table, the “top” refers to the lowest energy structure. We assigned as “successful” when at least one of the generated conformers considering the top 1, 3, 10 and all conformers are within the threshold of an RMSD of 0.5 Å (**Table 3.11** and **Figure 3.16**) along with the cutoffs of 1.0, 1.5, and 2.0

Å (**Table 3.12**). For Auto3D, when all the conformers are considered, the success of at least one conformer to be within 0.5 Å RMSD from the model structure when the initial structure is given as the ideal structure is 43.5% with an average ensemble size of 62. This value is 28% when only top 10 conformers are considered. We also analyzed the coverage of QM optimized “model” and “ideal” structures within the top 3 and 10 (i.e., the 3 or 10 lowest energy) structures predicted by the conformer generations. With the RMSD cutoffs from 0.5 Å to 2 Å, success ranges from 15.3-18.5% to 78.1-84.7% in model and ideal structures. Thus, if one would produce only three lowest energy structures out from Auto3D, the success of covering both model and ideal structure within 0.5 Å of RMSD cutoff is 15.3-18.5%. In the consideration of the top 10 conformers, these numbers reach 27.3-39.4%.

When the same analysis is performed by DeepConf with different parameter settings (parsets), we produced somewhat comparable results to Auto3D in fewer conformer (ensemble) sizes. The average number of conformers generated by parset2 was as few as 18. The success of capturing model and ideal structures within the top three/ten structures was slightly greater than Auto3D. In parset5, the success can increase up to 40.6% and 67.5% considering top 10 and all ensembles, respectively within only an average size of 32 conformers.

Table 3.11: Overall performance of DeepConf and Auto3D in reproduction of model/ideal structures from ideal structures. The number values in rmsd columns show the success percentage (%) with the corresponding criteria.

				wrt / QM optimized model				wrt / QM optimized ideal				cumulative (model+ideal)
Method	Parameter	RMSD	size	success %	mean	min	max	success %	mean	min	max	success %
DeepConf	parset2	top 1		9.8				9.6				9.7
		top 3		21.6	1.25	0.06	3.59	20.3	1.21	0.03	3.13	21.0
		top 10		31.7	0.92	0.06	2.89	43.1	0.70	0.03	2.47	37.4
		all	18	35.5	0.81	0.06	2.53	54.7	0.54	0.03	2.24	45.1
	parset4	top 1		10.3				5.8				8.1
		top 3		20.6	1.31	0.03	3.92	12.9	1.46	0.06	3.37	16.7
		top 10		32.7	1.01	0.03	3.92	25.9	1.04	0.04	3.09	29.3
		all	322	56.5	0.54	0.03	2.38	71.0	0.39	0.02	2.01	63.8
	parset5*	top 1		13.4				7.4				23.7
		top 3		23.0	1.23	0.05	3.96	18.7	1.28	0.04	3.18	22.7
		top 10		34.6	0.93	0.04	3.00	40.6	0.72	0.04	2.34	37.1
		all	36	46.6	0.69	0.02	2.32	67.5	0.43	0.03	1.59	56.2
	parset6	top 1		4.8				8.4				6.6
		top 3		9.3	1.51	0.12	3.93	19.6	1.06	0.12	3.08	14.5
		top 10		21.0	1.09	0.12	3.10	33.3	0.76	0.08	2.54	27.1
		all	21	23.5	0.93	0.10	3.10	36.7	0.66	0.11	2.35	30.1
	parset7	top 1		4.8				8.4				6.6
		top 3		9.1	1.57	0.08	3.85	17.1	1.15	0.10	3.10	13.1
		top 10		18.0	1.19	0.08	3.56	33.0	0.82	0.06	2.52	25.5
		all	207	37.1	0.71	0.07	2.27	58.8	0.49	0.06	2.00	47.9
Auto3D	auto3d	top 1		7.5				7.5				7.5
		top 3		15.0	1.36	0.06	4.11	18.2	1.24	0.03	3.39	16.6
		top 10		28.2	0.98	0.06	3.66	40.5	0.78	0.03	2.66	34.4
		all	62	43.5	0.69	0.06	2.45	65.4	0.47	0.03	1.80	54.4
RDKit	MMFF94	top 1		6.8				5.9				6.4
		top 3		15.5	1.33	0.05	3.36	16.4	1.28	0.05	3.38	15.9
		top 10		29.6	0.95	0.05	3.14	37.8	0.76	0.05	2.58	33.7
		all	20	35.8	0.78	0.05	2.88	47.8	0.59	0.05	2.05	41.8
	UFF	top 1		8.0				5.2				6.6
		top 3		14.8	1.27	0.06	3.58	13.9	1.39	0.05	3.41	14.3
		top 10		24.6	0.93	0.06	3.07	30.3	0.79	0.05	2.50	27.4
		all	20	29.4	0.82	0.06	2.78	38.7	0.65	0.05	2.42	34.1
* Subset of 282 compounds												

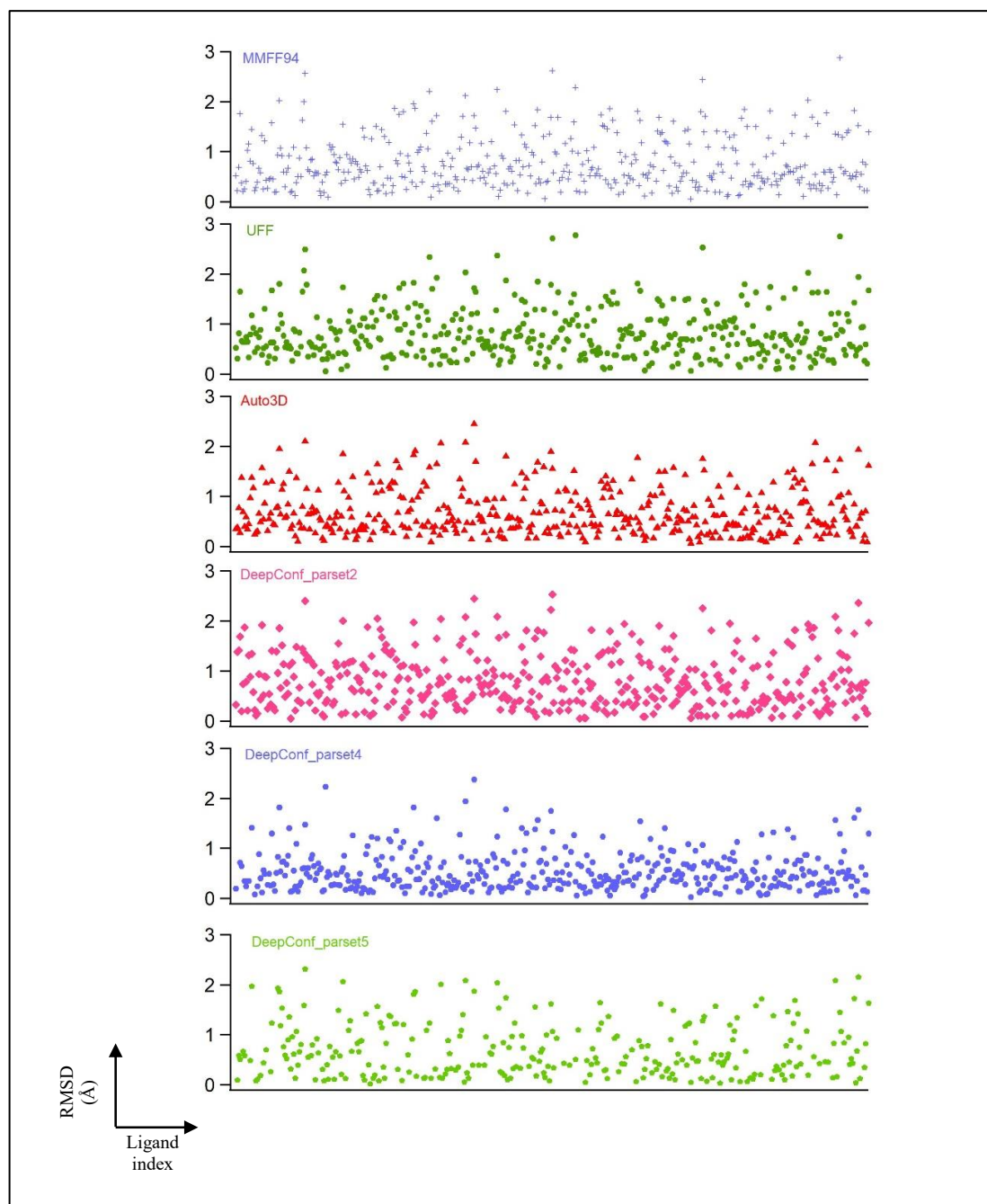


Figure 3.16: Distribution of minimum RMSD values of all conformers generated by different algorithms.

Table 3.12: Overall performance of Confgen and Auto3D in reproduction of model/ideal structures from ideal structures using several cut-off values set for success percentage (%).

Method	Parset	RMSD	size	wrt/ QM optimized model							wrt/ QM optimized ideal						
				success %				mean	min	max	success %				mean	min	max
				≤0.5 Å	≤1.0 Å	≤1.5 Å	≤2.0 Å				≤0.5 Å	≤1.0 Å	≤1.5 Å	≤2.0 Å			
Auto3D	sdf input	all	62	43.5	77.2	92.3	98.9	0.69	0.06	2.45	65.4	92.9	99.3	100.0	0.47	0.03	1.80
		top 1		7.5	18.2	31.2	53.5				7.5	17.3	33.7	60.1			
		top 3		15.0	35.1	55.4	79.3	1.36	0.06	4.11	18.2	38.3	63.1	85.9	1.24	0.03	3.39
		top 10		28.2	57.6	77.4	92.9	0.98	0.06	3.66	40.5	68.8	89.7	96.4	0.78	0.03	2.66
DeepConf	parset6	all	21	23.5	59.9	82.9	94.3	0.93	0.10	3.10	36.7	84.7	94.1	97.0	0.66	0.11	2.35
		top 1		4.8	14.4	23.9	45.3				8.4	22.6	41.7	68.6			
		top 3		9.3	27.3	48.5	74.0	1.51	0.12	3.93	19.6	49.7	77.9	92.9	1.06	0.12	3.08
		top 10		21.0	50.1	73.3	92.3	1.09	0.12	3.10	33.3	76.3	91.8	97.7	0.76	0.08	2.54
	parset7	all	207	37.1	77.0	93.8	98.4	0.71	0.07	2.27	58.8	95.7	99.3	99.8	0.49	0.06	2.00
		top 1		4.8	14.4	23.9	45.3				8.4	22.6	41.7	68.6			
		top 3		9.1	25.3	44.4	69.7	1.57	0.08	3.85	17.1	45.1	72.0	89.5	1.15	0.10	3.10
		top 10		18.0	45.3	67.0	88.2	1.19	0.08	3.56	33.0	70.4	88.8	97.0	0.82	0.06	2.52
	parset2	all	18	35.5	68.3	87.2	97.3	0.81	0.06	2.53	54.7	88.6	96.4	99.3	0.54	0.03	2.24
		top 1		9.8	20.7	36.2	57.2				9.6	17.5	29.4	56.5			
		top 3		21.6	37.8	62.6	82.9	1.25	0.06	3.59	20.3	40.1	62.6	87.2	1.21	0.03	3.13
		top 10		31.7	58.8	81.1	94.8	0.92	0.06	2.89	43.1	76.8	90.7	98.4	0.70	0.03	2.47
	parset5	all	36	46.6	74.9	89.8	97.2	0.69	0.02	2.32	67.5	92.2	98.9	99.6	0.43	0.03	1.59
		top 1		13.4	24.0	37.8	60.8				7.4	16.6	30.0	50.5			
		top 3		23.0	40.6	62.5	82.3	1.23	0.05	3.96	18.7	35.3	57.2	82.0	1.28	0.04	3.18
		top 10		34.6	58.3	77.7	91.9	0.93	0.04	3.00	40.6	71.0	89.0	97.2	0.72	0.04	2.34
	parset4	all	322	56.5	88.3	97.0	99.5	0.54	0.03	2.38	71.0	96.7	99.5	99.8	0.39	0.02	2.01
		top 1		10.3	22.9	38.3	57.9				5.8	15.0	26.6	53.3			
		top 3		20.6	41.1	59.1	76.6	1.31	0.03	3.92	12.9	29.7	47.4	72.9	1.46	0.06	3.37
		top 10		32.7	57.5	75.9	89.0	1.01	0.03	3.92	25.9	50.2	74.8	90.7	1.04	0.04	3.09
RDKit	MMFF94	all	20	35.8	72.2	87.9	97.7	0.78	0.05	2.88	47.8	87.5	95.9	99.8	0.59	0	2.05
		top 1		6.8	18.2	32.1	60.8				5.9	15.9	27.3	54.0			
		top 3		15.5	35.3	56.7	83.1	1.33	0.05	3.36	16.4	38.5	58.5	83.8	1.28	0.05	3.38
		top 10		29.6	59.7	78.4	94.5	0.95	0.05	3.14	37.8	72.7	89.1	98.4	0.76	0.05	2.58
	UFF	all	20	29.4	72.2	89.1	97.7	0.82	0.06	2.78	38.7	85.0	96.6	99.5	0.65	0.05	2.42
		top 1		8.0	22.8	38.7	60.6				5.2	13.7	23.7	46.7			
		top 3		14.8	40.8	62.0	84.1	1.27	0.06	3.58	13.9	34.9	53.3	76.3	1.39	0.05	3.41
		top 10		24.6	62.0	82.7	96.1	0.93	0.06	3.07	30.3	71.8	90.9	97.9	0.79	0.05	2.50

When a similar analysis using RDKit using MMFF94 was performed, 29.6% success was observed when considering the top 10 conformers with respect to QM optimized model structures. RDKit with UFF force field was even lower than that with 24.6% success. In most cases, Auto3D and our DeepConf have surpassed the conformations generated by conventional methods.

In addition, we also generated the conformers via DeepConf without any optimization and picked the minimum energy structure among these conformers. This way of getting conformers is much faster since no optimization is involved, and only single point energies (SPE) are used in ranking the conformers. Parset6 and parset7 use the same values as parset2 with the only difference of no optimization (i.e., SPE only) in both and clustering is also off in the latter. These two settings show quite successful performance with respect to their tremendously low computational cost.

Apart from our definition of success rates, we also compared the results in analogy to earlier studies, which reported the mean, minimum and maximum RMSD values without any success conditions by means of cut-off in RMSD. Friedrich et al. [194] have reported benchmarking of numerous conformer generation algorithms using PLATINUM diverse dataset [195] (for ~2859 molecules) and observed mean RMSD of 0.79 Å for RDKit using MMFF94 and 0.82 using UFF when maximum ensemble size is set to 50. These results are in perfect agreement with our predictions of 0.78 Å and 0.82 Å (in the ensemble sizes of 20), respectively, for RDKit using MMFF94 and UFF for prediction of QM optimized model structures starting from QM optimized ideal structures. In the same study, the best performance among 22 different algorithms (of ConfGenX by Schrodinger) were found to have the mean (of the closest structures) RMSD of 0.63 Å and 0.54 Å for maximum cluster sizes of 50 and 250, respectively. In another study [195], they have also reported that RDKit gives the performance of 0.98 Å when the cluster size is 10, which is also in perfect agreement with our finding, which is 0.93-0.95 Å for the top 10 structure evaluation. The lowest value they could get was 0.59 Å when they generated up to 500 conformers using the same method. One of the most recent studies by Seidel et al. [142] reported that RDKit utilized with ETKDGv3 in the same force fields (MMFF94 and UFF) has a performance of mean RMSD of 0.74 Å and 0.71 Å for a maximum ensemble size of

50. Their introduced method, so-called “CONFORGE”, yielded 0.67 Å for the same value. As a conclusion, to the best of our knowledge none of the studies could bring this value below 0.5 Å no matter what maximum ensemble size is used.

When the same performance analysis was performed for Auto3D, we observed the mean RMSD was as low as 0.69 Å for model structures and 0.47 Å (**Table 3.11**) for ideal structures in an average ensemble size of 62 (For individual ensemble sizes and RMSD values, refer to Supporting Information). Similarly, our DeepConf with different parameter settings (**Table 3.12**) could produce mean RMSD as low as 0.54 Å for model and 0.39 Å for ideal structures. Here, “parset2” uses 180° (nfold=2) for each torsion point and produces the mean RMSD of 0.81 Å (model) and 0.54 Å (ideal) in just 18 average ensemble size. Moreover, by increasing the “nfold”, one can easily boost up the performance of DeepConf. Here, in our test of parset4, where nfold=4 (i.e., 90° rotations on each torsion), we could improve the RMSD values to 0.54 Å and 0.39 Å for model and ideal structures, respectively. One might argue that the ensemble size is also increased tremendously (up to 322 in parset4). This is mostly because we have set the convergence criteria to 0.02 eV/Å (maximum force) in all parameter settings to reduce computational cost. When lower convergence criteria were applied, the ensemble size would drop to tens rather than hundreds.

We also evaluated the performance of our DeepConf in the case of no-optimization involvement and thus based on single point energy (SPE) calculations since sometimes, it might be desired to generate the conformers as fast as possible without waiting for optimization process. Parset6 and parset7 use the same values as parset2 with the only difference of no optimization (i.e., SPE only) in both and clustering is also off in the latter.

Parset5 uses similar clustering to parset2, but instead of picking the lowest energy structure from each clusters, it picks 2 additional structures. Therefore, for each of the “nfold” clusters a total of 3 structures are optimized. By doing this, mistakes by the K-means clustering are attempted to overcome. This improves the performance of reproducing QM optimized model and ideal structures. Within 36 total suggested local minima, we observed a success of 46.6% in model and 67.1% in ideal structures based on

the $\text{RMSD} \leq 0.5 \text{ \AA}$. Mean RMSD values of among all conformers generated are observed as 0.69 and 0.43 $\text{\AA}$ for model and ideal structures, respectively.

We also assessed the performance of our DeepConf in the case of no-optimization involvement and thus based on single point energy (SPE) calculations since sometimes, it might be desired to generate the conformers as fast as possible without waiting for optimization process. These two settings show quite successful performance with respect to their tremendously low computational cost.

With the assumption that model structures and ideal structures represent two different minima for any ligand, RMSD values of the conformers produced by DeepConf, Auto3D or RDKit against bioactive conformations are a good indicator to show the success of the conformer generation algorithms. However, this is only one part of the problem and the energy difference between model and ideal structure also needs to be distinguished by the algorithm. For this reason, we have also analyzed the results by comparing relative energies among conformers (**Table 3.13**).

In this step, we have first compared the absolute energies produced by DeepConf and Auto3D to their QM optimized analogues (For, explicit energies and differences of individual ligands, refer Supporting Information). The energy differences QM optimized model and ideal structures, span in a range from -14.96 kcal/mol to 27.20 kcal/mol (**Table 3.13**). As the ideal structure is not explicitly meant to be the global minimum and our procedure for selecting ligands only assured that QM optimized model and ideal structures represent two different minima; a model structure can have lower energy than the ideal structure. Although, the relative energy between the model and the ideal structures range -14.96 kcal to 27.20 kcal/mol, the mean energy difference, among all ligands is just 2.75 kcal/mol, clearly indicating the energetically close structures.

QM optimized structures (model or ideal) represent a single minimum whereas conformer generation algorithms generate many other minima alongside model and ideal structures when the algorithm is initiated with QM optimized ideal structures. These conformers are already reported as ensemble size.

To calculate the mean unsigned difference (MUD) values between the conformer generation algorithm's predicted structures and the QM-optimized structures, we used the energy values of structures with the minimum RMSD to the QM-optimized reference. This approach was applied even if the algorithm identified lower-energy structures, as we focused on conformers within an RMSD cutoff of 0.5 Å, which are assumed to represent the same conformer identified by both QM and ML methods. This is because our reference points are QM optimized model and ideal structures, and we have not thoroughly explored all the conformers with QM calculations due to computational cost. The MUD values of DeepConf in predicting both QM optimized model and ideal structures (i.e., $\langle \Delta E_{model}^{ML-QM} \rangle$ and $\langle \Delta E_{ideal}^{ML-QM} \rangle$), are 2.65 and 2.66 kcal/mol, respectively, with a span from -6 kcal/mol to 24 kcal/mol for the model structures, and in a range of -7 to 12 kcal/mol on ideal structures. Similarly, the MUD value of ideal structures in Auto3D is 2.87 kcal/mol ranging between -5 kcal/mol to 17 kcal/mol. The MUD values of Auto3D predicted model structures deviate from the QM optimized model structures by 3.32 kcal/mol in a range of -10 kcal/mol and 18 kcal/mol, indicating slightly less confidence. Unfortunately, RDKit predictions using either MMFF94 or UFF do not give absolute energy values, so we could not directly compare RDKit predicted absolute energies to QM optimized energies.

Table 3.13: The performance of conformer generation algorithms on reproducing QM optimized energies. Each kT refers to 0.592 kcal/mol and MUD is the abbreviation of Mean Unsigned Difference (in kcal/mol) over all evaluated ligands.

			ΔE^{ML-QM} (kcal/mol)			Success Energy cutoff			
Method	Parameter		MUD	min	max	kT	2kT	3kT	Count
DeepConf	parset2	model	2.65	-6.7	24.14	14.5%	35.0%	44.4%	117
		ideal	2.66	-7.44	12.2	14.5%	27.3%	38.5%	
	parset5*	model	2.91	-14.7	11.3	13.0%	23.0%	41.0%	100
		ideal	2.68	-7.5	12.1	17.0%	27.0%	40.0%	
Auto3D	auto3d	model	3.32	-10.1	18.5	11.0%	25.5%	37.9%	145
		ideal	2.87	-5.0	16.7	13.1%	24.8%	35.9%	
* Subset of 282 compounds									

We also considered how the conformer generation algorithms perform predicting the energy gap between the two minima (QM optimized model and ideal) since this gap is related to the relative populations (**Table 3.14** and **Figure 3.17**). In this analysis, $\Delta E_{model-ideal}^{QM}$ (i.e., the energy difference between QM optimized model and ideal structures) as the reference relative energy and $\Delta E_{model-ideal}^{conf}$ in which conformer generation algorithms' best solutions (the ones with minimum RMSD with ≤ 0.5 Å) were selected as the sample. Thus, the success on the relative energies corresponds to $\Delta\Delta E = \Delta_{model-ideal}^{conf} - \Delta_{model-ideal}^{QM}$. The algorithm is assumed to be successful for a ligand when $\Delta\Delta E$ is less than kT energy of 0.592 kcal/mol. It should be noted that the success of the relative energies suffers from the low performance of RMSD values. DeepConf with parset2 finds 30.7% within kT energy margin between the model and ideal structures while Auto3D finds 23.4%. Interestingly, MMFF94 based RDKit predictions are also quite comparable to Auto3D especially when the margin is increased to larger kT values. Yet, DeepConf predicted success is larger than all the other three methods.

Table 3.14: The performance of conformer generation algorithms on reproducing relative energies between model and ideal structures. Each kT refers to 0.592 kcal/mol and MUD is the abbreviation of Mean Unsigned Difference (in kcal/mol) over all evaluated ligands.

Method	Parameter	$\Delta\Delta E$ (kcal/mol)	Success Energy cutoff			count
		MUD	kT	2kT	3kT	
DeepConf	parset2	2.9	30.7%	44.4%	48.7%	117
	parset5*	1.8	26.0%	42.0%	58.0%	100
Auto3D	auto3d	2.6	23.4%	37.2%	46.9%	145
RDKit	MMFF94	3.3	22.3%	37.9%	48.5%	103
	UFF	3.0	16.5%	28.2%	43.5%	85
* Subset of 282 compounds						

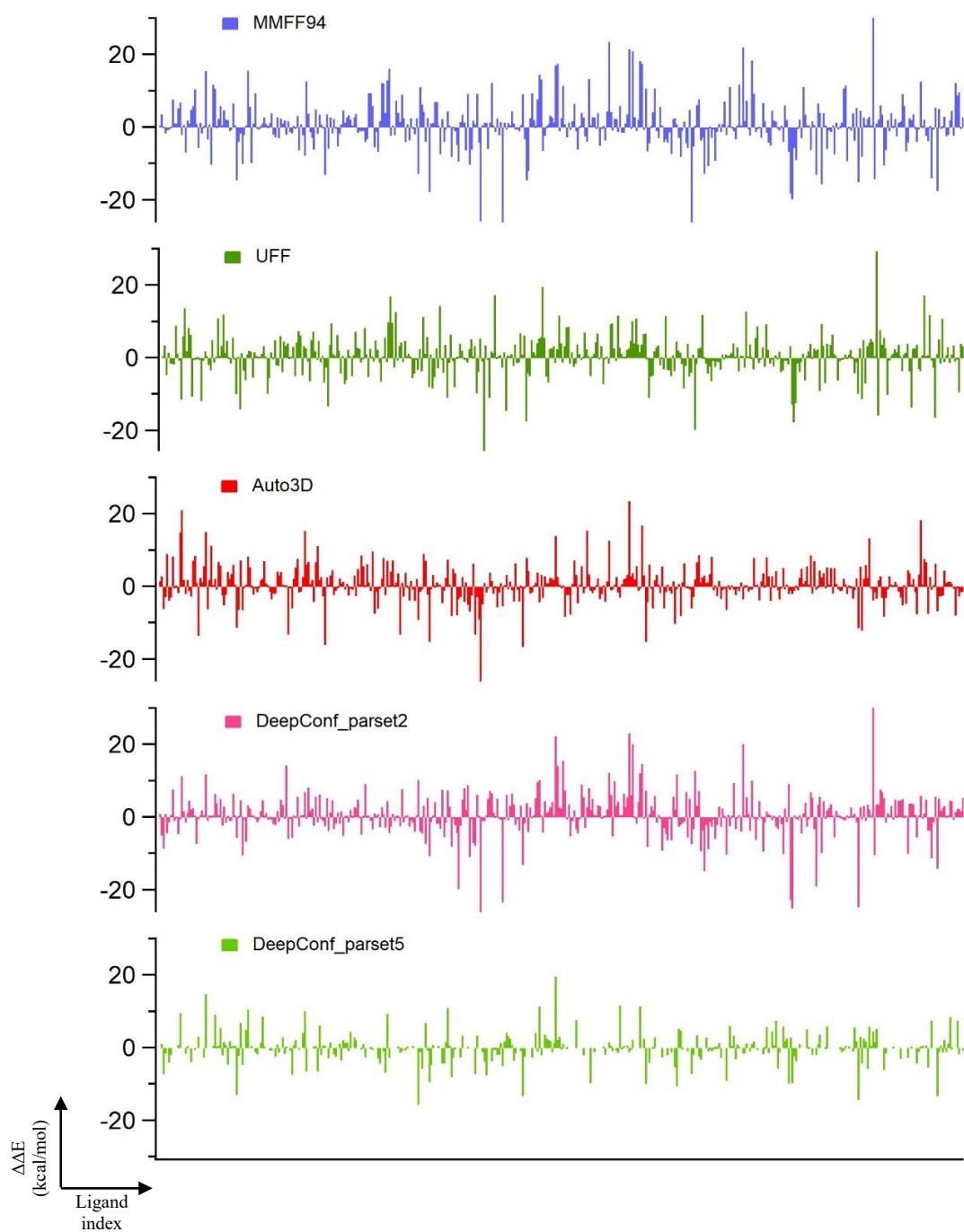


Figure 3.17: Distribution of relative energy differences ($\Delta\Delta E = \Delta_{model-ideal}^{conf} - \Delta_{model-ideal}^{QM}$) of conformers generated by different algorithms.

Table 3.15: The performance of conformer generation algorithms on reproducing the true global and local minima for selected 21 bioactive compounds.

Global to G16 global success %																	
	G16		DeepConf: parset2					DeepConf: parset5					Auto3D				
id	size	Bioactive rank	size	top1	top3	top10	all	size	top1	top3	top10	all	size	top1	top3	top10	all
048	48	10	27	1.6	0.7	0.4	0.4	76	1.3	1.3	0.3	0.3	35	1.8	1.0	0.7	0.4
05B	43	3	3	1.6	1.6	1.6	1.6	11	1.8	1.8	1.6	1.6	13	1.0	1.0	0.7	0.7
0C7	13	1	21	0.7	0.7	0.7	0.7	71	3.2	3.1	2.3	0.9	99	0.6	0.6	0.6	0.6
0FT	70	2	16	1.8	1.8	0.8	0.8	47	2.0	1.8	1.2	0.8	227	1.2	1.2	0.4	0.4
0KQ	54	11	8	2.5	1.5	1.0	1.0	24	1.9	1.9	0.9	0.9	62	2.2	2.0	0.9	0.3
0KR	54	39	8	2.1	1.8	0.8	0.8	22	2.2	1.4	1.4	0.8	74	1.9	1.9	1.3	0.5
0VG	29	1	12	2.7	2.7	2.2	2.2	43	2.8	2.6	2.4	1.6	69	2.7	2.7	2.5	1.1
0YO	12	2	7	1.4	0.1	0.1	0.1	7	1.5	0.2	0.1	0.1	4	1.5	0.1	0.1	0.1
15V	5	2	15	0.1	0.1	0.1	0.1	46	0.2	0.2	0.2	0.2	60	0.9	0.2	0.2	0.2
19R	33	4	27	1.4	1.3	0.9	0.9	93	1.5	1.4	0.6	0.6	124	1.2	1.2	0.3	0.3
1A2	75	11	8	1.9	1.4	1.4	1.4	23	2.5	0.7	0.7	0.7	62	2.2	0.4	0.4	0.4
1C7	58	2	14	0.8	0.8	0.8	0.8	46	2.9	1.8	0.9	0.8	118	2.1	0.7	0.2	0.2
1GB	42	2	15	1.6	1.6	1.6	1.6	47	2.1	2.1	2.1	1.6	105	2.5	2.2	2.2	1.0
1IF	18	15	15	0.9	0.6	0.6	0.6	42	0.4	0.4	0.4	0.4	43	2.6	1.0	0.6	0.6
21G	27	6	15	2.7	1.8	0.6	0.6	34	0.5	0.5	0.5	0.3	22	0.5	0.5	0.5	0.5
2M3	29	4	16	1.5	1.5	1.5	1.5	48	2.2	2.2	1.3	1.3	117	2.2	2.1	0.9	0.9
31T	50	5	14	1.8	1.7	1.0	1.0	35	3.1	1.7	1.0	0.3	65	0.5	0.1	0.1	0.1
33M	13	7	4	1.8	0.3	0.3	0.3	4	1.8	0.5	0.5	0.5	4	1.8	0.5	0.5	0.5
342	83	2	30	4.2	4.2	2.7	2.7	96	4.5	3.9	2.3	2.2	164	3.0	3.0	2.6	2.5
36Z	9	1	6	1.5	0.0	0.0	0.0	4	1.0	0.6	0.6	0.6	4	1.9	0.1	0.1	0.1
38W	14	3	22	2.7	2.3	0.6	0.6	53	1.1	1.1	0.9	0.9	26	1.1	1.1	1.1	0.6
mean	37.1	6.3	14.4	1.8	1.4	0.9	0.9	41.5	1.9	1.5	1.1	0.8	71.3	1.7	1.1	0.8	0.6
success %	RMSD ≤	0.5		5	19	24	24		14	24	29	33		5	33	48	57
	RMSD ≤	1.0		19	38	67	67		14	33	57	76		19	48	76	90
	RMSD ≤	1.5		33	52	81	81		33	52	76	81		43	71	86	95

Local to G16 local success %									
	DeepConf: parset2			DeepConf: parset5			Auto3D		
id	top3	top10	all	top3	top10	all	top3	top10	all
048	67	90	44	67	50	26	67	80	37
05B	33	33	33	67	50	45	67	60	46
0C7	33	50	24	67	40	11	67	40	7
0FT	0	20	19	33	40	21	0	10	17
0KQ	67	38	38	67	90	42	33	70	40
0KR	33	38	38	67	50	32	67	70	35
0VG	33	20	17	33	30	7	67	50	14
0YO	67	43	43	67	43	43	33	25	25
15V	0	10	7	0	0	4	0	0	5
19R	67	50	19	67	70	13	33	80	12
1A2	0	13	13	33	20	9	33	40	15
1C7	67	40	29	67	70	20	67	70	20
1GB	67	40	27	33	40	15	67	40	10
1IF	0	0	0	0	0	0	0	0	2
21G	67	50	33	67	70	35	67	70	41
2M3	67	20	13	67	20	4	0	30	6
31T	67	40	29	33	30	26	33	30	23
33M	33	25	25	33	25	25	0	0	0
342	33	30	10	33	10	3	33	20	5
36Z	0	0	0	0	0	0	0	0	0
38W	33	50	23	33	50	13	67	50	19
mean	40	33	23	44	38	19	38	40	18

In our previous evaluation, success was measured based on finding the "model" (bioactive) conformation starting from the "ideal" structure. Both Auto3D and DeepConf showed comparable performance, surpassing conventional conformer generation methods and reported literature values. However, this analysis only

considered two conformers per ligand, leaving the question open as to how efficiently these algorithms identify the true "global" and other "local" minima.

To address this, we conducted an additional study on a subset of 21 randomly selected ligands from the 439 PDB ligands. We performed classical MD simulations for each ligand, extracting 100 frames per ligand and optimizing them using QM methods. After eliminating duplicate structures, we obtained QM reference conformers for each ligand. The lowest-energy structure among these was designated as the "global" minimum, while the remaining conformers were classified as local minima.

Table 3.15 presents the results of this expanded analysis, including the number of QM-optimized conformers per ligand and the ranking of bioactive conformations among them. Interestingly, bioactive conformers were predominantly ranked among the lowest 10 energy structures, with an average rank of 6.3 across the dataset. On average, 37.1 unique conformers were generated per ligand from the QM optimizations of 100 MD frames.

Following the methodology of Table 3.5 and Table 3.11, we analyzed the ability of DeepConf and Auto3D to predict the global and local minima among these ligands. When only the lowest-energy structure of each algorithm is considered (Top 1), Auto3D and DeepConf (parset 2) identify the QM global minimum with 5% success, while DeepConf (parset 5) achieves 14% success. When the top 3, top 10, or all available conformers are considered, Auto3D reaches 57% success, whereas DeepConf (parset 5) reaches 33% success. The mean RMSD values for the top 1, top 3, and top 10 conformers are similar for both methods. However, when all conformers are considered, Auto3D achieves a lower mean RMSD (0.6 Å), surpassing DeepConf. The local-to-G16 local success rates for conformers generated by DeepConf and Auto3D were also evaluated. The ensemble sizes for Auto3D, parset 2, and parset 5 are 71.3, 14.4, and 41.5, respectively. 38% of Auto3D's top 3 conformers match local minima in the QM dataset, while parset 5 achieves 44% success in this regard.

We also investigated the effect of energy minimization steps and convergence thresholds (fmax) on both Auto3D and DeepConf (Table 3.16). While reducing fmax lowered the ensemble sizes, it did not improve success rates or RMSD values. In fact, overly stringent minimization criteria often worsened results due to the reduction in the number of explored conformers.

Table 3.16: Effect of maximum step / convergence threshold on conformer generation performances.

	auto3d_f002			auto3d_f00001			confgen_f002			confgen_f00002		
	Auto3D global to G16 global			Auto3D global to G16 global								
ID	#auto3d numconf	min top	min in all	#auto3d numconf	min top	min in all	parset5	mintop	min in all	parset5	mintop	min in all
38W	26	1.1	0.6	13	1.1	0.8	53	1.1	0.9	23	1.1	1.1
2WB	28	0.8	0.6	15	0.8	0.8						
2M3	117	2.2	0.9	68	2.8	0.7	48	2.2	1.3	34	2.5	1.1
2KW	35	2.6	1.0	6	2.6	1.1	48	2.2	1.0			
21G	22	0.5	0.5	20	0.4	0.4	34	0.5	0.3			
1GB	105	2.5	1.0	28	1.1	1.1	47	2.1	1.6	27	2.1	1.6
0YO	4	1.5	0.1	2	1.5	1.5	7	1.5	0.1	10	1.4	0.1
0VG	69	2.7	1.1	25	2.7	1.9	43	2.8	1.6			
0C7	99	0.6	0.6	31	0.5	0.2				36	0.5	0.2
05B	13	1.0	0.7	9	0.9	0.7	11	1.8	1.6	10	1.6	1.5
mean	51.8	1.6	0.7	21.7	1.4	0.9	36.4	1.8	1.0	23.3	1.5	0.9
%success<0.5		1	2		2	2		1	2		1	2
%success<1		3	9		4	6		1	3		1	2
%success<1.5		6	10		7	9		3	5		3	4

Finally, we also analyzed the computational performance of different parsets and Auto3D (Table 3.17).

Table 3.17: Computational performance of Auto3D and DeepConf with different parameter settings for **one ligand** with **five** torsion

	parset1	parset2	parset3	parset4	parset5	parset6	parset7	auto3d
Number of SPE calculations	10240	320	320	10240	320	320	320	32
Number of optimizations	1024	32	64	1024	32	32	0	32
Number of clusters	1024	32	32	1024	64	32	32	0
Wall time (minutes)*	40	2	3.5	80	8	4	0.2	1.2
*: The calculations are performed on the CPU only using a single machine with Intel(R) Xeon(R) Platinum 8480+ 2.0GHz, 256GB Memory, 2x56=112 cores								

3.2.3. Summary

In conclusion, we introduce DeepConf, a novel low-energy conformer generation algorithm that utilizes ANI-ML potentials at DFT-level accuracy for efficiently reproducing bioactive conformations. Our method demonstrates robustness in challenging scenarios, such as when the initial structure is far from equilibrium or when fewer, high-quality conformers are required. By focusing on single-bond rotations, we provide the first comprehensive assessment of ANI-ML potentials using DeepConf alongside Auto3D for bioactive conformation generation. Additionally, we propose new evaluation methods that can serve as a guideline for future conformer generation algorithms. Both DeepConf and Auto3D based on ANI-ML potentials achieve a mean RMSD of less than 0.5 Å, outperforming conventional approaches. With advanced features like geometry optimization, single-point energy-based conformer generation, and compatibility with various ML potentials, DeepConf offers a powerful tool for bioactive conformation evaluation. We should also emphasize that ANI-ML potentials have an intrinsic error margin of ~ 1.5 kcal/mol, which affects both DeepConf and Auto3D. This error contributes to challenges in correctly ranking global and local minima, particularly in cases where multiple low-energy conformers exist.

4. CONCLUSIONS

In this study, we investigate drug repurposing as a time-efficient and cost-effective therapeutic strategy against COVID-19, caused by the SARS-CoV-2 virus, which has drastically disrupted global healthcare systems and socioeconomic structures. Drug repurposing involves the identification of novel therapeutic indications for already FDA-approved compounds, leveraging their established pharmacokinetic and safety profiles to accelerate the drug development pipeline. This approach is particularly valuable in the context of rapidly spreading infectious diseases such as COVID-19.

We performed a structure-based virtual screening of approximately 2,500 FDA-approved drugs targeting the SARS-CoV-2 main protease (Mpro), a critical enzyme involved in viral replication. The workflow incorporates molecular docking, molecular dynamics (MD) simulations, and end-state free energy methods to refine candidate selection. The initial docking step was carried out using the GOLD software, and the predicted binding poses were subsequently rescored using ANI machine learning potentials, which offer quantum mechanical accuracy at a reduced computational cost.

Furthermore, the study extends to high-throughput screening of these drugs against Mpro variants associated with different SARS-CoV-2 strains, including beta, delta, and omicron lineages. The integration of ANI potentials allows us to capture subtle electronic effects that classical force fields often fail to resolve. Following docking and rescoring, we performed extensive MD simulations and free energy calculations (e.g., MM/PBSA, ANI_LIE) to estimate binding affinities and assess complex stability.

The outcomes of this multi-tiered computational protocol will provide insights into drug-Mpro interactions at the atomic level and guide experimental validation efforts. This comprehensive approach aims to identify potent candidates for immediate therapeutic intervention and contributes to the broader field of structure-based drug repurposing through the integration of advanced ML-based scoring functions.

REFERENCES

- [1] Ferrara P, Gohlke H, Price DJ, Klebe G, Brooks CL. Assessing scoring functions for protein–ligand interactions. *J Med Chem*. 2004;47:3032–47., (n.d.).
- [2] L. Ferreira, R. Dos Santos, G. Oliva, A. Andricopulo, Molecular Docking and Structure-Based Drug Design Strategies, *Molecules* 20 (2015) 13384–13421. <https://doi.org/10.3390/molecules200713384>.
- [3] E. Lionta, G. Spyrou, D. Vassilatis, Z. Cournia, Structure-Based Virtual Screening for Drug Discovery: Principles, Applications and Recent Advances, *Curr Top Med Chem* 14 (2014) 1923–1938. <https://doi.org/10.2174/1568026614666140929124445>.
- [4] Wenjia Wang, Applications of biophysical methods in small-molecule modulators targeting protein function, University of Groningen, 2021. <https://doi.org/10.33612/diss.156123664>.
- [5] L. Ferreira, R. Dos Santos, G. Oliva, A. Andricopulo, Molecular Docking and Structure-Based Drug Design Strategies, *Molecules* 20 (2015) 13384–13421. <https://doi.org/10.3390/molecules200713384>.
- [6] R.L.M. van Montfort, P. Workman, Structure-based drug design: aiming for a perfect fit, *Essays Biochem* 61 (2017) 431–437. <https://doi.org/10.1042/EBC20170052>.
- [7] S.V.K.R.S.R.V.R.S.R. Shalini Ramesh, Computer-Aided Drug Designing – Modality of Diagnostic System, 1st ed., 2021.
- [8] M. Batool, B. Ahmad, S. Choi, A Structure-Based Drug Discovery Paradigm, *Int J Mol Sci* 20 (2019) 2783. <https://doi.org/10.3390/ijms20112783>.
- [9] E. Valeur, S.M. Guéret, H. Adihou, R. Gopalakrishnan, M. Lemurell, H. Waldmann, T.N. Grossmann, A.T. Plowright, Neue Modalitäten für schwierige Zielstrukturen in der Wirkstoffentwicklung, *Angewandte Chemie* 129 (2017) 10428–10459. <https://doi.org/10.1002/ange.201611914>.
- [10] R. Lappano, M. Maggiolini, G protein-coupled receptors: novel targets for drug discovery in cancer, *Nat Rev Drug Discov* 10 (2011) 47–60. <https://doi.org/10.1038/nrd3320>.
- [11] M. Batool, B. Ahmad, S. Choi, A Structure-Based Drug Discovery Paradigm, *Int J Mol Sci* 20 (2019) 2783. <https://doi.org/10.3390/ijms20112783>.
- [12] A. ODA, Development and Validation of Programs for Ligand-binding-pocket Search, *YAKUGAKU ZASSHI* 131 (2011) 1429–1435. <https://doi.org/10.1248/yakushi.131.1429>.

- [13] E.H.B. Maia, L.C. Assis, T.A. de Oliveira, A.M. da Silva, A.G. Taranto, Structure-Based Virtual Screening: From Classical to Artificial Intelligence, *Front Chem* 8 (2020). <https://doi.org/10.3389/fchem.2020.00343>.
- [14] R. Nussinov, C.-J. Tsai, The Different Ways through Which Specificity Works in Orthosteric and Allosteric Drugs, *Curr Pharm Des* 18 (2012) 1311–1316. <https://doi.org/10.2174/138161212799436377>.
- [15] E. Harigua-Souiai, I. Cortes-Ciriano, N. Desdouits, T.E. Malliavin, I. Guizani, M. Nilges, A. Blondel, G. Bouvier, Identification of binding sites and favorable ligand binding moieties by virtual screening and self-organizing map analysis, *BMC Bioinformatics* 16 (2015) 93. <https://doi.org/10.1186/s12859-015-0518-z>.
- [16] S.H. Oliveira, F.A. Ferraz, R. V Honorato, J. Xavier-Neto, T.J. Sobreira, P.S. de Oliveira, KVFinder: steered identification of protein cavities as a PyMOL plugin, *BMC Bioinformatics* 15 (2014) 197. <https://doi.org/10.1186/1471-2105-15-197>.
- [17] A. Armon, D. Graur, N. Ben-Tal, ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information, *J Mol Biol* 307 (2001) 447–463. <https://doi.org/10.1006/jmbi.2000.4474>.
- [18] T. Simões, D. Lopes, S. Dias, F. Fernandes, J. Pereira, J. Jorge, C. Bajaj, A. Gomes, Geometric Detection Algorithms for Cavities on Protein Surfaces in Molecular Graphics: A Survey, *Computer Graphics Forum* 36 (2017) 643–683. <https://doi.org/10.1111/cgf.13158>.
- [19] R.A. Laskowski, SURFNET: A program for visualizing molecular surfaces, cavities, and intermolecular interactions, *J Mol Graph* 13 (1995) 323–330. [https://doi.org/10.1016/0263-7855\(95\)00073-9](https://doi.org/10.1016/0263-7855(95)00073-9).
- [20] J. Yu, Y. Zhou, I. Tanaka, M. Yao, Roll: a new algorithm for the detection of protein pockets and cavities with a rolling probe sphere, *Bioinformatics* 26 (2010) 46–52. <https://doi.org/10.1093/bioinformatics/btp599>.
- [21] P. Schmidtke, A. Bidon-Chanal, F.J. Luque, X. Barril, MDpocket: open-source cavity detection and characterization on molecular dynamics trajectories, *Bioinformatics* 27 (2011) 3276–3285. <https://doi.org/10.1093/bioinformatics/btr550>.
- [22] T. Simões, D. Lopes, S. Dias, F. Fernandes, J. Pereira, J. Jorge, C. Bajaj, A. Gomes, Geometric Detection Algorithms for Cavities on Protein Surfaces in Molecular Graphics: A Survey, *Computer Graphics Forum* 36 (2017) 643–683. <https://doi.org/10.1111/cgf.13158>.
- [23] L. Vangeel, W. Chiu, S. De Jonghe, P. Maes, B. Slechten, J. Raymenants, E. André, P. Leyssen, J. Neyts, D. Jochmans, Remdesivir, Molnupiravir and

Nirmatrelvir remain active against SARS-CoV-2 Omicron and other variants of concern, *Antiviral Res* 198 (2022). <https://doi.org/10.1016/j.antiviral.2022.105252>.

[24] E. Yuriev, P.A. Ramsland, Latest developments in molecular docking: 2010–2011 in review, *Journal of Molecular Recognition* 26 (2013) 215–239. <https://doi.org/10.1002/jmr.2266>.

[25] V. Salmaso, S. Moro, Bridging Molecular Docking to Molecular Dynamics in Exploring Ligand-Protein Recognition Process: An Overview, *Front Pharmacol* 9 (2018). <https://doi.org/10.3389/fphar.2018.00923>.

[26] X.-Y. Meng, H.-X. Zhang, M. Mezei, M. Cui, Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery, *Current Computer Aided-Drug Design* 7 (2011) 146–157. <https://doi.org/10.2174/157340911795677602>.

[27] T. A, B. VA, Molecular Docking: From Lock and Key to Combination Lock, *J Mol Med Clin Appl* 2 (2018). <https://doi.org/10.16966/2575-0305.106>.

[28] Z. Wang, H. Sun, X. Yao, D. Li, L. Xu, Y. Li, S. Tian, T. Hou, Comprehensive evaluation of ten docking programs on a diverse set of protein–ligand complexes: the prediction accuracy of sampling power and scoring power, *Physical Chemistry Chemical Physics* 18 (2016) 12964–12975. <https://doi.org/10.1039/C6CP01555G>.

[29] R.G. Coleman, M. Carchia, T. Sterling, J.J. Irwin, B.K. Shoichet, Ligand Pose and Orientational Sampling in Molecular Docking, *PLoS One* 8 (2013) e75992. <https://doi.org/10.1371/journal.pone.0075992>.

[30] B.J. Neves, M. Mottin, J.T. Moreira-Filho, B.K. de P. Sousa, S.S. Mendonca, C.H. Andrade, Best Practices for Docking-Based Virtual Screening, in: *Molecular Docking for Computer-Aided Drug Design*, Elsevier, 2021: pp. 75–98. <https://doi.org/10.1016/B978-0-12-822312-3.00001-1>.

[31] I.A. Guedes, C.S. de Magalhães, L.E. Dardenne, Receptor–ligand molecular docking, *Biophys Rev* 6 (2014) 75–87. <https://doi.org/10.1007/s12551-013-0130-2>.

[32] G.M. Morris, D.S. Goodsell, R.S. Halliday, R. Huey, W.E. Hart, R.K. Belew, A.J. Olson, Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function, *J Comput Chem* 19 (1998) 1639–1662. [https://doi.org/10.1002/\(SICI\)1096-987X\(19981115\)19:14<1639::AID-JCC10>3.0.CO;2-B](https://doi.org/10.1002/(SICI)1096-987X(19981115)19:14<1639::AID-JCC10>3.0.CO;2-B).

[33] R. Huey, G.M. Morris, A.J. Olson, D.S. Goodsell, A semiempirical free energy force field with charge-based desolvation, *J Comput Chem* 28 (2007) 1145–1152. <https://doi.org/10.1002/jcc.20634>.

[34] B.S.P. Bentham Science Publisher, Scoring Functions for Protein-Ligand Docking, *Curr Protein Pept Sci* 7 (2006) 407–420. <https://doi.org/10.2174/138920306778559395>.

- [35] R.D. Taylor, P.J. Jewsbury, J.W. Essex, protein-small molecule docking methods, *J Comput Aided Mol Des* 16 (2002) 151–166. <https://doi.org/10.1023/A:1020155510718>.
- [36] P. Ferrara, H. Gohlke, D.J. Price, G. Klebe, C.L. Brooks, Assessing Scoring Functions for Protein–Ligand Interactions, *J Med Chem* 47 (2004) 3032–3047. <https://doi.org/10.1021/jm030489h>.
- [37] J. Li, A. Fu, L. Zhang, An Overview of Scoring Functions Used for Protein–Ligand Interactions in Molecular Docking, *Interdiscip Sci* 11 (2019) 320–328. <https://doi.org/10.1007/s12539-019-00327-w>.
- [38] R. Wang, Y. Lu, S. Wang, Comparative Evaluation of 11 Scoring Functions for Molecular Docking, *J Med Chem* 46 (2003) 2287–2303. <https://doi.org/10.1021/jm0203783>.
- [39] R. Wang, Y. Lu, S. Wang, Comparative Evaluation of 11 Scoring Functions for Molecular Docking, *J Med Chem* 46 (2003) 2287–2303. <https://doi.org/10.1021/jm0203783>.
- [40] C. Shen, Y. Hu, Z. Wang, X. Zhang, J. Pang, G. Wang, H. Zhong, L. Xu, D. Cao, T. Hou, Beware of the generic machine learning-based scoring functions in structure-based virtual screening, *Brief Bioinform* 22 (2021). <https://doi.org/10.1093/bib/bbaa070>.
- [41] P.C. Nair, M.E. Sobhia, CoMFA based de novo design of pyridazine analogs as PTP1B inhibitors, *J Mol Graph Model* 26 (2007) 117–123. <https://doi.org/10.1016/j.jmgm.2006.10.002>.
- [42] C.M. Venkatachalam, X. Jiang, T. Oldfield, M. Waldman, LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites, *J Mol Graph Model* 21 (2003) 289–307. [https://doi.org/10.1016/S1093-3263\(02\)00164-X](https://doi.org/10.1016/S1093-3263(02)00164-X).
- [43] G. Jones, P. Willett, R.C. Glen, A.R. Leach, R. Taylor, Development and validation of a genetic algorithm for flexible docking 1 Edited by F. E. Cohen, *J Mol Biol* 267 (1997) 727–748. <https://doi.org/10.1006/jmbi.1996.0897>.
- [44] Z. Shentu, M. Al Hasan, C. Bystroff, M.J. Zaki, Context shapes: Efficient complementary shape matching for protein–protein docking, *Proteins: Structure, Function, and Bioinformatics* 70 (2008) 1056–1073. <https://doi.org/10.1002/prot.21600>.
- [45] J. Pei, Q. Wang, Z. Liu, Q. Li, K. Yang, L. Lai, PSI-DOCK: Towards highly efficient and accurate flexible ligand docking, *Proteins: Structure, Function, and Bioinformatics* 62 (2006) 934–946. <https://doi.org/10.1002/prot.20790>.
- [46] V.D. Mouchlis, A. Afantitis, A. Serra, M. Fratello, A.G. Papadimitriou, V. Aidinis, I. Lynch, D. Greco, G. Melagraki, Advances in De Novo Drug Design:

From Conventional to Machine Learning Methods, *Int J Mol Sci* 22 (2021) 1676. <https://doi.org/10.3390/ijms22041676>.

[47] G. Schneider, K. Baringhaus, *De Novo Design: From Models to Molecules*, in: *De Novo Molecular Design*, Wiley, 2013: pp. 1–55. <https://doi.org/10.1002/9783527677016.ch1>.

[48] V.D. Mouchlis, A. Afantitis, A. Serra, M. Fratello, A.G. Papadiamantis, V. Aidinis, I. Lynch, D. Greco, G. Melagraki, *Advances in De Novo Drug Design: From Conventional to Machine Learning Methods*, *Int J Mol Sci* 22 (2021) 1676. <https://doi.org/10.3390/ijms22041676>.

[49] G. Schneider, U. Fechner, Computer-based de novo design of drug-like molecules, *Nat Rev Drug Discov* 4 (2005) 649–663. <https://doi.org/10.1038/nrd1799>.

[50] K. Kawai, N. Nagata, Y. Takahashi, De Novo Design of Drug-Like Molecules by a Fragment-Based Molecular Evolutionary Approach, *J Chem Inf Model* 54 (2014) 49–56. <https://doi.org/10.1021/ci400418c>.

[51] D. Tian, Y. Liu, C. Liang, L. Xin, X. Xie, D. Zhang, M. Wan, H. Li, X. Fu, H. Liu, W. Cao, An update review of emerging small-molecule therapeutic options for COVID-19, *Biomedicine and Pharmacotherapy* 137 (2021). <https://doi.org/10.1016/j.biopha.2021.111313>.

[52] S. Das, A. Abraham, A. Konar, Swarm Intelligence Algorithms in Bioinformatics, in: 2008: pp. 113–147. https://doi.org/10.1007/978-3-540-76803-6_4.

[53] M. Sorokina, P. Merseburger, K. Rajan, M.A. Yirik, C. Steinbeck, COCONUT online: Collection of Open Natural Products database, *J Cheminform* 13 (2021) 2. <https://doi.org/10.1186/s13321-020-00478-9>.

[54] H. Chen, O. Engkvist, Y. Wang, M. Olivecrona, T. Blaschke, The rise of deep learning in drug discovery, *Drug Discov Today* 23 (2018) 1241–1250. <https://doi.org/10.1016/j.drudis.2018.01.039>.

[55] C.K. Schissel, S. Mohapatra, J.M. Wolfe, C.M. Fadzen, K. Bellovoda, C.-L. Wu, J.A. Wood, A.B. Malmberg, A. Loas, R. Gómez-Bombarelli, B.L. Pentelute, Interpretable Deep Learning for De Novo Design of Cell-Penetrating Abiotic Polymers, (2020). <https://doi.org/10.1101/2020.04.10.036566>.

[56] E. Putin, A. Asadulaev, Y. Ivanenkov, V. Aladinskiy, B. Sanchez-Lengeling, A. Aspuru-Guzik, A. Zhavoronkov, Reinforced Adversarial Neural Computer for *de Novo* Molecular Design, *J Chem Inf Model* 58 (2018) 1194–1204. <https://doi.org/10.1021/acs.jcim.7b00690>.

- [57] N. Ståhl, G. Falkman, A. Karlsson, G. Mathiason, J. Boström, Deep Reinforcement Learning for Multiparameter Optimization in *de novo* Drug Design, *J Chem Inf Model* 59 (2019) 3166–3176. <https://doi.org/10.1021/acs.jcim.9b00325>.
- [58] M. Popova, O. Isayev, A. Tropsha, Deep reinforcement learning for *de novo* drug design, *Sci Adv* 4 (2018). <https://doi.org/10.1126/sciadv.aap7885>.
- [59] M. Hartenfeller, G. Schneider, Enabling future drug discovery by *de novo* design, *WIREs Computational Molecular Science* 1 (2011) 742–759. <https://doi.org/10.1002/wcms.49>.
- [60] M.A. Ator, J.P. Mallamo, M. Williams, Overview of Drug Discovery and Development, *Curr Protoc Pharmacol* 35 (2006). <https://doi.org/10.1002/0471141755.ph0909s35>.
- [61] M. Schenone, V. Dančák, B.K. Wagner, P.A. Clemons, Target identification and mechanism of action in chemical biology and drug discovery, *Nat Chem Biol* 9 (2013) 232–240. <https://doi.org/10.1038/nchembio.1199>.
- [62] J. Hughes, S. Rees, S. Kalindjian, K. Philpott, Principles of early drug discovery, *Br J Pharmacol* 162 (2011) 1239–1249. <https://doi.org/10.1111/j.1476-5381.2010.01127.x>.
- [63] S.J.Y. Macalino, V. Gosu, S. Hong, S. Choi, Role of computer-aided drug design in modern drug discovery, *Arch Pharm Res* 38 (2015) 1686–1701. <https://doi.org/10.1007/s12272-015-0640-5>.
- [64] D.C. Swinney, J. Anthony, How were new medicines discovered?, *Nat Rev Drug Discov* 10 (2011) 507–519. <https://doi.org/10.1038/nrd3480>.
- [65] R. Santos, O. Ursu, A. Gaulton, A.P. Bento, R.S. Donadi, C.G. Bologa, A. Karlsson, B. Al-Lazikani, A. Hersey, T.I. Oprea, J.P. Overington, A comprehensive map of molecular drug targets, *Nat Rev Drug Discov* 16 (2017) 19–34. <https://doi.org/10.1038/nrd.2016.230>.
- [66] G. Koscielny, P. An, D. Carvalho-Silva, J.A. Cham, L. Fumis, R. Gasparyan, S. Hasan, N. Karamanis, M. Maguire, E. Papa, A. Pierleoni, M. Pignatelli, T. Platt, F. Rowland, P. Wankar, A.P. Bento, T. Burdett, A. Fabregat, S. Forbes, A. Gaulton, C.Y. Gonzalez, H. Hermjakob, A. Hersey, S. Jupe, Ş. Kafkas, M. Keays, C. Leroy, F.-J. Lopez, M.P. Magarinos, J. Malone, J. McEntyre, A. Munoz-Pomer Fuentes, C. O'Donovan, I. Papatheodorou, H. Parkinson, B. Palka, J. Paschall, R. Petryszak, N. Pratanwanich, S. Sarntivijal, G. Saunders, K. Sidiropoulos, T. Smith, Z. Sondka, O. Stegle, Y.A. Tang, E. Turner, B. Vaughan, O. Vrousitou, X. Watkins, M.-J. Martin, P. Sanseau, J. Vamathevan, E. Birney, J. Barrett, I. Dunham, Open Targets: a platform for therapeutic target identification and validation, *Nucleic Acids Res* 45 (2017) D985–D994. <https://doi.org/10.1093/nar/gkw1055>.

- [67] H.J. Jung, H.J. Kwon, Target deconvolution of bioactive small molecules: the heart of chemical biology and drug discovery, *Arch Pharm Res* 38 (2015) 1627–1641. <https://doi.org/10.1007/s12272-015-0618-3>.
- [68] D. Jones, H. Kim, X. Zhang, A. Zemla, G. Stevenson, W.F.D. Bennett, D. Kirshner, S.E. Wong, F.C. Lightstone, J.E. Allen, Improved Protein–Ligand Binding Affinity Prediction with Structure-Based Deep Fusion Inference, *J Chem Inf Model* 61 (2021) 1583–1592. <https://doi.org/10.1021/acs.jcim.0c01306>.
- [69] X.S. Zheng, T.-F. Chan, H.H. Zhou, Genetic and Genomic Approaches to Identify and Study the Targets of Bioactive Small Molecules, *Chem Biol* 11 (2004) 609–618. <https://doi.org/10.1016/j.chembiol.2003.08.011>.
- [70] G. Sliwoski, S. Kothiwale, J. Meiler, E.W. Lowe, Computational Methods in Drug Discovery, *Pharmacol Rev* 66 (2014) 334–395. <https://doi.org/10.1124/pr.112.007336>.
- [71] T. Katsila, G.A. Spyroulias, G.P. Patrinos, M.-T. Matsoukas, Computational approaches in target identification and drug discovery, *Comput Struct Biotechnol J* 14 (2016) 177–184. <https://doi.org/10.1016/j.csbj.2016.04.004>.
- [72] K. Martinez-Mayorga, A. Madariaga-Mazon, J.L. Medina-Franco, G. Maggiora, The impact of chemoinformatics on drug discovery in the pharmaceutical industry, *Expert Opin Drug Discov* 15 (2020) 293–306. <https://doi.org/10.1080/17460441.2020.1696307>.
- [73] F.I. Saldívar-González, J.L. Medina-Franco, Approaches for enhancing the analysis of chemical space for drug discovery, *Expert Opin Drug Discov* 17 (2022) 789–798. <https://doi.org/10.1080/17460441.2022.2084608>.
- [74] E. López-López, J. Bajorath, J.L. Medina-Franco, Informatics for Chemistry, Biology, and Biomedical Sciences, *J Chem Inf Model* 61 (2021) 26–35. <https://doi.org/10.1021/acs.jcim.0c01301>.
- [75] J. Gasteiger, Chemistry in Times of Artificial Intelligence, *ChemPhysChem* 21 (2020) 2233–2242. <https://doi.org/10.1002/cphc.202000518>.
- [76] K. Martinez-Mayorga, A. Madariaga-Mazon, J.L. Medina-Franco, G. Maggiora, The impact of chemoinformatics on drug discovery in the pharmaceutical industry, *Expert Opin Drug Discov* 15 (2020) 293–306. <https://doi.org/10.1080/17460441.2020.1696307>.
- [77] K.M. Merz, R. Amaro, Z. Cournia, M. Rarey, T. Soares, A. Tropsha, H.A. Wahab, R. Wang, Editorial: Method and Data Sharing and Reproducibility of Scientific Results, *J Chem Inf Model* 60 (2020) 5868–5869. <https://doi.org/10.1021/acs.jcim.0c01389>.
- [78] S. Gupta, A.K. Singh, P.P. Kushwaha, K.S. Prajapati, M. Shuaib, S. Senapati, S. Kumar, Identification of potential natural inhibitors of SARS-CoV2 main

protease by molecular docking and simulation studies, *J Biomol Struct Dyn* 39 (2021). <https://doi.org/10.1080/07391102.2020.1776157>.

[79] W.P. Walters, Virtual Chemical Libraries, *J Med Chem* 62 (2019) 1116–1124. <https://doi.org/10.1021/acs.jmedchem.8b01048>.

[80] R. Lu, X. Zhao, J. Li, P. Niu, B. Yang, H. Wu, W. Wang, H. Song, B. Huang, N. Zhu, Y. Bi, X. Ma, F. Zhan, L. Wang, T. Hu, H. Zhou, Z. Hu, W. Zhou, L. Zhao, J. Chen, Y. Meng, J. Wang, Y. Lin, J. Yuan, Z. Xie, J. Ma, W.J. Liu, D. Wang, W. Xu, E.C. Holmes, G.F. Gao, G. Wu, W. Chen, W. Shi, W. Tan, Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding, *The Lancet* 395 (2020) 565–574. [https://doi.org/10.1016/S0140-6736\(20\)30251-8](https://doi.org/10.1016/S0140-6736(20)30251-8).

[81] Z. Gong, G. Hu, Q. Li, Z. Liu, F. Wang, X. Zhang, J. Xiong, P. Li, Y. Xu, R. Ma, S. Chen, J. Li, Compound Libraries: Recent Advances and Their Applications in Drug Discovery, *Curr Drug Discov Technol* 14 (2017). <https://doi.org/10.2174/1570163814666170425155154>.

[82] L. Mousavizadeh, S. Ghasemi, Genotype and phenotype of COVID-19: Their roles in pathogenesis, *Journal of Microbiology, Immunology and Infection* 54 (2021) 159–163. <https://doi.org/10.1016/j.jmii.2020.03.022>.

[83] M. Su, Q. Yang, Y. Du, G. Feng, Z. Liu, Y. Li, R. Wang, Comparative Assessment of Scoring Functions: The CASF-2016 Update, *J Chem Inf Model* 59 (2019) 895–913. <https://doi.org/10.1021/acs.jcim.8b00545>.

[84] P. Eastman, M.S. Friedrichs, J.D. Chodera, R.J. Radmer, C.M. Bruns, J.P. Ku, K.A. Beauchamp, T.J. Lane, L.-P. Wang, D. Shukla, T. Tye, M. Houston, T. Stich, C. Klein, M.R. Shirts, V.S. Pande, OpenMM 4: A Reusable, Extensible, Hardware Independent Library for High Performance Molecular Simulation, *J Chem Theory Comput* 9 (2013) 461–469. <https://doi.org/10.1021/ct300857j>.

[85] G. Jones, P. Willett, R.C. Glen, A.R. Leach, R. Taylor, Development and validation of a genetic algorithm for flexible docking¹¹ Edited by F. E. Cohen, *J Mol Biol* 267 (1997) 727–748. <https://doi.org/https://doi.org/10.1006/jmbi.1996.0897>.

[86] E. Akkus, O. Tayfuroglu, M. Yildiz, A. Kocak, Accurate Binding Free Energy Method from End-State MD Simulations, *J Chem Inf Model* 62 (2022) 4095–4106. <https://doi.org/10.1021/acs.jcim.2c00601>.

[87] M. Temel, O. Tayfuroglu, A. Kocak, The performance of ANI-ML potentials for ligand-n(H₂O) interaction energies and estimation of hydration free energies from end-point MD simulations, *J Comput Chem* 44 (2023) 559–569. <https://doi.org/10.1002/jcc.27022>.

[88] E. Akkus, O. Tayfuroglu, M. Yildiz, A. Kocak, Revisiting MMPBSA by Adoption of MC-Based Surface Area/Volume, ANI-ML Potentials, and Two-

Valued Interior Dielectric Constant, *J Phys Chem B* 127 (2023) 4415–4429. <https://doi.org/10.1021/acs.jpcc.3c00834>.

[89] H.S. Tasci, E. Akkus, M. Yildiz, A. Kocak, Computational analysis of substrate recognition of Sars-Cov-2 Mpro main protease, *Comput Biol Chem* 107 (2023) 107960. <https://doi.org/https://doi.org/10.1016/j.compbiolchem.2023.107960>.

[90] A. Kocak, I. Erol, M. Yildiz, H. Can, Computational insights into the protonation states of catalytic dyad in BACE1–acyl guanidine based inhibitor complex, *J Mol Graph Model* 70 (2016) 226–235. <https://doi.org/10.1016/j.jmgm.2016.10.013>.

[91] A. Kocak, M. Yıldız, Molecular Dynamics Studies of the Norovirus-Host Cell Interaction Mediated by H-type 1 Antigen, *Trak Univ J Nat Sci ASAP* (2019). <https://doi.org/10.23902/trkjnat.508120>.

[92] M. Yildiz, A. Kocak, 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 (2019) 962–974. <https://doi.org/10.1089/cmb.2018.0163>.

[93] A. Kocak, M. Yildiz, Docking, molecular dynamics and free energy studies on aspartoacylase mutations involved in Canavan disease, *J Mol Graph Model* 74 (2017) 44–53. <https://doi.org/10.1016/j.jmgm.2017.03.011>.

[94] A. Kocak, HBGA binding modes and selectivity in noroviruses upon mutation: a docking and molecular dynamics study, *J Mol Model* 25 (2019) 369. <https://doi.org/10.1007/s00894-019-4261-7>.

[95] S.G. Kankinou, M. Yildiz, A. Kocak, Exploring potential Plasmodium kinase inhibitors: a combined docking, MD and QSAR studies, *J Biomol Struct Dyn* (n.d.) 1–11. <https://doi.org/10.1080/07391102.2023.2249111>.

[96] U. Uzuner, E. Akkus, A. Kocak, S. Çelik Uzuner, Exploring epigenetic drugs as potential inhibitors of SARS-CoV-2 main protease: a docking and MD simulation study, *J Biomol Struct Dyn* (n.d.) 1–12. <https://doi.org/10.1080/07391102.2023.2236714>.

[97] T.B. Kimber, Y. Chen, A. Volkamer, Deep learning in virtual screening: Recent applications and developments, *Int J Mol Sci* 22 (2021). <https://doi.org/10.3390/ijms22094435>.

[98] N. Berdigaliyev, M. Aljofan, An overview of drug discovery and development, *Future Med Chem* 12 (2020) 939–947. <https://doi.org/10.4155/fmc-2019-0307>.

[99] C.N. Cavasotto, J.I. Di Filippo, Artificial intelligence in the early stages of drug discovery, *Arch Biochem Biophys* 698 (2021) 108730. <https://doi.org/https://doi.org/10.1016/j.abb.2020.108730>.

- [100] T.O.A.E.-Y. Zhang, Types of Machine Learning Algorithms, in: IntechOpen, Rijeka, 2010: p. Ch. 3. <https://doi.org/10.5772/9385>.
- [101] M.A. Khamis, W. Gomaa, W.F. Ahmed, Machine learning in computational docking, *Artif Intell Med* 63 (2015) 135–152. <https://doi.org/https://doi.org/10.1016/j.artmed.2015.02.002>.
- [102] D.-L. Ma, D.S.-H. Chan, C.-H. Leung, Drug repositioning by structure-based virtual screening, *Chem Soc Rev* 42 (2013) 2130–2141. <https://doi.org/10.1039/C2CS35357A>.
- [103] T. Cheng, Q. Li, Z. Zhou, Y. Wang, S.H. Bryant, Structure-Based Virtual Screening for Drug Discovery: a Problem-Centric Review, *AAPS J* 14 (2012) 133–141. <https://doi.org/10.1208/s12248-012-9322-0>.
- [104] L. Zhang, H.-X. Ai, S.-M. Li, M.-Y. Qi, J. Zhao, Q. Zhao, H.-S. Liu, Virtual screening approach to identifying influenza virus neuraminidase inhibitors using molecular docking combined with machine-learning-based scoring function, *Oncotarget*; Vol 8, No 47 (2017).
- [105] L. Zhang, M. Qiao, H. Gao, B. Hu, H. Tan, X. Zhou, C.M. Li, Investigation of mechanism of bone regeneration in a porous biodegradable calcium phosphate (CaP) scaffold by a combination of a multi-scale agent-based model and experimental optimization/validation, *Nanoscale* 8 (2016) 14877–14887. <https://doi.org/10.1039/C6NR01637E>.
- [106] L. Zhang, S. Zhang, Using game theory to investigate the epigenetic control mechanisms of embryo development: Comment on: “Epigenetic game theory: How to compute the epigenetic control of maternal-to-zygotic transition” by Qian Wang et al., *Phys Life Rev* 20 (2017) 140–142. <https://doi.org/https://doi.org/10.1016/j.plrev.2017.01.007>.
- [107] J.S. Smith, O. Isayev, A.E. Roitberg, Data Descriptor: ANI-1, A data set of 20 million calculated off-equilibrium conformations for organic molecules, *Sci Data* 4 (2017). <https://doi.org/10.1038/sdata.2017.193>.
- [108] J.S. Smith, A.E. Roitberg, O. Isayev, Transforming Computational Drug Discovery with Machine Learning and AI, *ACS Med Chem Lett* 9 (2018) 1065–1069. <https://doi.org/10.1021/acsmedchemlett.8b00437>.
- [109] J.S. Smith, B. Nebgen, N. Lubbers, O. Isayev, A.E. Roitberg, Less is more: Sampling chemical space with active learning, *Journal of Chemical Physics* 148 (2018). <https://doi.org/10.1063/1.5023802>.
- [110] J.S. Smith, B.T. Nebgen, R. Zubatyuk, N. Lubbers, C. Devereux, K. Barros, S. Tretiak, O. Isayev, A. Roitberg, Outsmarting Quantum Chemistry Through Transfer Learning, 2018. <https://doi.org/10.26434/chemrxiv.6744440.v1>.

- [111] D. Kaul, An overview of coronaviruses including the SARS-2 coronavirus – Molecular biology, epidemiology and clinical implications, *Curr Med Res Pract* 10 (2020) 54–64. <https://doi.org/10.1016/j.cmrp.2020.04.001>.
- [112] World Health Organization. Weekly epidemiological update on COVID-19., (n.d.).
- [113] L. Torre-Fuentes, J. Matías-Guiu, L. Hernández-Lorenzo, P. Montero-Escribano, V. Pytel, J. Porta-Etessam, U. Gómez-Pinedo, J.A. Matías-Guiu, *ACE2, TMPRSS2*, and Furin variants and SARS-CoV-2 infection in Madrid, Spain, *J Med Virol* 93 (2021) 863–869. <https://doi.org/10.1002/jmv.26319>.
- [114] N. Vankadari, Structure of Furin Protease Binding to SARS-CoV-2 Spike Glycoprotein and Implications for Potential Targets and Virulence, *J Phys Chem Lett* 11 (2020) 6655–6663. <https://doi.org/10.1021/acs.jpcclett.0c01698>.
- [115] M.A. Marra, S.J.M. Jones, C.R. Astell, R.A. Holt, A. Brooks-Wilson, Y.S.N. Butterfield, J. Khattra, J.K. Asano, S.A. Barber, S.Y. Chan, A. Cloutier, S.M. Coughlin, D. Freeman, N. Girn, O.L. Griffith, S.R. Leach, M. Mayo, H. McDonald, S.B. Montgomery, P.K. Pandoh, A.S. Petrescu, A.G. Robertson, J.E. Schein, A. Siddiqui, D.E. Smailus, J.M. Stott, G.S. Yang, F. Plummer, A. Andonov, H. Artsob, N. Bastien, K. Bernard, T.F. Booth, D. Bowness, M. Czub, M. Drebot, L. Fernando, R. Flick, M. Garbutt, M. Gray, A. Grolla, S. Jones, H. Feldmann, A. Meyers, A. Kabani, Y. Li, S. Normand, U. Stroher, G.A. Tipples, S. Tyler, R. Vogrig, D. Ward, B. Watson, R.C. Brunham, M. Krajden, M. Petric, D.M. Skowronski, C. Upton, R.L. Roper, The Genome Sequence of the SARS-Associated Coronavirus, *Science* (1979) 300 (2003) 1399–1404. <https://doi.org/10.1126/science.1085953>.
- [116] D.R. Owen, C.M.N. Allerton, A.S. Anderson, L. Aschenbrenner, M. Avery, S. Berritt, B. Boras, R.D. Cardin, A. Carlo, K.J. Coffman, A. Dantonio, L. Di, H. Eng, R. Ferre, K.S. Gajiwala, S.A. Gibson, S.E. Greasley, B.L. Hurst, E.P. Kadar, A.S. Kalgutkar, J.C. Lee, J. Lee, W. Liu, S.W. Mason, S. Noell, J.J. Novak, R.S. Obach, K. Ogilvie, N.C. Patel, M. Pettersson, D.K. Rai, M.R. Reese, M.F. Sammons, J.G. Sathish, R.S.P. Singh, C.M. Stepan, A.E. Stewart, J.B. Tuttle, L. Updyke, P.R. Verhoest, L. Wei, Q. Yang, Y. Zhu, An oral SARS-CoV-2 M^{pro} inhibitor clinical candidate for the treatment of COVID-19, *Science* (1979) 374 (2021) 1586–1593. <https://doi.org/10.1126/science.abl4784>.
- [117] G.R. Painter, R.A. Bowen, G.R. Bluemling, J. DeBergh, V. Edpuganti, P.R. Gruddanti, D.B. Guthrie, M. Hager, D.L. Kuiper, M.A. Lockwood, D.G. Mitchell, M.G. Natchus, Z.M. Sticher, A.A. Kolykhalov, The prophylactic and therapeutic activity of a broadly active ribonucleoside analog in a murine model of intranasal venezuelan equine encephalitis virus infection, *Antiviral Res* 171 (2019) 104597. <https://doi.org/10.1016/j.antiviral.2019.104597>.
- [118] J. Boucau, R. Uddin, C. Marino, J. Regan, J.P. Flynn, M.C. Choudhary, G. Chen, A.M. Stuckwisch, J. Mathews, M.Y. Liew, A. Singh, Z. Reynolds, S.L. Iyer, G.C. Chamberlin, T.D. Vyas, J.M. Vyas, S.E. Turbett, J.Z. Li, J.E. Lemieux, A.K. Barczak, M.J. Siedner, Characterization of Virologic Rebound Following

Nirmatrelvir-Ritonavir Treatment for Coronavirus Disease 2019 (COVID-19), *Clinical Infectious Diseases* 76 (2023) e526–e529. <https://doi.org/10.1093/cid/ciac512>.

[119] Y. Hu, E.M. Lewandowski, H. Tan, X. Zhang, R.T. Morgan, X. Zhang, L.M.C. Jacobs, S.G. Butler, M. V Gongora, J. Choy, X. Deng, Y. Chen, J. Wang, Naturally occurring mutations of SARS-CoV-2 main protease confer drug resistance to nirmatrelvir., *BioRxiv* (2022). <https://doi.org/10.1101/2022.06.28.497978>.

[120] J. Boucau, R. Uddin, C. Marino, J. Regan, J.P. Flynn, M.C. Choudhary, G. Chen, A.M. Stuckwisch, J. Mathews, M.Y. Liew, A. Singh, Z. Reynolds, S.L. Iyer, G.C. Chamberlin, T.D. Vyas, J.M. Vyas, S.E. Turbett, J.Z. Li, J.E. Lemieux, A.K. Barczak, M.J. Siedner, Characterization of Virologic Rebound Following Nirmatrelvir-Ritonavir Treatment for Coronavirus Disease 2019 (COVID-19), *Clinical Infectious Diseases* 76 (2023) e526–e529. <https://doi.org/10.1093/cid/ciac512>.

[121] H. Mitsuya, K. Maeda, D. Das, A.K. Ghosh, Development of Protease Inhibitors and the Fight with Drug-Resistant HIV-1 Variants, in: 2008: pp. 169–197. [https://doi.org/10.1016/S1054-3589\(07\)56006-0](https://doi.org/10.1016/S1054-3589(07)56006-0).

[122] K. Palacio-Rodríguez, I. Lans, C.N. Cavasotto, P. Cossio, Exponential consensus ranking improves the outcome in docking and receptor ensemble docking, *Sci Rep* 9 (2019) 5142. <https://doi.org/10.1038/s41598-019-41594-3>.

[123] R. Hilgenfeld, From SARS to MERS: crystallographic studies on coronaviral proteases enable antiviral drug design, *FEBS Journal* 281 (2014) 4085–4096. <https://doi.org/10.1111/febs.12936>.

[124] M. Macchiagodena, M. Pagliai, P. Procacci, Identification of potential binders of the main protease 3CL^{pro} of the COVID-19 via structure-based ligand design and molecular modeling, *Chem Phys Lett* 750 (2020) 137489. <https://doi.org/10.1016/J.CPLETT.2020.137489>.

[125] C. Ma, Z. Xia, M.D. Sacco, Y. Hu, J.A. Townsend, X. Meng, J. Choza, H. Tan, J. Jang, M. V. Gongora, X. Zhang, F. Zhang, Y. Xiang, M.T. Marty, Y. Chen, J. Wang, Discovery of Di- and Trihaloacetamides as Covalent SARS-CoV-2 Main Protease Inhibitors with High Target Specificity, *J Am Chem Soc* 143 (2021) 20697–20709. <https://doi.org/10.1021/jacs.1c08060>.

[126] D.R. Owen, C.M.N. Allerton, A.S. Anderson, L. Aschenbrenner, M. Avery, S. Berritt, B. Boras, R.D. Cardin, A. Carlo, K.J. Coffman, A. Dantonio, L. Di, H. Eng, R. Ferre, K.S. Gajiwala, S.A. Gibson, S.E. Greasley, B.L. Hurst, E.P. Kadar, A.S. Kalgutkar, J.C. Lee, J. Lee, W. Liu, S.W. Mason, S. Noell, J.J. Novak, R.S. Obach, K. Ogilvie, N.C. Patel, M. Pettersson, D.K. Rai, M.R. Reese, M.F. Sammons, J.G. Sathish, R.S.P. Singh, C.M. Steppan, A.E. Stewart, J.B. Tuttle, L. Updyke, P.R. Verhoest, L. Wei, Q. Yang, Y. Zhu, An oral SARS-CoV-2 M^{pro} inhibitor clinical candidate for the treatment of COVID-19, *Science* (1979) 374 (2021) 1586–1593. <https://doi.org/10.1126/science.abl4784>.

- [127] N. Kitamura, M.D. Sacco, C. Ma, Y. Hu, J.A. Townsend, X. Meng, F. Zhang, X. Zhang, M. Ba, T. Szeto, A. Kukuljac, M.T. Marty, D. Schultz, S. Cherry, Y. Xiang, Y. Chen, J. Wang, Expedited Approach toward the Rational Design of Noncovalent SARS-CoV-2 Main Protease Inhibitors, *J Med Chem* 65 (2022) 2848–2865. <https://doi.org/10.1021/acs.jmedchem.1c00509>.
- [128] C. Ma, M.D. Sacco, B. Hurst, J.A. Townsend, Y. Hu, T. Szeto, X. Zhang, B. Tarbet, M.T. Marty, Y. Chen, J. Wang, Boceprevir, GC-376, and calpain inhibitors II, XII inhibit SARS-CoV-2 viral replication by targeting the viral main protease, *Cell Res* 30 (2020) 678–692. <https://doi.org/10.1038/s41422-020-0356-z>.
- [129] J. Qiao, Y.-S. Li, R. Zeng, F.-L. Liu, R.-H. Luo, C. Huang, Y.-F. Wang, J. Zhang, B. Quan, C. Shen, X. Mao, X. Liu, W. Sun, W. Yang, X. Ni, K. Wang, L. Xu, Z.-L. Duan, Q.-C. Zou, H.-L. Zhang, W. Qu, Y.-H.-P. Long, M.-H. Li, R.-C. Yang, X. Liu, J. You, Y. Zhou, R. Yao, W.-P. Li, J.-M. Liu, P. Chen, Y. Liu, G.-F. Lin, X. Yang, J. Zou, L. Li, Y. Hu, G.-W. Lu, W.-M. Li, Y.-Q. Wei, Y.-T. Zheng, J. Lei, S. Yang, SARS-CoV-2 M^{pro} inhibitors with antiviral activity in a transgenic mouse model, *Science* (1979) 371 (2021) 1374–1378. <https://doi.org/10.1126/science.abf1611>.
- [130] B. Boras, R.M. Jones, B.J. Anson, D. Arenson, L. Aschenbrenner, M.A. Bakowski, N. Beutler, J. Binder, E. Chen, H. Eng, H. Hammond, J. Hammond, R.E. Haupt, R. Hoffman, E.P. Kadar, R. Kania, E. Kimoto, M.G. Kirkpatrick, L. Lanyon, E.K. Lendy, J.R. Lillis, J. Logue, S.A. Luthra, C. Ma, S.W. Mason, M.E. McGrath, S. Noell, R.S. Obach, M.N. O’ Brien, R. O’Connor, K. Ogilvie, D. Owen, M. Pettersson, M.R. Reese, T.F. Rogers, R. Rosales, M.I. Rossulek, J.G. Sathish, N. Shirai, C. Steppan, M. Ticehurst, L.W. Updyke, S. Weston, Y. Zhu, K.M. White, A. García-Sastre, J. Wang, A.K. Chatterjee, A.D. Mesecar, M.B. Frieman, A.S. Anderson, C. Allerton, Preclinical characterization of an intravenous coronavirus 3CL protease inhibitor for the potential treatment of COVID19, *Nat Commun* 12 (2021) 6055. <https://doi.org/10.1038/s41467-021-26239-2>.
- [131] W. Dai, B. Zhang, X.-M. Jiang, H. Su, J. Li, Y. Zhao, X. Xie, Z. Jin, J. Peng, F. Liu, C. Li, Y. Li, F. Bai, H. Wang, X. Cheng, X. Cen, S. Hu, X. Yang, J. Wang, X. Liu, G. Xiao, H. Jiang, Z. Rao, L.-K. Zhang, Y. Xu, H. Yang, H. Liu, Structure-based design of antiviral drug candidates targeting the SARS-CoV-2 main protease, *Science* (1979) 368 (2020) 1331–1335. <https://doi.org/10.1126/science.abb4489>.
- [132] L. Zhang, D. Lin, X. Sun, U. Curth, C. Drosten, L. Sauerhering, S. Becker, K. Rox, R. Hilgenfeld, Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α -ketoamide inhibitors, *Science* (1979) 368 (2020) 409–412. <https://doi.org/10.1126/science.abb3405>.
- [133] Z. Jin, X. Du, Y. Xu, Y. Deng, M. Liu, Y. Zhao, B. Zhang, X. Li, L. Zhang, C. Peng, Y. Duan, J. Yu, L. Wang, K. Yang, F. Liu, R. Jiang, X. Yang, T. You, X. Liu, X. Yang, F. Bai, H. Liu, X. Liu, L.W. Guddat, W. Xu, G. Xiao, C. Qin, Z. Shi, H. Jiang, Z. Rao, H. Yang, Structure of M^{pro} from SARS-CoV-2 and

discovery of its inhibitors, *Nature* 582 (2020) 289–293. <https://doi.org/10.1038/s41586-020-2223-y>.

[134] B.-X. Quan, H. Shuai, A.-J. Xia, Y. Hou, R. Zeng, X.-L. Liu, G.-F. Lin, J.-X. Qiao, W.-P. Li, F.-L. Wang, K. Wang, R.-J. Zhou, T.T.-T. Yuen, M.-X. Chen, C. Yoon, M. Wu, S.-Y. Zhang, C. Huang, Y.-F. Wang, W. Yang, C. Tian, W.-M. Li, Y.-Q. Wei, K.-Y. Yuen, J.F.-W. Chan, J. Lei, H. Chu, S. Yang, An orally available Mpro inhibitor is effective against wild-type SARS-CoV-2 and variants including Omicron, *Nat Microbiol* 7 (2022) 716–725. <https://doi.org/10.1038/s41564-022-01119-7>.

[135] L. Subissi, I. Imbert, F. Ferron, A. Collet, B. Coutard, E. Decroly, B. Canard, SARS-CoV ORF1b-encoded nonstructural proteins 12–16: Replicative enzymes as antiviral targets, *Antiviral Res* 101 (2014) 122–130. <https://doi.org/10.1016/j.antiviral.2013.11.006>.

[136] B.-X. Quan, H. Shuai, A.-J. Xia, Y. Hou, R. Zeng, X.-L. Liu, G.-F. Lin, J.-X. Qiao, W.-P. Li, F.-L. Wang, K. Wang, R.-J. Zhou, T.T.-T. Yuen, M.-X. Chen, C. Yoon, M. Wu, S.-Y. Zhang, C. Huang, Y.-F. Wang, W. Yang, C. Tian, W.-M. Li, Y.-Q. Wei, K.-Y. Yuen, J.F.-W. Chan, J. Lei, H. Chu, S. Yang, An orally available Mpro inhibitor is effective against wild-type SARS-CoV-2 and variants including Omicron, *Nat Microbiol* 7 (2022) 716–725. <https://doi.org/10.1038/s41564-022-01119-7>.

[137] M. Macchiagodena, M. Pagliai, P. Procacci, Identification of potential binders of the main protease 3CLpro of the COVID-19 via structure-based ligand design and molecular modeling, *Chem Phys Lett* 750 (2020) 137489. <https://doi.org/10.1016/j.cplett.2020.137489>.

[138] M. Su, Q. Yang, Y. Du, G. Feng, Z. Liu, Y. Li, R. Wang, Comparative Assessment of Scoring Functions: The CASF-2016 Update, *J Chem Inf Model* 59 (2019) 895–913. <https://doi.org/10.1021/acs.jcim.8b00545>.

[139] R. Kumar, Effects of Stereoisomers on Drug Activity, *Am J Biomed Sci Res* 13 (2021) 220–222. <https://doi.org/10.34297/AJBSR.2021.13.001861>.

[140] A.A. Spasov, I.N. Iezhitsa, P.M. Vassiliev, A.A. Ozerov, R. Agarwal, Pharmacology of Drug Stereoisomers, 76 (2022). <https://doi.org/10.1007/978-981-19-2320-3>.

[141] Z. Liu, T. Zubatiuk, A. Roitberg, O. Isayev, Auto3D: Automatic Generation of the Low-Energy 3D Structures with ANI Neural Network Potentials, *J Chem Inf Model* 62 (2022) 5373–5382. <https://doi.org/10.1021/acs.jcim.2c00817>.

[142] T. Seidel, C. Permann, O. Wieder, S.M. Kohlbacher, T. Langer, High-Quality Conformer Generation with CONFORGE: Algorithm and Performance Assessment, *J Chem Inf Model* 63 (2023) 5549–5570. https://doi.org/10.1021/ACS.JCIM.3C00563/ASSET/IMAGES/LARGE/CI3C00563_0014.JPEG.

- [143] E. Perola, P.S. Charifson, Conformational Analysis of Drug-Like Molecules Bound to Proteins: An Extensive Study of Ligand Reorganization upon Binding, *J Med Chem* 47 (2004) 2499–2510. <https://doi.org/10.1021/JM030563W>.
- [144] N. Foloppe, R. Hubbard, Towards Predictive Ligand Design With Free-Energy Based Computational Methods?, *Curr Med Chem* 13 (2006) 3583–3608. <https://doi.org/10.2174/092986706779026165>.
- [145] W.L. Jorgensen, L.L. Thomas, Perspective on free-energy perturbation calculations for chemical equilibria, *J Chem Theory Comput* 4 (2008) 869–876. <https://doi.org/10.1021/ct800011m>.
- [146] H. Fu, C. Chipot, X. Shao, W. Cai, Standard Binding Free-Energy Calculations: How Far Are We from Automation?, *J Phys Chem B* 127 (2023) 10459–10468. <https://doi.org/10.1021/acs.jpcc.3c04370>.
- [147] N. Lu, D.A. Kofke, Accuracy of free-energy perturbation calculations in molecular simulation. II. Heuristics, *Journal of Chemical Physics* 115 (2001) 6866–6875. <https://doi.org/10.1063/1.1405449>.
- [148] E.A. Rifai, V. Ferrario, J. Pleiss, D.P. Geerke, Combined Linear Interaction Energy and Alchemical Solvation Free-Energy Approach for Protein-Binding Affinity Computation, *J Chem Theory Comput* 16 (2020) 1300–1310. <https://doi.org/10.1021/acs.jctc.9b00890>.
- [149] M. Badaoui, P.J. Buigues, D. Berta, G.M. Mandana, H. Gu, T. Földes, C.J. Dickson, V. Hornak, M. Kato, C. Molteni, S. Parsons, E. Rosta, Combined Free-Energy Calculation and Machine Learning Methods for Understanding Ligand Unbinding Kinetics, *J Chem Theory Comput* 18 (2022) 2543–2555. <https://doi.org/10.1021/acs.jctc.1c00924>.
- [150] G. Hummer, A. Szabo, Calculation of free-energy differences from computer simulations of initial and final states, *Journal of Chemical Physics* 105 (1996) 2004–2010. <https://doi.org/10.1063/1.472068>.
- [151] I. Geronimo, M. De Vivo, Alchemical Free-Energy Calculations of Watson-Crick and Hoogsteen Base Pairing Interconversion in DNA, *J Chem Theory Comput* 18 (2022) 6966–6973. <https://doi.org/10.1021/acs.jctc.2c00848>.
- [152] P. Mikulskis, S. Genheden, U. Ryde, A large-scale test of free-energy simulation estimates of protein-Ligand binding affinities, *J Chem Inf Model* 54 (2014) 2794–2806. <https://doi.org/10.1021/ci5004027>.
- [153] P. Procacci, Alchemical determination of drug-receptor binding free energy: Where we stand and where we could move to, *J Mol Graph Model* 71 (2017) 233–241. <https://doi.org/10.1016/j.jmgm.2016.11.018>.

- [154] L. Wang, Y. Wu, Y. Deng, B. Kim, L. Pierce, G. Krilov, D. Lupyan, S. Robinson, M.K. Dahlgren, J. Greenwood, D.L. Romero, C. Masse, J.L. Knight, T. Steinbrecher, T. Beuming, W. Damm, E. Harder, W. Sherman, M. Brewer, R. Wester, M. Murcko, L. Frye, R. Farid, T. Lin, D.L. Mobley, W.L. Jorgensen, B.J. Berne, R.A. Friesner, R. Abel, Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field, *J Am Chem Soc* 137 (2015) 2695–2703. <https://doi.org/10.1021/ja512751q>.
- [155] S. Wan, A.P. Bhati, S.J. Zasada, P. V. Coveney, Rapid, accurate, precise and reproducible ligand-protein binding free energy prediction: Binding free energy prediction, *Interface Focus* 10 (2020). <https://doi.org/10.1098/rsfs.2020.0007>.
- [156] B.O. Brandsdal, F. Österberg, M. Almlöf, I. Feierberg, V.B. Luzhkov, J. Åqvist, Free energy calculations and ligand binding, *Adv Protein Chem* 66 (2003) 123–158. [https://doi.org/10.1016/S0065-3233\(03\)66004-3](https://doi.org/10.1016/S0065-3233(03)66004-3).
- [157] J.M.J. Swanson, R.H. Henchman, J.A. McCammon, Revisiting Free Energy Calculations: A Theoretical Connection to MM/PBSA and Direct Calculation of the Association Free Energy, *Biophys J* 86 (2004) 67–74. [https://doi.org/10.1016/S0006-3495\(04\)74084-9](https://doi.org/10.1016/S0006-3495(04)74084-9).
- [158] J.D. Chodera, D.L. Mobley, M.R. Shirts, Alchemical free energy calculations with gromacs Many questions in biology and pharmaceutical chemistry involve free energy differences, *Building* (2008).
- [159] E.A. Rifai, M. van Dijk, D.P. Geerke, Recent Developments in Linear Interaction Energy Based Binding Free Energy Calculations, *Front Mol Biosci* 7 (2020). <https://doi.org/10.3389/fmolb.2020.00114>.
- [160] O. Nicolotti, M. Convertino, F. Leonetti, M. Catto, S. Cellamare, A. Carotti, Estimation of the Binding Free Energy by Linear Interaction Energy Models, *Mini-Reviews in Medicinal Chemistry* 12 (2012) 551–561. <https://doi.org/10.2174/138955712800493843>.
- [161] J. Kalayan, A. Chakravorty, J. Warwicker, R.H. Henchman, Total free energy analysis of fully hydrated proteins, *Proteins: Structure, Function and Bioinformatics* 91 (2023) 74–90. <https://doi.org/10.1002/prot.26411>.
- [162] F.M. Ytreberg, R.H. Swendsen, D.M. Zuckerman, Comparison of free energy methods for molecular systems, *Journal of Chemical Physics* 125 (2006) 184114. <https://doi.org/10.1063/1.2378907>.
- [163] D. Branduardi, F.L. Gervasio, M. Parrinello, From A to B in free energy space, *Journal of Chemical Physics* 126 (2007) 54103. <https://doi.org/10.1063/1.2432340>.

- [164] S. Boresch, H.L. Woodcock, Convergence of single-step free energy perturbation, *Mol Phys* 115 (2017) 1200–1213. <https://doi.org/10.1080/00268976.2016.1269960>.
- [165] H.H. Loeffler, J. Michel, C. Woods, FESetup: Automating Setup for Alchemical Free Energy Simulations, *J Chem Inf Model* 55 (2015) 2485–2490. <https://doi.org/10.1021/acs.jcim.5b00368>.
- [166] D. Wu, X. Zheng, R. Liu, Z. Li, Z. Jiang, Q. Zhou, Y. Huang, X.N. Wu, C. Zhang, Y.Y. Huang, H. Bin Luo, Free energy perturbation (FEP)-guided scaffold hopping, *Acta Pharm Sin B* 12 (2022) 1351–1362. <https://doi.org/10.1016/j.apsb.2021.09.027>.
- [167] T.B. Steinbrecher, M. Dahlgren, D. Cappel, T. Lin, L. Wang, G. Krilov, R. Abel, R. Friesner, W. Sherman, Accurate Binding Free Energy Predictions in Fragment Optimization, *J Chem Inf Model* 55 (2015) 2411–2420. <https://doi.org/10.1021/acs.jcim.5b00538>.
- [168] E. King, E. Aitchison, H. Li, R. Luo, Recent Developments in Free Energy Calculations for Drug Discovery, *Front Mol Biosci* 8 (2021). <https://doi.org/10.3389/fmolb.2021.712085>.
- [169] L.F. Song, K.M. Merz, Evolution of alchemical free energy methods in drug discovery, *J Chem Inf Model* 60 (2020) 5308–5318. <https://doi.org/10.1021/acs.jcim.0c00547>.
- [170] V. Limongelli, Ligand binding free energy and kinetics calculation in 2020, *Wiley Interdiscip Rev Comput Mol Sci* 10 (2020). <https://doi.org/10.1002/wcms.1455>.
- [171] G. König, B.R. Brooks, W. Thiel, D.M. York, On the convergence of multi-scale free energy simulations, *Mol Simul* 44 (2018) 1062–1081. <https://doi.org/10.1080/08927022.2018.1475741>.
- [172] E.A. Rifai, M. Van Dijk, N.P.E. Vermeulen, A. Yanuar, D.P. Geerke, A Comparative Linear Interaction Energy and MM/PBSA Study on SIRT1-Ligand Binding Free Energy Calculation, *J Chem Inf Model* 59 (2019) 4018–4033. <https://doi.org/10.1021/acs.jcim.9b00609>.
- [173] H.H. Loeffler, S. Bosisio, G. Duarte Ramos Matos, D. Suh, B. Roux, D.L. Mobley, J. Michel, Reproducibility of Free Energy Calculations across Different Molecular Simulation Software Packages, *J Chem Theory Comput* 14 (2018) 5567–5582. <https://doi.org/10.1021/acs.jctc.8b00544>.
- [174] W.M. Menzer, B. Xie, D.D.L. Minh, On Restraints in End-Point Protein–Ligand Binding Free Energy Calculations, *J Comput Chem* 41 (2020) 573–586. <https://doi.org/10.1002/jcc.26119>.

- [175] M. Aldeghi, J.P. Bluck, P.C. Biggin, Absolute alchemical free energy calculations for ligand binding: A beginner's guide, Humana Press Inc., 2018. https://doi.org/10.1007/978-1-4939-7756-7_11.
- [176] G. Poli, T. Seidel, T. Langer, Conformational sampling of small molecules with iCon: Performance assessment in comparison with OMEGA, *Front Chem* 6 (2018) 349617. <https://doi.org/10.3389/FCHEM.2018.00229/BIBTEX>.
- [177] P.C.D. Hawkins, A.G. Skillman, G.L. Warren, B.A. Ellingson, M.T. Stahl, Conformer generation with OMEGA: Algorithm and validation using high quality structures from the protein databank and cambridge structural database, *J Chem Inf Model* 50 (2010) 572–584. https://doi.org/10.1021/CI100031X/SUPPL_FILE/CI100031X_SI_004.TXT.
- [178] K.S. Watts, P. Dalal, R.B. Murphy, W. Sherman, R.A. Friesner, J.C. Shelley, ConfGen: A conformational search method for efficient generation of bioactive conformers, *J Chem Inf Model* 50 (2010) 534–546. https://doi.org/10.1021/CI100015J/SUPPL_FILE/CI100015J_SI_001.PDF.
- [179] I.J. Chen, N. Foloppe, Conformational Sampling of Druglike Molecules with MOE and Catalyst: Implications for Pharmacophore Modeling and Virtual Screening, *J. Chem. Inf. Model.* 48 (2008) 1773.
- [180] P. Pracht, F. Bohle, S. Grimme, Automated exploration of the low-energy chemical space with fast quantum chemical methods, *Physical Chemistry Chemical Physics* 22 (2020) 7169–7192. <https://doi.org/10.1039/C9CP06869D>.
- [181] T.B. Leite, D. Gomes, M.A. Miteva, J. Chomilier, B.O. Villoutreix, P. Tuffery, Frog: A Free Online Drug 3D Conformation Generator, *Nucleic Acids Res.* 35 (2007) W568.
- [182] M.A. Miteva, F. Guyon, P. Tuffery, Frog2: Efficient 3D Conformation Ensemble Generator for Small Compounds, *Nucleic Acids Res.* 38 (2010) W622.
- [183] N.M. O'Boyle, T. Vandermeersch, C.J. Flynn, A.R. Maguire, G.R. Hutchison, Confab - Systematic Generation of Diverse Low-Energy Conformers, *J. Cheminf.* 3 (2011) 8.
- [184] N.M. O'Boyle, M. Banck, C.A. James, C. Morley, T. Vandermeersch, G.R. Hutchison, Open Babel: An Open chemical toolbox, *J Cheminform* 3 (2011) 1–14. <https://doi.org/10.1186/1758-2946-3-33/TABLES/2>.
- [185] S. Riniker, G. A. Landrum, Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation, *J Chem Inf Model* 55 (2015) 2562–2574. <https://doi.org/10.1021/acs.jcim.5b00654>.
- [186] N.O. Friedrich, C. De Bruyn Kops, F. Flachsenberg, K. Sommer, M. Rarey, J. Kirchmair, Benchmarking Commercial Conformer Ensemble Generators, *J Chem Inf Model* 57 (2017) 2719–2728.

https://doi.org/10.1021/ACS.JCIM.7B00505/SUPPL_FILE/CI7B00505_SI_003.TXT.

[187] J.S. Smith, R. Zubatyuk, B. Nebgen, N. Lubbers, K. Barros, A.E. Roitberg, O. Isayev, S. Tretiak, The ANI-1ccx and ANI-1x data sets, coupled-cluster and density functional theory properties for molecules, *Sci Data* 7 (2020). <https://doi.org/10.1038/s41597-020-0473-z>.

[188] J.S. Smith, O. Isayev, A.E. Roitberg, Data Descriptor: ANI-1, A data set of 20 million calculated off-equilibrium conformations for organic molecules, *Sci Data* 4 (2017). <https://doi.org/10.1038/sdata.2017.193>.

[189] X. Gao, F. Ramezanghorbani, O. Isayev, J.S. Smith, A.E. Roitberg, TorchANI: A Free and Open Source PyTorch-Based Deep Learning Implementation of the ANI Neural Network Potentials, *J. Chem. Inf. Model.* 60 (2020) 3408.

[190] J.S. Smith, B. Nebgen, N. Lubbers, O. Isayev, A.E. Roitberg, Less is more: Sampling chemical space with active learning, *Journal of Chemical Physics* 148 (2018). <https://doi.org/10.1063/1.5023802>.

[191] D. Hao, X. He, A.E. Roitberg, S. Zhang, J. Wang, A. E. Roitberg, S. Zhang, J. Wang, Development and Evaluation of Geometry Optimization Algorithms in Conjunction with ANI Potentials, *J Chem Theory Comput* 18 (2022) 978–991. <https://doi.org/10.1021/acs.jctc.1c01043>.

[192] Z. Liu, T. Zubatyuk, A. Roitberg, O. Isayev, Auto3D: Automatic Generation of the Low-Energy 3D Structures with ANI Neural Network Potentials, *J Chem Inf Model* 62 (2022) 5373–5382. <https://doi.org/10.1021/acs.jcim.2c00817>.

[193] C. Devereux, J.S. Smith, K.K. Davis, K. Barros, R. Zubatyuk, O. Isayev, A.E. Roitberg, Extending the Applicability of the ANI Deep Learning Molecular Potential to Sulfur and Halogens, *J Chem Theory Comput* 16 (2020) 4192–4202.
https://doi.org/10.1021/ACS.JCTC.0C00121/ASSET/IMAGES/MEDIUM/CT0C00121_M009.GIF.

[194] N.O. Friedrich, C. De Bruyn Kops, F. Flachsenberg, K. Sommer, M. Rarey, J. Kirchmair, Benchmarking Commercial Conformer Ensemble Generators, *J Chem Inf Model* 57 (2017) 2719–2728.
https://doi.org/10.1021/ACS.JCIM.7B00505/SUPPL_FILE/CI7B00505_SI_003.TXT.

[195] N.O. Friedrich, A. Meyder, C. De Bruyn Kops, K. Sommer, F. Flachsenberg, M. Rarey, J. Kirchmair, High-Quality Dataset of Protein-Bound Ligand Conformations and Its Application to Benchmarking Conformer Ensemble Generators, *J Chem Inf Model* 57 (2017) 529–539.
https://doi.org/10.1021/ACS.JCIM.6B00613/SUPPL_FILE/CI6B00613_SI_003.TXT.

BIOGRAPHY

İrem Nur ZENGİN completed her undergraduate education in Molecular Biology and Genetics at Gebze Technical University. She is currently pursuing her M.Sc. in the Department of Chemistry at the same institution. Her research focuses on computational chemistry, molecular modeling, and drug design. During her graduate studies, she has worked on analyzing ligand–protein interactions using free energy calculations and AI-assisted methods.

PUBLICATIONS AND PRESENTATIONS FROM THE THESIS

Zengin, İ. N., Koca, M. S., Tayfuroglu, O., Yildiz, M., & Kocak, A. (2024). Benchmarking ANI potentials as a rescoring function and screening FDA drugs for SARS-CoV-2 Mpro. *Journal of Computer-Aided Molecular Design*, 38, 15. <https://doi.org/10.1007/s10822-024-00554-4>

Tayfuroglu, O., Zengin, I. N., Koca, M. S., & Kocak, A. (2025). DeepConf: Leveraging ANI-ML Potentials for Exploring Local Minima with Application to Bioactive Conformations. *Journal of Chemical Information and Modeling*, 65(6), 2818–2833. <https://doi.org/10.1021/acs.jcim.4c0205>

APPENDICES

APPENDIX-A: SELECTED 439 COMPOUNDS FOR BIOACTIVE CONFORMATION USING THE CONFORMER ALGORITHMS. THREE-DIGIT PDB CODES OF THE LIGANDS ARE FOLLOWED BY “M”. BOTH MODEL AND IDEAL STRUCTURES WERE OPTIMIZED AT WB97X/6-31G(D) LEVEL.

QM optimized absolute energy (eV)					model to ideal				
PDB id (M###)	model	ideal	$\Delta E(\text{kcal})$	RMSD ($\text{\AA}$)	PDB id (M###)	model	ideal	$\Delta E(\text{kcal})$	RMSD ($\text{\AA}$)
M048	-35013.06119	-35013.0377	-0.5416	2.35	MH DU	-	-	-8.4523	2.29
M05B	-36738.45331	-36738.24223	-4.8675	3.11	MH FY	24221.40445	24221.03792	-7.092	2.27
M06V	-33751.56273	-33751.83343	6.24261	3.25	MH GF	47580.09495	47579.78741	-0.39136	2.77
M0BZ	-30604.61273	-30604.60568	-0.1627	2.05	MH M2	-	-	-0.4753	2.24
M0C5	-30414.31157	-30414.29825	-0.3071	2.08	MH NY	37419.27366	37419.25305	5.36145	2.11
M0C6	-43110.31197	-43110.30151	-0.2412	2.04	MH RV	-	-	0.59044	2.68
M0C7	-42274.12343	-42274.11433	-0.2099	2.32	MH WZ	64850.06147	64850.08707	-3.0538	2.3
M0FT	-33481.21534	-33480.93328	-6.5045	2.39	MH Z8	-	-	-1.1928	2
M0KQ	-39465.13357	-39465.15389	0.46854	2.35	MH ZW	26550.62244	26550.57071	-0.05173	2.19
M0KR	-30100.47237	-30100.47458	0.05092	2.19	MH ZW	-	-43279.9272	2.1391	1.96
M0LC	-37020.99388	-37020.93004	-1.4723	2.39	MH ZW	43279.83443	-	-0.1327	2.23
M0VG	-32104.53473	-32104.12248	-9.5066	2.07	MH ZW	38907.30797	38907.30222	-0.00575	2.38
M0XF	-31543.90412	-31543.25533	-14.961	2.1	MH ZW	-	-	-7.4043	2.38
M0YO	-33589.85976	-33589.77373	-1.9838	2.03	MH ZW	32428.82236	32428.50128	-1.0004	2.03
M15V	-41341.64028	-41341.30359	-7.7641	2.21	MH ZW	35375.92911	35375.88573	-0.04338	2.12
M19R	-25929.10389	-25928.58968	-11.858	2.44	MH ZW	-	-35592.955	0.04937	2.12
M1A2	-38545.76528	-38545.79904	0.77845	2.1	MH ZW	35592.95286	-	-1.9844	2.19
M1C7	-35254.3571	-35254.16346	-4.4655	2.41	MH ZW	32072.30519	32072.21914	-0.08605	2.22
M1GB	-31653.95439	-31653.78059	-4.0079	2.65	MH ZW	-	-	-3.3737	2.22
M1IF	-36662.58613	-36662.59519	0.20893	2.08	MH ZW	36219.38248	36219.23618	-0.0134	3.22
M1LC	-28940.73672	-28940.77505	0.884	2.18	MH ZW	39262.82273	39262.82215	-0.00058	2.12
M1RC	-42626.42912	-42626.90302	10.9283	2.31	MH ZW	-	-	-2.7981	2.12
M21G	-38019.16584	-38019.12035	-1.049	2.23	MH ZW	40113.06732	40112.94598	0.05621	2.19
M2KW	-52061.52075	-52061.45376	-1.5448	2.31	MH ZW	26958.29204	26958.29448	-0.00244	2.25
M2M3	-30135.33614	-30135.31793	-0.42	1.95	MH ZW	-36343.7935	-	0.34348	2.25
					MH ZW	-	36343.80839	-2.1093	2.93
					MH ZW	-32135.313	-	-0.6134	2.44
					MH ZW	-	32135.22153	-0.6134	2.44
					MH ZW	26157.63588	26157.60929	0.36186	2.01
					MH ZW	-	-	-0.1748	2
					MH ZW	36759.29635	36759.31204	2.58292	2.35
					MH ZW	-	-	-1.6941	2.85
					MH ZW	38774.00779	38774.00021	-0.00758	2.35
					MH ZW	-	-	-1.6941	2.85
					MH ZW	30903.84466	30903.95667	-0.11201	2.85
					MH ZW	-	-	-1.6941	2.85
					MH ZW	26452.38004	26452.30657	-0.07347	2.85

APPENDIX-A: CONTINUED

M2PU	-33088.54726	-33088.47764	-1.6055	2.13	MJRE	-	-	-10.606	2.17
M2WB	-30592.4351	-30592.6676	5.36162	2.4		39439.41648	39438.95656		
M31O	-47770.27585	-47770.09948	-4.0672	3.27	MJUL	-	-	-2.3817	2.19
M31T	-35243.70269	-35243.58002	-2.8288	2.36		34698.10642	34698.00314		
M33M	-27859.92381	-27859.92427	0.01054	2.14	MJW4	-	-	0.74752	2.6
M342	-33790.71739	-33790.52835	-4.3593	3.27		34281.03501	34281.06742		
M36Z	-22823.97163	-22823.97351	0.04344	2.11	MJX4	-	-34619.037	-5.9548	2.85
M38W	-33113.85696	-33113.74081	-2.6785	2.95		34619.29523			
M3BK	-60085.89498	-60085.59087	-7.0127	2.67	MJXV	-	-	-1.6211	2.16
M3N5	-27545.71588	-27545.64852	-1.5534	2.25		31652.31368	31652.24338		
M3NV	-44511.76699	-44511.52885	-5.4916	2.1	MK3O	-	-	-0.5951	2.05
M3RF	-31471.04314	-31470.9508	-2.1295	2.17		31377.41235	31377.38655		
M3RJ	-35050.86227	-35050.81037	-1.1969	2.59	MK69	-	-30982.3027	-1.4672	2.04
M3RQ	-48762.38217	-48762.40973	0.63565	2.32		30982.36633			
M3RR	-49832.19313	-49832.22225	0.6715	2.27	MKB4	-	-	0.42514	2.06
M40J	-66688.75623	-66688.46798	-6.6471	2.6		40846.35105	40846.36949		
M472	-30506.31899	-30506.3195	0.01156	2.38	MKE2	-	-	-2.3219	2.69
M4CD	-39106.34565	-39106.91043	13.024	1.91		31190.46941	31190.36872		
M4EP	-25968.38726	-25968.20198	-4.2726	2.41	MKE5	-	-	-2.3575	2.48
M4H1	-22879.35831	-22879.3597	0.03201	1.99		31190.44251	31190.34028		
M4HN	-32682.67499	-32682.935	5.99593	2.25	MKEB	-	-32260.0113	-3.082	2.47
M4HW	-30838.89469	-30838.79923	-2.2013	2.66		32260.14495			
M4XO	-33948.40913	-33948.35305	-1.2932	3.69	MKEJ	-	-	-10.611	2.75
M4ZJ	-36349.93256	-36349.63331	-6.9008	2.31		44457.26297	44456.80285		
M52R	-36685.35867	-36685.15691	-4.6528	2.77	MKEQ	-	-	-3.0641	2.52
M52V	-33310.10332	-33310.08191	-0.4936	3.13		31190.09198	31189.95911		
M53K	-31678.37567	-31678.36231	-0.3082	2.28	MKEV	-	-	-11.241	2.5
M53R	-31170.91093	-31170.93082	0.45883	3.1		45526.81972	45526.33228		
M53S	-32240.54676	-32240.56706	0.46802	3.19	MKEW	-	-42626.6291	-2.6805	2.54
M53V	-33310.09237	-33310.08099	-0.2624	3.11		42626.74534			
M558	-45646.98088	-45646.87582	-2.4227	2.23	MKEY	-	-	-1.5266	2.84
M565	-33625.16516	-33624.98245	-4.2133	2.43		44457.42145	44457.35526		
M58C	-33655.43249	-33655.284	-3.4243	2.59	MKEZ	-	-	-2.8507	2.74
M58D	-28073.91378	-28073.95117	0.86226	2.03		32192.31005	32192.18644		
					MKF1	-35061.5494	-	-6.84	2.3
							35061.25279		
					MKF4	-	-	-6.9557	2.3
						34897.18822	34896.88659		
					MKFO	-	-	0.50754	2.58
						41844.69733	41844.71934		
					MKHM	-	-	1.32415	2
						35205.89965	35205.95707		
					MKHU	-	-36287.847	4.40834	2.19
						36287.65584			
					MKID	-	-	0.65496	2.12
						55648.55264	55648.58104		
					MKQE	-	-	-1.2221	2.15
						34993.17821	34993.12522		
					MKR8	-26788.6782	-	-1.6872	2.1
							26788.60503		
					MKRF	-	-	-4.7999	1.95
						36050.94639	36050.73824		
					MKRP	-	-	0.96751	2.17
						30649.91811	30649.96007		
					MKRW	-	-	0.0032	1.98
						33919.72263	33919.72277		
					MKU8	-	-	0.0965	2.29
						34881.62432	34881.62851		
					MKW2	-36600.7364	-	4.67568	2.71
							36600.93916		
					MKWN	-	-	4.53973	2.28
						40315.14291	40315.33977		
					ML9G	-	-	2.98742	2.19
						36538.64985	36538.77939		
					MLG2	-	-30938.7601	4.17818	2.6
						30938.57892			
					MLII	-	-	1.44091	2.44
						30511.34303	30511.40551		

APPENDIX-A: CONTINUED

M58G	-24450.4612	-24450.48078	0.45144	2.03	MM06	-	-45024.4369	3.57602	2.57
M59I	-32481.24151	-32481.22591	-0.3596	2.3		45024.28183			
M59L	-31003.12921	-31002.91178	-5.0139	2.68	MM3Z	-	-	-3.9409	2.76
M59Y	-26529.41624	-26529.50255	1.99042	2.08		22284.25072	22284.07983		
M5BL	-32214.69351	-32214.78408	2.08876	2.05	MM9T	-	-	-0.202	2.94
M5CG	-38719.4719	-38719.31824	-3.5433	2.01		34731.62079	34731.61203		
M5L7	-28601.51263	-28601.6391	2.91666	2.28	MMBW	-	-	-0.8506	2.35
M5LY	-55626.83672	-55626.65038	-4.2971	2.41		34998.14944	34998.11256		
M5LZ	-55626.83385	-55626.67261	-3.7182	2.29	MMVK	-	-	0.89624	2.22
M5O7	-33020.93881	-33020.83269	-2.447	2.72		31342.54507	31342.58394		
M5OQ	-32147.71916	-32147.7128	-0.1467	2.48	MMVW	-	-	0.8919	2.34
M5PZ	-55747.04992	-55747.29234	5.59027	3		26212.24011	26212.27879		
M5QQ	-37997.75156	-37997.74465	-0.1592	2.86	MMWR	-39040.0951	-	0.09503	2.18
M5UA	-29296.99815	-29297.16691	3.89159	2.13			39040.09922		
M62G	-42078.58407	-42078.43944	-3.3354	2.5	MN5Y	-	-	-0.5627	2.88
M639	-34549.22231	-34549.13109	-2.1036	3.91		32090.74949	32090.72509		
M64A	-31045.18141	-31044.96913	-4.8953	2.39	MN8M	-	-	-1.3338	2.37
M658	-35749.15163	-35749.28222	3.01143	2.38		28880.93702	28880.87918		
M69E	-34150.70154	-34150.75662	1.27014	2.09	MNI7	-	-	3.13666	2.2
M6A7	-39465.86907	-39465.83699	-0.7398	2.24		33731.78592	33731.92194		
M6DM	-35309.76091	-35309.66238	-2.272	2.48	MNN8	-	-	-2.114	2.03
M6IT	-49435.62916	-49435.58126	-1.1045	1.97		37414.90216	37414.81049		
M6XP	-42085.31432	-42085.14797	-3.836	2.76	MNPF	-28496.6916	-	13.5362	2.14
M6YE	-34397.16747	-34396.8769	-6.7006	2.63			28497.27859		
M704	-43071.72312	-43071.75983	0.84653	2.84	MNQQ	-	-	2.25095	2.09
M73Z	-35965.07112	-35965.03544	-0.8229	2.3		44087.94278	44088.04039		
M74E	-35997.76729	-35997.732	-0.8137	2.24	MO5M	-	-	-4.1634	2.92
M74F	-36997.83531	-36997.84869	0.30848	2.72		34881.29265	34881.11211		
M756	-31713.52676	-31713.44132	-1.9701	2.72	MO5P	-	-	-5.6394	2.95
M75Q	-57285.40329	-57285.39139	-0.2746	2.07		32181.46303	32181.21849		
M75T	-30331.53042	-30331.56835	0.87473	2.29	MO8M	-	-	-8.307	2.57
						33470.66419	33470.30396		
					MO9U	-	-	2.37357	2.37
						37026.89955	37027.00248		
					MOA5	-	-	0.67109	2.15
						23157.61857	23157.64767		
					MOAY	-	-	6.48903	2.32
						37343.59852	37343.87991		
					MOGO	-	-	-0.1156	2.19
						32192.58353	32192.57852		
					MOSI	-	-40606.7795	5.68821	2.3
						40606.53284			
					MOVX	-	-	0.79631	2.14
						42079.89471	42079.92924		
					MOZ8	-	-	0.32844	3.33
						29704.59162	29704.60587		
					MP07	-	-	0.37096	2.86
						29964.56674	29964.58283		
					MP0W	-	-	3.12859	2.36
						28406.27506	28406.41073		
					MP0X	-	-	0.7599	2.51
						44126.94449	44126.97745		
					MPDZ	-	-40129.2188	-1.4164	2.03
						40129.28022			
					MPFN	-	-	1.13762	2.05
						21918.90094	21918.95027		
					MPM1	-	-	-0.9787	2.99
						35976.92406	35976.88162		
					MPQ6	-	-	-1.0819	2.42
						42112.44897	42112.40206		
					MPQI	-	-	6.81063	2.41
						46825.93573	46826.23106		

APPENDIX-A: CONTINUED

M76N	-38086.01886	-38086.60946	13.6195	2.1	MPU6	-	-	-0.2858	2.25
M778	-45150.44179	-45150.39172	-1.1545	2.4		29475.89201	29475.87962		
M77V	-28868.29547	-28868.31617	0.47736	2.23	MPUU	-	-	-3.562	2.73
M7BF	-30974.81198	-30974.70481	-2.4714	2.76		43149.86847	43149.71401		
M7H2	-49036.76619	-49036.75392	-0.2829	2.9	MPUX	-	-32176.0001	-0.28	2.43
M7JV	-25934.31514	-25934.22281	-2.1292	2.14		32176.01224			
M7KW	-27389.92781	-27389.84066	-2.0097	2.83	MPV5	-	-	-2.9675	2.1
M7LL	-28007.0872	-28007.19249	2.42801	2.35		37220.47015	37220.34147		
M7OG	-32645.87802	-32645.7833	-2.1844	2.24	MPY1	-	-	3.17317	2.49
M7OP	-36565.79028	-36565.85024	1.38284	2.86		23777.43097	23777.56857		
M7OY	-49071.59215	-49071.61466	0.51925	2.92	MPZ4	-23204.3463	-	-0.3326	2.01
M7U2	-33969.56248	-33969.50726	-1.2735	2.48			23204.33188		
M7W7	-23385.80456	-23385.76541	-0.9027	2	MPZK	-27383.0168	-	-0.3134	2.25
M82G	-32523.44616	-32523.40579	-0.931	2.56			27383.00321		
M84N	-21814.63025	-21814.61548	-0.3407	2.03	MQOI	-	-	2.24682	2.28
M86E	-54577.98128	-54577.67551	-7.0513	2.45		44759.34482	44759.44225		
M890	-27003.52053	-27003.3979	-2.828	2.55	MQ22	-	-	-10.558	2.38
M89C	-27543.95899	-27543.85084	-2.494	2.52		34553.84942	34553.39157		
M8F0	-29764.20966	-29764.16608	-1.0051	2.1	MQ27	-	-30103.4594	0.0078	2.08
M8F7	-28088.73299	-28088.72097	-0.2772	2.15		30103.45906			
M8F9	-29763.91609	-29763.95221	0.83286	2.19	MQ4M	-	-	-2.6482	2.21
M8FX	-46089.01106	-46088.97574	-0.8146	2.13		43149.08703	43148.97219		
M8GR	-34827.33405	-34827.35525	0.48876	2.09	MQAQ	-	-	0.40716	2.23
M8GX	-35263.62594	-35263.68557	1.37503	2.02		32852.40873	32852.42639		
M8IP	-23843.35197	-23843.38549	0.77287	2.31	MQC2	-	-	2.43056	2.4
M8J7	-25579.25956	-25579.26416	0.10615	2.19		26337.74212	26337.84752		
M8LN	-45515.92005	-45515.69947	-5.0867	3.62	MQCR	-	-	-9.6783	2.65
M8NR	-32789.74315	-32789.7563	0.30338	2.15		37822.13558	37821.71589		
M8P3	-24429.76206	-24429.81665	1.25895	2.17	MQD4	-	-	-5.9965	3.11
M8PT	-35755.71287	-35756.05103	7.79794	2.34		43586.66507	43586.40503		
M8R4	-33890.44605	-33890.48783	0.96349	2.3	MQF9	-37893.7518	-	0.91986	2.2
M8X2	-35690.36048	-35690.15272	-4.7911	2.87			37893.79169		
M8X5	-35794.83883	-35794.64254	-4.5264	2.46	MQFX	-	-	0.00804	2.29
						38328.42834	38328.42868		
					MLZ	-29070.286	-	-0.4666	1.84
							29070.26576		
					MQML	-	-	0.23383	2.17
						43689.68991	43689.70005		
					MQOQ	-	-	-1.0843	2.83
						45863.35298	45863.30596		
					MQQ3	-	-	-0.0318	2.8
						30122.14939	30122.14801		
					MQQ9	-	-	-0.7291	2.2
						25368.36963	25368.33801		
					MQRO	-	-	0.41781	2.07
						30239.31136	30239.32948		
					MQTC	-	-	1.98993	2.85
						37523.62254	37523.70883		
					MQUQ	-	-	-1.3808	2.28
						43149.89121	43149.83133		
					MQVC	-	-	-0.0721	2.27
						25223.70043	25223.69731		
					MQW1	-	-	-1.6308	2.48
						43586.54921	43586.47849		
					MQXY	-	-	-0.054	2.09
						28127.53278	28127.53044		
					MR41	-	-	-2.5674	2.11
						41545.03625	41544.92492		
					MR7F	-	-44251.7894	1.47563	2.74
						44251.72541			
					MR8F	-	-	2.34504	2.02
						28279.73678	28279.83847		
					MRG4	-	-	-1.6343	2.07
						26855.19652	26855.12565		
					MRH8	-	-	-0.495	2.63
						36745.23855	36745.21708		

APPENDIX-A: CONTINUED

M8XR	-33918.37946	-33918.34646	-0.761	3.03	MRI3	-	-	1.40196	2.16
M8XX	-31778.25783	-31777.82775	-9.9177	2.73		34406.19367	34406.25447		
M8ZP	-34769.78562	-34769.81488	0.67487	1.97	MRIZ	-	-	12.351	2.32
M93F	-31902.8095	-31902.86797	1.34816	2.05		30670.68981	30671.22541		
M985	-30107.86465	-30107.85758	-0.1631	2.33	MRJ9	-	-30671.2254	15.641	2.37
M9IP	-26958.91549	-26958.94765	0.74145	2.31		30670.54714			
M9J2	-47770.32864	-47770.15084	-4.1003	3.05	MRKZ	-	-	0.92114	2.29
M9L0	-37253.59253	-37253.41374	-4.1231	2.26		33115.18472	33115.22467		
M9LX	-33342.03193	-33342.02908	-0.0657	2.15	MRLR	-	-	5.90789	2.75
M9S5	-33120.61644	-33120.60584	-0.2445	2.67		46761.61224	46761.86844		
M9TO	-30941.77582	-30941.71623	-1.3743	2.82	MROD	-	-	-0.0705	2.18
M9WI	-33511.58689	-33511.60096	0.32453	4.22		34009.55821	34009.55515		
M9XT	-29304.83385	-29304.7776	-1.2973	2.34	MRQB	-	-	-1.006	2.16
MA6Y	-29666.15632	-29666.19131	0.80701	2.46		26497.61051	26497.56689		
MA82	-37345.02668	-37345.13001	2.38287	2.33	MRTM	-	-	-3.1403	2.71
MABZ	-42875.91583	-42875.9225	0.15375	2.08		26623.04133	26622.90515		
MAIE	-21345.77452	-21345.85725	1.90781	2.48	MS4O	-	-28390.0509	-0.5213	2.18
MAJL	-31248.85399	-31248.78031	-1.6992	2.36		28390.07351			
MAQI	-30089.11878	-30089.28299	3.7866	2.37	MS9H	-	-	-2.9807	2.1
MAQV	-32287.50965	-32287.25436	-5.8872	2.12		32149.61918	32149.48992		
MB1I	-32300.37611	-32300.37545	-0.0152	2.32	MSDL	-	-	-3.0316	2.06
MB2D	-34520.36056	-34520.21894	-3.2657	2.2		33482.08054	33481.94908		
MB6A	-29099.17233	-29099.13994	-0.7471	2.1	MSGH	-	-	0.15632	2.13
MB7V	-44026.96882	-44026.83599	-3.0631	2.39		28964.62464	28964.63142		
MBA8	-44184.41892	-44184.86868	10.3718	2.84	MSOZ	-42429.5271	-	-2.7477	3.34
MBGE	-29534.37761	-29534.55769	4.15272	2.16			42429.40795		
MBI3	-34223.03559	-34222.97289	-1.446	2.65	MSUU	-	-	-1.6474	2.48
MBJM	-27664.13561	-27663.98361	-3.5052	2.17		26294.17515	26294.10371		
MBVI	-30603.76152	-30603.75384	-0.1771	2.11	MSWF	-28144.8678	-	-0.7738	2.29
MBX2	-35221.22219	-35221.22456	0.05466	2.37			28144.83425		
MBYB	-38454.97422	-38455.15183	4.09577	2.61	MSZV	-	-	7.73232	2.16
MC1W	-55616.07387	-55616.02007	-1.2407	2.08		37153.56595	37153.90126		
MC1Y	-24351.32795	-24351.33976	0.27252	2.64	MT8Z	-	-	-3.4807	0.91
						31503.33478	31503.18385		
					MT9I	-	-	-3.3275	1.82
						35674.31367	35674.16937		
					MTGM	-	-	8.15033	2.22
						36717.39561	36717.74905		
					MTL8	-	-	-3.9427	2.02
						24923.97819	24923.80722		
					MTLQ	-	-	0.00285	2.48
						25987.94015	25987.94028		
					MTPB	-	-	0.28025	3.42
						28492.89863	28492.91078		
					MTU7	-	-	0.51021	2.14
						48496.04757	48496.06969		
					MTUQ	-	-	-1.4136	2.19
						45188.80453	45188.74323		
					MTVT	-	-33351.2934	0.89812	2.18
						33351.25445			
					MU03	-	-	2.24058	2.07
						28177.27265	28177.36982		
					MU06	-	-	-1.2224	2.57
						36301.26202	36301.20901		
					MU1U	-	-	0.55674	1.99
						42608.29982	42608.32396		
					MU26	-	-	0.21578	1.91
						31204.88866	31204.89802		
					MU75	-	-	0.44877	2.01
						36301.22058	36301.24004		
					MU86	-	-	0.15346	2.02
						31113.82671	31113.83337		
					MU8I	-	-	0.30519	2.31
						31113.83784	31113.85107		
					MU8Y	-31113.9092	-31113.8411	-1.5704	2.08

APPENDIX-A: CONTINUED

MC3Z	-32141.68857	-32141.71402	0.58703	3.55	MUDV	-	-	0.1021	2.22
MC8H	-37309.90982	-37309.57921	-7.6241	2.67		45612.47155	45612.47598		
MCD9	-42064.65906	-42064.65824	-0.019	2.39	MUES	-	-34243.4504	-1.6733	2.12
MCF9	-32573.39211	-32573.74302	8.09205	2.46		34243.52297			
MCIW	-42064.57572	-42064.57989	0.09635	2.55	MUGI	-	-	-0.0477	2.27
MCL3	-30505.05303	-30505.07149	0.42567	2.33		38137.62179	38137.61972		
MCQF	-44609.83651	-44609.87913	0.98288	2.55	MUNQ	-	-	-0.0559	2.32
MCRH	-36489.06015	-36489.08852	0.65413	2.32		32122.47037	32122.46795		
MCT5	-32225.43308	-32225.54597	2.60331	2.62	MUTV	-	-	-1.5215	2.33
MCX9	-36986.8048	-36986.64045	-3.7901	3.23		29953.89336	29953.82738		
MCXZ	-45199.48566	-45199.32334	-3.7432	3.15	MUU0	-	-	14.3927	2.13
MD05	-70643.69685	-70643.59137	-2.4324	2.49		37683.09666	37683.72079		
MD1F	-35319.62888	-35319.43605	-4.4468	2.36	MUVF	-	-	-2.0306	2.01
MD97	-33795.77689	-33796.20744	9.92875	2.35		36232.90901	36232.82095		
MD9D	-32235.45172	-32235.58747	3.13046	2.98	MUWB	-35253.4289	-	6.58871	2.32
MD9M	-29616.14749	-29616.42222	6.3354	2.12			35253.71461		
MD9V	-28143.14936	-28143.34763	4.5722	2.35	MUWM	-	-	-1.149	2.72
MDAF	-31614.04545	-31613.60907	-10.063	2.29		34414.64698	34414.59715		
MDVF	-30631.50813	-30631.58739	1.82778	2.41	MUYM	-	-	-4.3755	2.76
MDXB	-36487.79384	-36488.97335	27.2003	2.11		33187.05327	33186.86352		
MDXV	-33350.35613	-33350.38268	0.61208	2.31	MVHS	-59384.708	-	-2.7186	3.2
ME3B	-33162.26858	-33162.30222	0.77562	2.58			59384.59011		
ME75	-31144.32966	-31144.47004	3.23725	2.38	MVJQ	-	-	-3.1147	2.05
ME82	-30299.74009	-30299.73567	-0.1019	3.25		29047.19513	29047.06006		
MECV	-38446.48419	-38446.26616	-5.028	2.33	MVM3	-	-	0.00354	2.37
MEDF	-33748.33167	-33748.17171	-3.6888	2.92		36480.05613	36480.05629		
MEHW	-29625.81971	-29625.81978	0.00173	2.09	MVN8	-	-	-12.876	2.86
MEJE	-31825.324	-31825.32412	0.00263	2.09		57859.77182	57859.21345		
MEK2	-54017.25275	-54017.25076	-0.0458	2.01	MVOR	-	-	7.82595	2.14
MEKB	-31116.0266	-31116.07869	1.20118	2.12		34457.79621	34458.13558		
MEO8	-31510.27204	-31510.2054	-1.5366	2.13	MVOZ	-	-	-3.6989	2.11
MER9	-29152.69216	-29153.30386	14.1059	2.74		28629.27091	28629.11051		
MEW2	-42001.38998	-42001.35716	-0.7568	2.35	MVPS	-	-	-0.836	3.63
						32897.78887	32897.75262		
					MVU3	-	-	-4.3394	2.12
						25142.76182	25142.57364		
					MVX2	-	-	-2.4366	2.23
						34551.50292	34551.39726		
					MW08	-	-	8.03529	2.05
						35576.74367	35577.09211		
					MW6Y	-	-	0.83959	2.24
						34149.82262	34149.85903		
					MW7C	-	-	-0.3871	2.7
						32135.31265	32135.29586		
					MW7D	-	-	-0.0133	2.57
						32135.29129	32135.29071		
					MW7G	-	-27955.5033	-2.3548	2.5
						27955.60541			
					MW7H	-	-	-0.1367	2.59
						29019.67727	29019.67134		
					MW7I	-	-	-2.3434	2.92
						32135.38626	32135.28464		
					MW7L	-32135.2774	-	-0.9124	2.34
							32135.23783		
					MW8U	-	-	-3.9164	3.18
						47092.95899	47092.78916		
					MWAP	-	-	-0.4201	2.44
						32309.53548	32309.51726		
					MWAU	-	-	-1.0585	2.36
						32746.10953	32746.06363		
					MWDD	-	-	-3.9434	1.28
						28882.24803	28882.07703		
					MWJK	-35186.3477	-	2.06041	2.61
							35186.43704		
					MWNH	-	-	4.91776	2
						37099.30197	37099.51523		

APPENDIX-A: CONTINUED

MEWN	-35996.03916	-35995.97603	-1.4558	1.95	MWRO	-	-	-2.5346	2.78
						34009.69208	34009.58217		
MEXJ	-39268.36399	-39268.317	-1.0836	2.65	MWUZ	-	-	-2.0489	2.09
						54123.53378	54123.44493		
MEZQ	-33232.71478	-33232.70302	-0.2712	2.02	MWXJ	-	-	-3.6411	2.06
						25693.12355	25692.96565		
MF29	-32472.7208	-32472.57372	-3.3919	2.14	MWY0	-	-	0.29442	2.14
						33699.49094	33699.50371		
MF44	-16684.9345	-16684.95756	0.53175	2.2	MWY5	-36576.7134	-	3.87756	2.4
							36576.88155		
MF6J	-32313.72311	-32313.8029	1.8401	2.55	MX3K	-	-	-0.503	2.06
						23314.59201	23314.57019		
MF7D	-36911.57705	-36911.62523	1.11083	2.03	MX3L	-	-35723.3488	-2.4866	2.34
						35723.45663			
MF80	-29600.016	-29600.11681	2.3246	2.65	MX45	-	-34255.2229	-0.1147	2.21
						34255.22788			
MFBY	-35978.00581	-35977.94129	-1.4879	2.03	MX7G	-	-	-0.3466	2.37
						35055.05965	35055.04462		
MFEZ	-31482.77786	-31483.26102	11.142	3.1	MX8G	-	-	-2.7964	2.78
						47560.94689	47560.82563		
MFFS	-40469.31948	-40469.38292	1.46303	2.19	MXH9	-	-	3.95062	2.41
						40534.15037	40534.32169		
MG2Y	-66688.57811	-66688.53595	-0.9722	2.58	MXHW	-	-	1.22749	2.39
						50530.66659	50530.71982		
MG3X	-34554.78165	-34554.72004	-1.4207	2.21	MXJX	-	-	13.3598	2.84
						35996.03118	35996.61051		
MG5C	-34528.59804	-34528.54768	-1.1614	2.44	MXMZ	-	-	0.07676	2.81
						30213.18016	30213.18349		
MG8J	-32676.83677	-32676.86687	0.69397	1.99	MXRO	-	-	-4.9317	2.32
						28419.06313	28418.84928		
MG9K	-34668.87866	-34668.84719	-0.7258	2.28	MXT3	-	-	10.6628	2.44
						33216.59735	33217.05974		
MGB3	-23991.33238	-23991.05827	-6.3211	2.02	MXXZ	-	-	-1.737	2.13
						33304.84685	33304.77152		
MGB7	-26005.59883	-26005.45766	-3.2554	2.38	MY4T	-23046.1722	-	-0.0261	2.56
							23046.17107		
MGD5	-45515.91483	-45515.59797	-7.307	3.22	MY5J	-	-	-0.659	2.14
						40349.76945	40349.74087		
MGPW	-37069.99159	-37069.77116	-5.0833	2.28	MY81	-	-	-2.4369	2.18
						29541.43438	29541.32871		
MGYO	-27499.54064	-27499.71323	3.97983	2.04	MYE5	-	-	0.04278	2.36
						27935.52039	27935.52224		
MGYW	-37611.21578	-37611.25718	0.95455	2.96	MYJ4	-31985.866	-	-0.3685	2.07
							31985.85002		
MH3E	-31541.49526	-31541.41173	-1.9264	2.49	MYNS	-	-	-3.4158	3.99
						35181.77721	35181.62909		
MH3H	-29605.53846	-29605.48068	-1.3324	2.42	MYTD	-	-	-2.014	2.29
						29268.76693	29268.67959		
MH3N	-32682.50595	-32682.37549	-3.0086	2.96	MYTU	-	-	1.03177	2.08
						37569.38844	37569.43318		
MH3P	-92800.75988	-92800.77214	0.28275	2.5	MZ18	-	-34066.3251	5.83921	2.85
						34066.07189			
MH3Q	-28595.23551	-28595.20073	-0.8021	2.36	MZKW	-	-	-3.7203	2.27
						48780.89405	48780.73273		
MH5F	-28259.81383	-28259.19302	-14.316	2.25	MZM4	-	-	1.00068	2.02
						34784.14342	34784.18681		
MH5I	-42635.04926	-42634.77935	-6.2241	2.58	MZP2	-	-	-0.3248	2.41
						31158.40833	31158.39425		
MH7P	-45538.52458	-45538.53329	0.20074	2.34	MZUI	-	-	-1.7692	2.94
						47804.32469	47804.24797		
MHCJ	-38246.55446	-38246.41845	-3.1364	3.04					