Evaluating and Improving the Viability of Machine Learning to Solve Chemical Problems

by

Dakota Lee Folmsbee

Bachelor of Science, Clarkson University, 2016

Submitted to the Graduate Faculty of the

Kenneth P. Dietrich School of Arts and Sciences

in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2022

UNIVERSITY OF PITTSBURGH

KENNETH P. DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Dakota Lee Folmsbee

It was defended on

February 11, 2022

and approved by

Kenneth Jordan, Professor, Department of Chemistry

Peng Liu, Associate Professor, Department of Chemistry

David R. Koes, Associate Professor, Department of Computational & Systems Biology, School of Medicine

Dissertation Director:
Geoffrey R. Hutchison, Associate Professor, Department of Chemistry

Evaluating and Improving the Viability of Machine Learning to Solve Chemical Problems

Dakota Lee Folmsbee, PhD

University of Pittsburgh, 2022

While improvements in computer processing have allowed for increasingly faster quantum mechanical (QM) calculations, the need for alternative techniques to accelerate computer-accelerated material design continues to grow. Screening methods have tackled this through methods that search chemical space more efficiently but often use faster, albeit less accurate methods for evaluation due to the large number of calculations conducted. Machine learning (ML) has shown promise as a potential surrogate for time-consuming quantum mechanical calculations, such as density functional and first-principles method, that would lend these screening methods a fast and accurate approach to evaluation.

This work sets out to determine the viability of ML methods through multiple tests. The ranking of thermally accessible conformations was conducted to establish ML's capacity to differentiate small energy differences compared to other established methods. The performance of ML methods was found to be equivalent to that of semi-empirical methods in both accuracy and evaluation time, demonstrating promise for future improvements of ML models. Next, ML's understanding of chemical physics was tested by analyzing the short and long-range interactions that occur with bond compressing and stretching as well as the effect of steric hindrance of dihedral angles. The work demonstrated the extent the training set has on the

model as short and long-range interactions not present in the set became apparent in the testing of the models. Additionally, the inclusion of torsion sampling in the ANI-2 training exemplifies why more robust training sets are needed for more accurate ML methods.

Current work on ML indicates a strong need for additional diversity in training data. Initial work done on comparing experimental crystallographic geometry and gas-phase computed conformer torsional preferences examine the possible use of a quantum-based ETKDG, QTDG, for future conformer training set generation for expanding existing training sets. Future work on expanding data sets is crucial for ML performance as ML methods are very reliant on the scope of the training set. Incomplete training sets that do not appropriately represent chemical space diminish the applicability of ML to solve chemical problems.

Table of Contents

List of Tables

List of Figures

Preface

Coming to Pitt, I never thought I would be doing computational chemistry. My undergraduate research had focused on synthetic organic chemistry, and while I wanted a change of pace, I was not expecting to forgo the lab entirely. The combination of chemistry with computational methods has since become a passion as I researched rapid prediction and screening methods for inverse materials design. My work here at Pitt has grown me as both a chemist and a person, and I look forward to continuing to build upon this foundation.

I would like to thank those who helped me along my journey to get me here today. First and foremost, I want to thank my advisor, Dr. Geoffrey Hutchison, for their help and support throughout my time at Pitt and for always having an open door for questions. I want to also thank my committee, Dr. Kenneth Jordan, Dr. Peng Liu, and Dr. David Koes for their time and support. I also want to extend a big thanks to past and present members of the Hutchison group for all of their help and friendships over the years. I especially would like to thank Dr. Nate Miller, Dr. Christopher Petroff, and Danielle Hiener for always being available to listen to my practice presentations and read through my drafts. I also want to thank Luke Langkamp for his assistance on data set compilation during his undergraduate research. Outside of the lab, I would like to thank my friends in the chemistry department at Pitt who were always willing to get coffee, grab lunch, or just get a breath of fresh air when projects were not going our way.

I want to extend my gratitude to my family who has been incredibly supportive of me over the years. I want to thank my parents, Judy and Scott Folmsbee, for doing everything they could to support me in my endeavors. I would also like to thank my brother, Colton Folmsbee, who was always there for laughs and a smile when I needed it most. Finally, I

want to thank my amazing wife, Alyla Ballantine, for her love and support that helped me get through the stresses of both undergraduate and graduate school.

Pittsburgh, February 2022 D. Folmsbee

1.0 Introduction

Interest in computer-accelerated material design continues to grow as recent improvements in computer processing have allowed for increasingly faster quantum mechanical (QM) calculations. Although major advancements in computational power have been made, accurate methods such as Density Functional Theory (DFT) and Coupled Cluster (CC) can take several hours to days for calculations, often making brute force methods prohibitively expensive. This has highlighted the need for alternative techniques to accelerate the search through chemical space.

Searching chemical space for molecules with optimal properties is a considerable challenge. Chemical space is enormous, estimated to contain approximately 10^{60} possible drug-like molecules[1]. This vast size makes exhaustive search methods extremely time-consuming even when using faster, albeit less accurate methods like Force Fields (FF) or semi-empirical methods. The inefficiency of brute force searches has stimulated the exploration of screening methods as an alternative, more efficient approach.

Screening methods aim to efficiently search chemical space by reducing the number of time-consuming calculations required to find an optimal or near-optimal subset of candidates. A common technique used in drug discovery is high-throughput virtual screening in which a large database of known chemical space is explored for potential candidates[2]. This lowers the overall computational cost required to find optimal candidates as time-consuming calculations can be limited to a small set of candidates in the final evaluation step. High-throughput virtual screening typically makes use of a previously constructed database for screening, limiting the search area to a predefined subset of chemical space. Chemical space limitations of a database can be overcome through the augmentation of the database using methods like genetic algorithms (GA).

1.1 Genetic Algorithms

The GA uses concepts from evolutionary biology (genotypes, fitness, and natural selection) to optimize populations through generations for increasingly desirable features. In successive generations, the "better" candidates are retained with mutation and crossover events occurring to allow for the exchange and introduction of genes. Over time, top candidates survive successive generations while less ideal candidates are eliminated. The process of generational optimization allows for the exploitation of desirable chemical space as the GA will explore promising areas while avoiding known unfavorable areas. This results in an evolutionary method that learns important characteristics to pass down to successive generations producing promising candidates in a fraction of the time of exhaustive search methods[3].

While the GA provides a significant speedup when compared to brute force methods, the speed and accuracy of the fitness evaluation step during the selection of candidates after each generation can often determine the speed of chemical space exploration. Time-consuming conventional quantum mechanical methods such as DFT and CC can provide accurate results for each generation but slow down the screening process. Semi-empirical and FF methods can provide a significant speedup, but at the cost of accuracy. Fast evaluation methods such as machine learning (ML) can aid in the acceleration of the discovery process by providing a rapid, yet accurate evaluation method.

1.2 Machine Learning

ML is increasingly being used to tackle today's problems in numerous different fields from financial trading to autonomous driving. The power of ML methods stems from a model's ability to learn from patterns within the data without explicit instruction. This makes

ML ideal for analysis and predictions from large amounts of data computationally more efficient. The use of ML in chemistry has been proposed as a surrogate for time-consuming quantum mechanical calculations with the idea that once trained ML models will provide rapid predictions with the potential of density functional or first-principles methods accuracy[4–14].

For the application of ML in chemistry, molecules need to be translated from the standard depictions used in the classroom and laboratory into a representation that an ML model can interpret and from which patterns can be inferred. Representations of molecules need to capture the underlying physics such that models can infer patterns and provide physics-based predictions.

Early representations looked to accomplish this through the use of existing cheminformatics representations like SMILES[15–17] (Simplified Molecular Input Line Entry System) and InChi[18]. Variational Auto Encoders (VAE) have been used to map SMILES strings to a latent space that can be searched for optimal candidates[19,20]. When the optimal candidates are found, they can be decoded back into SMILES format. The decoding back into SMILES can cause significant issues with methods like these as there is no guarantee that the decoded SMILES will be syntactically valid or properly represent a molecular structure. This has spurred the creation of DeepSmiles[21], which aims to avoid this issue of invalid decoded molecules by changing the syntax rules to avoid common errors which can occur when modifying SMILES, allowing ML methods to more consistently generate syntactically valid SMILES strings.

Other representations consider the inclusion of local and global connectivity information with representations like coulomb matrix[22] (CM), which uses molecular information in the Hamiltonian, such as coordinates and nuclear charges, for the modeling of atomization energies. The CM representation is seen below:

$$M_{IJ} = \begin{cases} 0.5 Z_I^{2.4} & \text{for } I = J, \\ \frac{Z_I Z_J}{|R_I - R_J|} & \text{for } I \neq J. \end{cases} \tag{1.1}$$

consists of a square matrix (M_{IJ}) where the off-diagonal elements (Z_I, Z_J, R_{IJ}, and R_J) are the Coulomb nuclear repulsion between the atom pairs. A further modification of the CM representation is the Bag of Bonds[23] (BoB) representation. The BoB representation reorganizes the CM representation so that the atoms and pair-wise interactions are sorted into bags (e.g., C, C-C, and C-N). This is done in a bag-of-words text mining descriptor style and each bag is filled with $Z_I Z_J/|R_I - R_J|$ to represent the connectivity of the molecule. Further adaptation was done to make the Bond Angle-ML[24] (BAML) representation. This many-body expansion of BoB sought to include further connectivity information through the inclusion of supplementary bags containing angles and torsions.

Another approach to representations is the use of graph-based representations derived from cheminformatics. These are built upon the 2D representations of the molecular graph in which the atoms are represented by vertices and bonds by edges in the graph. Extended-connectivity fingerprints[25] (ECFP) are a molecular graph model that expands out along the bonds from each heavy atom for 2-3 steps to observe the local connectivity of each heavy atom. The extensions for each heavy atom are stored as a fragment that is hashed as a fingerprint. After iterating over the molecule, these fragment fingerprints are combined to describe the molecule. Kearnes *et al.*[26] proposed expanding upon the base molecular graph by including atom and pairwise properties in the representation. This additional information provides distances as well as atom and bond type to make a simple molecular graph convolution model. This graph convolution model can be further expanded to include more atom features such as formal or partial charges, hybridization, as well as additional pair features such as whether the atom pair is in the same ring.

An additional representation becoming increasingly common is the atom-centered symmetry function[4] (ACSF). The ACSF representation takes into account the connectivity information through a description of local atomic environments with radial functions seen in Eqn. 1.2 and angular symmetry functions seen in Eqn. 1.3. In recent years modifications of ACSFs for deep neural networks (DNN) have been increasingly more common with the

adoption of the representation in the ANI family of DNNs[6–8,27]. ACSFs have also been used for non-DNNs with FCHL[28,29] showing promising results for ML methods like kernel ridge regression.

$$G_i^1 = \sum_{j \neq i}^{\text{all}} e^{-\eta(R_{ij}-R_s)^2} f_c(R_{ij}) \tag{1.2}$$

$$G_i^2 = \sum_{j,k \neq i}^{\text{all}} (1 + \lambda \cos\theta_{ijk})^{\zeta} \times e^{-\eta(R_{ij}^2+R_{ik}^2+R_{jk}^2)} f_c(R_{ij}) f_c(R_{ik}) f_c(R_{jk}) \tag{1.3}$$

OrbNet[11,30] is another graph-based approach, which uses nodes and edges to relay the atomic orbital information of the molecular system instead of representing the molecule atomwise. The nodes correspond to diagonal symmetry-adapted atomic orbitals (SAAOs) while edges correspond to off-diagonal SAAOs. OrbNet differs from the representations and models previously discussed as it uses the idea of delta-learning.

Methods like OrbNet utilize delta-learning in the form of a Δ-ML to improve the performance of lower-cost methods. Conventionally, Δ-ML models utilize the ML component to augment a rapid but less accurate method, often semi-empirical, such that the performance becomes on par with more accurate time-consuming methods. This is done through the use of the faster calculation method performing a baseline calculation. This baseline calculation is passed on to the ML model which has been trained on the more accurate method's calculation. Instead of learning directly from the molecule like previous methods, the ML learns the difference between the two models and acts as a correction. Once trained, the Δ-ML model ideally provides a more accurate prediction at the cost of only the inaccurate method and the ML model which are faster than the time-consuming method[30].

An increasingly important aspect of ML models is the training sets used. Various training sets have been used to train models for chemistry, many of which are subsets and augmentations of the GDB-17[31] data set, a large data set of organic small molecules constructed through enumeration. Two commonly used subsets are the QM7[22,32] and QM9[33]

data sets consisting of optimized molecules containing H, C, N, O, and F with up to 7 or 9 heavy atoms respectively. These are relatively small training sets, with QM9 being the larger at 134k molecules, and have been used to demonstrate early performance for various property predictions[5,34]. Another popular batch of data sets based on the GDB-17 set are the ANI-1[35] and ANI-2[8] data sets. These sets contain both equilibrium and non-equilibrium structures generated from normal-mode sampling for molecules up to eight heavy atoms containing H, C, N, and O. The ANI-2 set augments the previous information by adding the elements of F, Cl, and S as well as additional torsion sampling.

The increased interest in molecular ML as a surrogate for time-consuming conventional quantum mechanical methods has spurred an evaluation of the viability of the various ML methods. Early work centered around testing of ML representations and methods for calculating different thermochemical properties[22–24] with benchmark studies failing to find a one size fits all solution[5,34], but showing early promise of representations based on molecular graph information. More recent deep neural networks (DNN) have sought to provide methods capable of evaluating potential energy surfaces for dynamics and performing geometry optimizations[6–8,27,36]. While these have been promising studies showing adequate performance for non-equilibrium geometries close to the equilibrium geometry, more work needs to be done to determine their effectiveness at handling short and long-range interactions.

1.3 Dissertation Overview

The viability of ML stems not just from the ability of the method to perform adequately for optimized geometries but to also appropriately handle thermally accessible conformations. ML methods need to properly perform well for both short and long-range interactions before they can be considered suitable replacements for conventional methods. This work looks to examine the extent to which ML can be considered a surrogate for time-consuming quantum

calculations and how to make improvements such that ML can aid in the materials discovery process.

When testing the viability of ML, evaluating the method's ability to accurately distinguish between thermally accessible conformations is an important test. Most molecules have multiple geometrically distinct conformers and ML methods need to be able to provide energy predictions that appropriately differentiate between them. This work looks to determine what ML methods if any, can sufficiently differentiate among conformations and how they compare to conventional methods.

With the knowledge of the ML method's physical understanding of conformational geometries, we set out to learn the extent of physics common ML models comprehend. An evaluation of potential energy curves was performed to determine if ML methods could predict whether a bond was at equilibrium, as well as whether the ML methods understand both short and long-range interactions that occur with bond compressing and stretching. In addition to the physics of bond interactions, further analysis was done to determine whether ML methods could find preferred dihedral angles and the effect of steric hindrance.

After analyzing the pitfalls of current state-of-the-art ML methods, an important next step is analyzing the shortcomings of training data and determining how to make adequate improvements. Increasing training set diversity is a key issue in the quest to create ML methods that can act as a surrogate for conventional methods. The most commonly used training sets lack the diversity of atom species, making applications like protein binding and DNA sequencing impossible for ML methods trained on just these sets. Modifications to training sets can also be made to include more conformations and non-equilibrium geometries to improve the performance of ML methods in differentiating between conformational geometries. While training set generation can be performed with conformational search methods like experimental-torsion distance geometry with basic knowledge[37] (ETKDG), it is important to consider whether a quantum-based alternative may be the desired direction for ML methods that will be expected to predict properties at the level of accuracy of quantum mechanical

calculations.

Our group has demonstrated the proficiency of GAs in chemistry for screening chemical space for inverse material design[3]. These GAs have incorporated faster, but less accurate semi-empirical evaluation methods for the fitness function in place of more accurate methods in an attempt to efficiently traverse chemical space. The addition of ML methods for evaluation looks promising as a way to improve the accuracy of the evaluation in the fitness function while maintaining a relatively short evaluation time. The combination of methods like GA and ML aims to provide an efficient search of chemical space with the accuracy of time-consuming quantum calculations, but more work needs to examine the full capabilities of ML in the field of chemistry.

2.0 Assessing Conformer Energies using Electronic Structure and Machine Learning Methods

This chapter is adapted from:

> Folmsbee, D., Hutchison, G. Assessing conformer energies using electronic structure and machine learning methods. *ChemRxiv:13151069.v1* **2020**.
> DOI: 10.26434/chemrxiv.11920914.v2

which is also published as:

> Folmsbee, D., Hutchison, G. Assessing conformer energies using electronic structure and machine learning methods. *Int. J. Quantum Chem.* **2020**, 2007486.
> DOI: 10.1002/qua.26381.

It is a collaborative effort in which G. H. and I carried out calculations and data analysis, generated figures, and wrote the manuscript; G. H. conceived and directed the project.

2.1 Summary

We have performed a large-scale evaluation of current computational methods, including conventional small-molecule force fields, semiempirical, density functional, *ab initio* electronic structure methods, and current machine learning (ML) techniques to evaluate relative single-point energies. Using up to 10 local minima geometries across ~700 molecules, each optimized by B3LYP-D3BJ with single-point DLPNO-CCSD(T) triple-zeta energies, we consider over 6,500 single points to compare the correlation between different methods for both relative energies and ordered rankings of minima. We find promise from current ML methods and

recommend methods at each tier of the accuracy-time tradeoff, particularly the recent GFN2 semiempirical method, the B97-3c density functional approximation, and RI-MP2 for accurate conformer energies. The ANI family of ML methods shows promise, particularly the ANI-1ccx variant trained in part on coupled-cluster energies. Multiple methods suggest continued improvements should be expected in both performance and accuracy.

2.2 Introduction

For almost all molecules, multiple geometrically-distinct conformers exist. Understanding and predicting thermodynamically accessible ensembles of molecular conformers is a key task underlying much of computational chemistry.[38–40] In principle, for each rotatable bond, the number of possible minima increases exponentially. Consequently, most conformer sampling methods[41] use classical small-molecule force fields to evaluate energies because of their fast performance, despite potentially poor correlation with quantum mechanical methods.[42]

Multiple efforts have evaluated the success of wavefunction and density functional first-principles methods to compare the energetics of different conformers.[43–49] While experimental crystal structures and bioactive docked conformers are not always the lowest energy conformer, recent efforts have demonstrated only small energy differences when using quantum chemical methods instead of force fields.[50,51]

Even for simple molecules such as 1,1'-biphenyl, use of large basis set coupled cluster methods are needed to accurately place the dihedral angle and barrier.[52] Other works have documented the need for accurate treatment of non-covalent interactions to model conformers in π-conjugated oligomers.[53]

One common assumption is the presumed balance between increasing desired thermochemical accuracy and increased computational time. That is, more computationally intensive methods produce more accurate geometries and thermochemical properties. For example,

the rise of composite *ab initio* thermochemical recipes such as G3,[54] G4,[55] and W1[56,57] to W4[58] seeks to provide highly accurate thermochemical predictions by separate estimates of basis set extrapolation and electron correlation. Still, such methods are largely limited to small molecules due to the high computational cost.[59] As mentioned above, efforts for conformer sampling have often focused on classical force fields or multi-level approaches using semiempirical methods.[41,60,61]

In our previous paper,[42] we considered both the single-point energies and geometry optimizations of a range of common computational chemistry methods, including classical force fields, semiempirical quantum chemistry, and dispersion-corrected density functional methods. In general, due to the large differences in the potential energy surfaces predicted by force fields and quantum methods, we found poor correlation between both single point energies at the same geometry and optimized geometries using different methods.

In this work, in order to expand our range of computational methods, we only consider the relative single point energies from the same set of density-functional optimized geometries, comparing multiple current methods to a high-quality coupled cluster baseline. We consider the mean absolute relative errors in energies (MARE), as well as the correlation of relative energies, reflected in the R^2 coefficient of determination, and the ranking of single-point energies reflected in the Spearman ρ correlation. The use of correlation coefficients and the Spearman correlation intend to consider whether methods exhibit systematic errors that may not affect linear correlation or ranking of energetic stabilities.

While we find increased accuracy typically still requires exponential increases in computational time, several methods stand out as widely useful methods for ranking conformer energies. Future improvements in standard computational methods and machine learning surrogates suggest that both increased accuracy and efficiency are expected from further method development.

2.3 Computational Methods

Calculations were performed using Open Babel version 3.0[62] for all force field calculations (MMFF94[63–67] and UFF[68,69]), `OpenMOPAC` for PM7,[70] `xtb` version 6.2[71] for GFN0[72] GFN1[73] and GFN2 calculations,[74] and `Orca` 4.0.1[75] for all density functional and *ab initio* calculations, unless otherwise indicated. For density functional methods, the D3(BJ)[76–79] dispersion correction scheme was used as indicated, except for ωB97X-D3[80] which uses a similar approach. For *ab initio* methods, `Orca` 4.0.1 was used for MP2[81] and DLPNO-CCSD(T)[82,83] with "TightPNO" using the cc-pVTZ basis set.[84,85] Energies are read from all output files using the `cclib`[86] version 1.6.2, and `pybel` version 3.0.[87]

Machine learning methods included "bag-of-features" representations and ANI-1x[7], ANI-1ccx[?] , and ANI-2x[8] models. The Bag-of-Features representations chosen were Bag of Bonds[23](BOB), Bond Angle Torsion[24](BAT), and Bond Angle Torsion Typed (BATTY). BOB represents atoms and pair-wise interactions into sorted bags with BAT being a many-body expansion to include angles and torsions. Both of these representations were implemented using `chemreps`.[88] The BATTY representation takes inspiration from BAT in order to include minimal atom typing in all bond, angle, and torsion bags while excluding nonbonding interaction and nuclear charge bags in the final representation, as discussed below. `scikit-learn`[89] was used for kernel ridge regression of Bag-of-Features representations.

For this work, all timings are single-core CPU times using a 2.60 GHz Intel Skylake CPU (Intel Xeon Gold 6126) with 192GB RAM per node, through the University Pittsburgh Center for Research Computing.

Python scripts and Jupyter notebooks were used to compile all data into `pandas`[90] data frames, using `numpy`[91] and `scipy`[92] functions for analysis. 3DMol.js was used for interactive molecular visualization of conformers.[93] Plotly was used for interactive plots.[94]

All scripts and data, including molecular geometries, are provided through GitHub (`https://github.com/ghutchis/conformer-benchmark`) with the intent that additional

computational methods can be added to these benchmark comparisons.

2.4 Test Set Selection

As in our previous work,[42] a dataset consisting of experimental crystal structures of 700 small molecules capable of multiple conformer geometries was provided to us by Ebejer[95] and were derived from the work of Hawkins et al.[60] along with ligands from the Astex Diverse Set.[96] These compounds have been repeatedly used to evaluate the quality of conformer generation.[60,95] Approximately half (320 molecules) consist solely of carbon, hydrogen, nitrogen, and oxygen (CHON) atoms, while the remainder are more complex drug-like compounds and ligands from the Protein Data Bank (PDB).[60] A list of Simplified Molecular Input Line Entry Specification (SMILES)[97] for all 700 molecules can be found in the Supporting Information.

For *ab initio* calculations using the cc-pVTZ basis sets, relativistic effective core potentials were not available for molecules containing iodine. Thus, for comparisons with DLPNO-CCSD(T) and RI-MP2 methods, such species were omitted. Similarly, the ANI-1x and ANI-1cxx methods only support molecules containing CHON atoms and evaluations were only performed on the subset of molecules supported. The ANI-2x method supports additional elements, but not bromine or iodine and thus evaluations were similarly only performed on the supported subset for that method.

For bag of feature ML testing, the training set was five conformers of each molecule, with the remaining conformers as test/validation. Any molecule with fewer than five conformers had the conformers added to the training set and was omitted from the test set.

2.5 Results

In this work, we focus on the evaluation of single point atomization energy calculations on a subset of ~700 organic molecules. Conformers were initially created from a set of 250 diverse poses with maximal heavy-atom root mean squared deviation (RMSD) using Open Babel, and at most 10 poses were selected based on the lowest heat of formation calculated by PM7, followed by full geometry optimization using B3LYP-D3BJ with the def2-SVP basis set.[42]

Using this set of DFT-optimized minima, in this work, single point atomization energies were computed using the DLPNO-CCSD(T)[82,83] method using the cc-pVTZ basis set.[84,85] This approach has been found to be a highly accurate method for calculating thermochemical properties and with a significantly lower computational cost for medium to large organic molecules, compared to canonical CCSD(T) methods.[82,98,99] Using only the set of molecules in which all standard (i.e., not machine-learning based) methods completed leaves 6511 entries. Of those, 9 molecules (out of 690) had 2 or fewer poses and were also removed, leaving 681 unique molecules and ~6500 entries for comparison.

To our knowledge, this is the most extensive computational validation set, both in terms of the number of compounds, geometries, and computational methods for studying low energy molecular conformers. We provide all data and analysis scripts as open data and open source to allow future reuse via a GitHub repository.

As illustrated in Figure 2.1, each method is correlated with DLPNO-CCSD(T) / cc-pVTZ energies for each molecule (e.g., `astex_1hwi` in Figure 2.1). Since each molecule has several conformers, three metrics are compiled, the mean absolute relative energy (MARE) compared to the DLPNO-CCSD(T) atomization energies, the Pearson R^2 correlation, and the Spearman correlation ρ. The MARE metric gives an absolute measure of the energetic errors, but since different methods use different energy scales (e.g., heats of formation for PM7 and force fields), the statistical correlations use linear regression (R^2) and relative ordering (Spearman ρ) to remove sources of systematic energy differences. For each metric across each method,

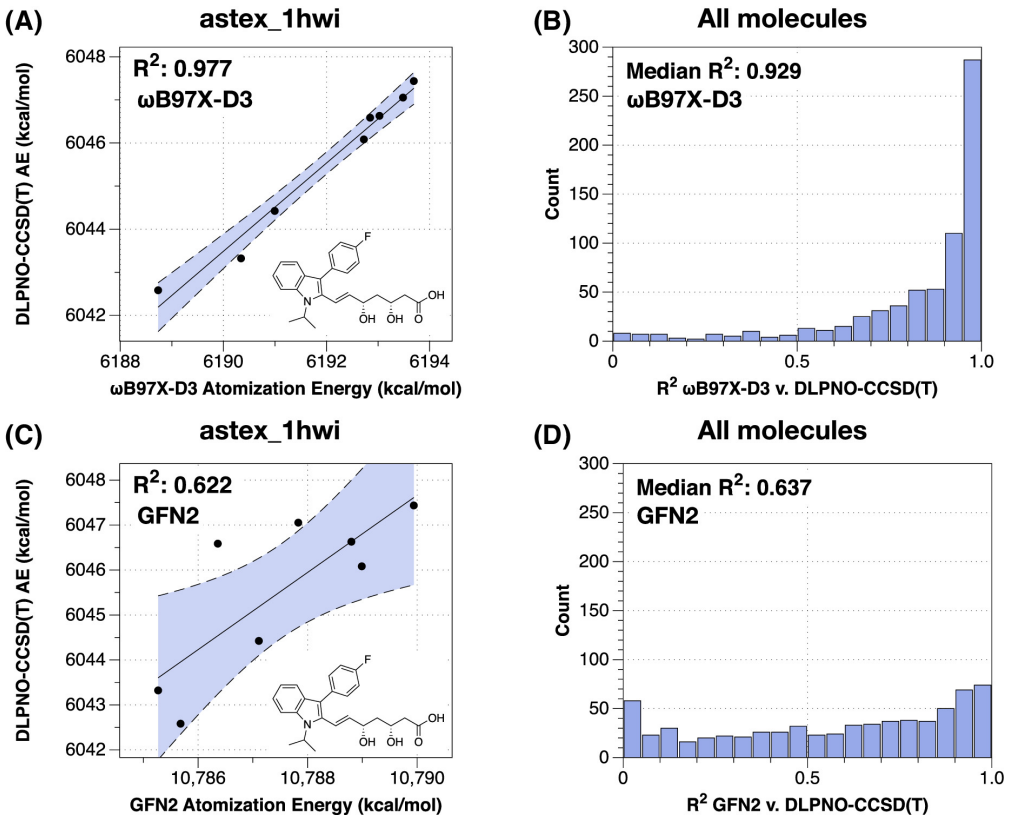

Figure 2.1: Example analysis of ωB97X-D3 and GFN2 methods, starting with (A) correlation between ωB97X-D3 and DLPNO-CCSD(T) energies for a single molecule, (B) histogram of R^2 correlations across all molecules, (C) correlation between GFN2 and DLPNO-CCSD(T) energies, and (D) corresponding histogram of R^2 correlations across all molecules.

the median value was compiled as illustrated in Figure 2.1, to represent the overall quality of a given method.

Since the metrics are unlikely to reflect normal distributions (e.g., Figure 2.1 shows highly non-Gaussian distributions), determining confidence intervals cannot be established from analytical formulas. Consequently, we used bootstrap sampling to establish 95% confidence values for the medians, as reported below. For ease of discussion, we have given the confidence ranges in all tables and figures, but indicate $\pm$ errors using the average of the upper and lower bounds. In general, the asymmetry between upper and lower bounds are smalll.

By considering a large number of diverse organic molecules with many poses per molecule,

we seek to sample a wide variety of conformer energy preferences (e.g., intramolecular hydrogen and halogen bonding, π-π stacking, electrostatic interactions, etc.). While using optimized low-energy conformers may under-estimate the accuracy of methods for high-energy structures,[44] we believe the current work is a challenging but useful comparison. In general, such high-energy geometries reflect steric repulsion more than the diverse types of interactions driving low-energy geometries.

Moreover, many computational predictions rely on Boltzmann-weighted averages of multiple thermally accessible conformers, including NMR prediction,[38,39] reactions, and even understanding the effects of dipole moments on solvent viscosity.[100] Consequently, deriving accurate relative energies of molecular conformers is a crucial task, as discussed below.

2.5.1 Comparison of Single Points vs. DLPNO-CCSD(T)

For comparison, we considered a wide variety of currently available computational methods:

- **Common classical organic force fields**: MMFF94,[63–67] UFF,[68] GAFF[101]
- **Semiempirical wave function:** PM7[70]
- **Density functional tight binding**: GFN0,[72] GFN1,[73] GFN2[74]
- **Low-cost density functional approximation**s: PBEh-3c,[102] B97-3c[103]
- **Dispersion-corrected density functionals**: B3LYP,[104–107] PBE[108,109], ωB97X-D[80] with dispersion correction (using def2-TZVP basis set[110,111])
- **Møller-Plesset RI-MP2**[81] with the cc-pVTZ basis set[84,85]

In the case of B3LYP and PBE dispersion-corrected functionals, we also considered both the commonly-used double-zeta def2-SVP and triple-zeta def2-TZVP basis sets to understand the effects of basis set size. For B3LYP, PBE, and ωB97X, we also considered the accuracy with and without dispersion correction.

Table 2.1: Overall statistics across all molecules studied and all methods. Columns indicate median mean absolute relative error (MARE in kcal/mol), median R^2 correlation, median Spearman correlation, and median single-core CPU time in seconds. MARE, R^2, and Spearman correlation are relative to the DLPNO-CCSD(T)/cc-pVTZ baseline. Ranges indicate 95% confidence intervals for the median metrics established by bootstrap sampling.

Method	MARE		R^2		Spearman ρ		CPU Time	
DLPNO-CCSD(T)	0		1		1		21901.38	276.95
RI-MP2	0.11	[0.11-0.13]	0.96	[0.96-0.97]	0.95	[0.95-0.96]	2118.83	42.26
ωB97X-D3	0.16	[0.15-0.17]	0.93	[0.91-0.94]	0.92	[0.9-0.93]	2524.83	35.67
B3LYP (TZ)	0.17	[0.15-0.19]	0.92	[0.91-0.93]	0.92	[0.9-0.93]	1672.79	20.67
B97-3c	0.2	[0.18-0.22]	0.9	[0.89-0.92]	0.9	[0.88-0.92]	137.45	2.16
PBE (TZ)	0.21	[0.19-0.23]	0.88	[0.87-0.9]	0.89	[0.88-0.9]	358.65	6.94
PBEh-3c	0.21	[0.18-0.23]	0.88	[0.86-0.9]	0.88	[0.87-0.9]	453.04	9.46
B3LYP (SVP)	0.23	[0.21-0.26]	0.87	[0.84-0.89]	0.88	[0.87-0.89]	330.94	4.35
PBE (SVP)	0.26	[0.24-0.3]	0.83	[0.81-0.86]	0.85	[0.84-0.88]	149.03	2.24
ANI-1ccx	0.44	[0.36-0.52]	0.64	[0.57-0.71]	0.71	[0.64-0.77]	1.45	0.0
GFN2	0.39	[0.33-0.43]	0.64	[0.59-0.68]	0.72	[0.68-0.75]	2.6	0.07
GFN1	0.35	[0.31-0.41]	0.62	[0.58-0.66]	0.7	[0.66-0.73]	2.66	0.05
ANI-2x	0.41	[0.36-0.48]	0.62	[0.56-0.69]	0.68	[0.65-0.72]	3.45	0.01
ANI-1x	0.45	[0.38-0.54]	0.59	[0.52-0.66]	0.65	[0.57-0.72]	1.46	0.0
BATTY/n	0.42	[0.38-0.48]	0.47	[0.41-0.54]	0.5	[0.4-0.6]	0.14	2.16e-05
GFN0	0.44	[0.39-0.49]	0.4	[0.35-0.48]	0.53	[0.46-0.56]	0.07	0.0
GAFF	1.64	[1.42-1.83]	0.35	[0.29-0.41]	0.48	[0.44-0.54]	0.01	5.73e-05
MMFF94	0.7	[0.58-0.85]	0.33	[0.29-0.38]	0.47	[0.43-0.52]	0.0	4.4e-05
BOB	1.92	[1.72-2.16]	0.32	[0.28-0.39]	0.1	[0.0-0.2]	0.14	3.92e-05
PM7	0.62	[0.56-0.71]	0.32	[0.27-0.36]	0.33	[0.27-0.41]	0.06	0.0
BAT	1.18	[1.03-1.3]	0.31	[0.28-0.38]	0.2	[0.1-0.3]	0.18	1.32e-05
UFF	5.03	[4.4-5.61]	0.29	[0.24-0.34]	0.32	[0.24-0.41]	0.0	8.61e-06

2.5.2 Basis Set Effects

For the frequently-used B3LYP-D3BJ and PBE-D3BJ density functional methods, we considered both the def2-SVP and def2-TZVP basis sets. In both cases, the triple zeta basis set significantly improved correlation with the DLPNO-CCSD(T)/cc-pVTZ baseline, for example, the median R^2 scores improved from 0.868 $\pm$0.064 to 0.920$\pm$0.025 for B3LYP-D3BJ and from 0.835$\pm$0.025 to 0.885$\pm$0.018 for PBE-D3BJ. There were comparable improvements in median Spearman rank correlation and reduced mean absolute relative errors, all statistically significant (i.e. p-values far below 0.001). The increased basis sets also roughly doubled the CPU time required.

While the PBE method is still significantly faster than B3LYP, the newer B97-3c proves to be faster than either with comparable accuracy (i.e., roughly intermediate to the TZ results for B3LYP-D3BJ and PBE-D3BJ). Additionally, the time required for B3LYP-D3BJ/def2-TZVP calculations is only slightly less than RI-MP2/cc-pVTZ results, which exhibit significantly improved accuracy relative to DLPNO-CCSD(T)/cc-pVTZ (i.e., median $R^2 = 0.964 \pm 0.006$ and median MARE of 0.115 $\pm$0.011 kcal/mol for RI-MP2).

Thus increasing basis set size for these density functional methods, at least from double zeta to triple zeta, does improve accuracy, albeit at a significant computational cost. In general, the B97-3c method provides accuracy comparable to popular dispersion-corrected DFT methods such as B3LYP-D3BJ with faster performance, and RI-MP2 provides greater accuracy at a very similar speed.

2.5.3 Dispersion Corrections

Since the bonding is consistent across multiple conformers, the ranking of small energy differences is known to be dominated by non-bonding interactions.[112,113] Density functional methods are known to incorrectly account for dispersion interactions, which has led to a variety of empirical corrections.[76–79,114–117] Comparing un-corrected PBE, B3LYP, and ωB97X single-point energies to DLPNO-CCSD(T) illustrate a significant effect. The uncorrected median R^2 values drop by ~0.12, and the median Spearman correlations drop by ~0.08. For example, the median R^2 of B3LYP / TZ drops from 0.920 $\pm$0.012 to 0.706$\pm$0.050 without the D3BJ dispersion correction.

On the time-scale of a density functional calculation, these empirical dispersion corrections require only a minuscule time, yet significantly improve the accuracy of the relative energies. Thus, even though this work is concerned with intramolecular interactions in conformers, dispersion-corrected density functional calculations should always be used. Continued efforts, such as the improved D3 methods[118] or the new D4 method[114,115] will hopefully improve

Table 2.2: Effect of dispersion correction for DFT methods. Values in brackets indicate 95% confidence intervals from bootstrap sampling.

Method	Median R^2				Median Spearman ρ			
	Dispersion		No Dispersion		Dispersion		No Dispersion	
DLPNO-CCSD(T)	1		—		1		—	
ωB97X	0.93	[0.91-0.94]	0.88	[0.86-0.9]	0.92	[0.9-0.93]	0.89	[0.87-0.9]
B3LYP (TZ)	0.92	[0.91-0.93]	0.71	[0.66-0.76]	0.92	[0.9-0.93]	0.78	[0.75-0.81]
PBE (TZ)	0.89	[0.87-0.9]	0.75	[0.71-0.79]	0.89	[0.88-0.9]	0.81	[0.77-0.83]
B3LYP (SVP)	0.87	[0.84-0.89]	0.73	[0.67-0.76]	0.88	[0.87-0.89]	0.78	[0.76-0.81]
PBE (SVP)	0.84	[0.81-0.86]	0.75	[0.7-0.79]	0.86	[0.84-0.88]	0.81	[0.77-0.83]

their accuracy further.

2.5.4 Comparison of Timing

As discussed above, a frequent concern for conformer screening is the relative computational performance. In general, classical molecular force field methods have been preferred since they allow the generation of hundreds of conformers per compound in seconds. While traditional high-level *ab initio* methods are considered a "gold standard" for thermochemical energies, the time required for a single point energy evaluation may be high. For this work, all timings are single-core CPU times using a 2.60 GHz Intel Skylake CPU (Intel Xeon Gold 6126) with 192GB RAM per node.

As indicated in Figure 2.2, hybrid density functional methods such as B3LYP-D3BJ require significant single-core computational time for single-point energies of medium-sized organic molecules (median 26 $\pm$0.3 minutes) compared to GGA methods such as PBE or approximate density functional tight binding methods such as GFN1 / GFN2 (median 2.6$\pm$0.06 s yields ~600x speedup). Conventional density functional methods nevertheless represent a meaningful mid-point relative to DLPNO-CCSD(T) method, which may be faster than traditional coupled cluster methods but are still five to ten times slower than B3LYP (i.e., hours per single point energy).

Consequently, an important consideration is also the typical trade-off in computational chemistry between thermochemical accuracy and computational time. Since traditional

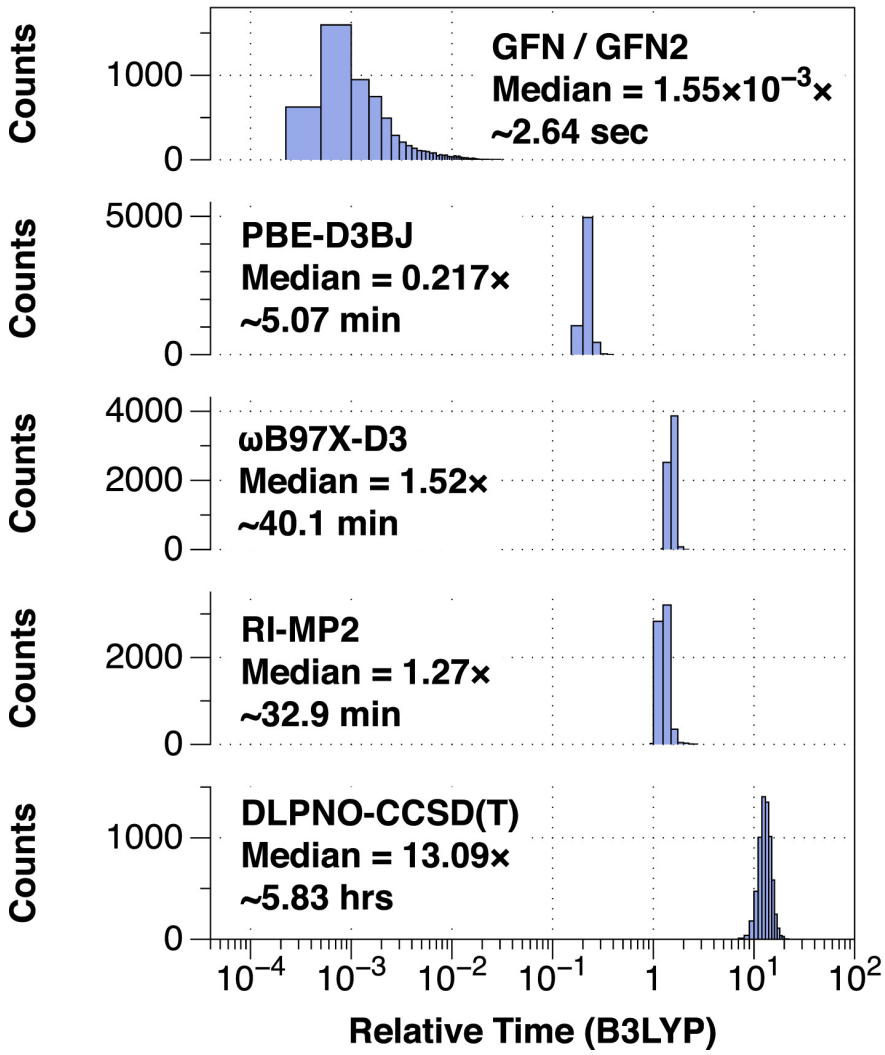

Figure 2.2: Histograms of relative timings for key methods considered, normalized to B3LYP-D3BJ single points on the same molecule, using ORCA 4.0.1. Median relative times and median wall clock times for single-core runs are included for reference.

MP2 and coupled-cluster methods exhibit high computational complexity, much research ignored them for medium to large organic molecules due to the time required. Particularly in computational screening and conformer generation, fast molecular force fields such as MMFF94 and UFF, as well as semiempirical quantum chemical methods such as AM1,[119] PM3,[120] PM6,[121] and PM7[70] were considered "good enough" to generate structures for further refinement with density functional and other methods. More recent methods, particularly

the ANI machine learning methods and the GFN family of density functional tight binding appear to significantly improve on accuracy with only modest increases in the time required.

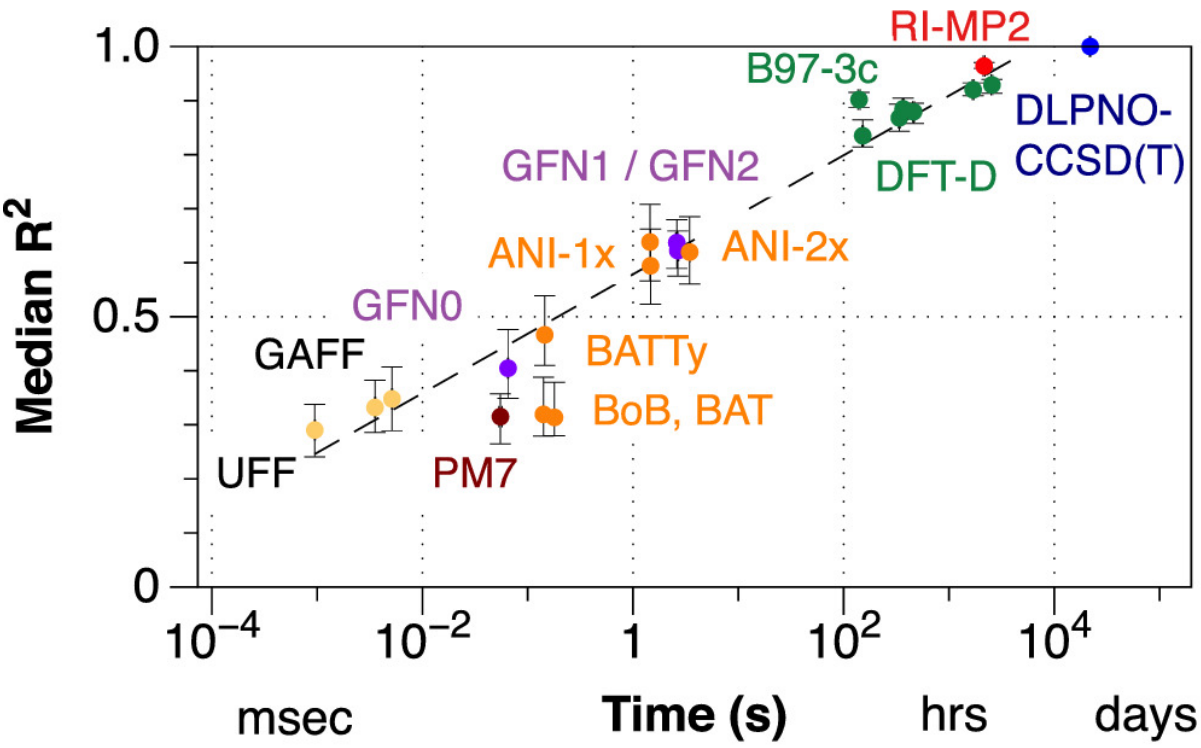

Figure 2.3: Comparison of single-core computational time required for energy evaluation (in log scale) to median R^2 found when compared to DLPNO-CCSD(T) energies. Error bars indicate 95% confidence intervals of time and medain R^2 from bootstrap sampling. Dashed line indicates approximate "best current method" threshold defined from force fields through RI-MP2 methods.

We find that consistent with common assumptions, even recent methods roughly adhere to the requirement of significant increases in computational (time) cost to gain increased thermochemical accuracy, as illustrated in Figure 2.3 with R^2. Similar trends are found for MARE and Spearman ρ metrics. Since multiple studies have demonstrated the need for accurate treatment of noncovalent interactions including intramolecular electrostatic and dispersion effects for understanding conformer relative energies, it is not surprising that this benchmark illustrates the significant accuracy advantage of modern dispersion-corrected density functional and wavefunction methods.

2.5.5 Use of Machine Learning Methods as Surrogates: ANI and Bag-of-Features

One possible solution to the trade-off between accuracy and computational cost would be the growing use of machine learning (ML) methods in chemistry, particularly as a surrogate for thermochemical parameters such as atomization energies.[5,22,122] Typically, these ML methods use deep neural networks (DNN) and have been trained to density functional calculations, particularly hybrid B3LYP or ωB97X atomization energies[123 6] although recent efforts have included training on coupled-cluster quality data as well.[?]

In principle, since the evaluation of the DNN is fast, the time required for the prediction of an ML method is dominated by the time to generate the input descriptors – still only a small fraction of that required for a quantum calculation. Therefore, if an ML method could reproduce density functional or coupled-cluster energies at semiempirical or force field computational cost, it would dramatically change the conventional accuracy/time tradeoff.

While evaluation of DNN methods would be significantly faster on processing units (GPUs), and may not be optimized for CPU evaluation, we note that many quantum chemistry methods are also accelerated on GPUs. Thus we retain the single-core CPU timings in Table 2.1 and Figure 2.3 but note that the actual speed of ML methods such as ANI would be faster when evaluated on a modern GPU.

ANI methods Table 2.1 and Figure 2.3 show the ANI family ML methods, ANI-1x, ANI-1ccx, and ANI-2x, performing similarly to GFN tight binding semiempirical methods in both accuracy and speed. ANI-1cxx outperforms the ANI-1x model that does not contain dispersion corrections while performing slightly better than the ANI-2 model. The inclusion of dispersion correction for DFT methods is clearly beneficial as they improve upon their non-dispersion corrected counterparts, as seen in Table 2.2.

In principal, it is possible to perform *post hoc* addition of a D3 dispersion correction to both ANI-1x and ANI-2x. Table 2.3 shows potentially improved performance over their

non-dispersion corrected counterparts, although the differences are not statistically significant. Moreover, since the D3 dispersion correction for ωB97X-D3 cannot be calculated by standard tools, applying such a *post hoc* correction is challenging. For our set, one could calculate the dispersion correction from the ωB97X-D3 calculations performed on the same molecule, but without such density functional calculations, applying dispersion correction would be impossible.

While the newer D4 correction[114,115] can be calculated using the DFTD4 program,[124] we find adding D4 corrections worsen the median R^2 and Spearman metrics, although again the differences are not statistically significant. The variance of applying D3 and D4 corrections to the ANI models illustrates the challenge in current machine learning methods. Since they inherently add some error on top of the underlying data used for training the model, use of coupled-cluster or other highly accurate dispersion-corrected training is needed.

Table 2.3: Comparison of post hoc dispersion correction for ANI machine learning methods. Values in brackets indicate 95% confidence intervals from bootstrap sampling.

Method	Median R^2						Median Spearman ρ					
	No Dispersion		D3		D4		No Dispersion		D3		D4	
ANI-1ccx	0.64	[0.57-0.7]	–		–		0.71	[0.64-0.77]	–		–	
ANI-1x	0.59	[0.52-0.66]	0.63	[0.57-0.71]	0.57	[0.48-0.67]	0.65	[0.57-0.72]	0.71	[0.65-0.75]	0.62	[0.56-0.71]
ANI-2x	0.62	[0.56-0.68]	0.66	[0.61-0.7]	0.6	[0.54-0.66]	0.69	[0.64-0.72]	0.71	[0.67-0.73]	0.66	[0.62-0.7]

Bag of Feature methods The performance of the bag-of-features models, while faster than the ANI symmetry function models, were more comparable to the accuracy of force field methods. The inclusion of additional information to the descriptor such as three and four-body interactions and atom typing were beneficial to the bag-of-features models, the accuracy pales in comparison to the ANI symmetry function models.

Standard bag-of-features have at minimum a bag of nuclear charges and a bag of two-body interactions as seen in BOB and further bags are added that contain additional information such as angles and torsions with BAT. This approach was taken for the BATTY representation with the modification of using minimal atom typing (i.e., sp, sp^2, sp^3 hybridization) to sort bags. Unlike other bag-of-features representations, the performance of BATTY was increased

by removing the bags of nuclear charges and excluding the nonbonding interactions from the two-body interactions bag to create a bag of simple bonds. Since relative conformer energies are strongly dominated by non-bonded interactions, this finding is surprising, although perhaps separating bonding and two-body non-bonded interactions facilitate ML training. A recent example, BAND-NN, took the approach of separating the bonding and nonbonding information similarly to classical force fields and finds an improvement in performance.[125]

ML commonly employs techniques to normalize the data, improving the model's training.[126,127] In this work, we used physically-motivated normalization techniques for the bag-of-features representations. Four molecular properties, the number of atoms, bonds, electrons, and the molecular mass, were chosen for normalizing the atomization energy. BATTY saw improvements in performance when normalizing by the number of atoms (i.e., BATTY/n) and the number of bonds (BATTY/b) across Spearman, R^2, and MARE. The other bag-of-feature representations experienced a slight improvement in R^2 when normalizing by the number of atoms but not an improvement in the MARE. Normalizing the atomization energy for bag-of-features methods does provide minor improvements, but not enough to compete with the ANI-1 and ANI-2 methods.

ML methods, despite training on density functional and coupled-cluster energies, are still not as accurate as conventional quantum methods for predicting conformer energies. At present, the ANI family is comparable to the semiempirical GFN methods for accuracy on this task.

Table 2.4: Effects of normalization descriptors on machine learning methods (e.g. BATTY/n refers to BATTY with number of atom normalization). Numbers in brackets indicate 95% confidence intervals for the median MARE, R^2, and Spearman ρ metrics.

Method	Normalization	MARE		R^2		Spearman ρ	
BOB	—	1.92	[1.73-2.16]	0.32	[0.27-0.39]	0.1	[0.0-0.2]
BOB	Atoms	1.94	[1.76-2.17]	0.36	[0.32-0.42]	0.1	[0.0-0.2]
BOB	Mass	2.2	[1.93-2.41]	0.32	[0.27-0.37]	0.1	[-0.1-1.0]
BOB	Electrons	2.06	[1.75-2.28]	0.32	[0.28-0.37]	0	[-0.1-1.0]
BOB	Bonds	5.09	[4.46-5.78]	0.27	[0.24-0.32]	0	[-0.1-1.0]
BAT	—	1.18	[1.05-1.3]	0.31	[0.28-0.37]	0.2	[0.1-0.3]
BAT	Atoms	1.36	[1.19-1.49]	0.34	[0.28-0.4]	0.1	[0.1-0.3]
BAT	Mass	1.4	[1.27-1.55]	0.31	[0.27-0.37]	0.2	[0.1-0.3]
BAT	Electrons	1.28	[1.16-1.45]	0.32	[0.28-0.38]	0.15	[0.1-0.3]
BAT	Bonds	1.62	[1.45-1.81]	0.35	[0.3-0.4]	0.1	[0.0-0.2]
BATTY	—	0.51	[0.47-0.6]	0.4	[0.34-0.44]	0.4	[0.3-0.5]
BATTY	Atoms	0.42	[0.38-0.48]	0.47	[0.4-0.54]	0.5	[0.4-0.6]
BATTY	Mass	0.69	[0.61-0.75]	0.41	[0.35-0.48]	0.4	[0.3-0.5]
BATTY	Electrons	0.63	[0.55-0.71]	0.42	[0.35-0.48]	0.4	[0.3-0.5]
BATTY	Bonds	0.42	[0.37-0.5]	0.48	[0.41-0.54]	0.5	[0.4-0.6]

2.6 Discussion

2.6.1 Effects of Conformer Energy Ranges on Accuracy Metrics

Previous work has suggested that the poor correlations found between force field and semiempirical methods are derived from the small number of low-energy conformers considered in this benchmark.[44] Certainly, one might imagine that when considering multiple geometries with only small differences in energies, random errors are magnified. Figure 2.4 illustrates a histogram of the ranges in DLPNO-CCSD(T) energies across the molecules considered. Despite the small ranges in energies, there is little correlation between the energy range of a molecule and the accuracy metrics of a particular method. This suggests no bias from the small energy windows used in this benchmark set.

Figure 2.5 indicates there is no correlation between R^2 and the energy window of

the conformers. The ML methods have a relatively even distribution of R^2 across the energy window indicating that random errors in the model may have more of an impact on performance than the size of the energy window.

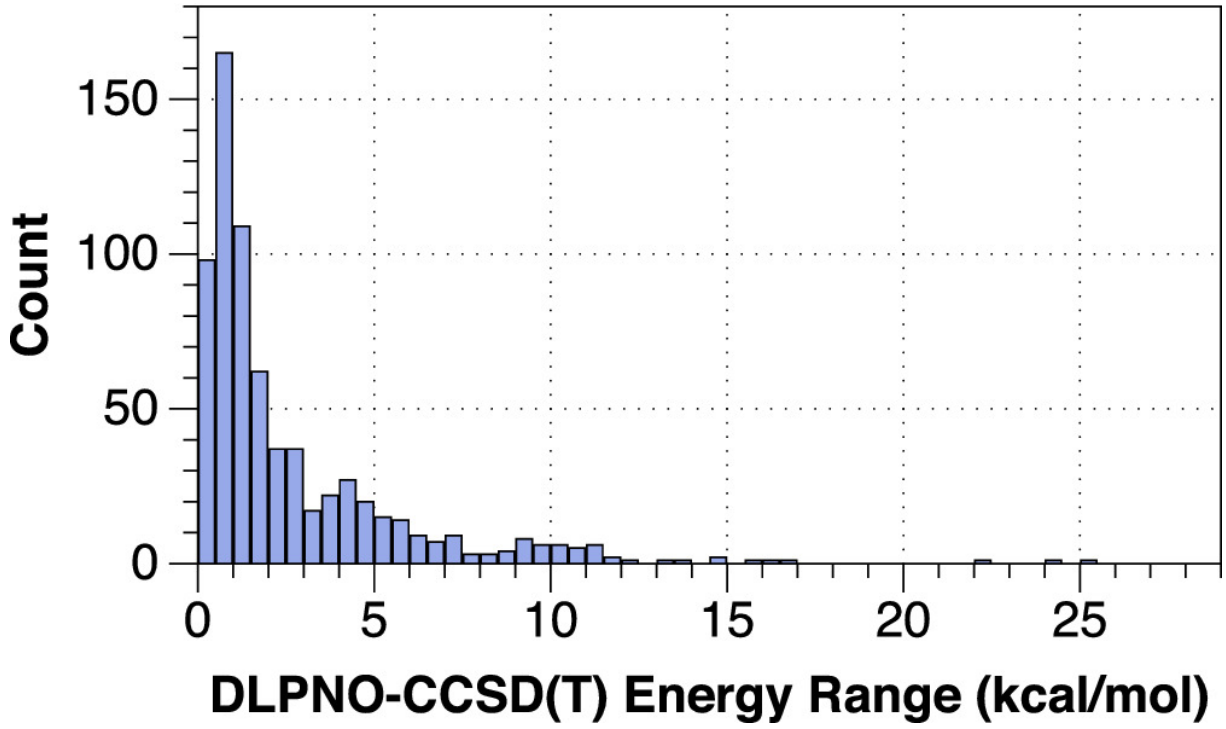

Figure 2.4: Histogram of relative DLPNO-CCSD(T) energy ranges across multiple conformers.

2.6.2 Connection Between Accuracy Metrics: MARE, R^2, Spearman

In principle, the mean absolute relative errors in energies (MARE) consider both random and systematic errors of a method, while the R^2 and Spearman correlation metrics remove systematic errors through linear correlation (R^2) or ranking (Spearman ρ). However, for the comparisons here, there is a strong connection between all three metrics, as illustrated in Figure 2.6. Methods with smaller MARE have almost a linear correlation with increased median R^2. The three classical force field methods have essentially the same median R^2 metric despite differences in MARE, likely due to systematic errors in the methods. Similarly, while increasing the data in the bag-of-features descriptors from BOB to BAT decreases the

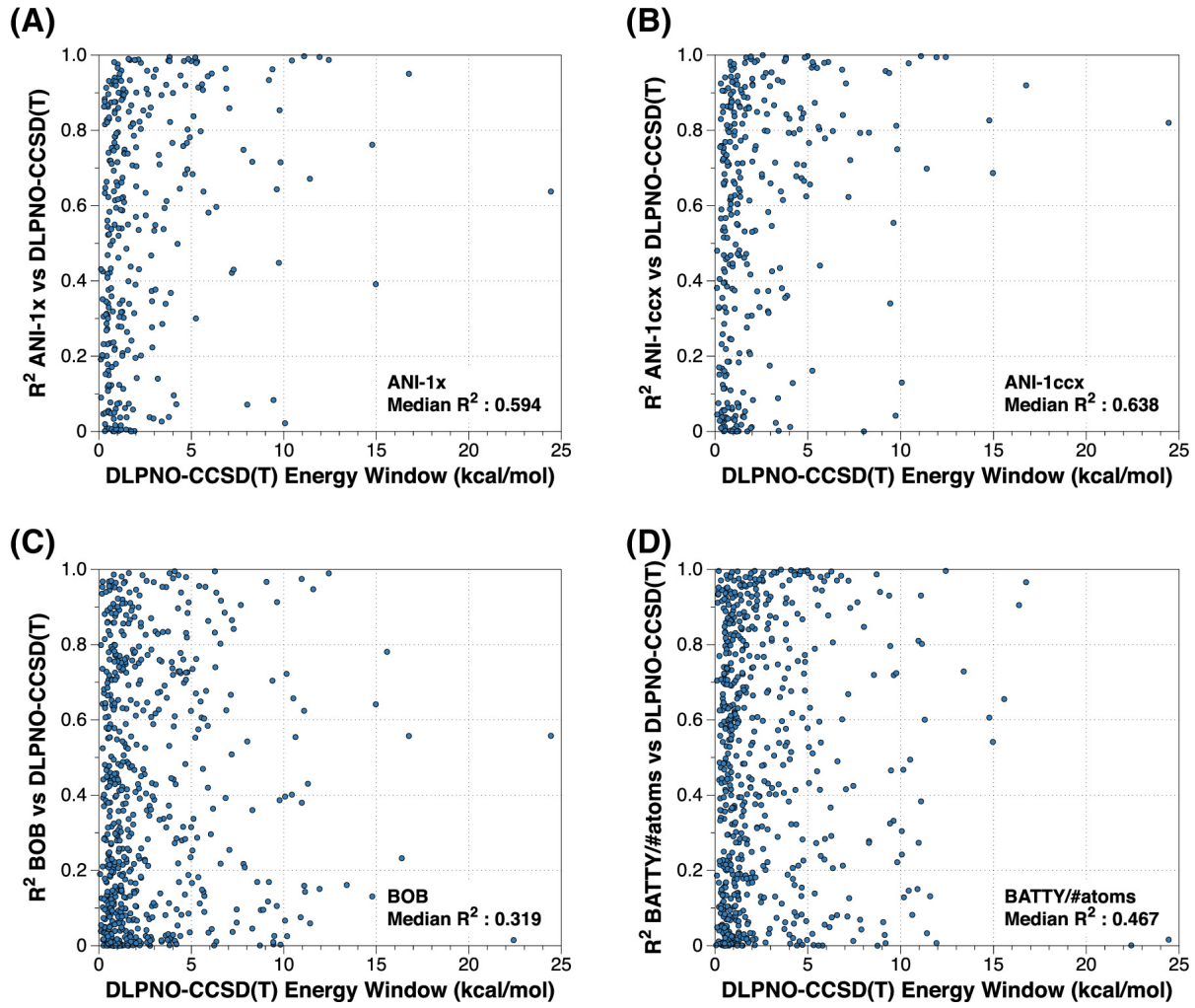

Figure 2.5: Examples of the relation of energy windows to R^2 for the ML methods (A) ANI-1x, (B) ANI-1ccx, (C) BOB, and (D) BATTY/# atoms.

median MARE from 1.92 kcal/mol to 1.18 kcal/mol, the accuracy as judged by the median R^2 remains essentially constant (0.31 and 0.32, respectively).

2.6.3 Dipole Moment Ranges

Since we generally find very small energy differences between the conformers considered in this work, one might wonder whether such differences have meaningful consequences. Due to Boltzmann statistics, many properties are dominated by the lowest energy geometry, even with

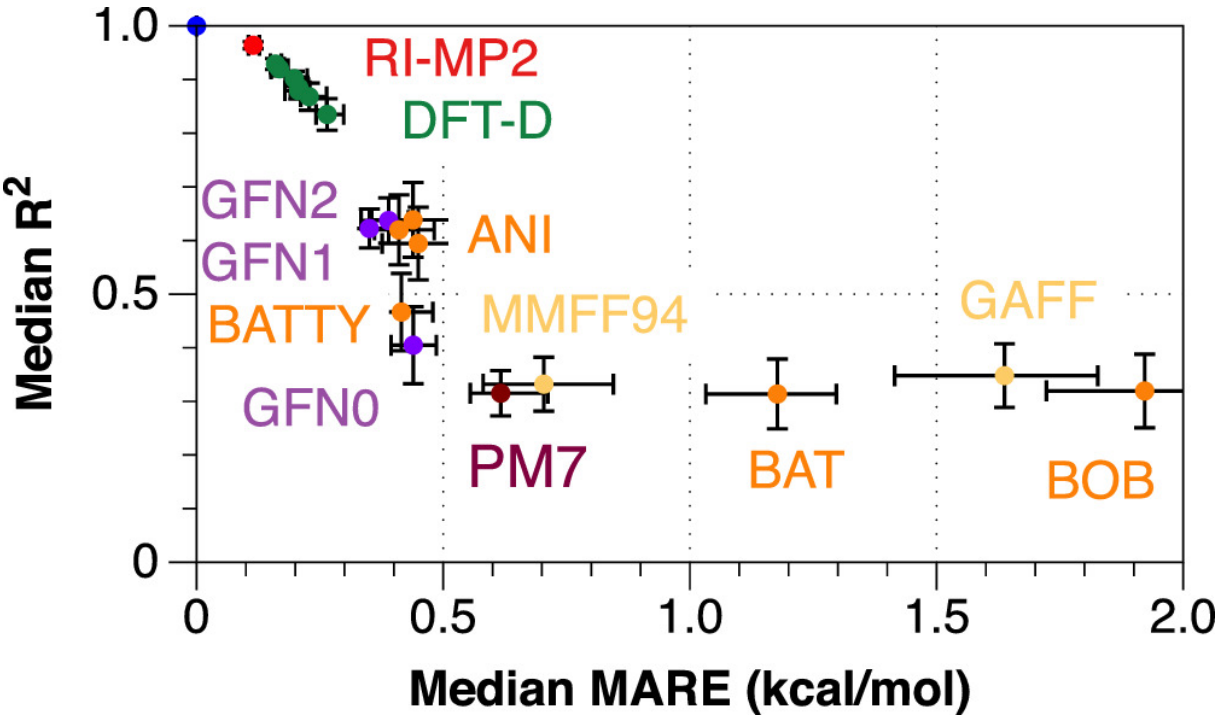

Figure 2.6: Correlation between mean absolute relative energies (MARE) and median R^2 correlation. Since the R^2 metric minimizes systematic errors, the high degree of correlation between the two metrics indicate most methods exhibit relatively random / non-systematic errors. Error bars indicate 95% confidence intervals from bootstrap sampling.

small energy windows to other geometries. One recent example comes from understanding the effects of dipole moments on solvent viscosity.[100] Finding all conformers with proper weighting is thus crucial to predicting the dipole moment of an ensemble of different conformers.

We find over the set of molecules considered, over 140 molecules have a range of 3 D or more, and 75 molecules have a range of 4 Debye or above across multiple conformers in the study. Figure 2.8 illustrates the example of `omegacsd_CNBPCT`, with two conformers that are close in energy yet span dramatically different dipole moments. Using B3LYP-D3BJ (TZ), the computed dipole moments range from 1.41D to 9.78D. The molecule contains two carbonyl bonds, either parallel (high dipole moment) or anti-parallel (low dipole moment) depending on the rotation of several bonds and the more polar conformer is predicted to be more stable by B3LYP-D3BJ, possibly due to an intramolecular hydrogen bond. On the

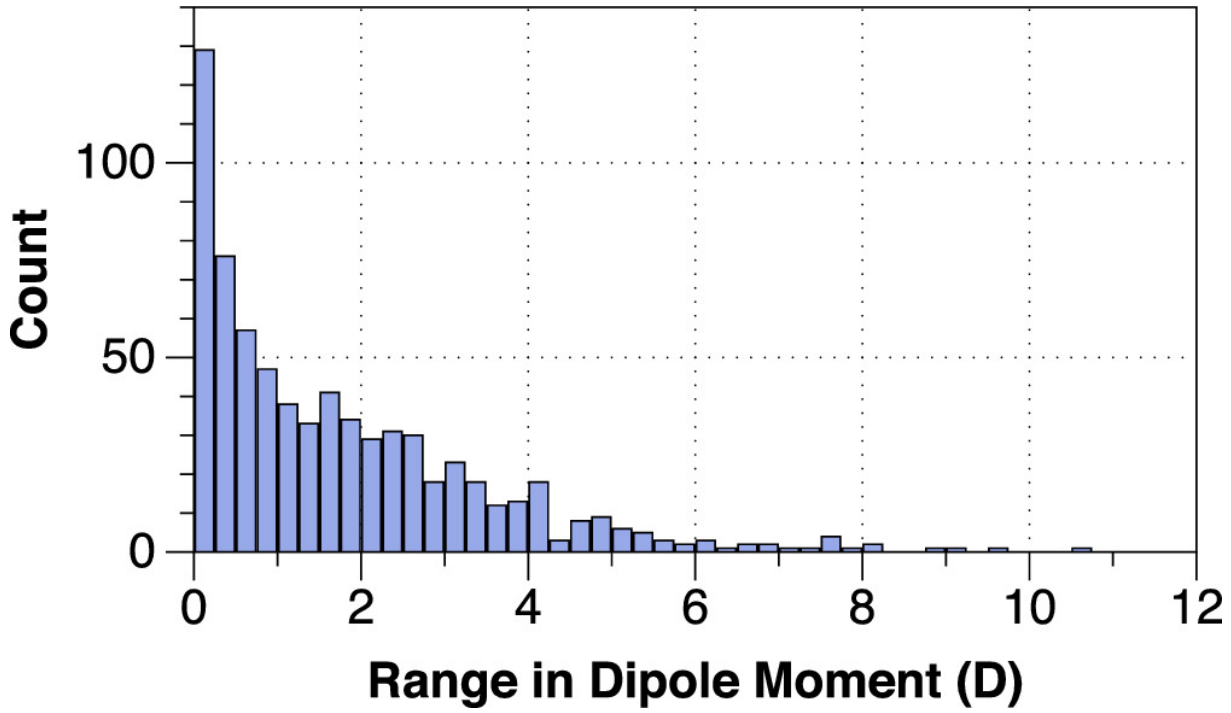

Figure 2.7: Histogram of the range of B3LYP-computed dipole moments in Debye across the conformers considered in this work. While most molecules show only small differences in polarity across conformers, many have over 3-4 Debye ranges.

other hand, using DLPNO-CCSD(T) cc-pVTZ, the conformers differ by only 0.3 kcal/mol, with the anti-parallel, less polar conformer more stable than the other.

Such polarity differences are examples in which small differences in conformer energies can have significant effects on molecular properties. Since experimental properties reflect a Boltzmann-weighted average of multiple thermally accessible conformers, even small differences in conformer energies have large effects on populations involved in property prediction, as recently discussed with conformer and polarity effects on solvent viscosity.[100]

Machine Learning Batch Evaluation An advantage for ML and force field predictions is the ability to batch evaluate by loading all conformers of a molecule at once and evaluating them as a batch opposed to evaluating one at a time, as with conventional quantum chemistry methods. Table 2.5 indicates the median sequential times from Table 2.1 and median time

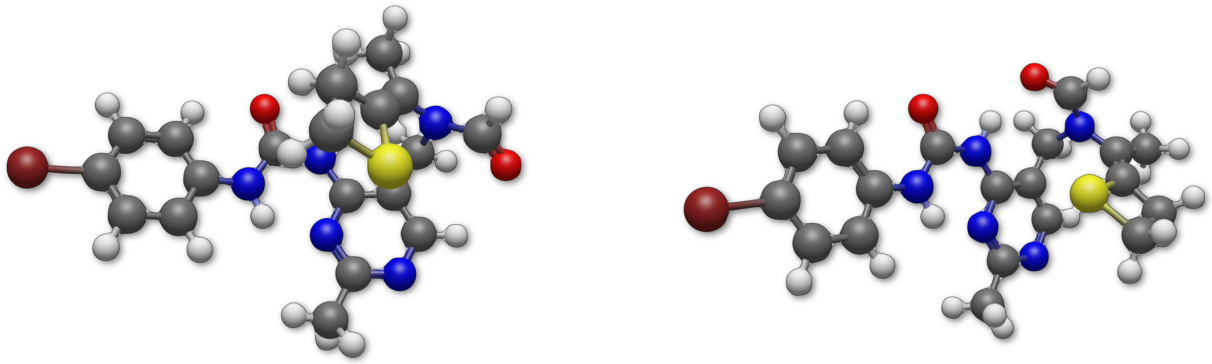

Figure 2.8: Example of conformational diversity in dipole moment in the molecule omegacsd_CNBPCT reflecting anti-parallel carbonyl (*left* - rmsd45) or parallel carbonyl groups (*right* - rmsd92), with B3LYP-D3BJ def2-TZVP computed dipole moments ranging from 1.41D to 9.78D, respectively. The two geometries differ by only 1.3 kcal/mol at the B3LYP-D3BJ def2-TZVP level, with the more polar conformer (right) stabilized by an intramolecular hydrogen bond. Using DLPNO-CCSD(T) cc-pVTZ, the less polar conformer (left) is more stable by 0.3 kcal/mol.

per single point in batch evaluation. Speedups range ~70-170 times faster for both force field and ANI methods. We note that while the ANI methods improve performance in batch evaluation, traditional force field methods do as well, with similar speedups.

Table 2.5: Comparison of single-core median sequential time to median batch time (in seconds), and relative speedups for batch evaluation.

Method	Median Time		Median Batch Time		Speedup
MMFF94	0.0	4.4e-05	5.05e-05	6.02e-07	70.89
GAFF	0.01	5.73e-05	3.29e-05	4.49e-07	160.64
UFF	0.0	8.61e-06	4.32e-05	4.46e-07	21.88
ANI-1x	1.46	0.0	0.01	0.0	113.15
ANI-1ccx	1.45	0.0	0.01	0.0	111.5
ANI-2x	3.45	0.01	0.02	9.7e-05	172.44

2.7 Conclusions

The current work extends previous efforts to consider the accuracy of modern computational chemistry methods to rank the energies of drug-like conformers. Since such energy differences are small, this poses a challenging benchmark even for density functional methods. Use of dispersion-corrections for density functionals are required – the slim time required is offset with dramatically increased accuracy. While triple-zeta and larger basis sets also provide higher accuracy, likely because of better treatment of non-covalent interactions, the large number of possible conformers forces trade-offs in accuracy and computational time required.

Current ML methods show great promise, particularly the ANI-1ccx method trained in part on coupled-cluster energies,[?] since they provide accuracy comparable to the semiempirical GFN2 method and can be performed in batch and accelerated on GPUs. Despite claims of reaching and exceeding DFT accuracy, we do not find these methods yet meet the accuracy of modern dispersion-corrected methods. Nevertheless, we expect these methods will provide increased accuracy in the future. An important caveat is the need to train on accurate data, such as dispersion-corrected density functional, MP2, or coupled-cluster calculations.

We expect continued improvement from other methods, particularly multiple efforts to improve classical force fields,[128–131] inclusion of polarizable atomic charges,[132–139] novel force fields from experimental data, density functional and other quantum methods,[140–145] and continued development of approximate semiempirical quantum methods.[74]

At present, we can highly recommend methods at each tier of the accuracy-time tradeoff, particularly the recent GFN2 semiempirical method, the B97-3c density functional approximation, and RI-MP2 for accurate conformer energies. Previous efforts to use a hierarchy of methods are still useful, for example, the use of GFN2 methods to refine initial conformer ensembles, followed by refinement of a smaller set of low-energy geometries with more accurate methods. Batch evaluation with ANI methods are also efficient, although they do not yet span the range of elements supported by semiempirical methods such as GFN2 or density

functional methods.

The current benchmark reflects conformational preferences in a vacuum as judged by enthalpy differences alone. Since free energy differences drive experimental conformers, introducing entropic considerations will be needed for further work.[52] Moreover, much chemistry is performed in solution, thus work on understanding conformer energies in solvation is also critical.[146,147]

3.0 Evaluation of Thermochemical Machine Learning for Potential Energy Curves and Geometry Optimization

This chapter is adapted from:

> Folmsbee, D., Koes, D., Hutchison, G. Evaluation of Thermochemical Machine Learning for Potential Energy Curves and Geometry Optimization. *ChemRxiv:13151069.v1* **2020**. DOI: 10.26434/chemrxiv.13029437.v1

which is also published as:

> Folmsbee, D., Koes, D., Hutchison, G. Evaluation of Thermochemical Machine Learning for Potential Energy Curves and Geometry Optimization. *J. Phys. Chem. A.* **2021**, 2007486. DOI: 10.1021/acs.jpca.0c10147.

It is a collaborative effort in which I tested the machine learning models and carried out the data analysis, generated the figures, and wrote the bulk of the manuscript; D. K. provided the libmolgrid-based convolutional neural net method and helped with manuscript edits; G. H. conceived and directed the project.

3.1 Summary

While many machine learning methods, particularly deep neural networks, have been trained for density functional and quantum chemical energies and properties, the vast majority of these methods focus on single-point energies. In principle, such ML methods, once trained, offer thermochemical accuracy on par with density functional and wave function methods but at speeds comparable to traditional force fields or approximate semiempirical methods. So far,

most efforts have focused on optimized equilibrium single-point energies and properties. In this work, we evaluate the accuracy of several leading ML methods across a range of bond potential energy curves and torsional potentials. Methods were trained on the existing ANI-1 training set, calculated using the ωB97X / 6-31G(d) single points at non-equilibrium geometries. We find that across a range of small molecules, several methods offer both qualitative accuracy (e.g., correct minima, both repulsive and attractive bond regions, anharmonic shape, and single minima) and quantitative accuracy in terms of the mean absolute percent error near the minima. At the moment, ANI-2x, FCHL, and a new libmolgrid-based convolutional neural net show good performance.

3.2 Introduction

Machine learning (ML) methods have been proposed as surrogates for time-consuming quantum mechanical calculations, such as density functional and first-principles methods, for their rapid prediction potential once trained[4–14]. For ML to be a successful surrogate, the methods need to be able to perform property predictions adequately for optimized geometries, capture not just the well of the potential energy curve but also the anharmonicity that force field methods fail to capture, and appropriately handle multiple conformations of the same molecule.

Numerous studies have shown the proficiency of ML methods to predict thermochemical parameters at already optimized geometries utilizing various types of representations and neural network structures[5,34]. Early representations, such as Coulomb Matrix[22] and bag-of-features[23,24], demonstrated success in property predictions with further iterations of representations such as FCHL[28,29] continuing to improve the property prediction at optimized geometries. These ML methods are typically trained on the QM7[22,32] or QM9[33,148] data sets consisting of optimized molecules with up to 7 or 9 heavy atoms respectively and help to

demonstrate ML's potential as a surrogate.

Additional deep neural network (DNN) methods, like ANI[6–8,27] and BAND NN[36], used training data beyond optimized single points to better evaluate the potential surface for dynamics and geometry optimizations. These methods utilize the ANI-1 data set[35], or ANI-2 data set in the case of ANI-2x, for training as they contain both equilibrium and non-equilibrium structures of up to eight heavy atoms containing H, C, N, and O with the non-equilibrium structures being generated from normal-mode sampling. The training set for ANI-2x adds the additional elements of F, Cl, and S while providing additional torsion sampling data.[8] The BAND NN model uses a subset of the ANI-1 data set with only non-equilibrium geometries with energies within 30 kcal/mol of the equilibrium energy. Although these methods have been shown to perform adequately in their respective papers, the range for bond stretch applications has been limited to the harmonic portion of the potential energy curve, rarely examining the potential energy curves further from equilibrium.

Recent work has expanded the knowledge on ML performance for predicting and ranking thermally accessible conformations[149]. Though ML was not tasked with large bond stretches as in this work, the ability of ML methods to rank conformational energy was only comparable to that of semiempirical methods. While this is not equivalent to the accuracy of density functional (DFT) or *ab initio* electronic structure methods, for ML methods to be an accurate surrogate for quantum chemical methods, continued advancements in ML models and training sets are needed to provide further performance improvements.

For ML to become a viable replacement for current methods, ML needs to achieve optimized geometries and predict properties without relying on force field (FF) methods. Most FFs have been refined for small molecules and biomolecules and can struggle with non-covalent and steric interactions for applications such as conjugated polymers. While these issues can be lessened with specific parameterization[150,151], geometries of FFs generally can be less than ideal[152]. ML trained on higher levels of theory ideally captures these non-covalent interactions and provides better initial optimized geometries.

With the rapid adoption of ML, there has been a growing desire to use ML in molecular dynamics (MD) applications to provide more accurate simulations than FFs at a much lower cost than time-consuming quantum mechanical calculations[149]. For ML to be reliable, it needs to properly predict geometric changes that occur in MD simulations from non-equilibrium bond stretching to torsional barriers. This work seeks to examine how well the current state of ML performs at these tasks, as well as to display the methods' understanding of chemical physics to help decide key needs for ML to improve as a surrogate for computationally expensive quantum calculations.

3.3 Methods

3.3.1 Molecules

A mixture of small and large molecules was chosen to evaluate ML performance on potential energy surfaces for a total of 17 bond stretches and 5 dihedral scans. The molecules examined were benzene (C-C and C-H stretching), methanol, methane, CO, H_2, ethylene, water, acetylene, hydrogen cyanide, N_2, ammonia, biphenyl, aspartame, sucrose, dialanine, and diglycine. Bond stretches were evaluated every 0.1Å while dihedrals were evaluated every 20° with the exception of biphenyl which was every 15°.

3.3.2 Computational Methods

The reference method, ωB97X[153], was performed using Orca 4.0.1[154] while the force field calculations, MMFF94[? ? ? ? ?] and GAFF[155], were performed using Open Babel version 3.0[156].

Machine learning methods and representations included the pre-trained models ANI-

1x[6,7], ANI-2x[8], BAND-NN[36], as well as FCHL[157], Bag of Bonds (BOB)[88], and Extended Connectivity Fingerprints (ECFP)[158,159]. Scikit-learn[89] was used for kernel ridge regression (KRR) and bayesian ridge regression (BRR) for BOB and random forest regression (RFR) with BOB and ECFP representations while FCHL used the custom KRR in QML.

We also trained a deep convolutional neural network (Colorful CNN), an approach that has been successfully used in protein-ligand binding affinity prediction[160,161]. The input molecule is represented as a voxelized grid of atomic densities as generated by the libmolgrid library[162]. Our network has six modules separated by pooling operations each with seven convolutional layers and was trained on the ANI-1x data set[163]. The trained Colorful CNN model can be found at `https://github.com/hutchisonlab/ml-benchmark`.

Due to method scaling efficiency for memory usage, a subset of the ANI-1 data set was taken for training representations using BOB/KRR and BOB/BRR. For consistency, ECFP/RFR and BOB/RFR were additionally trained on this subset. The subset consists of 5 non-equilibrium geometries for every molecule with up to 7 heavy atoms, as well as 5 non-equilibrium geometries for half of the molecules with 8 heavy atoms, to create a training set consisting of 33,496 molecules and 167,480 non-equilibrium geometries. All molecules from the test set were removed from the training set. This training set was additionally used for BOB/RFR and ECFP/RFR. An additional subset of the first 5000 non-equilibrium geometries was used for FCHL/KRR. Increasing the training set for FCHL/KRR had a negative impact on prediction performance so our results are with the model trained on 1000 different molecules for a total of 5000 non-equilibrium geometries.

3.4 Results and Discussion

To illustrate the qualitative performance of potential energy surface predictions, we analyzed both small and larger molecules outside of the ANI-1 data set used for training for each ML

method. We wish to focus on how the methods perform not only around the bond length at the energy minima, r_0, but also in the attractive and repulsive regimes to gain a better understanding of how ML methods would behave if given less ideal starting geometries for a task such as geometry optimization.

Table 3.1: Overview of machine learning performance sorted by median mean absolute percent error (MAPE).

Methods	**Median MAPE[1]**	**r_0[2]**	**Repulsive Wall[3]**	**Attractive Forces[4]**	**Minima after 2Å[5]**
ωB97X 6-31G(d)	0	17	17	17	0
ANI-2x	0.002	17	13	17	12
BOB/BRR	0.227	0	5	5	9
FCHL/KRR	0.255	10	16	15	13
Colorful CNN	0.2555	16	17	17	13
ANI-1x	0.265	16	11	17	5
BOB/KRR	0.313	1	9	11	13
BOB/RFR	43.881	2	3	0	8
BAND-NN	99.310	11	9	5	5
MMFF94	100.050	14	17	0	0
GAFF	100.133	13	17	0	0
ECFP/RFR	193.370	0	0	0	0

[1] Median mean absolute percent error over all 17 molecules from $r_0 \pm 0.25$Å.
[2] The number of molecules in which the lowest predicted energy point matches DFT.
[3] The number of times the method predicted a repulsive wall as the bond was compressed.
[4] The number of times the method predicted anharmonic attractive forces after r_0.
[5] The number of molecules predicted to have a local or global minima after 2Å.
[6] BAND-NN regularly would not predict energies for geometries with a bond stretch of 2Å or greater.

Each ML method was evaluated on the criteria demonstrated in Table 3.1 for bond stretches. The median mean absolute percent error (MAPE) was calculated from the energy values ranging from $r_0 \pm 0.25$Å for the molecules to determine how accurate and precise the ML predicted energies are. Since the ANI-1 training set samples harmonic displacements around the r_0 (e.g., Figure B.1) this range corresponds mostly to interpolation. Comparisons for repulsive short-range and attractive long-range interactions – extrapolations outside the training range are compiled in Table B.1. The r_0 evaluation criteria considered whether the method correctly predicted the DFT equilibrium bond length to be the lowest energy bond length. Additional evaluation criteria included the qualitative prediction of a repulsive wall, anharmonic long range interactions, and if there were incidences of additional minima past

2Å.

While methods like BOB/BRR and BOB/KRR had the second and fifth-lowest median MAPE, their ability to predict the geometry with the lowest energy, a repulsive wall, and attractive forces was quite poor compared to the other top methods based on MAPE. Other methods utilizing RFR also performed poorly, often predicting stepwise energy surfaces seen in Figure 3.1, thus being incapable of consistently predicting r_0, attractive, or repulsive forces. This is seen in Figure 3.1b when the bond breaking causes the only change in the ECFP representation and leads to the higher energy. Other ML methods such as ANI-1x, ANI-2x, FCHL, and Colorful CNN were able to accurately predict energies while also predicting the repulsive and attractive forces of the molecule. In short, while random forest methods may have accuracy at single-point properties, they prove inherently inaccurate for potential energy and should be avoided.

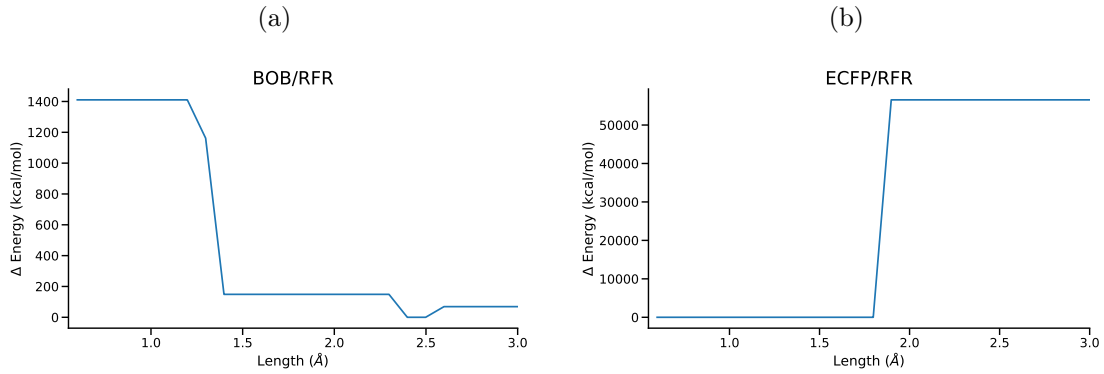

Figure 3.1: N_2 potential energy curves for ML methods utilizing random forest regression for predictions using (a) BOB and (b) ECFP for the ML descriptors.

A possible advantage for the ANI-1x and ANI-2x models is that some molecules in our test evaluation are found in the ANI-1x training set. In the training of the other methods, molecules in our test set were purposefully left out of the training set but may be present in the ANI-1x and ANI-2x model. For that reason, we will focus the remainder of our discussion on molecules outside of the ANI-1 training set, examining the best overall performers, ANI-1x, ANI-2x, FCHL, and Colorful CNN from Table 3.1. The performance of all methods is included

in the supplemental information.

Figure 3.2a displays the performance of ANI-1x, ANI-2x, Colorful CNN, and FCHL on the N-N bond stretch of N_2. While each of these ML methods predicts the correct r_0, there are issues in the prediction of the potential energy curve. ANI1-x, ANI-2x, and Colorful CNN fail to accurately depict the repulsive region with ANI-2x lowering in energy as the bond was compressed to 0.6Å. FCHL depicts the repulsive wall but inaccurately predicts the energy as the bond is compressed. All four methods accurately determined the attractive forces to about 2Å with ANI-2x matching ωB97X to 2.25Å.

The H-H stretch of H_2 in Figure 3.2b indicates one possible issue for ML. All four methods performed poorly with ANI-2x being the only method to obtain the correct r_0. This performance is likely due to the absence of H-H bonding data within the training set. H_2, while a unique bond, demonstrates the need to be careful when applying ML to molecules or chemistry completely outside the scope of the training set.

Figure 3.2c and 3.2d demonstrate the prediction capability of these ML methods on bond stretches for molecules larger than the training set. FCHL was only able to accurately capture the shape of the potential energy curve for dialanine, failing to capture the well of the potential energy curve for aspartame, perhaps from the difficulties training the entire ANI-1 set. ANI-1x, ANI-2x, and Colorful CNN retain both repulsive and attractive information while having accurate energies to that of ωB97X for both aspartame and dialanine. These methods do continue to exhibit difficulty in accurately predicting bond compression under 1Å as well as bond stretching after 2Å.

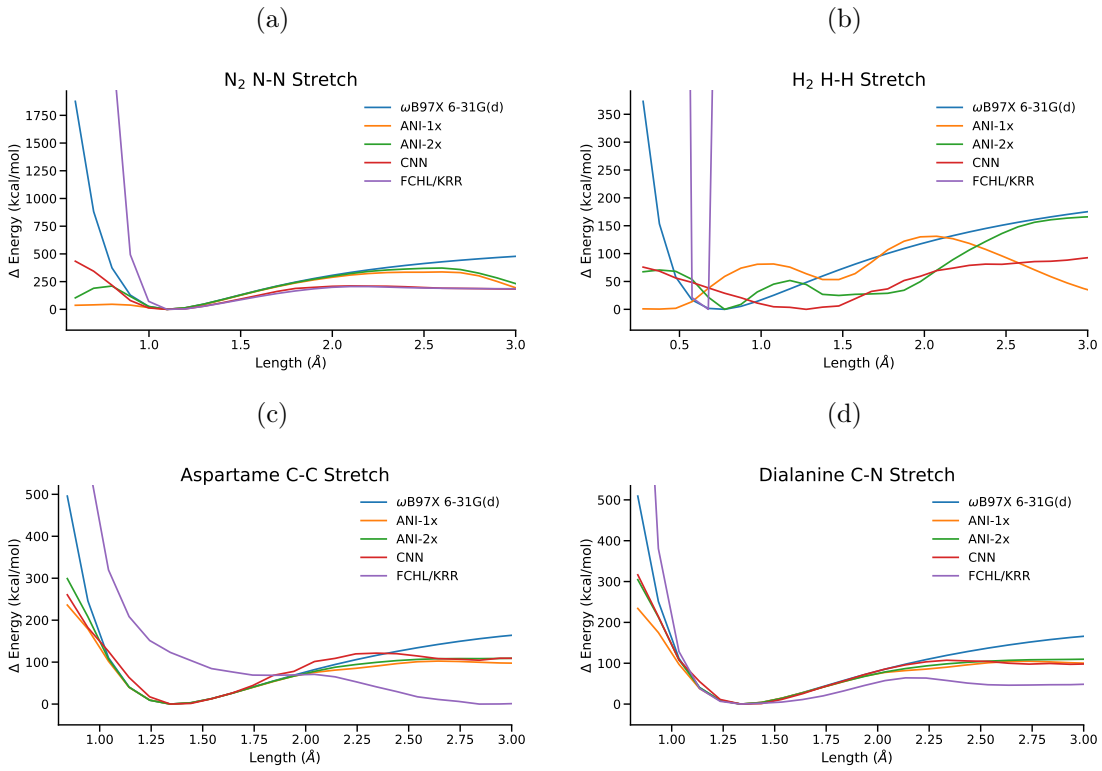

Figure 3.2: Bond stretch potential energy curves for (a) N_2, (b) H_2, (c) aspartame, (d) dialanine using total SCF energies in kcal/mol.

For bond stretches, ANI-1x, ANI-2x, Colorful CNN, and FCHL models show promise with initial training indicating these methods can accurately predict the bottom of the potential energy well. While force fields such as MMFF94 or GAFF can be used to obtain optimized geometries near this regime, ultimately ML methods should exhibit accuracy not only at single-point energy evaluation tasks, but at qualitatively and quantitatively accurate potential energy curves. Further training on long-range attractive forces might enable ML models to evaluate non-covalent interactions.

As an example, further evaluations were carried out on energy predictions from frozen-rotor dihedral angle scans performed with ωB97X 6-31G(d) for biphenyl and sucrose. Table 3.2 compiles the predicted lowest energy angle for these molecules as well as the barrier energies from −45° to 0° for biphenyl and 0° to −60° for sucrose.

ANI-1x and ANI-2x properly predict the lowest energy angle for biphenyl while Colorful

CNN predicts −45° to be a local, but not global, minima. FCHL improperly predicts rotation energies as seen in Figure 3.3a, predicting 0°, 180°, and −180° to be the lowest energy dihedrals. All of the methods over-predicted the height of the energy barrier for biphenyl.

For sucrose, all four methods correctly predicted the lowest energy angle. ANI-1x best captures the energy of the dihedral angles, seen in Figure 3.3b, with ANI-2x and Colorful CNN under-predicting the energy for most angles. Unlike with biphenyl, FCHL captures the shape of the torsion scan for sucrose but vastly over predicts the energies at each angle.

Table 3.2: The ML prediction of θ_0 and the barrier energy between the lowest and highest energy dihedrals for biphenyl and sucrose compared to the reference ωB97X 6-31G(d) method.

Methods	**Biphenyl**		**Sucrose**	
	θ_0 (°)	**Barrier Energy (kcal/mol)**	θ_0 (°)	**Barrier Energy (kcal/mol)**
ωB97X 6-31G(d)	-45	3.54	0	2.45×10^3
ANI-1x	-45	3.95	0	2.50×10^3
ANI-2x	-45	4.16	0	1.93×10^3
Colorful CNN	-135	5.49	0	9.46×10^2
FCHL/KRR	180	5.52	0	9.73×10^4

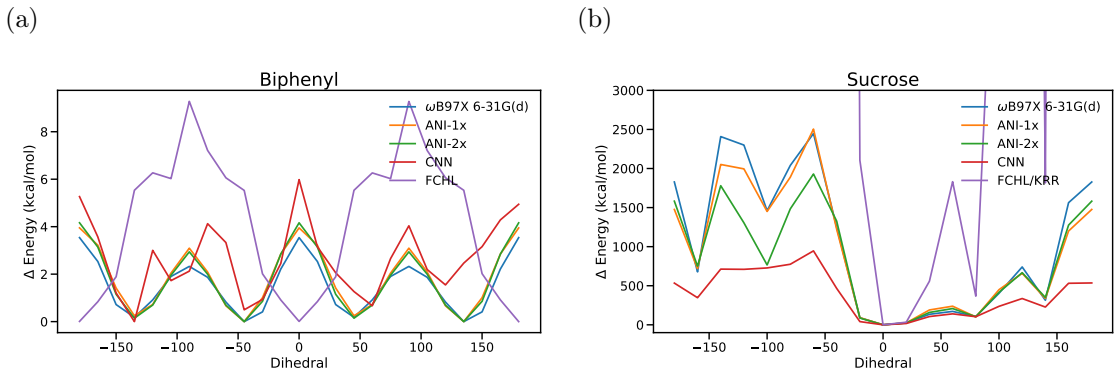

Figure 3.3: Dihedral energy predictions for (a) biphenyl and (b) sucrose in kcal/mol.

Dihedral scans demonstrate how small conformational changes in the molecule can affect the potential energy surface. The 2D torsion scans in Figures 3.4 and 3.5 compare ML performance to that of ωB97X and FFs, MMFF94 and GAFF. ANI-1x, ANI-2x, and Colorful CNN retain the resolution of some of the higher energy ϕ and ψ between −100° to 100°

while FCHL predicts these to be lower energy conformations similar to both FF methods. In lower energy conformations both BAND and BOB/KRR methods over-estimate these energy differences.

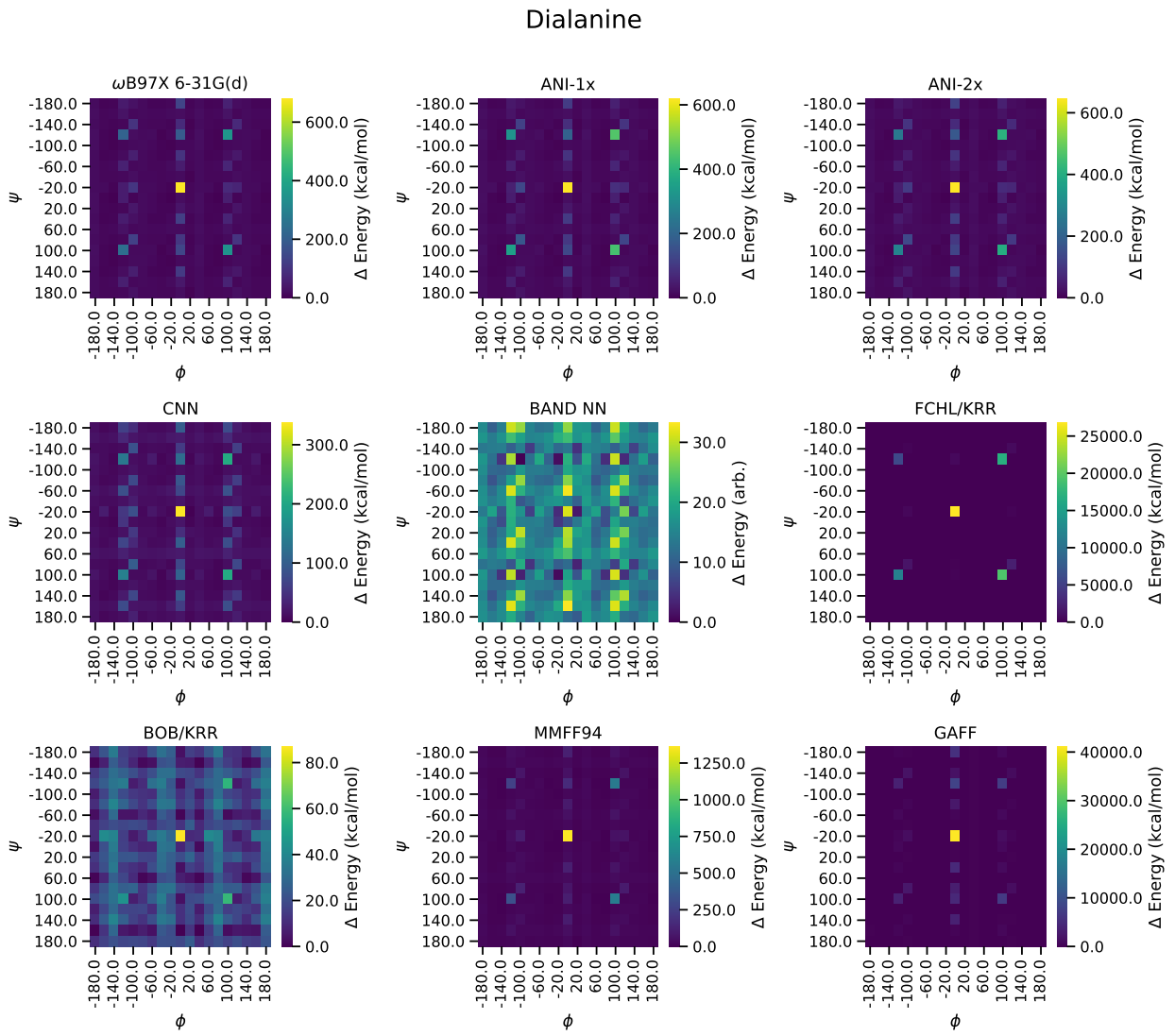

Figure 3.4: 2D torsion scans of dialanine in kcal/mol unless otherwise stated. Methods were tested at the geometries obtained with ωB97X 6-31G(d) from the torsion scan. Note that color schemes differ, due to large differences in energy scales.

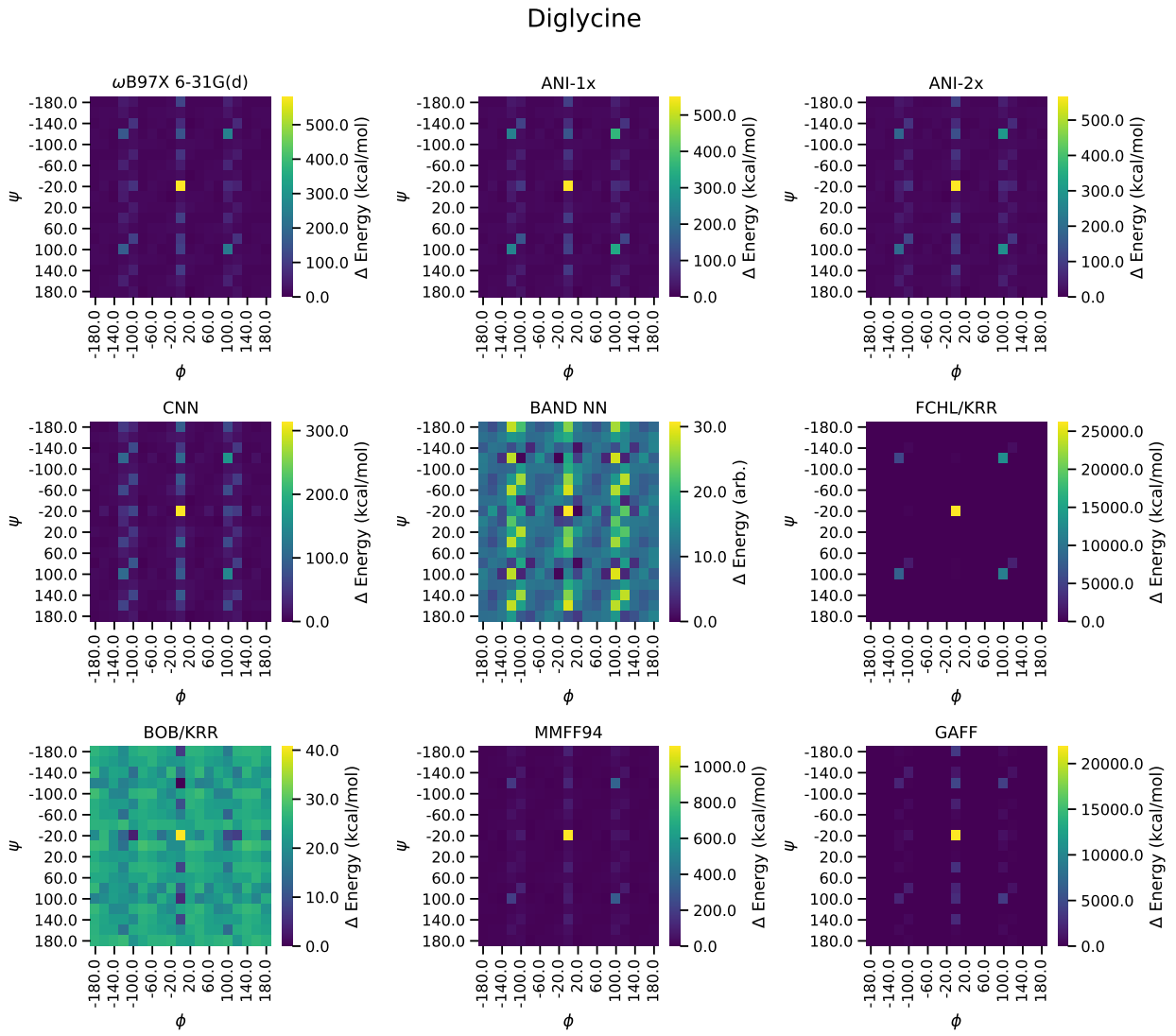

Figure 3.5: 2D torsion scans of diglycine in kcal/mol unless otherwise stated. Methods were tested at the geometries obtained with ωB97X 6-31G(d) from the torsion scan. Note that color schemes differ, due to large differences in energy scales.

The additional torsion training in ANI-2x provided a beneficial reduction in the MAE for both dialanine and diglycine, seen in Table 3.3, by roughly 35% from ANI-1x. Additional torsion sampling for methods Colorful CNN and FCHL should also provide a decrease in MAE for predicting dihedral angle energies. This could improve accuracy for the Colorful CNN method that is already qualitatively adequate.

A prevailing pitfall of ML methods stems from the training set. At the end of the day,

Table 3.3: Mean absolute error (MAE) in kcal/mol of 2D torsion scans for the top performing methods.

Methods	**Dialanine MAE ΔEnergy (kcal/mol)**	**Diglycine MAE ΔEnergy (kcal/mol)**
ANI-2x	1.89	1.71
ANI-1x	3.01	2.52
Colorful CNN	7.10	6.07
FCHL/KRR	252.17	200.86

the machine learning method is only as good as the training set. As seen with H_2, models struggle with chemical motifs outside of the training set. Current ML training sets largely consist of a subset of the molecules generated in the GDB-17[148] set, typically containing at least H, C, O, and N. While these training sets are a noble starting point for covering small organic molecules, they lack a diversity of atom species needed for applications such as protein binding and DNA sequencing. Additional data sets such as PubChemQC[164] could help to further expand the snapshot of chemical space ML methods are trained on.

3.5 Conclusions

Much work has focused on the use of machine learning methods as surrogates for computationally intensive density functional and quantum chemical methods. Often such efforts train and test on single-point energies of optimized structures. An important step is to evaluate ML methods across potential energy curves and surfaces for tasks such as geometry optimization.

ML methods such as ANI-2x, Colorful CNN, and FCHL perform decently near the well of the potential energy curve while struggling to properly predict repulsive regions and particularly long-range attractive forces. While this poor performance outside the domain of the training set is expected, these methods show promise with further improvements through the addition of stretched bonds in training data helping to improve model performance in this area. Increased torsion sampling for training ANI-2x improved the model's performance

over ANI-1x and should provide improvements for models like Colorful CNN and FCHL.

In general, there is still the issue of applying ML to the prediction of molecules too far outside the scope of the training set. The inclusion of additional elements and an increase in diversity of molecules in the training set from diverse data sets such as PubchemQC should alleviate some of these challenges.

We acknowledge the National Science Foundation (CHE-1800435) for support and the University of Pittsburgh Center for Research Computing through the computational resources provided.

4.0 Systematic Comparison of Experimental Crystallographic Geometries and Gas-Phase Computed Conformers

This chapter is adapted from a manuscript in preparation for submission; it is a work in progress. It is a collaborative effort with G. H. in which I performed data analysis on the torsions, generated the figures, and wrote large portions of the manuscript; G. H. performed GFN2 calculations, wrote part of the manuscript, and conceived and directed the project.

4.1 Summary

We have performed exhaustive torsion sampling on over 3 million compounds using the GFN2 method to compare potential bias between the crystallographic and gas-phase geometries. Many conformer sampling methods obtain torsional angle distributions from experimental crystallographic data, limiting the torsion preferences to molecules that must be stable, synthetically accessible, and able to be crystallized. In this work, we evaluate the differences in torsional preferences of experimental crystallographic geometries and gas-phase computed conformers to determine whether torsional angle distributions obtained from semi-empirical methods are suitable for conformer sampling. We find that differences in torsion preferences can be mostly attributed to crystallographic and gas-phase geometry differences or lack of available experimental crystallographic data. GFN2 demonstrated the ability to provide accurate and reliable torsional preferences that could provide a basis for a quantum-based ETKDG alternative method, QTDG, that does not rely on data from experimental crystal

structure elucidation.

4.2 Introduction

Most molecules exhibit some level of conformational flexibility – the existence of multiple low-energy geometries that differ mostly by changes in the torsional angles of both acyclic and ring bonds. Many methods have been developed to sample conformations, although benchmarks frequently focus on finding one geometry close to an experimental crystal structure. Consequently, most conformer sampling methods derive torsional angle distributions from experimental crystallographic data — not only to provide geometries close to such benchmarks but also as large diverse repositories of "ground truth" geometric properties such as bond lengths, angles, and dihedrals.

The challenge is that experimental crystallographic data is limited by the size of the data source and reflects some inherent biases. In order to be collected, the molecules must be stable, synthetically accessible, and were actually made and crystallized. While new cryo-electron microscopy (cryo-EM) techniques are improving dramatically and have less stringent requirements on crystals, generally growing a high quality crystal for small molecule crystallography is a time-consuming process. Moreover, it's known that compounds with crystal structures are generally smaller and exhibit fewer conformers than other compounds. Similarly, compounds containing elements outside the common organic subset (e.g., B, As, Se) or less common chemical motifs may be poorly represented in experimental crystallographic databases. Also, even for compounds found in an existing database, much chemistry is performed in solution and gas phases, where solid-state preferences may not directly apply. Finally, several works have noted challenges when deriving data from some crystallographic databases.

Consequently, finding unbiased alternative sources of accurate and reliable torsional

angle preferences could significantly expand the use of conformational sampling to new compounds. Typically, conformational sampling has been performed using small-molecule force fields, which have shown limited fidelity when compared to density functional and other first principals quantum chemical methods. The development of efficient semiempirical methods such as GFN2[165], as well as new machine learning methods such as ANI[6–8,27] and OrbNet[11,30], offer improved accuracy of torsional angles and non-bonded interactions with moderate computational cost. Moreover, several large-scale computational efforts including PubChemQC[166] and the QCArchive torsion scans have provided large amounts of high-quality density functional optimized geometries.

In this work, we outline an extensive effort to analyze conformers and torsional angle preferences of over 3 million compounds, using exhaustive sampling using the GFN2 method across both the experimental Crystallographic Open Database[167] (COD) and multiple sets of small molecules, including PubChemQC. We compare potential bias between the crystallographic and gas-phase geometries, including analysis with ωB97X-D3.

4.3 Methods

Molecules for this work were compiled from several sources, including the Crystallographic Open Database, PubChemQC, the Pitt Quantum Repository, and previous work on conformational flexibility, which used molecules from a subset of ZINC[168] and the Platinum ligand database[169]. For all sources, the largest substructure was retained (i.e., removing solvent or salts from the crystallographic unit cells). For compounds without an initial 3D coordinate set, Open Babel 3.1 was used to generate coordinates. As noted above, the total set of compounds included over 3 million unique molecules.

For each molecule, conformers were generated using the CREST program to exhaustively sample the potential energy surface, using the GFN2 method for energies and optimized

geometries. In some cases, CREST will produce fragments or chemical rearrangements (e.g., producing more stable compounds than the input) — these systems were excluded from analysis.

In this work, the lowest energy conformer by GFN2 energy was analyzed. Using previously-published torsional angle SMARTS patterns used in the ETKDG[37,170] methods for both acyclic and ring dihedrals, histograms of matching torsions were generated using RDKit[171] Python scripts (see supporting information).

4.4 Results and Discussion

To understand the differences in torsional preferences, we compared the experimental torsions from COD to the CREST generated conformers from COD and the CREST generated conformers of the combined sets of PubChemQC, Pitt Quantum Repository, and the subset of ZINC. While the compilation of torsions from COD is an excellent tool for understanding crystal structure preferences, the small data set limits the number of torsions for some structural motifs. Expanding our analysis to include a large number of structures can improve the understanding of uncommon motifs and determine if they are rightly infrequent motifs or just not abundantly present in the COD. While torsional information based on crystal structure is effective for poses of docked molecules, we expect to find differences in torsional preference compared to the gas-phase data. This work will examine the comparison of crystal structure and calculated gas-phase torsional preferences for both ring and acyclic containing torsions.

While there are expected differences in torsional preferences due to crystal structure and gas-phase geometry differences, there should still be some correlation between these two phases. To determine the degree of correlation, the r^2 of the kernel density estimation (KDE) for each acyclic torsion pattern was compiled in figure 4.1. The r^2 of the KDE plot was used

as a way to smooth the histogram and avoid issues when correlating areas with no torsions in the COD with the regions where some but few torsions were present in the combined set. The median r^2 was found to be 0.61 indicating while the geometry of the structures between phases differs, overall the torsional preferences are comparable. This was further shown when analyzing the patterns with an r^2 less than 0.2. The patterns in this regime had a median of 175 instances of the COD torsions in the patterns, too small of a number of data points to draw any significant conclusion from.

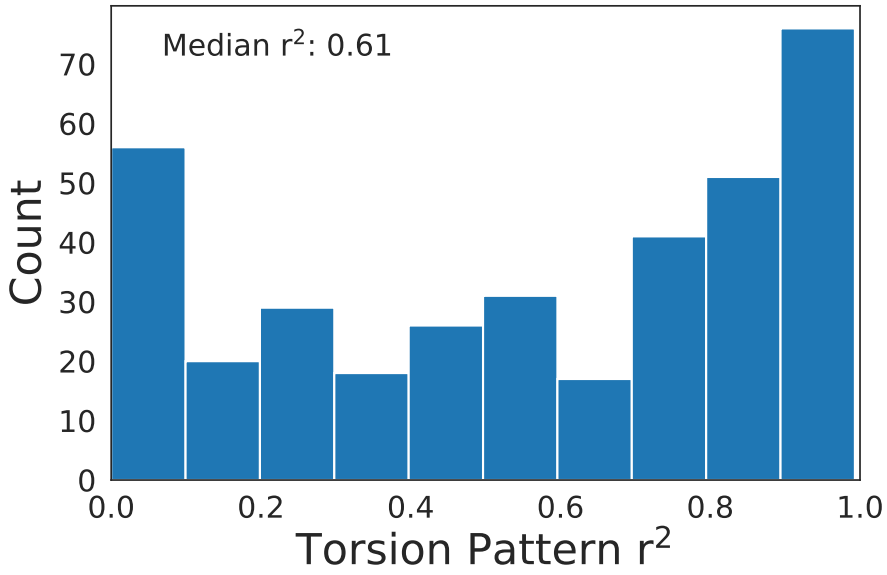

Figure 4.1: Correlation between experimental and gas-phase torsions for acyclic patterns.

Acyclic torsional preferences were analyzed to determine the qualitative correlation between the crystal structures and conformers. Torsion pattern 229 shown in figure 4.2 demonstrates the degree of correlation between the crystal structures and conformers while demonstrating the advantage of the additional data from all of the data sets provides. Pattern 229 has 179 torsions from the COD while the expanded data set boasts over 25k instances. This increase in data clarifies torsional preferences 90°to 180°as the conformers demonstrate clearer trends in this regime.

Qualitative analysis of these correlations also illustrates differences between the crystal

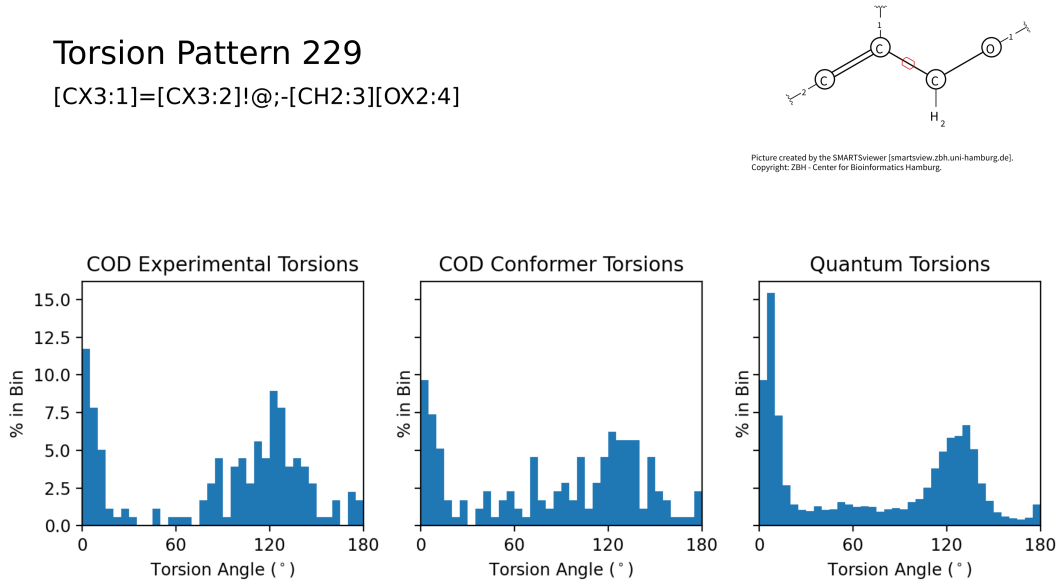

Figure 4.2: Increasing data clarifies existing torsional preferences.

structure and gas-phase that need to be explored. There are a few possibilities to consider when analyzing these differences. A difference can arise from either a torsional divergence between the two phases, inaccuracy of the GFN2 method, or a lack of data as discussed prior. While there are some geometry differences between the crystal structure and the gas-phase geometries, most of these are small expected differences. An example of these small differences is shown in figure 4.3. The experimental torsions show a mix of torsions in the range of 70°to 110°while the calculated torsions exhibit a dominant peak at 90°. Though both can have torsions in this range gas-phase geometries preferred a 90°angle.

To ensure the differences in figure 4.3 were due to phases, ωB97X-D3 calculations were performed to verify the GFN2 results. Molecules containing torsion pattern 270 and an angle of 90°were randomly selected from the combined set. The DFT calculations were found to be in agreement (within 1°) of the GFN2 torsions, indicating that such differences are indeed due to disparities between the crystal structure and the gas-phase geometries.

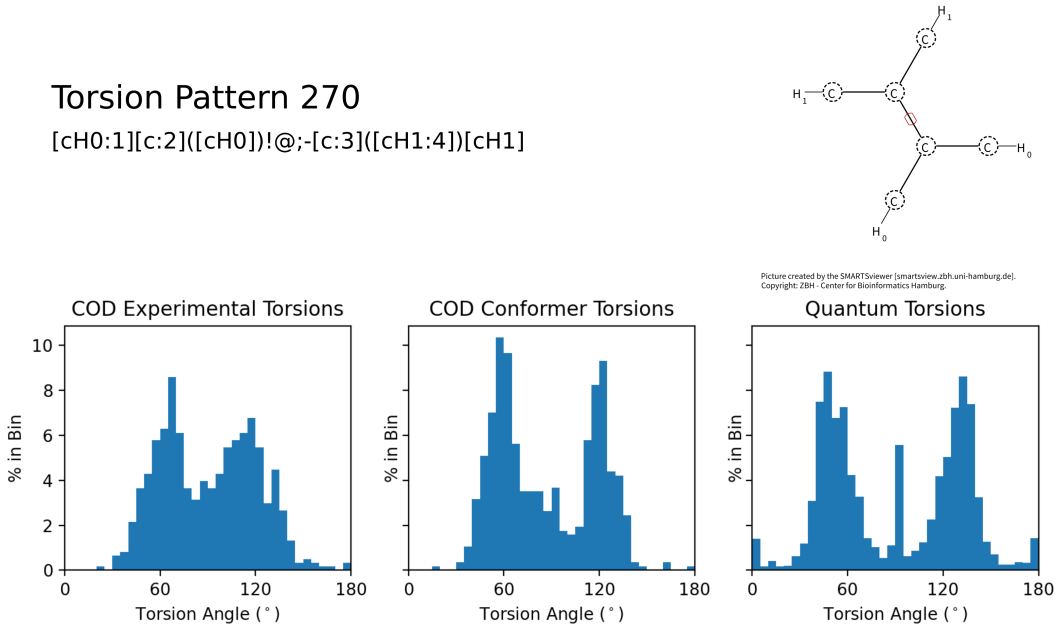

Figure 4.3: The differences in torsional preferences for experimental and gas-phase geometries.

Figure 4.4 demonstrates how the lack of data can impact qualitative assessments of torsional preferences. The experimental torsions show a mild preference around 60°and 120°with a lot of noise between, while the calculated torsions have more defined peaks at 50°, 90°, and 70°. The 10x increase in the number of torsions for the calculated set was able to provide enough to discern preferential angles.

In addition to acyclic preferences, we analyzed the preferences of torsional patterns that are part of ring structures. Ideally, torsional patterns in ring structures should correlate well between experimental and calculated gas-phase geometries due to the imposition of steric hindrances that impose constraints on the geometry. The correlation was analyzed in the same manner as the patterns that were acyclics by taking the r^2 of the kernel density estimation (KDE) for each ring torsion pattern and compiling them into figure 4.5. Compared to the median r^2 of 0.61 for the acyclic patterns, the ring patterns boasted a median r^2 of 0.83, indicating a decent correlation between the two sets.

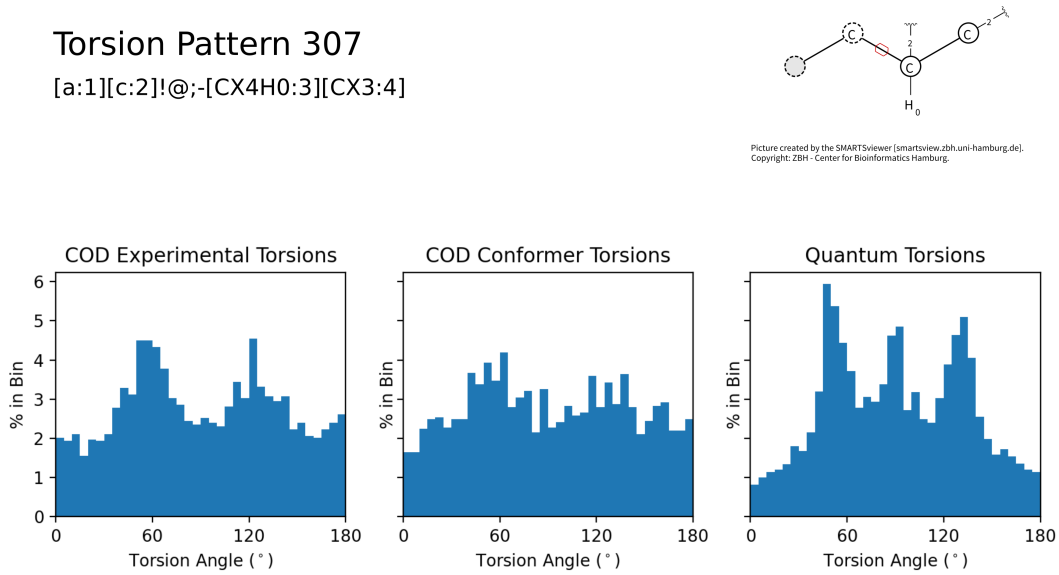

Figure 4.4: The increasing of data set size has on discerning effect on the preferences in torsion patterns.

Similar to the acyclic patterns, the increase in available data bolsters the angular preferences of the ring torsion patterns. Figure 4.6 exemplifies this by demonstrating how the increase in the number of torsions for that pattern affects the torsional preference. Though the experimental data suggests 60°and 180°are prominent angles, the additional data from the calculated sets firmly demonstrates that these are prominent angles for the torsion pattern. The addition of more data exhibits a more complete picture of torsional preferences that better represent chemical space as data is added.

These expected similarities for the ring patterns are due to the intermolecular constraints the ring structure imposes on the geometry. There is less of a chance of free rotation in these patterns, leaving fewer possible minimas. The correlation of rings demonstrates the accuracy of the quantum torsional preferences and indicates the ability to use this information as an additional method for determining torsional preferences of structural motifs not yet examined

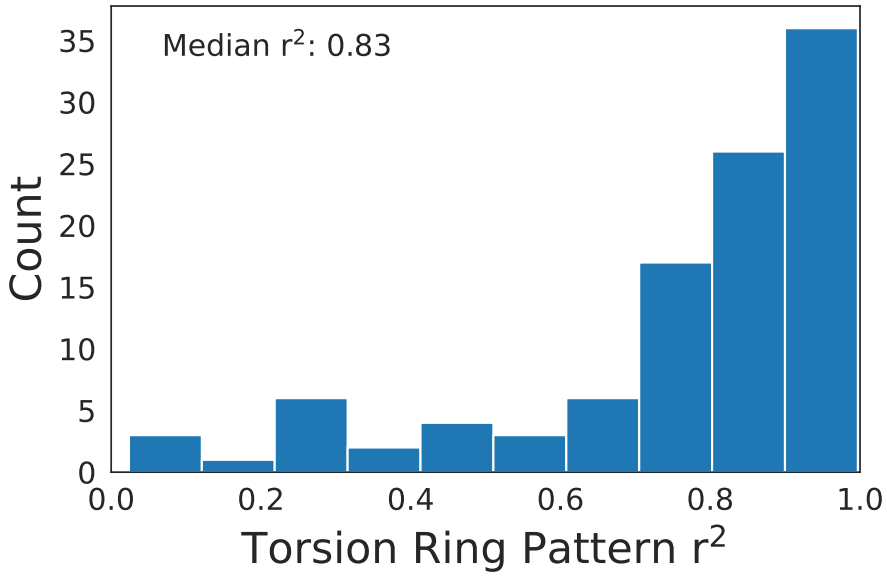

Figure 4.5: Correlation between experimental and gas-phase torsions for ring patterns.

through experimental means.

The outlined work demonstrates the desire for additional methods based on quantum calculations where crystal structure constraints may not be suitable or data may not be prevalent due to experimental limitations. Using the example of ETKDG[37], a quantum-based alternative, quantum torsion distance geometry (QTDG), could be useful for gas-phase applications. The QTDG method would no longer be constrained to structural preferences derived from what can be crystallized. This allows for an increase in the capability of the method as a larger amount of data would be available for a more diverse representation of chemical space.

4.5 Conclusions

The torsions of 3 million compounds were analyzed using exhaustive sampling with the GFN2 method across multiple small molecule sets to compare experimental crystallographic

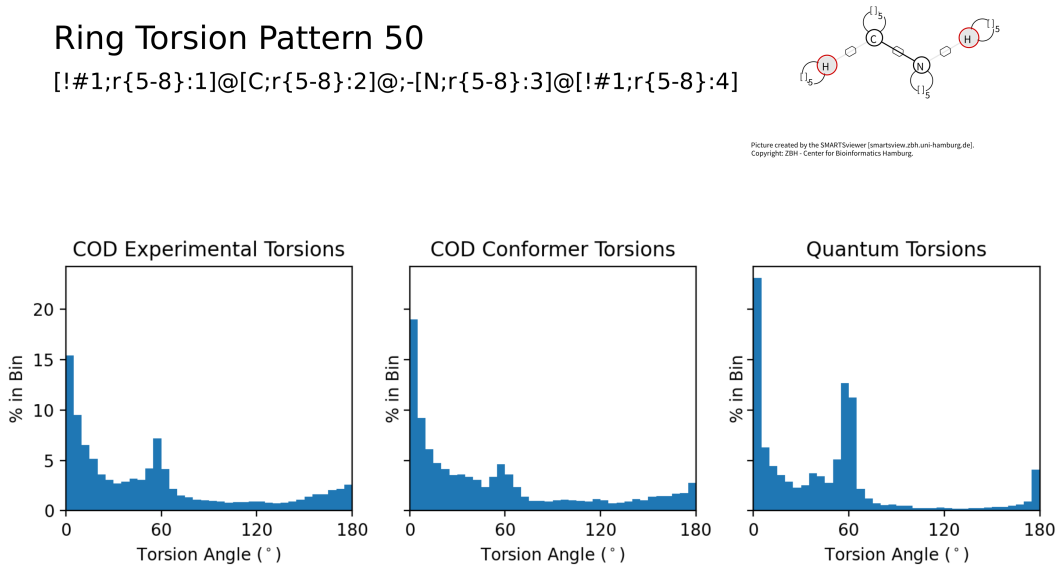

Figure 4.6: The additional data obtained from quantum calculations provides a more thorough understanding of torsional preferences.

geometries and gas-phase computed conformers. Though differences in torsional preferences were found, findings indicate differences are due to expected crystal structure and gas-phase differences along with a lack of available experimental crystallographic data. The use of quantum-based methods allows for the generation of torsional preferences in addition to the patterns currently based solely on experimental crystallographic geometries, allowing for a further understanding of chemical space that is not restricted to the ability of molecules to crystallize. This work has demonstrated the ability of GFN2 to provide accurate and reliable torsional preferences that could provide a basis for a quantum-based ETKDG method, QTDG. QTDG could provide an alternative to ETKDG for conformer generation of gas-phase applications with the ability to operate in a space that does not rely on experimental crystal structure elucidation.

5.0 Chemical Applications of Genetic Algorithms and Future Implications of Machine Learning

This chapter is adapted from a manuscript in preparation for submission; it is a work in progress. It is a collaborative effort with Danielle C. Hiener, Omri D. Abarbanel, Brianna L. Greenstein, and Geoffrey R. Hutchison. O. D. A and I wrote the machine learning sections of the manuscript; D. C. H. and B. L. G wrote the genetic algorithm sections of the manuscript; G. R. H. conceived the project.

5.1 Summary

Materials discovery and design has been given a large acceleration boost in recent years due to the increase in computational resources and algorithms available to researchers. This enables researchers to find potential chemical candidates with targeted properties for different applications and uses in the fraction of the time it took previously. In this review we discuss the strengths and weaknesses of two such tools: Genetic Algorithms (GA) and Machine Learning (ML), and how the combination of those two can overcome each other's shortcomings. We suggest a future where GA is used to improve ML models by expending the chemical space they are trained on, and where ML models are used in conjunction with GA as the algorithm's fitness evaluation step. While big strides have been made in both GA's and ML's use in chemistry we scarcely see them being used together, and we aspire that this review can open new and interesting research avenues.

5.2 Introduction

Computer-accelerated materials design is a fast growing field due to continuing improvements in computer processing units. These advancements have allowed for faster and more accurate quantum mechanical (QM) calculations to predict various molecular properties and behaviors. There is a direct relationship between accuracy and calculation time, where the more accurate methods, such as Density Functional Theory (DFT) and Coupled Cluster (CC), are orders of magnitude slower than simpler approaches such as Force Field (FF) or semi-empirical methods[172]. Even with considerable advancements in computation power, some calculations can take days or even weeks per calculation. This has motivated researchers to seek alternative techniques to expedite the search for new materials, taking advantage of recent discoveries and applications within computer science, data science, and statistics.

A promising method for accelerating material discovery is the utilization of a Genetic Algorithm (GA), which incorporates Darwinian evolution's notion of "survival of the fittest" into computational techniques. GAs are a type of evolutionary algorithm that are designed to optimize a specific property by narrowing the search space and allowing research to focus on only feasible candidates. A population of possible molecules is generated and the "fittest" molecules are selected to act as parents for a new generation of possible candidates, with each generation providing better performing molecules. By the end of this process, only the molecules with the best performance for optimizing a given property survives.

Another technique for materials discovery advancement is Machine Learning (ML), which has already been widely used in many applications such as speech recognition and product recommendations, due to its inherent ability to learn from past experiences in order to predict future instances. Its use within the realm of chemistry has already proved promising, with excellent results in applications such as drug discovery, solar cells, and polymers. Utilization of ML can accelerate the exploration within a search space for a given property and perform much faster than QM calculations.

GAs and ML can work towards the same goal — discover new materials faster by efficiently exploring the molecular search space. GAs do not require any previous knowledge, however they can be slow due to the evaluation of fitness, especially if QM calculations are required. ML can drastically improve the fitness evaluation speed, however requires training on known data which on itself can be computationally cumbersome. By working in tandem, GAs and ML can provide numerous ways of improving material discovery performance.

5.2.1 Search Space

A common challenge in chemical research is finding molecules with properties well optimized for a particular application. Chemical space is vast, with the subsection containing small organic molecules useful to drug discovery alone estimated to be comprised of more than 10^{60} possible structures.[1] This means that exhaustive searches of even this subsection would have a lower time bound of millions of hours when the necessary electronic property calculations are considered, each taking minutes to hours to complete. Strategies that efficiently find desirable molecular candidates without such searches are therefore necessary to allow researchers to find useful structures within reasonable time and resource limits.

5.2.2 Search Techniques

In contrast to the time and resource inefficiency of an exhaustive search of chemical space, alternative approaches that can reduce the number of calculations exist. High-throughput virtual screening typically makes use of a previously constructed database for screening, limiting the search area to a predefined subset of chemical space. It has been used to find molecular candidates for diverse applications, from organic materials and drug discovery to solar materials and topological insulators[2,173,174]. While this is a viable strategy, inverse design presents an alternative approach in which target features are determined first, then

molecular structures which optimize these features are found. Inverse design has proven useful through many implementations which use optimization, sampling, and search procedures to efficiently traverse chemical space to find ideal molecular targets. A more recent inverse design strategy is generative machine learning (ML), often implemented as a deep neural network via either a variational autoencoder or a generative adversarial model. Generative ML models determine the joint probability distribution of a data set and use this knowledge to predict new data that fits this distribution. In the field of chemistry, this allows generative ML models to find novel molecular candidates whose structure and properties align with this distribution.[175,176] Another inverse design strategy commonly used in chemistry is the genetic algorithm (GA), which applies key concepts from evolutionary biology (genotypes, fitness, and natural selection) to iteratively find solutions to an optimization problem by evolving generations of solutions with increasingly more desirable features.

Exploration vs. Exploitation In the realm of non-exhaustive-search optimization strategies, a key challenge is determining how a strategy will balance exploration of the total search space with exploitation of local areas of interest. This balance must be struck carefully, so as to give the strategy the best possible chance of discovering the desired global extremum of the search space in the least possible time. An optimization strategy too focused on exploration risks converging on an underdeveloped solution, while a strategy overly dominated by exploitation risks converging on a solution found in a local, rather than the global, extremum.

There are several common approaches to striking the exploration vs. exploitation balance. In simulated annealing, this balance is controlled by a temperature variable that gradually decreases over time, simulating the physical annealing process in which molten metals are slowly cooled from a molten state into a crystalline state. The algorithm works by moving from one candidate solution to another nearby, based on the probability of it being a better solution. Early on when the temperature variable is high, the algorithm accepts more candidate solutions which are evaluated to be worse than the present solution, allowing for

broader exploration of the search space. As the temperature falls, fewer worse solutions are accepted, allowing the algorithm to ultimately focus on a narrower search space of proven solutions.[177] The drawback of this approach is that because it first focuses on primarily on exploration and then focuses primarily on exploitation, it increases the likelihood that on a highly variable optimization surface, it is easier for the algorithm to become focused on a local rather than global extremum too early in the optimization process, resulting in an incompletely-optimized solution. In Bayesian optimization, an acquisition function is used to maintain the exploration vs. exploitation balance. This function selects the location of data points evaluated to update the prior into the posterior function and can take many forms. Generically, it uses predicted mean and variance values to guide selecting a data point by having one term that directs data point selection toward less well-known search space and another term that directs data point selection toward well-explored areas.[178,179] The quality of a Bayesian optimization method is determined by its specific acquisition function and its suitability for a given application, making it a more challenging optimization method to use efficiently. The computing time it takes to determine and calculate the next data point is also relatively high, scaling at n^3, meaning that if the acquisition function is not efficient in choosing the most meaningful points to explore, Bayesian optimization can quickly become an expensive method. In contrast to these two optimization methods, genetic algorithms maintain the balance between exploration and exploitation by two operators modelled on evolutionary theory, crossover and mutation. The crossover operator focuses on exploitation by combining the best known candidate solutions, while the mutation operator focuses on exploration by introducing random changes into the candidate solution pool to ensure broader search space examination.[180] Because they treat explorative and exploitative drives as separate, parallel processes, genetic algorithms are highly unlikely to become trapped in local extrema. These processes are also simple and time efficient, meaning that genetic algorithms are limited in their time complexity by the step of actually evaluating candidate solutions rather than in directing the path through search space.

5.3 Genetic Algorithms

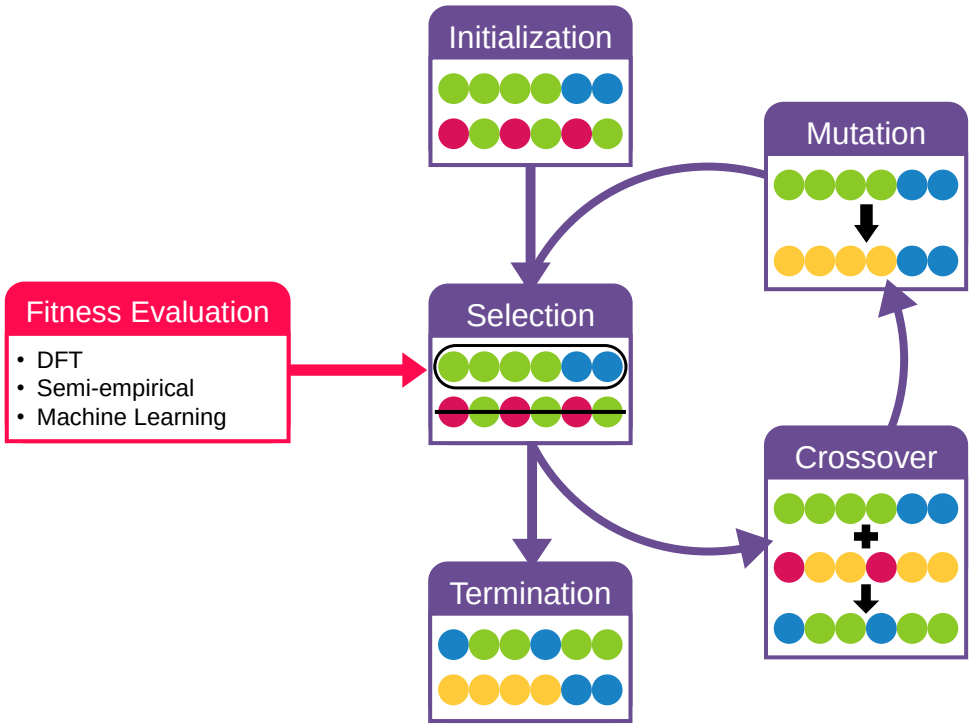

Figure 5.1: Schematic demonstrating the basic steps in the GA workflow, using simplified hexamers as an example of candidate molecules.

In a GA's evolutionary scheme, a population of possible solutions to an optimization problem are generated and then run through selection, crossover, and mutation operators to produce a new population of offspring (see Figure 5.1)[181]. Each successive population is known as a generation and can contain increasingly better solutions; generations are generated until some level of convergence is reached among the top solutions.

A key component to the success and exploration of a GA is the representation of each solution. Each solution must be encoded in a way that contains all necessary information and allows for crossover and mutation operations. GAs that efficiently sample chemical space frequently use string-based (SMILES, SMARTS, SELFIES) or graph-based representations

to encode molecular structure. SMILES is the most basic and intuitive chemical string representation that describes the molecule using atoms and bonds. Similarly, SMARTS is a language that can specify substructures within SMILES and is useful for finding particular patterns. To address the possibility of invalidity of some SMILES, SELFIES was developed to be more robust and handle entirely random strings[182]. To develop the two- or three-dimensional structure, the string representations are transformed into a molecular graph, where the nodes are atoms and edges are bonds. Recent work by Jan Jensen[183] showed that using a molecular graph directly as the representation for candidates in a GA can cross a large distance in chemical space with few generations. In another work by Jensen, graph, SMILES, DeepSMILES, and SELFIES representations were compared to traverse chemical space to rediscover target molecules with a GA[184]. Graph-based representation had the highest success rate and found the targets in fewer generations than other representations.

To evaluate the quality of each candidate solution, a fitness function is required that can score each solution in a ranking system by its ability to solve an optimization problem. Fitness functions that use time-consuming quantum mechanical calculations, such as density functional (DFT) and first-principles methods, can dramatically slow the search through chemical space due to their large resource requirements. Semiempirical methods such as GFN1-xTB[185] and GFN2-xTB[165] have exhibited acceptable accuracy when compared to DFT methods[186] while maintaining low computational cost, ideal for a GA fitness evaluation technique. Our group has employed these semiempirical methods to help accelerate the discovery process of the GA and allow for the scaling to massive search spaces[152,187]. For example, the evaluation of the power conversion efficiency for organic solar cells requires information about the electronic structure and optical properties, which is obtained through DFT and time-dependent DFT (TD-DFT) methods. Using these ab initio methods results in the fitness function taking days to evaluate, meaning a GA employing these techniques may need to run for months to complete a meaningful number of generations. While parallel processing can improve the speed, lower cost methods like the semi-empirical GFN2-xTB

and simplified TD-DFT (sTD-DFT) that takes minutes are vital to have a large impact on the efficiency.

Once each candidate in the population is evaluated, a selection operation is performed to choose candidates to act as parents to repopulate the next generation. There are various types of selection operators, most notably fitness proportionate, rank, and tournament selection. In fitness proportionate selection, such as roulette wheel selection, every individual has a chance of being selected as a parent, with a probability proportional to its fitness. This method applies a pressure for selection of more fit solutions. When the fitness scores are too similar like towards the end of the GA run, they will have the same probability. To overcome this, rank selection is a similar method that selects parents with a probability proportional to its rank among the population, not its fitness. An alternative method is k-way tournament selection, where k individuals are chosen at random and the best is selected as the parent.

In the re-population of the next generation, elitism is frequently used. This guarantees that a small number of the most fit solutions are in the next generation without undergoing mutation. This approach guarantees that the best solutions remain and will exist in the final generation.

To build the children for the next generation, the parents selected undergo crossover operation, which can vary depending on the optimization problem. In work by Hiener *et al.*, child co-polymers are designed with the monomeric units from one parent and the sequence of units from another[?] . To dissuade premature convergence, a mutation operator is introduced to ensure diversity within each population. Without the process of mutation, the solutions are limited to traits selected in the initial population, prohibiting the successful exploration through the search space. The percentage of solutions in each population allowed to undergo mutation is an important parameter for a successful genetic algorithm. In Greenstein *et al.*, non-fullerene acceptors underwent crossover by replacing the core or terminal units, or rearranging the electron-withdrawing/donating groups into a new sequence, with a mutation rate of 40%.

5.3.1 Recent Work

Recent work reveals the breadth of applications in which GA-based searches prove efficient tools for the exploration of chemical space. GAs have been used to aid drug discovery efforts, exploring areas of chemical space around known bioactive peptides with similarly active structures[1] as well as designing ligand candidates with good docking scores for binding to given protein targets[188]. They are also capable of discovering novel protein structures with enhanced functionalities like thermostable and solvent-tolerant metalloproteins[189] peptide scaffolds with tailored catalytic capabilities[190], and high antifreeze activity proteins[191]. GAs have proven capable of efficiently exploring conformational space, to find low energy structures on complex potential energy surfaces[192] as well as to generate geometrically diverse structures[152]. In nanomaterial research, GAs have been adapted to predict and optimize the atomic structures of bi- and trimetallic nanoparticles[193–195] as well as chemically diverse nanoclusters[196]. On the other end of the materials spectrum, GAs have successfully been implemented to explore large molecular systems such as predicting crystal structures[197] and designing MOF arrays for targeted applications.[198,199] In the realm of energetic materials, GAs have been used for a diverse set of chemical applications, including thermal energy storage systems[200], thermal heat batteries[201], organic photovoltaics[3,202], and high dielectric oligomers[187].

5.4 Machine Learning

ML has been extensively used in recent years in many applications as a surrogate to more computationally expensive quantum-mechanical (QM) methods such as DFT. ML has shown potential as a fast and accurate prediction method that could combine with a GA to provide more accurate rapid evaluations over semi-empirical methods. In the past, in order to train a ML model, a training set of molecules, specific to the model's objective, would have to be

created based on previous research and chemical intuition. Combining GA with ML opens the door to a more diverse and complex data sets that can improve a ML model's performance by increasing the search space.

5.4.1 Molecular Representations in Machine Learning

To apply ML methods to chemistry, molecules typically need to be translated from the familiar laboratory representation into a representation that the model can interpret. These representations need to convey the underlying physics of the molecule for models to infer patterns and learn. There are multiple ways of representing molecules in machine learning from character encoding with representations like SMILES[15–17] and InChi[18], to fragment-based encodings seen with Extended-Connectivity Fingerprints (ECFP)[25] and MinHash Fingerprints (MHFP)[203], to the inclusion of local and global information with representations like the Coulomb Matrix[22] and atom-centered symmetry functions (ACSF)[4].

Extended-connectivity fingerprints[25] (ECFP) are a molecular graph model that expands out along bonds from each heavy atom for 2-3 steps to observe the local connectivity of each heavy atom. The extensions from each heavy atom are stored as a fragment that is hashed as a fingerprint. After iterating over the molecule, these fragment fingerprints are combined to describe the molecule. MinHash fingerprints (NHFP) are similar to ECFP, in that they they both encode the local environments in a molecule. However, MHFP uses a different hashing algorithm, employing methods usually used in natural language processing and text mining, which can outperform ECFP in many cases[203].

Rupp *et al.*[22] proposed using molecular information in the Hamiltonian, such as coordinates and nuclear charges, for the modeling of atomization energies, leading to the creation of the

Coulomb matrix (CM) representation. The CM representation seen below:

$$M_{IJ} = \begin{cases} 0.5Z_I^{2.4} & \text{for } I = J, \\ \frac{Z_I Z_J}{|R_I - R_J|} & \text{for } I \neq J. \end{cases} \tag{5.1}$$

consists of a square matrix (M_{IJ}) in which the off-diagonal elements are the Coulomb nuclear repulsion between the atom pairs. This representation was further expanded upon to generate the Bag of Bonds[23] (BOB) representation. BOB is a reorganization of the CM representation in which the atoms and pair-wise interactions are sorted into bags (e.g. C, C-C, and C-N) in a bag-of-words text mining descriptor style and filled with $Z_I Z_J / |R_I - R_J|$. Further work was done to make the Bond Angle-ML[24] (BAML) representation, a many-body expansion of BOB through the inclusion of extra bags containing angles and torsions.

Another way of incorporating connectivity information of local chemical environments is through the use of molecular graphs, where each atom is a node and the bonds are the edges connecting the nodes. This representation emphasizes the bonding structure, giving a 2D topological map of the molecule[26,204]. A more advanced graph-based representation, ChemProp, is using a molecular graph as an import for a message-passing neural network to give a learned fingerprint representation for property prediction[205,206].

ACSFs are an additional way of including local chemical environments by describing local atomic environments with radial and angular symmetry functions. This approach for representing molecules has been adopted for use in multiple ML architectures as seen with ANI-1x[6,7], ANI-1ccx[27], and ANI-2x[8], and FCHL[28,29].

While a lot of work has gone into improving upon current representations through the creation of DeepSMILES[21] and BigSMILES[207], many state-of–of-the–the-art methods have moved to adaptions of ACSFs as seen in FCHL, ANI-1x, ANI-1ccx, and ANI-2x. Although results have shown that there is not a "one size fits all" descriptor or method[5,34], representations based on local environment connectivity have consistently demonstrated

improved accuracy over other representations[172,208].

5.4.2 Machine Learning Performance

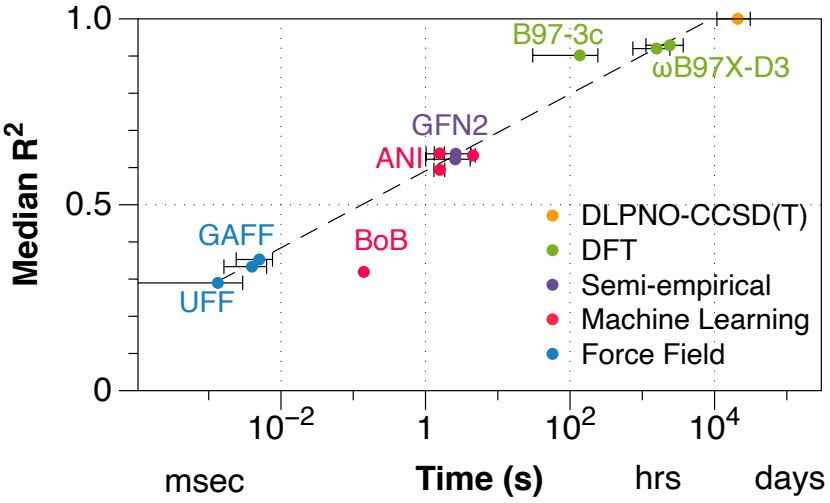

Figure 5.2: Performance of various computational chemistry methods is shown, plotting their median R^2 against timescale of calculations. Current ML methods fall near the middle, most comparable to semi-empirical methods

Much work has gone into the production of various ML representations and models for thermochemical applications with the goal to rival conventional quantum mechanical methods. Studies by Faber *et al.*[5] and MoleculeNet[34] compared representations and methods for various thermochemical properties. The consensus of these findings demonstrated the promise of ML models but exposed the lack of a one size fits all solution, with the best representation and method for one property evaluation being different than another.

Further testing done by our group tasked ML models with the ranking of conformers[172]. The work set out to determine the ability of ML models to distinguish between thermally accessible conformations and compared the results to conventional methods. ML models were found to not only perform roughly on par with but also take as long as semiempirical methods, see Figure 5.2. While performance is not yet a surrogate for the more computationally

expensive quantum-mechanical methods, these early ML models show promise with the possibility of future improvements.

Our group evaluated the capability of ML models for potential energy curves and geometry optimizations[208]. ANI-2x[8], Colorful CNN[161,162], and FCHL[29] were found to be among the best performers by consistently predicting the equilibrium bond length correctly and demonstrating an understanding of repulsive and attractive forces. While the methods performed remarkably around the equilibrium bond length, performance degraded at long-range interactions, a side effect of the limited displacement regime around the equilibrium bond length for most of the training sets.

One area where ML has been intensely used and researched is in molecular and material property prediction. For example, ML has been used to predict acid dissociation constants (pK_a)[209–211], ground state and excited state energies[212–217], and solubility in different solvents (LogS)[218,219], just to name a very selected few. Furthermore, in our group we implemented a ML model to predict the Marcus reorganization energies (λ) of polythiophene based oligomers[220]. This model showed a heteroscedastic behavior, where lower values of λ had higher prediction accuracy than higher λ. This is a fast developing research field, and many different models are continuously being published for different molecular properties at a seemingly accelerating rate.

Traditionally, more accurate computationally heavy methods, such as DFT or Coupled Cluster, have been used to calculate molecular properties, a process that can sometimes take several days or weeks for a single compound. This can hinder research and be wasteful due to the large resource requirements needed to run those calculations. And while processors have been gradually improving and getting faster and more efficient, predicting molecular properties using a trained ML model can give results in a fraction of the time it would take to do so using DFT, for example. While the accuracy of ML models, as of now, is closer to those of semi-empirical methods, we expect it to improve as better training sets are being made and better models are being trained. The so-called "Holy Grail" of ML is to have the

accuracy of DFT on a timescale of Force Field (FF) methods[172].

5.4.3 Issues with Machine Learning

Folmsbee *et al.*[208] recently demonstrated a key issue that can arise with ML. When testing ML models' performance on H2, a common model in quantum chemistry, the models expectedly performed poorly due to the training not containing information on hydrogen-hydrogen bonds. This expected inadequacy demonstrates a key issue for ML performance, the dependence of the training set space. The work also demonstrated other issues with training set constraints when examining both short and long-range interactions at distances outside of the training set finding similar issues. This concern can be easily addressed through the expansion of the training set to cover a more diverse representation of chemical space. There are larger datasets like that of PubchemQC[166] that could build upon the diversity of existing training sets by introducing additional atomic species and motifs while providing synthetically accessible training data.

5.5 Combining Genetic Algorithms and Machine Learning

GAs and ML have allowed for considerable advancements in materials discovery, yet on their own, both methods have drawbacks. GAs perform well screening large swaths of chemical space, yet can be time consuming during the property evaluation step. ML can quickly evaluate these properties, but are frequently limited to molecules within or similar to the training set. The full impact of these methods can be amplified using a hybrid approach of combining GA and ML.

5.5.1 Improving GAs with ML

One method is to utilize an ML model as the fitness function in a GA. Once the ML model is trained, dramatic speedups in evaluation time are possible. For example, researchers saw a 12,600-fold speedup in computation time by using an ML model to predict the stiffness and critical resolved shear stress for CoNiCrFeMn alloys as opposed to MD simulations[221]. This hybrid approach has already been applied to numerous classes of materials, such as aptamers for specific biomarkers[222], peptides for antimicrobial properties[223], organic molecules for maximum absorption wavelengths[224], and inorganic complexes for spin-state splitting[225]. Another approach to reduce evaluation time is with an ensemble-based ML-GA model. The combined method can use quantum chemical methods such as DFT or semi-empirical methods for evaluation within the GA over the first set of generations, while simultaneously training ML models. During each training run, multiple ML models will be saved as "snapshots" of the training up to a certain generation. This allows for the training of a smaller NN at the time of the snapshot instead of creating one NN from all previous data, decreasing training time of the NNs during the GA and potentially improving accuracy. The ensemble of these models will eventually take over evaluation once the variance between snapshots is below a set error threshold compared to the original calculation method. This greatly improves efficiency in the long run, as ML evaluation takes a fraction of the time of quantum-mechanical based methods, and the diversity among snapshot models reduces ML error.

5.5.2 Enhancing ML training sets with a GA

The advantages of combining these methods are not limited to enhancing GAs. Incorporating GAs in ML training can greatly enhance and diversify training datasets. Many existing works in quantum chemical ML use the QM9 dataset[31,33], a set containing ∼134k molecules with up to 9 heavy atoms consisting of C, O, N, and F, or a subset of the QM9[5,22–24,34]. There are

currently larger and more diverse quantum chemistry datasets like PubChemQC[166], a set containing 3.5 million molecules, that could expand training of ML models. The challenge when trying to use these datasets is the resources needed for training. This is especially apparent with methods like the bag of features descriptors, which require every bag to be the same as the largest bag for a given molecule in the dataset. With every descriptor vector of equal size to that of the largest molecule in the entire dataset, training on the entirety of a dataset the size of PubChemQC can require terabytes of memory usage for regression tasks.

Methods from previous work have utilized a GA to optimize a training set from the QM9 dataset for machine learning[226]. This optimization method can be performed for larger datasets reducing training time by decreasing the number of molecules required to be trained on while selecting a diverse representation of the chemical space of that large dataset. Methods such as clustering can likewise aid in the creation of a smaller subset of diverse molecules as candidates can be selected from the clusters to represent the chemical space of that cluster. Implementation of these optimization methods will allow for more efficient training on larger datasets.

Another major issue with ML that can be mitigated with GA is the diversity of training sets. The accuracy of ML predictions is limited to the quality of the dataset, and usually performs poorly on molecules too different than what it was trained on. One solution is to use a GA to design a diverse training set for ML by ensuring better representation of chemical space in order to increase accuracy of an ML predictor. A GA can exploit information such as atomic environments (eg. aromatics, bond types, chirality), number of atoms, and dipole moments by maximizing the dissimilarity from other compounds within a descriptor space to generate a diverse set of molecules.

5.6 Conclusions

As many subdisciplines in the field of chemistry continue to increase their use of computational modeling, efficient methods are needed for predicting and evaluating new molecules. GAs provide an established and promising solution because of their ability to quickly and thoroughly explore chemical space. By balancing the impulses of exploration and exploitation through separate, parallel operators they are able to avoid the expense of exhaustive searches. ML models have also become vital tools in computational chemistry to quickly calculate molecular properties. Although they currently rank closer to semi-empirical models in terms of accuracy, continuing improvements in model development and training set diversity continue to push ML models closer to DFT-level accuracy. By combining the strengths of both GAs and ML models, future work promises to further improve the efficiency and accuracy of computational chemistry calculations. Through ML fitness evaluation, ensemble models, and GA-generated and enhanced training sets, these two effective computational tools can draw from each other's strengths while mitigating each other's weaknesses.

6.0 Conclusions and Future Directions

6.1 Conclusions

Intrigue for computer-accelerated material design continues as improvements in computer processing and molecular screening techniques increase. Screening methods have opted for semi-empirical and FF methods over time-consuming conventional quantum mechanical methods as they can provide a significant speedup at the cost of accuracy, increasing the pace at which chemical space can be explored. Machine learning (ML) aims to remove this accuracy versus time trade-off and provide a method capable of being both fast and accurate. This dissertation has presented several evaluations of the viability of ML to solve these problems in chemistry while outlining necessary improvements for the method. This began by evaluating the ability of ML methods to accurately distinguish between thermally accessible conformations and how this performance compared to conventional methods tasked with this function. With the understanding of the degree to which ML was able to distinguish conformations, the extent of chemical physics understood by ML models was tested to determine if the methods could understand both short and long-range interactions that occur with bond compressing and stretching as well as the effect of steric hindrance of dihedral angles. Subsequent work was completed with the intent to generate a larger diversity of conformations to bolster ML training with a proposed alternative method for conformer generation. Lastly, the combination of genetic algorithms (GA) and ML was proposed with the aim to provide an efficient search of chemical space with the accuracy of time-consuming quantum calculations.

Most molecules have multiple geometrically distinct conformations, requiring evaluation methods to appropriately differentiate between them. Promising early ML methods, discussed

in Chapter 2, were tested in comparison with force field (FF), semi-empirical, density functional (DFT), and wavefunction methods to determine where ML methods performed relative to for both conformer ranking accuracy and evaluation time. The test set for this consisted of 700 small organic compounds, complex drug-like compounds, and ligands with multiple conformer geometries for a total of around 6500 entries. This set of DFT-optimized minima was evaluated using DLPNO-CCSD(T) for single-point atomization energies that would be used as the comparison for the other tested methods.

Despite the claims of ML methods reaching DFT accuracy, our findings indicate state-of-the-art methods perform on par for both accuracy and time with GFN2 semi-empirical methods. Batch evaluation of ML methods can allow for improved evaluation times compared to semi-empirical or DFT methods, though they lack the range of supported elements these methods have. While these may not be desired findings, we expect these ML methods will provide increased accuracy in the future as methods are improved as larger and more diverse training data becomes available.

With an understanding of ML performance for ranking thermally accessible conformations, a focus on the physical understanding of ML models was explored. Multiple studies had already focused on the performance of ML models around the equilibrium bond length, finding ML to perform adequately in the harmonic portion of the potential energy curve but neglected examination further from equilibrium[6–8,27,36]. Chapter 3 set out to examine whether ML could interpret both short and long-range interactions that occur with bond compressing and stretching as well as the effect of steric hindrances has on dihedral angles. This work analyzed the performance of nine common ML models on a total of 17 bond stretches and 5 dihedral scans and compared the results to the DFT method ωB97X.

The consistently top methods ANI-1x, ANI-2x, FCHL, and CNN best demonstrated the ability to accurately predict energies while also predicting the repulsive and attractive forces of the potential energy curves. While these ML methods performed adequately around the equilibrium bond length, ML methods struggled in the extremes of short and long-range

interactions, indicating the need for training data in this domain. Further need for more robust training sets was made clear during the evaluation of dihedral scans with ANI-2x outperforming the other ML methods with the help of the additional torsion sampling in the training set.

In general, applying ML to molecules outside the scope of the training set is still a present issue that affects the viability of ML to be a surrogate for DFT and other time-consuming methods. There is a further need for the inclusion of additional elements and an increase in the diversity of molecules in the training set for this goal to be achieved.

Taking into consideration the previous work of determining the improvements that need to be made to ML, Chapter 4 provides a basis for pursuing the creation of a better conformer training set. The focus of the work was to compare experimental crystallographic geometries and gas-phase computed conformers to determine similarity and if computed conformer torsional angle preferences could be used for the basis of an ETKDG alternative. This quantum information based ETKDG, coined QTDG, would need the computed torsions to demonstrate correlation with experimental results. To determine this, the torsions of 3 million compounds were analyzed using exhaustive sampling with the GFN2 method across multiple small molecule sets and compared to data from the COD.

Though the correlation for acyclic torsions was not perfect, this can be attributed to two main causes, expected differences between crystal structures and gas-phase, and the lack of data present in the COD for some torsion patterns. In the case of expected differences, ωB97X-D3 calculations were performed to verify the GFN2 results and were found to agree, indicating the differences were from disparities between the crystal structure and the gas-phase geometries. Differences in cyclic torsions exemplified instances in which the COD was incomplete in torsion data. While COD displayed torsional preferences, the additional data obtained from quantum calculations provided a more thorough understanding of the torsional preferences. In analyzing the data, GFN2 provided accurate and reliable torsional preferences that could provide a basis for quantum information based ETKDG, QTDG, for

further generation of conformational data sets.

6.2 Future Directions

The work outlined in Chapter 2 and Chapter 3 has helped demonstrate the need for larger and more diverse training sets. The augmentation of current training sets can be done through the use of existing data sets along with expansion through the means of proposed methods QTDG and genetic algorithms (GA). These additions are needed as ML has been shown to struggle with the prediction of molecules too far outside the scope of the training set.

Training sets for ML models are still rather limited in their diversity with common sets only containing H, C, N, O, and F with some limited to only optimized geometries[32,33]. The advantage of including both equilibrium and non-equilibrium structures[35] and torsion sampling[8] information in the training set was described in Chapter 3. While these inclusions improve the performance of ML, there were still important interactions and atom species missing from these sets. These training set shortcomings can be addressed by expanding the diversity of training data through the inclusion of already available data sets. Existing data sets like the Non-Covalent Interactions Atlas (NCIA) data sets can provide high-quality calculations on non-covalent systems while a data set like PubChemQC can provide 3 million electronic structure calculations of a diverse array of synthetically accessible organic molecules from PubChem.

Further expansion of training set data can be performed using both the QTDG method proposed in Chapter 4 along with a GA as discussed in Chapter 5. The QTDG method has the possibility of providing a quantum-based alternative to ETKDG for fast conformer searches that ideally have better agreement with quantum-based calculations like DFT, thus providing a training set with torsions that better represent ML's desired accuracy. The issue of data set diversity is a task that can be approached with a GA. A GA can design a diverse

training set for ML by maximizing the dissimilarity of compounds in the training set, ensuring areas of diverse chemical space are present in the training set. These different steps should be effective in the creation of a larger and more diverse training set that better represents the vastness of chemical space.

In addition to GAs being useful in the betterment of ML training sets, ML appears poised to aid the accuracy of GAs. GAs have already demonstrated the benefit of using lower-cost semi-empirical methods for fitness function evaluation for speeding up the GA discovery process, but this comes with an accuracy trade-off. ML could overcome this accuracy trade-off by providing DFT level accuracy predictions, allowing for a fast and accurate evaluation method. While current ML accuracy is on par with semi-empirical methods, further improvements could have the potential to provide GAs with a fast and accurate evaluation method for the fitness function, allowing them to accurately search chemical space for promising materials.

Appendix A: Supplementary Information for Assessing Conformer Energies using Electronic Structure and Machine Learning Methods

A.1 Supplementary Figures

Additional supporting information may be found at the GitHub repository for this article: `https://github.com/ghutchis/conformer-benchmark`

Appendix B: Supplementary Information for Evaluation of Thermochemical Machine Learning for Potential Energy Curves and Geometry Optimization

B.1 Supplementary Figures

Figures of all bond stretch potential energy curves, dihedral potential energy scans for all molecules and methods considered. All raw data, Python notebooks, and the trained Colorful CNN model can be found at `https://github.com/hutchisonlab/ml-benchmark`.

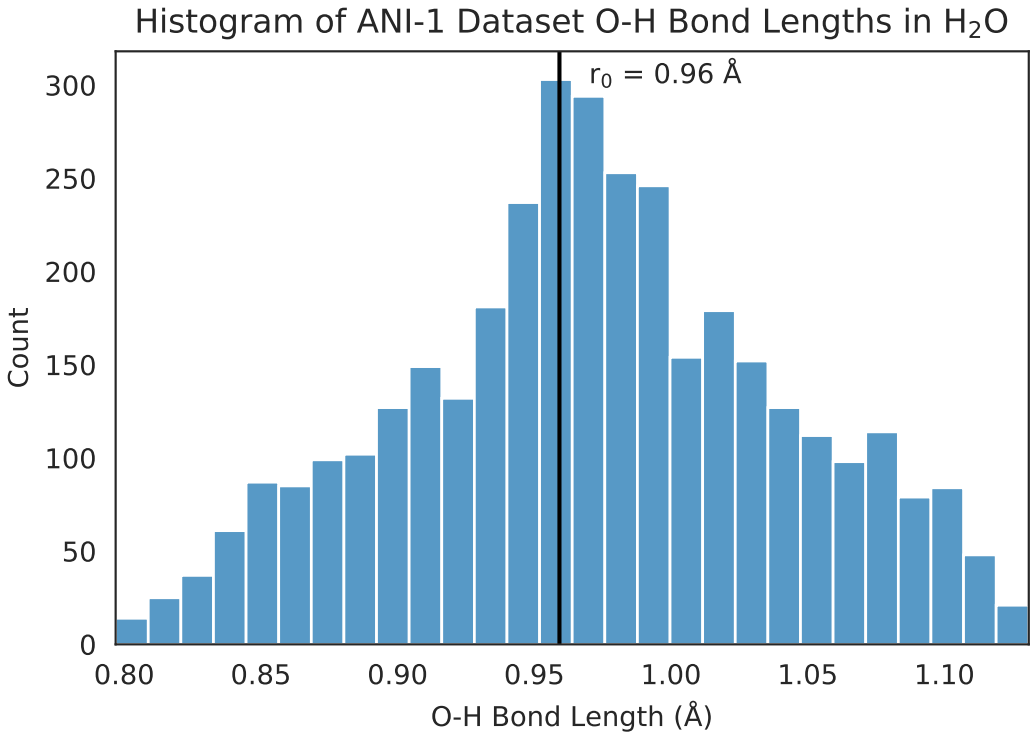

Figure B.1: Histogram of O-H bond lengths in ANI-1 data set for the normal-mode sampling of water.

Table B.1: Machine learning performance in median mean absolute percent error (MAPE) for multiple regions of the potential energy curves where r_0 is the equilibrium bond length.

Method	**MAPE** r_0**-0.75Å–**r_0**-0.25Å**	**MAPE** r_0**+/-0.25Å**	**MAPE** r_0**+0.25Å–**r_0**+1.25Å**	**MAPE** r_0**+1.25Å–**r_0**+2.25Å**
ANI-2x	0.294	0.002	0.039	0.127
BOB/BRR	1.395	0.223	0.638	0.933
FCHL/KRR	4.994	0.255	0.266	0.469
Colorful CNN	0.708	0.256	0.288	0.433
ANI-1x	0.740	0.265	0.308	0.677
BOB/KRR	1.787	0.343	0.657	0.833
BOB/RFR	45.558	43.881	36.496	37.168
BAND-NN	99.361	99.310	99.375	99.380
ECFP/RFR	197.747	193.370	103.809	104.613

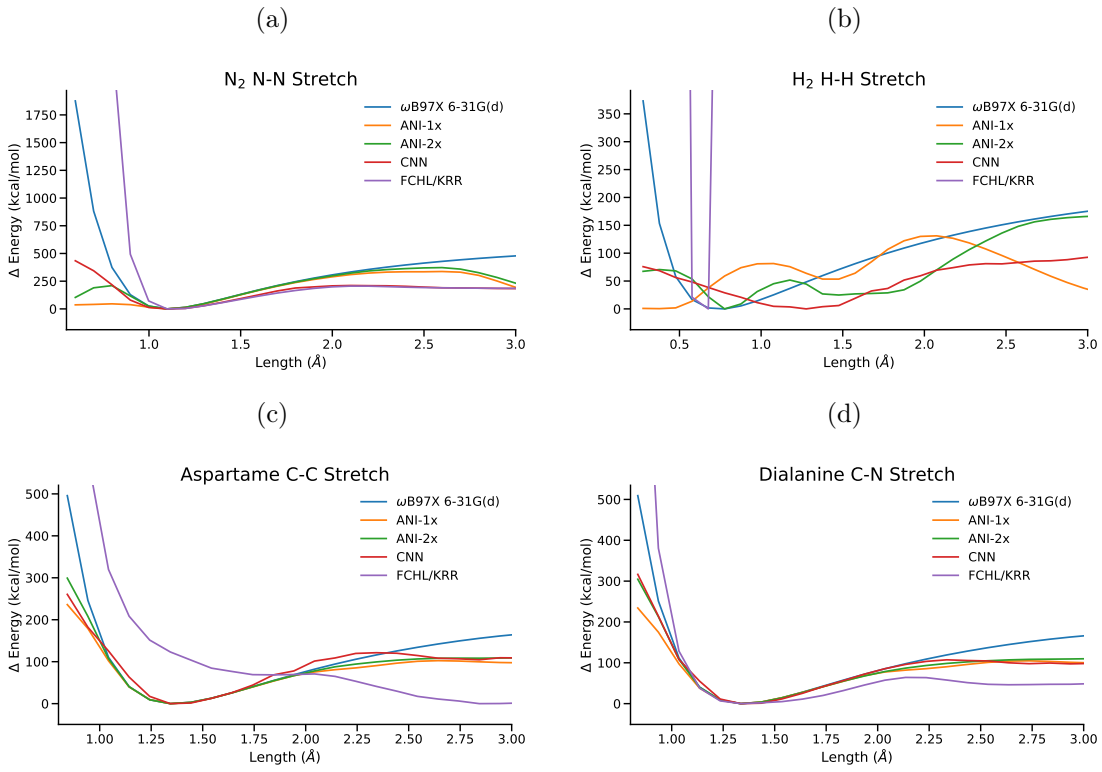

Figure B.2: Bond stretch potential energy curves for (a) N_2, (b) H_2, (c) aspartame, (d) dialanine using total SCF energies in kcal/mol.

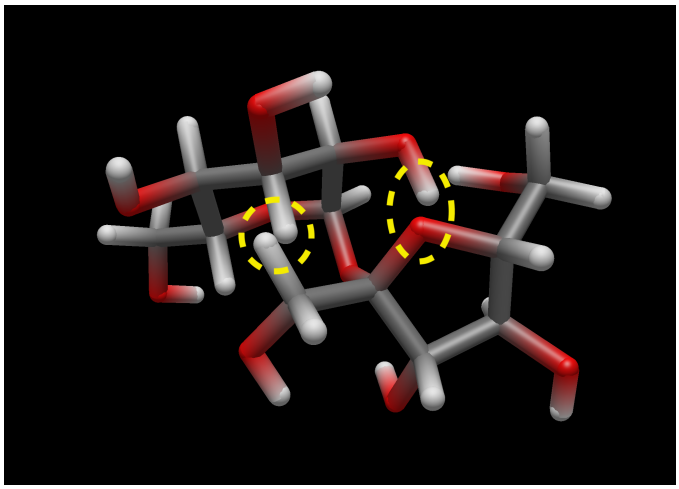

Figure B.3: Examples of steric clashes in sucrose dihedral angle scan. Yellow dashed circles highlight atoms with overlapping Van der Waalls radii.

(a) (b)

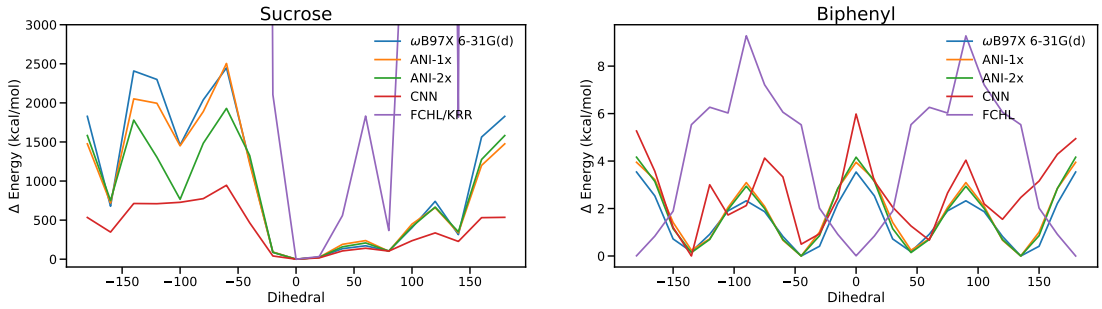

Figure B.4: Dihedral energy predictions for (a) biphenyl and (b) sucrose in kcal/mol.

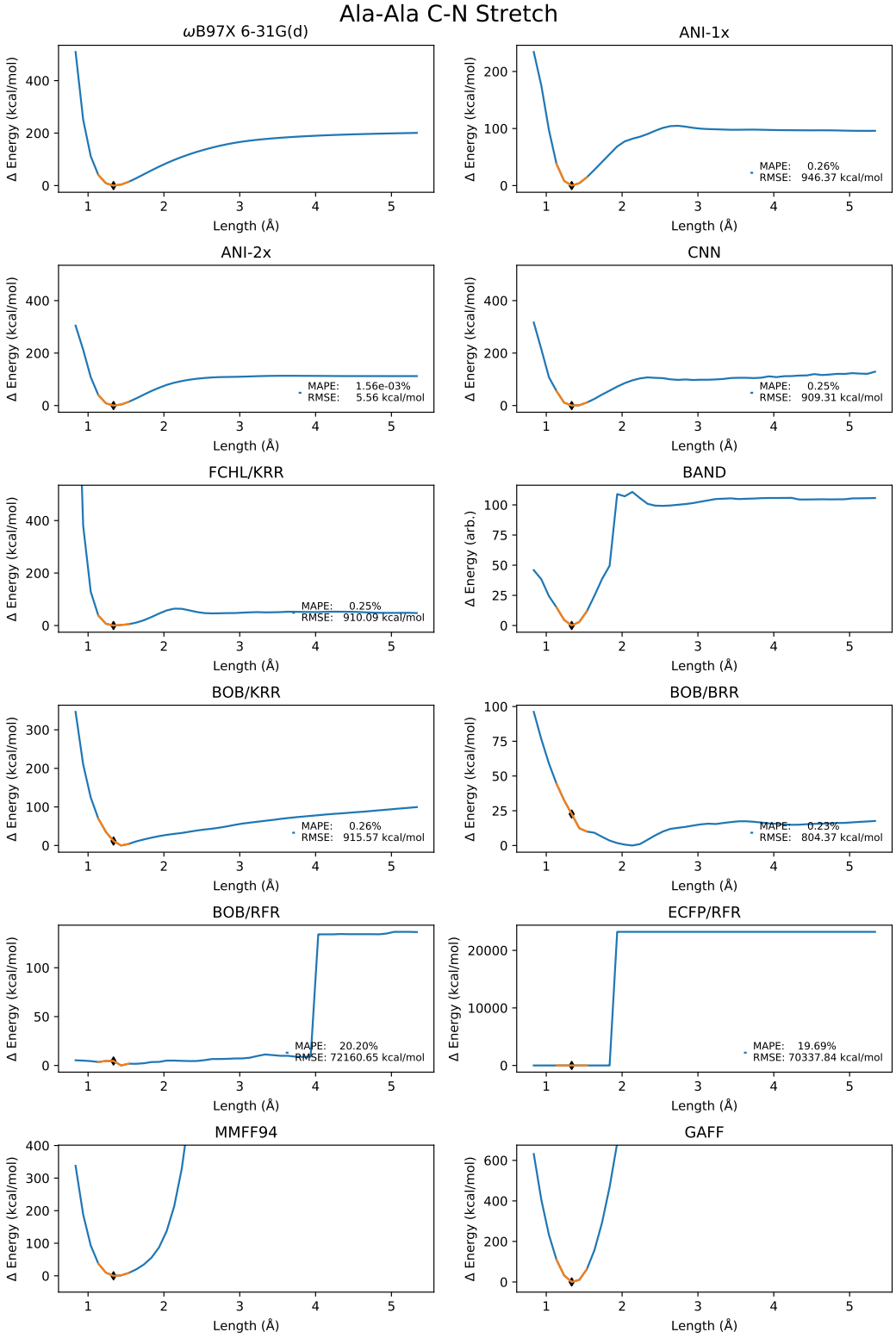

Figure B.5: Dialanine bond stretch for all methods

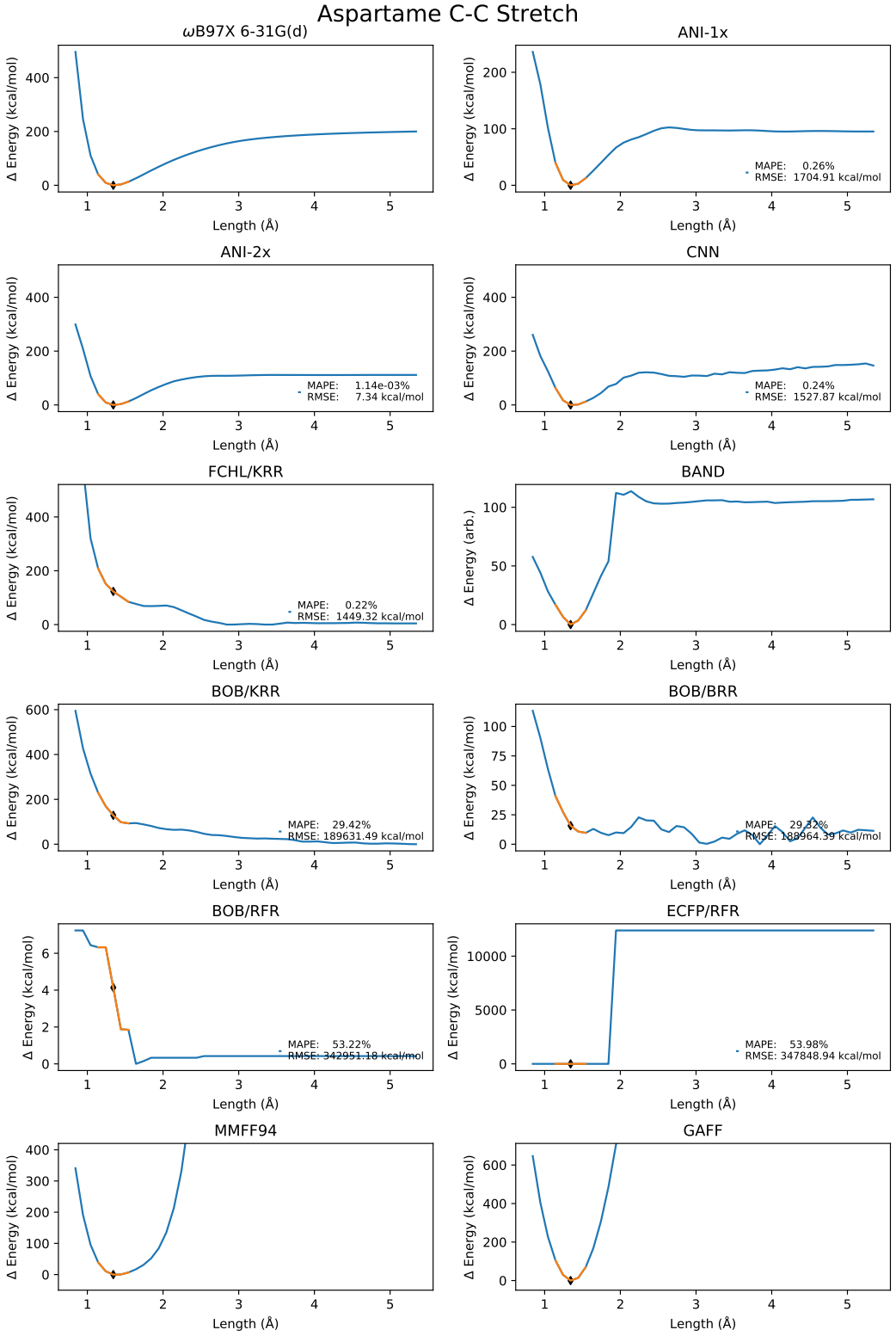

Figure B.6: Aspartame bond stretch for all methods

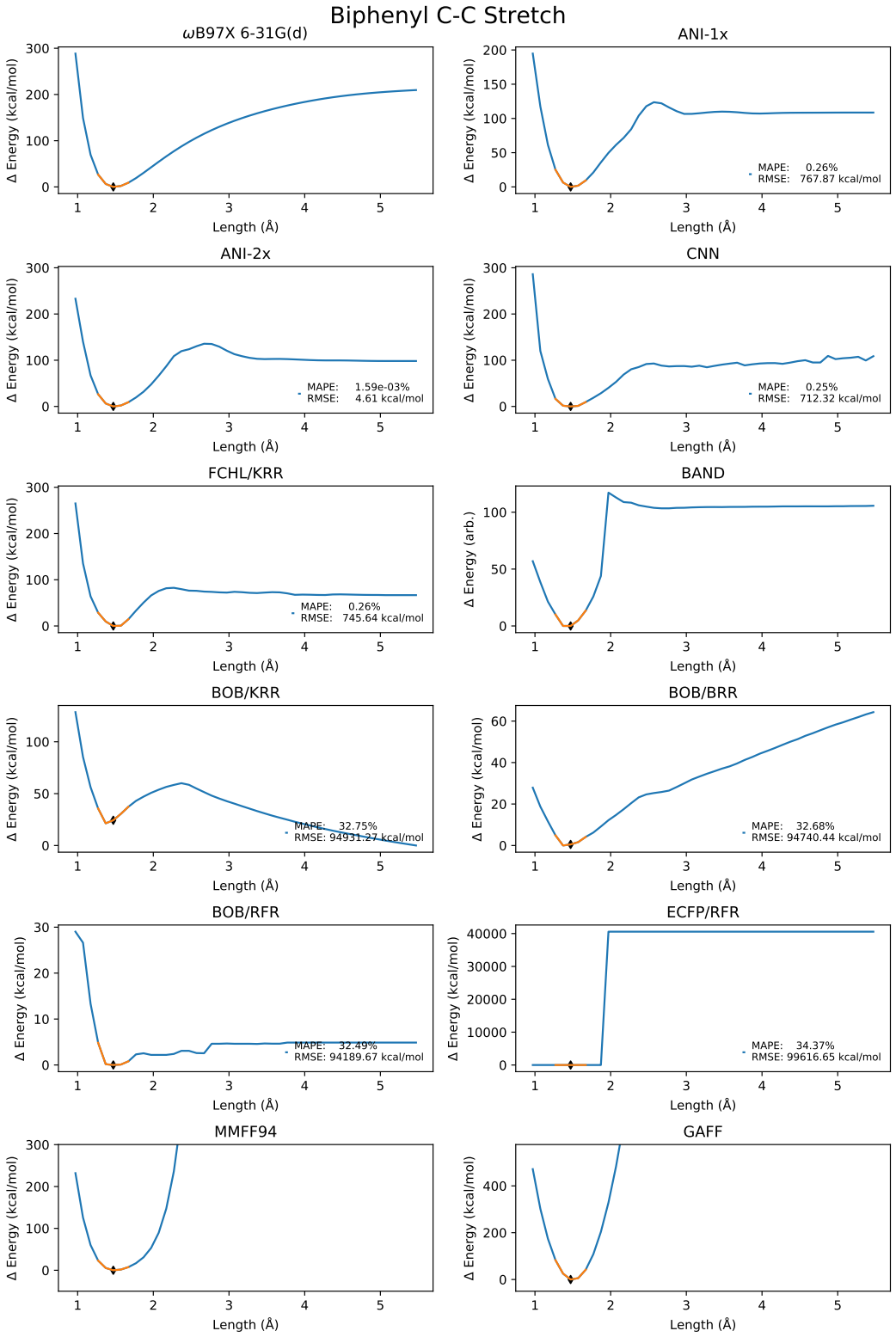

Figure B.7: Biphenyl bond stretch for all methods

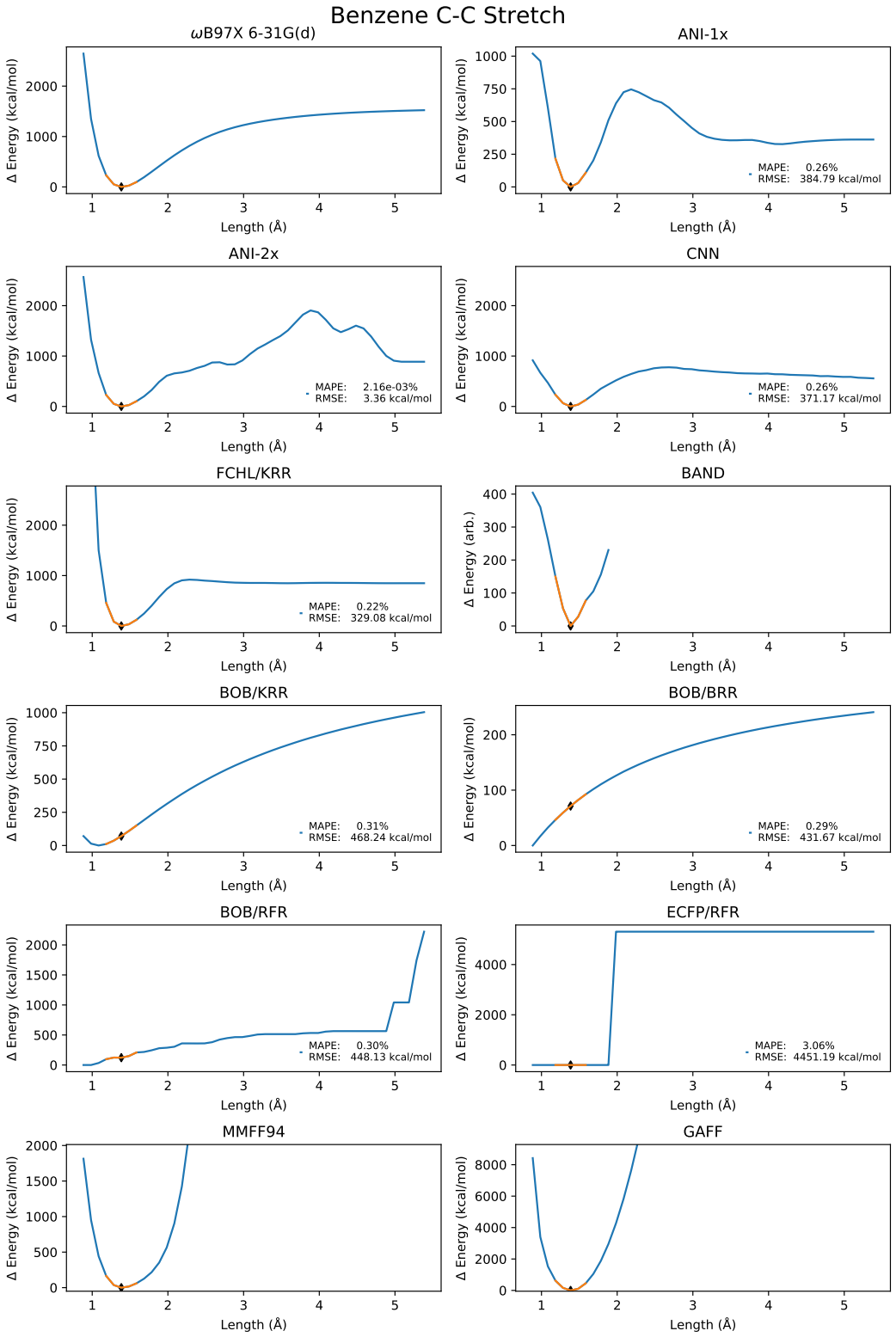

Figure B.8: Benzene C-C bond stretch for all methods

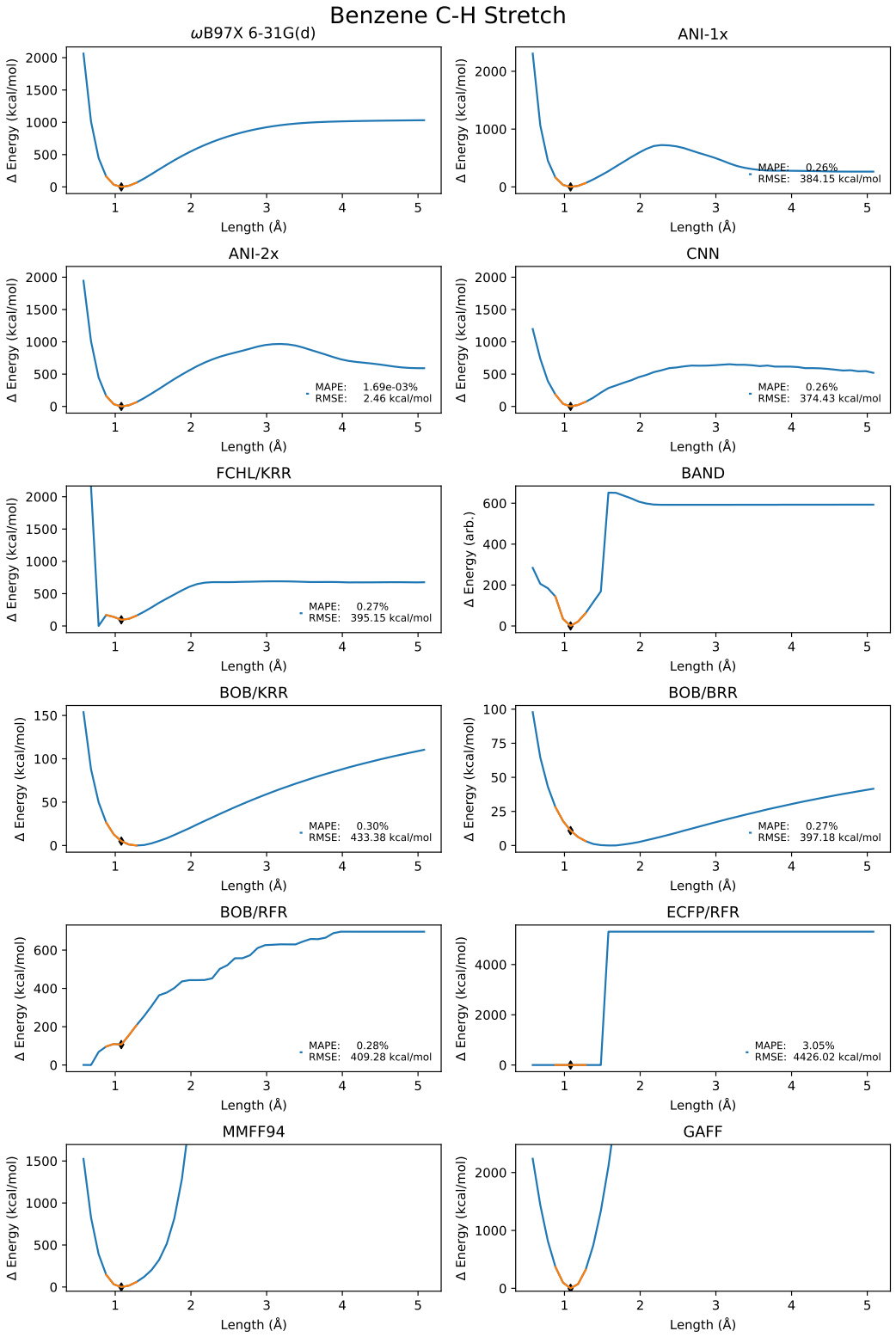

Figure B.9: Benzene C-H bond stretch for all methods

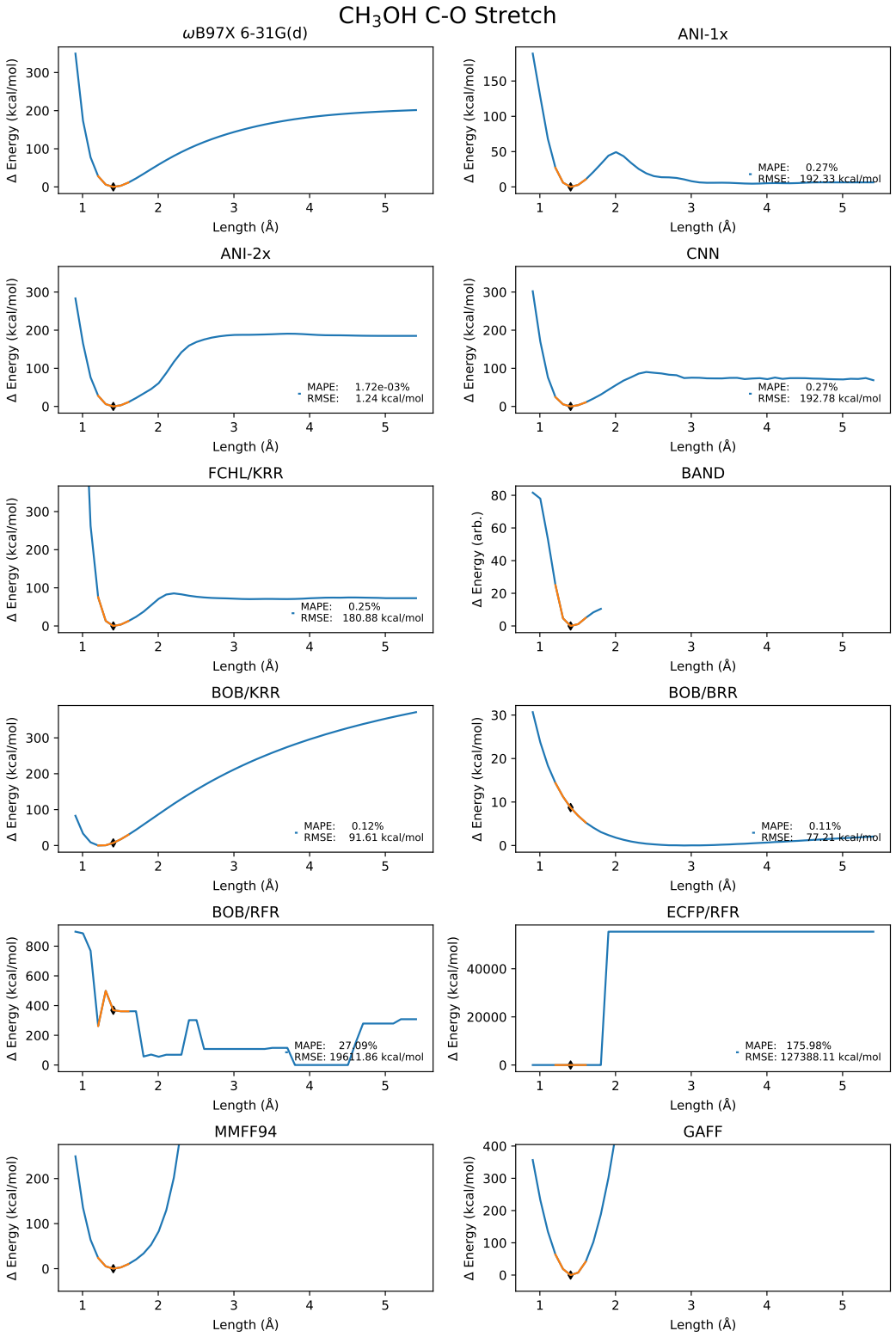

Figure B.10: Methanol bond stretch for all methods

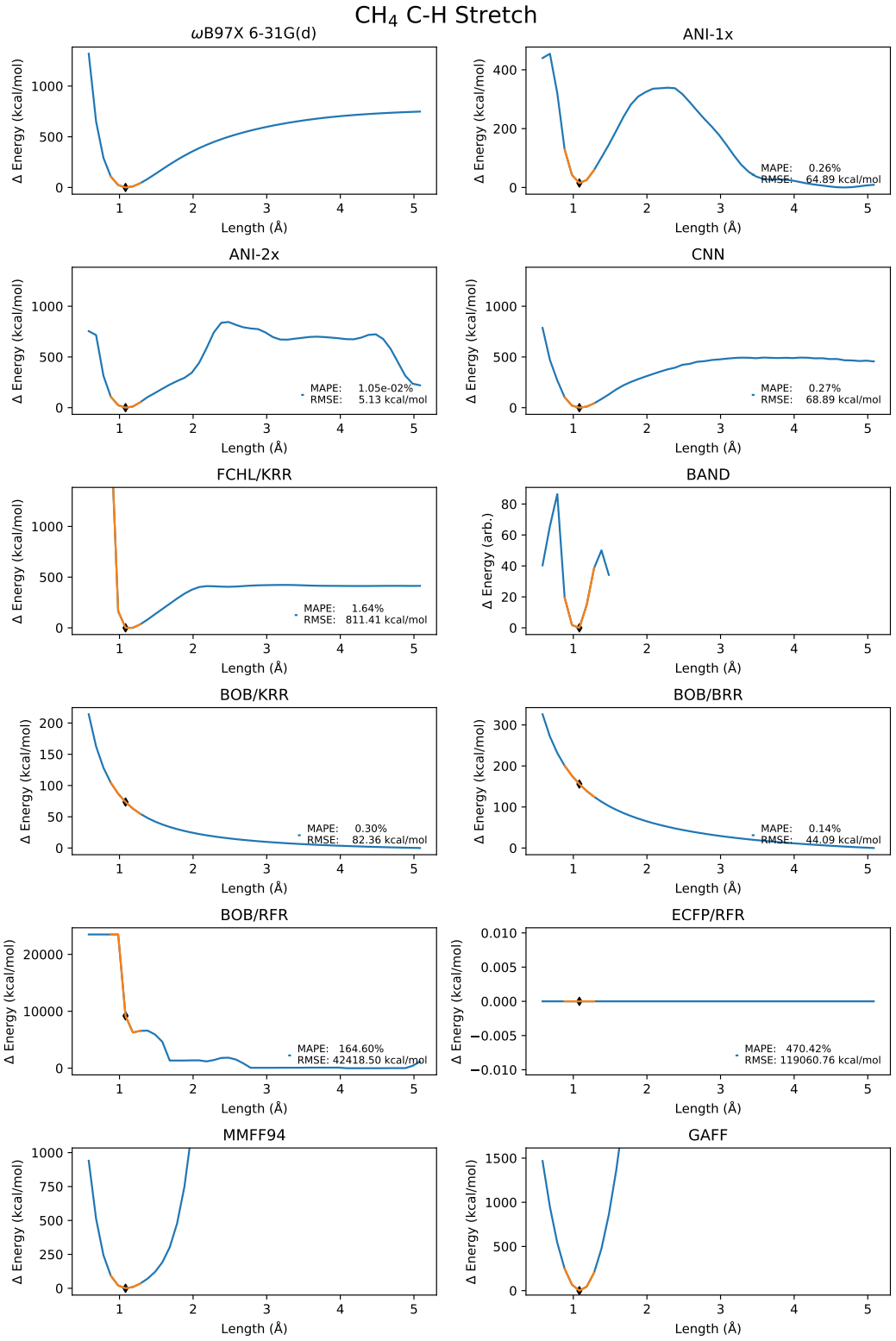

Figure B.11: Methane bond stretch for all methods

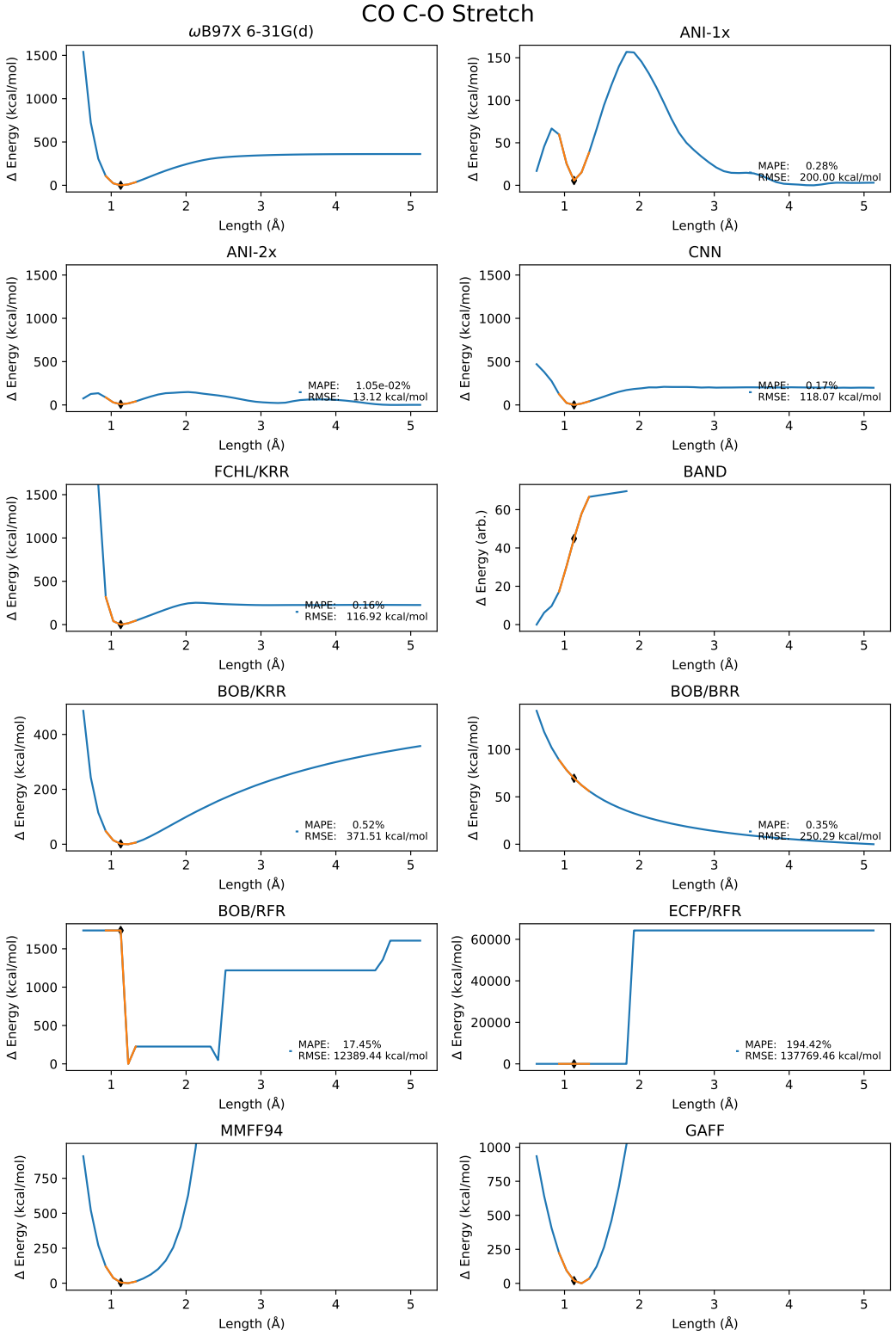

Figure B.12: Carbon monoxide bond stretch for all methods

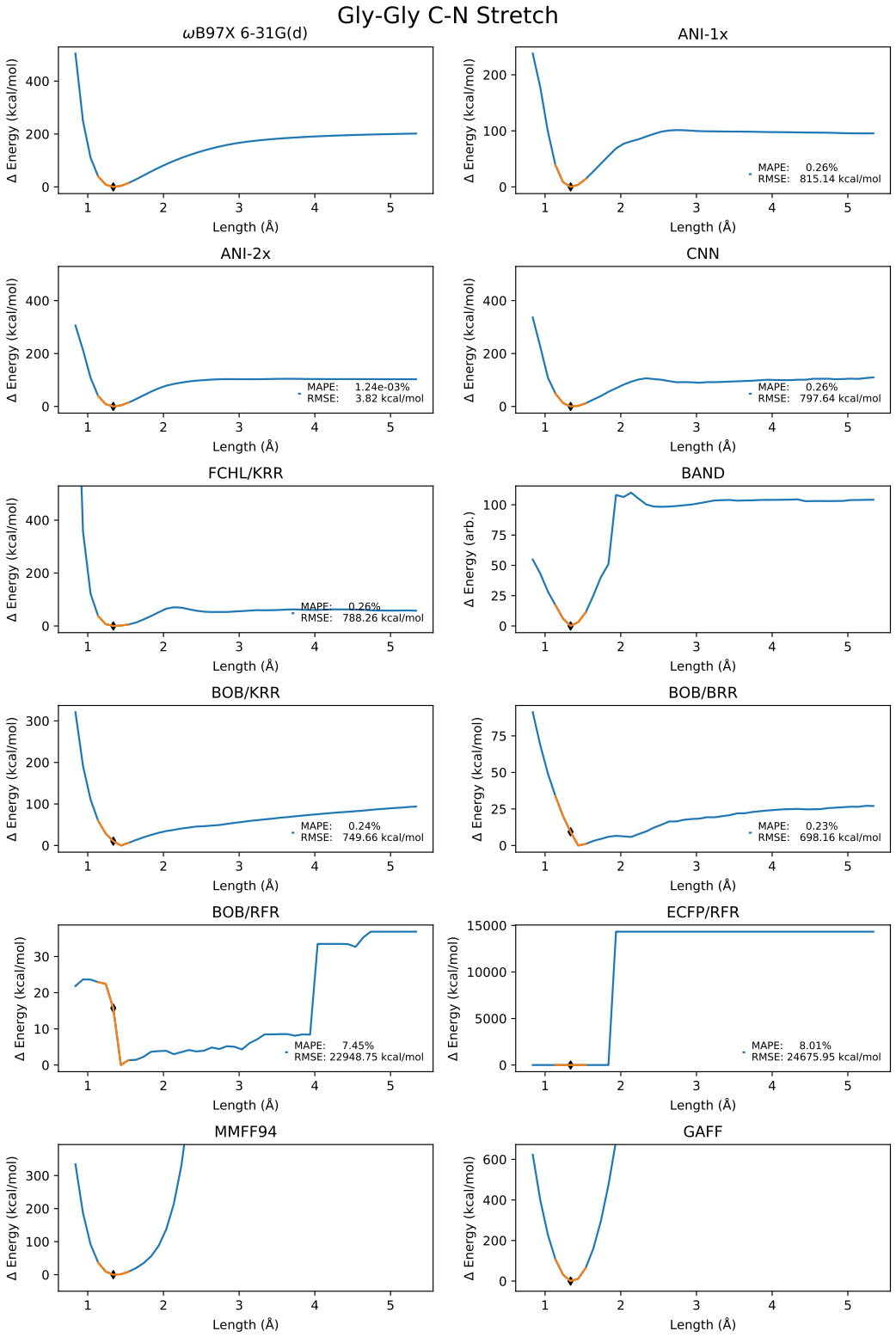

Figure B.13: Diglycine bond stretch for all methods

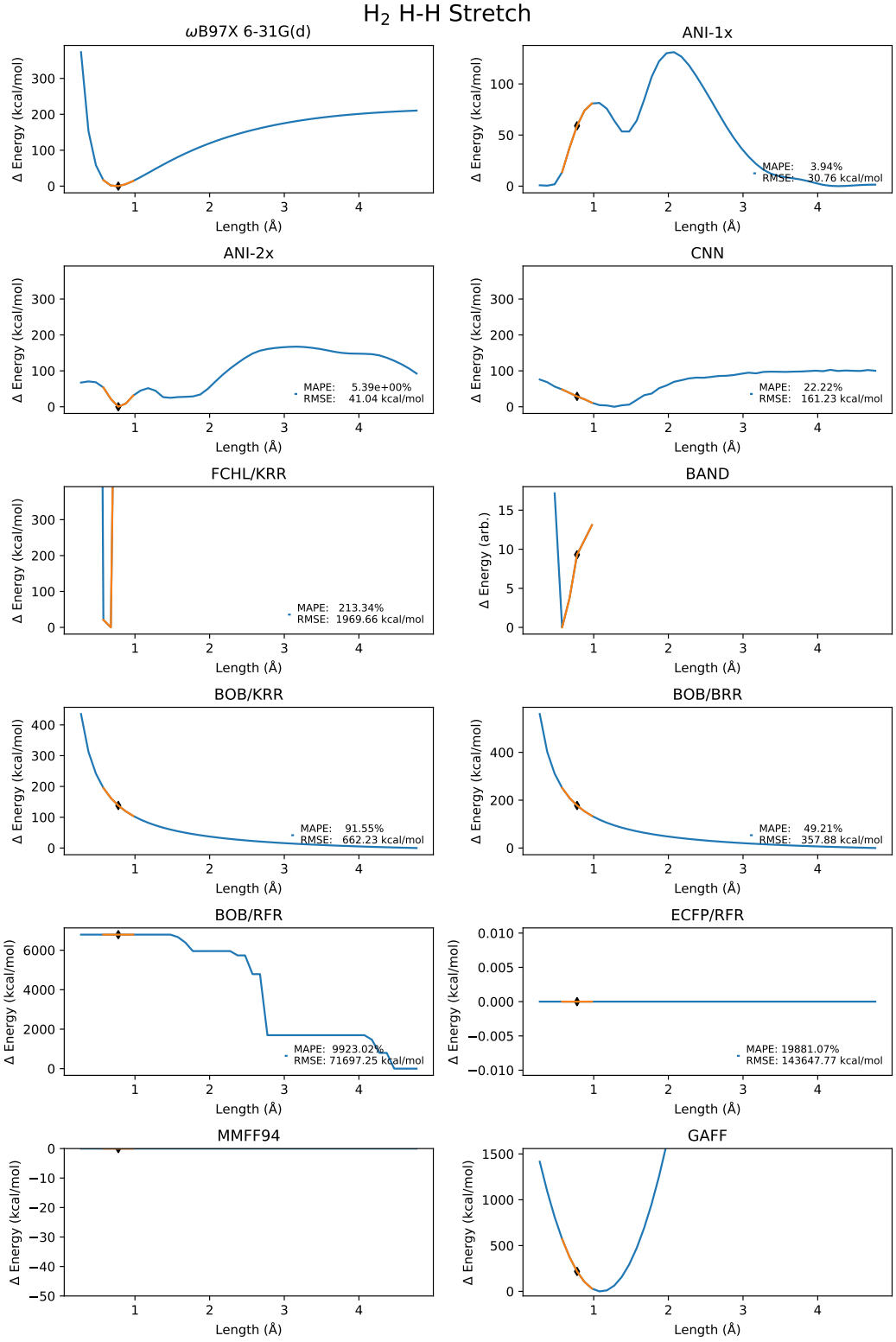

Figure B.14: H_2 bond stretch for all methods

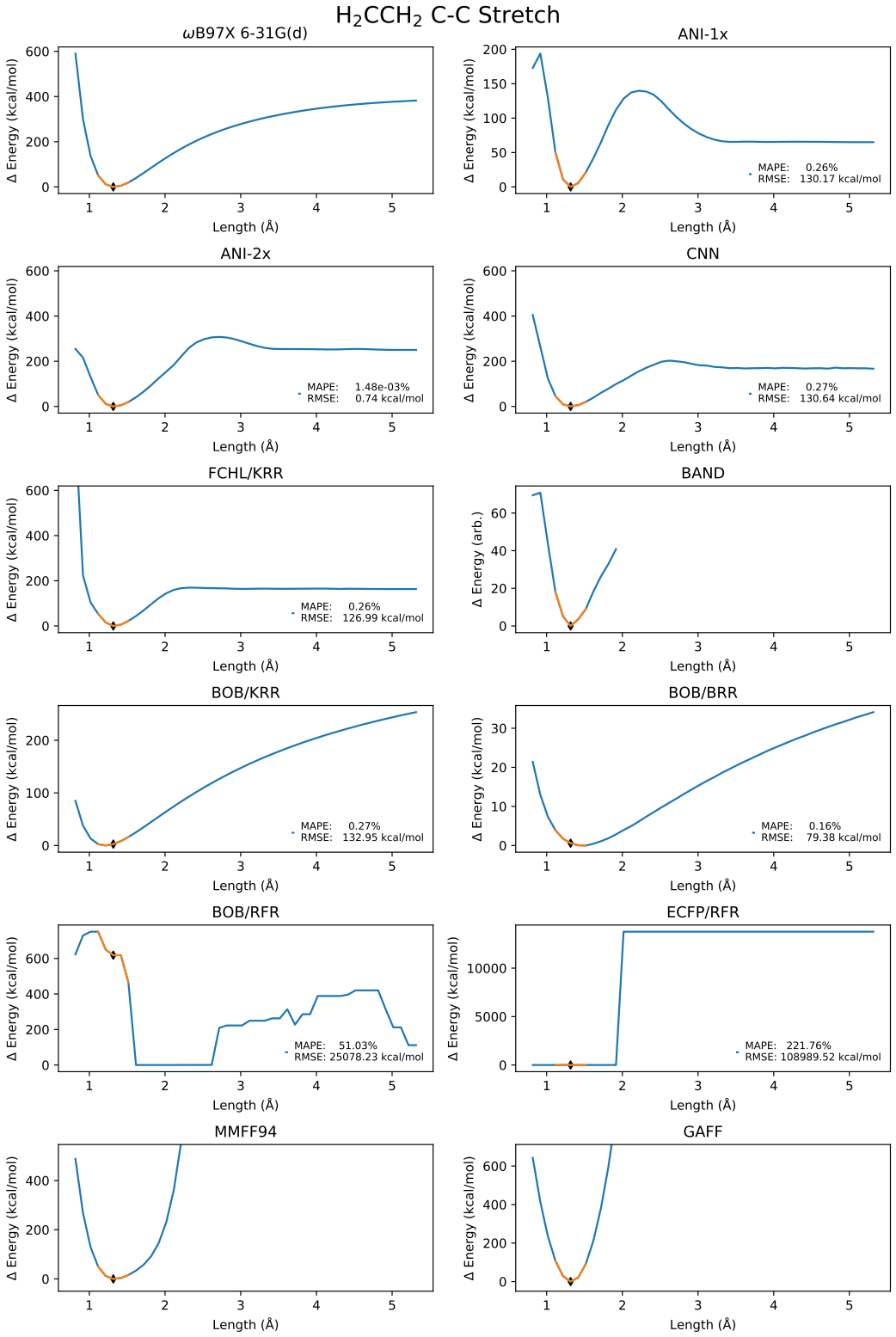

Figure B.15: Ethylene bond stretch for all methods

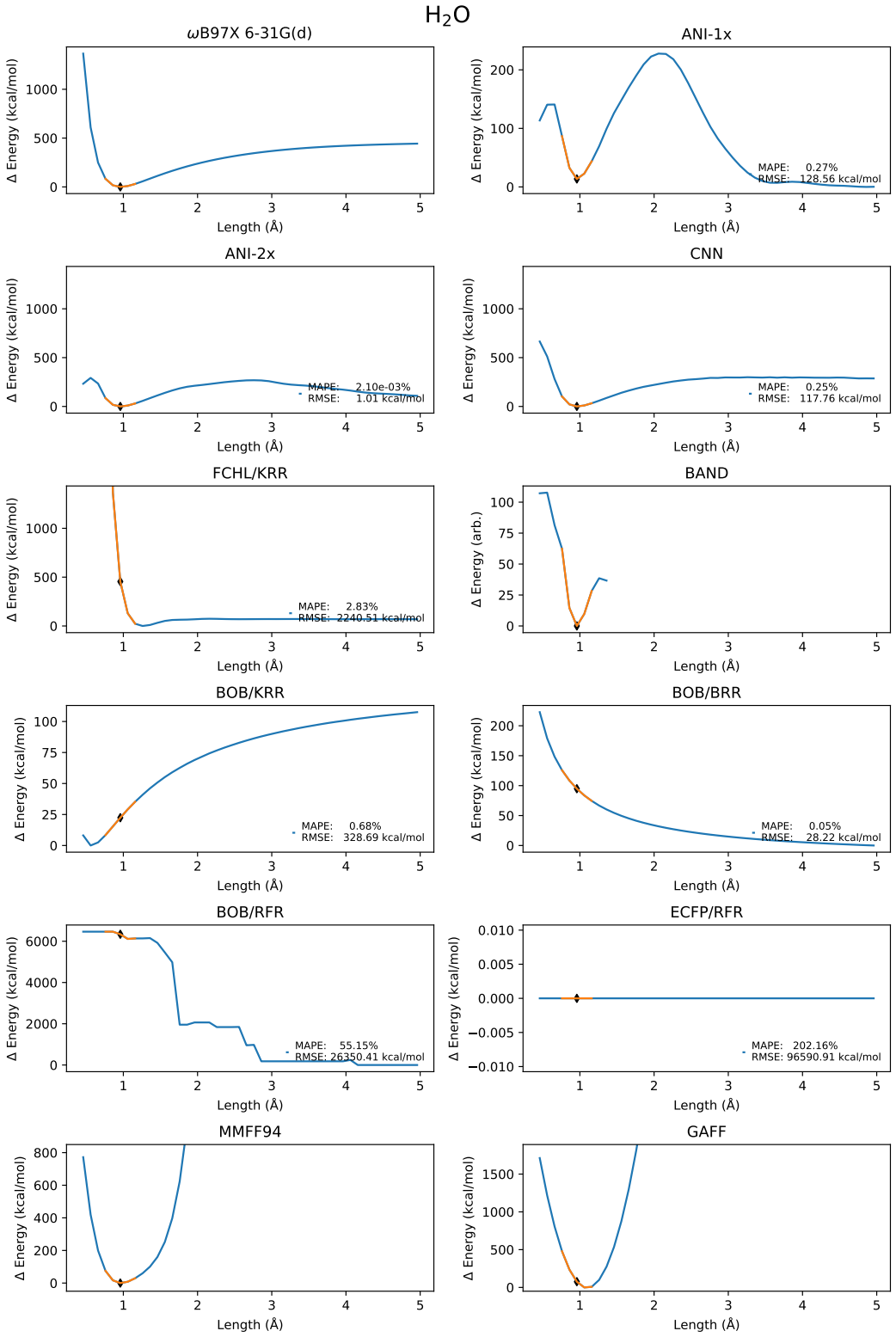

Figure B.16: Water bond stretch for all methods

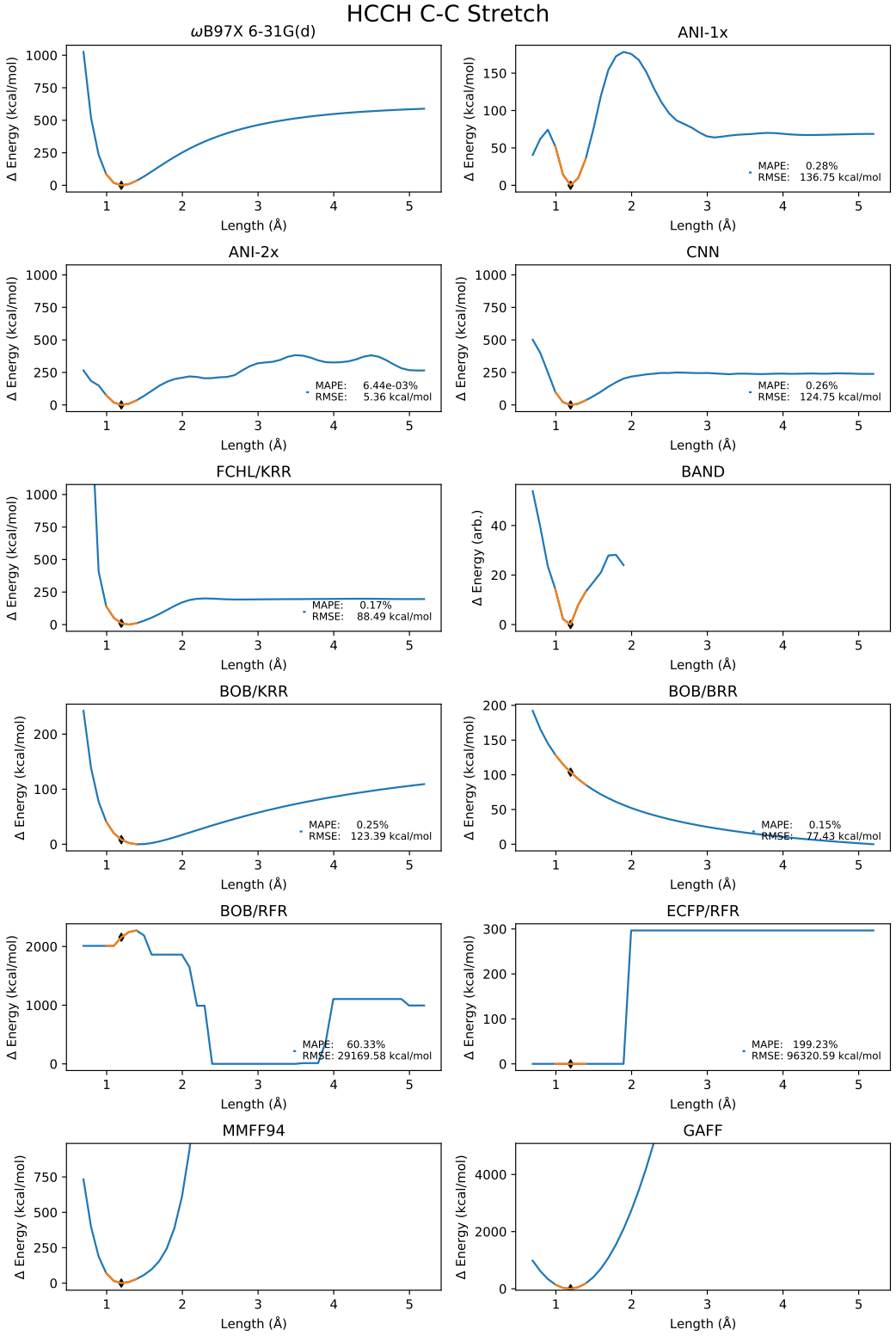

Figure B.17: Acetylene bond stretch for all methods

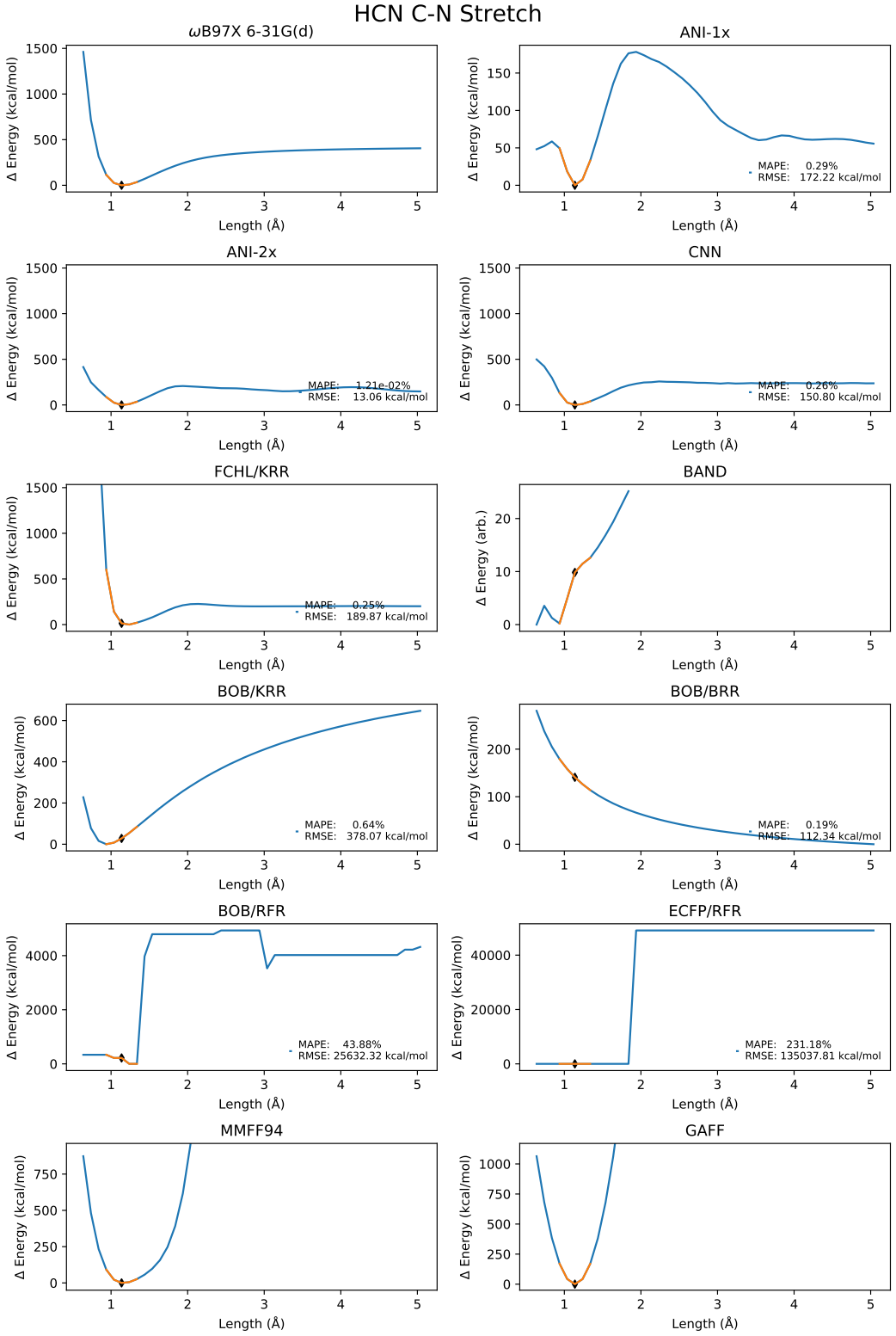

Figure B.18: Hydrogen cyanide bond stretch for all methods

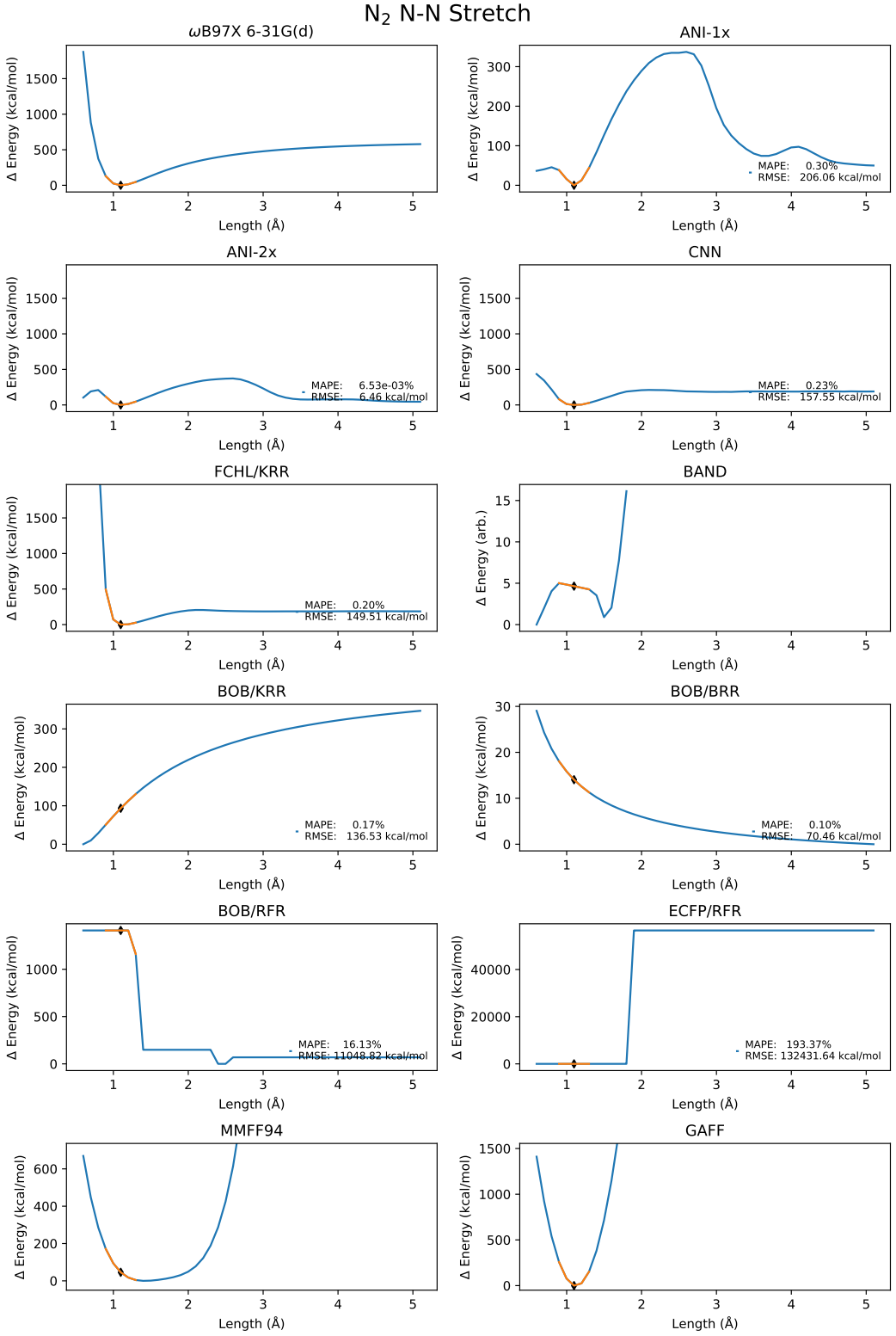

Figure B.19: N_2 bond stretch for all methods

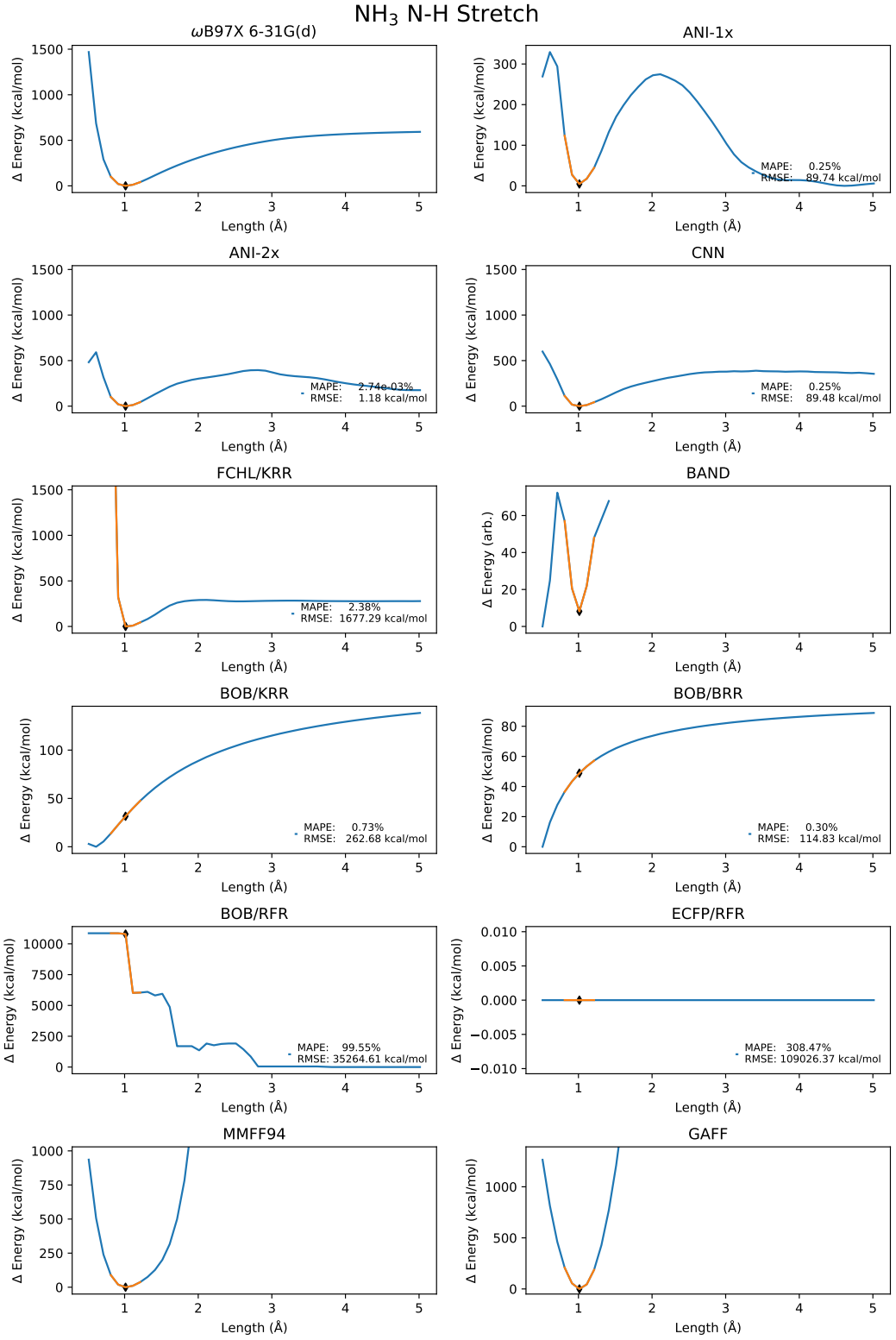

Figure B.20: Ammonia bond stretch for all methods

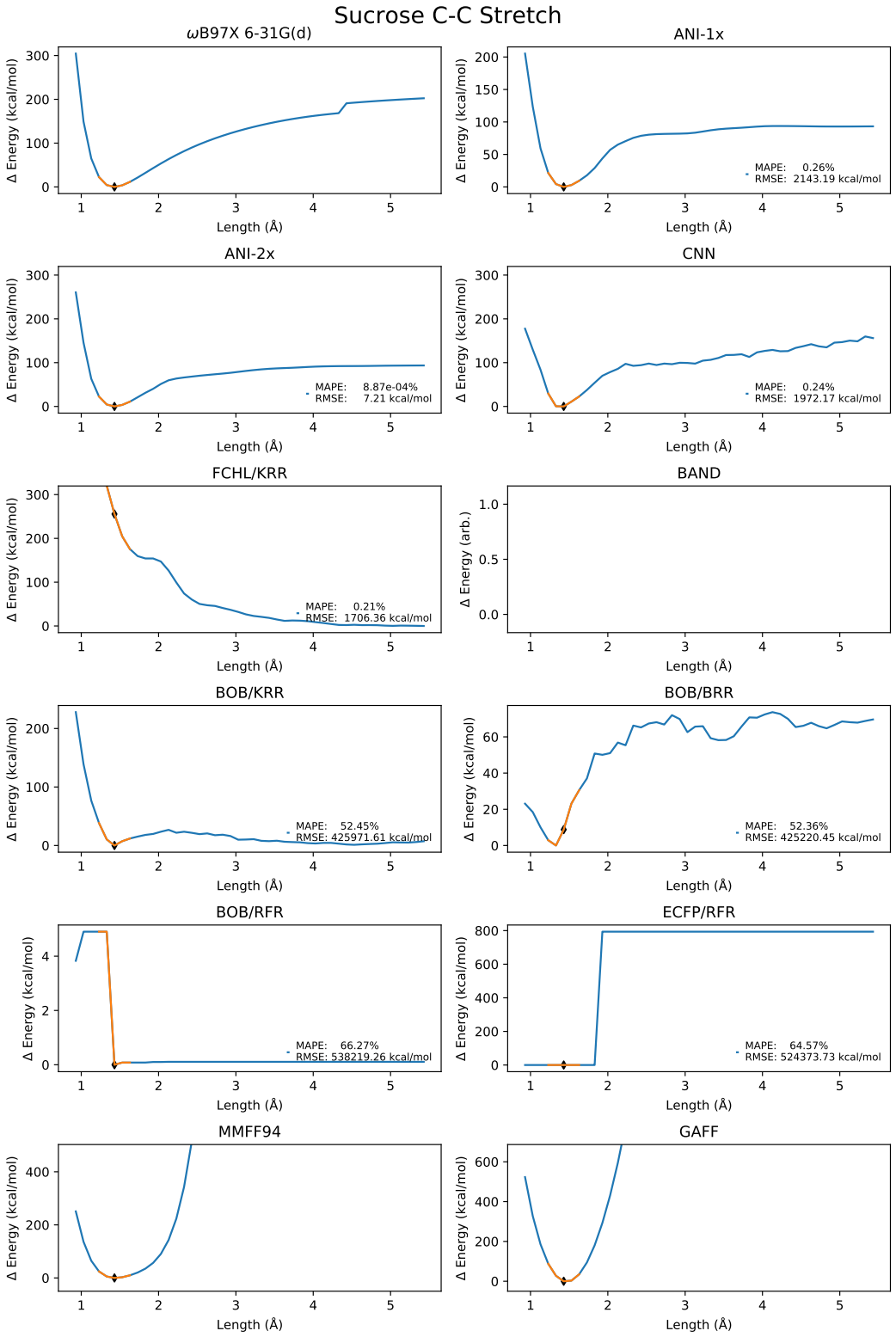

Figure B.21: Sucrose bond stretch for all methods

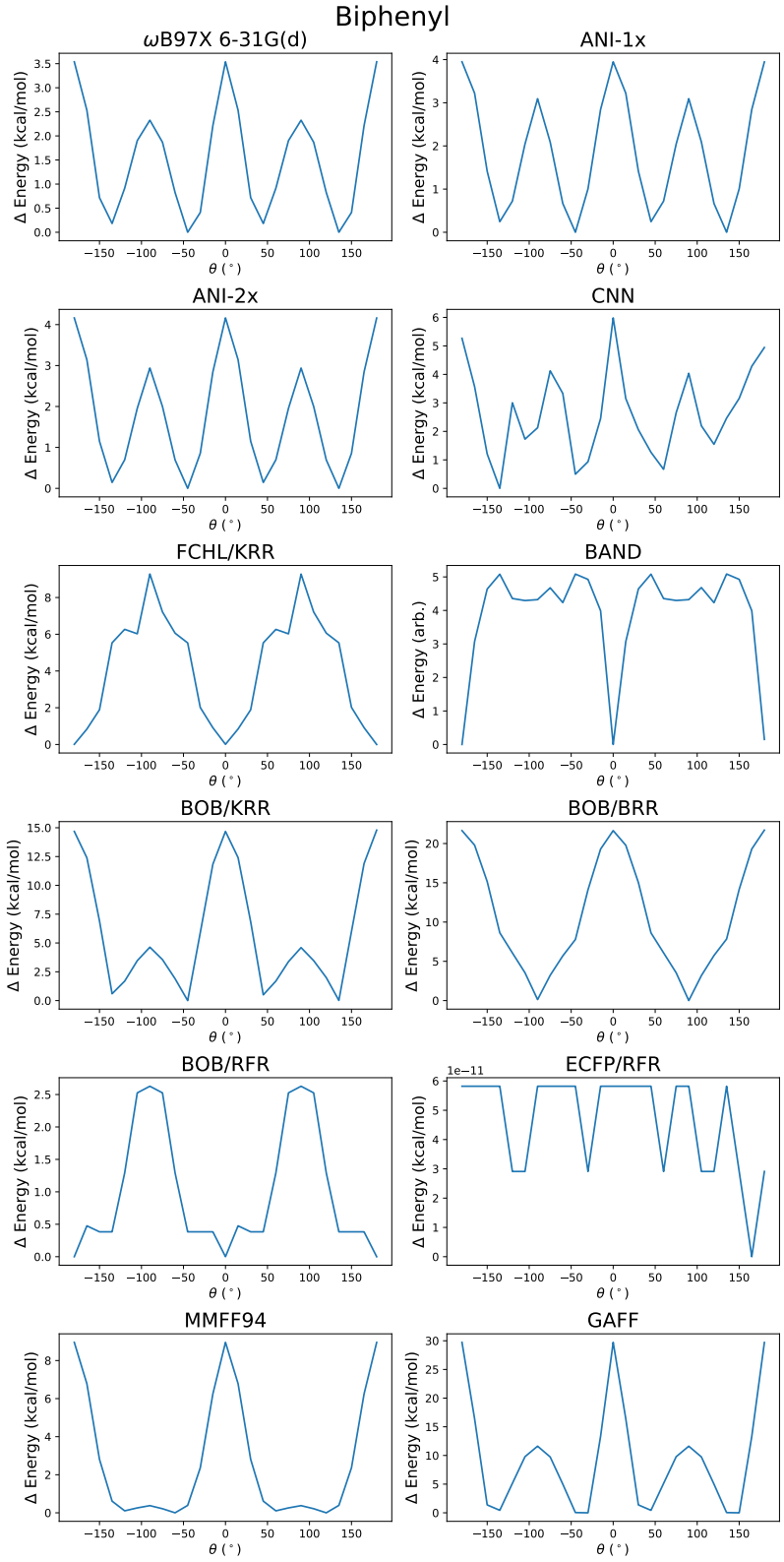

Figure B.22: Biphenyl torsion for all methods

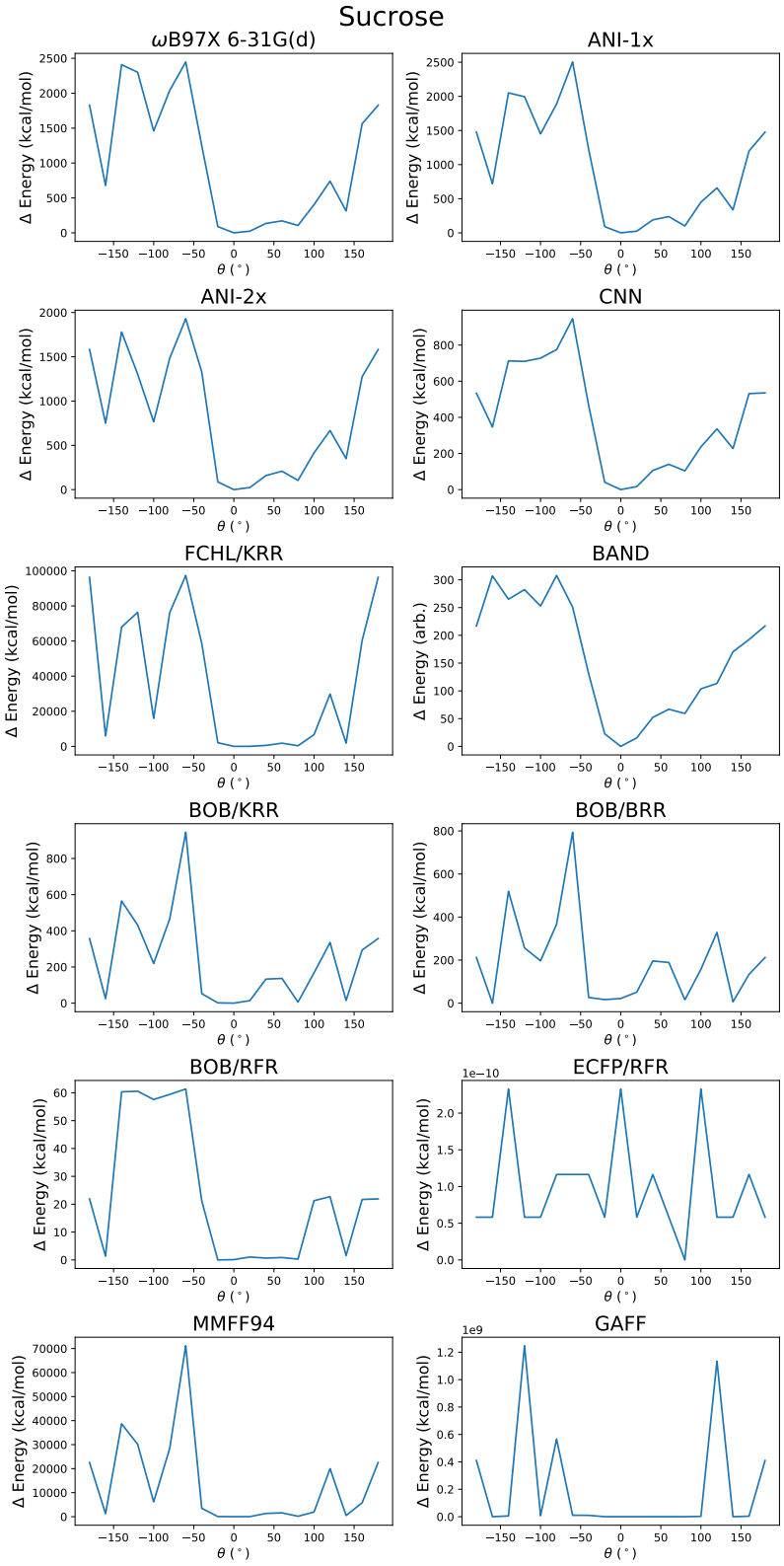

Figure B.23: Sucrose torsion for all methods

Appendix C: Supplementary Information for Systematic Comparison of Experimental Crystallographic Geometries and Gas-Phase Computed Conformers

C.1 Supplementary Figures

All raw data, Python notebooks, and torsion pattern figures can be found at `https://https://github.com/dlf57/quantum-torsions`.

Bibliography

[1] Reymond, J.-L.; van Deursen, R.; Blum, L. C.; Ruddigkeit, L. Chemical space as a source for new drugs. *Med. Chem. Commun.* **2010**, *1*, 30–38, DOI: 10.1039/C0MD00020E.

[2] Pyzer-Knapp, E. O.; Suh, C.; Gómez-Bombarelli, R.; Aguilera-Iparraguirre, J.; Aspuru-Guzik, A. What Is High-Throughput Virtual Screening? A Perspective from Organic Materials Discovery. *Annual Review of Materials Research* **2015**, *45*, 195–216, DOI: 10.1146/annurev-matsci-070214-020823.

[3] Kanal, I. Y.; Hutchison, G. R. Rapid Computational Optimization of Molecular Properties using Genetic Algorithms: Searching Across Millions of Compounds for Organic Photovoltaic Materials. 2017.

[4] Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Physical Review Letters* **2007**, *98*, DOI: 10.1103/physrevlett.98.146401.

[5] Faber, F. A.; Hutchison, L.; Huang, B.; Gilmer, J.; Schoenholz, S. S.; Dahl, G. E.; Vinyals, O.; Kearnes, S.; Riley, P. F.; von Lilienfeld, O. A. Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error. *Journal of Chemical Theory and Computation* **2017**, *13*, 5255–5264, DOI: 10.1021/acs.jctc.7b00577, PMID: 28926232.

[6] Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chemical Science* **2017**, *8*, 3192–3203, DOI: 10.1039/c6sc05720a.

[7] Smith, J. S.; Nebgen, B.; Lubbers, N.; Isayev, O.; Roitberg, A. E. Less is more: Sampling chemical space with active learning. *The Journal of Chemical Physics* **2018**, *148*, 241733, DOI: 10.1063/1.5023802.

[8] Devereux, C.; Smith, J. S.; Davis, K. K.; Barros, K.; Zubatyuk, R.; Isayev, O.; Roitberg, A. E. Extending the Applicability of the ANI Deep Learning Molecular Potential to Sulfur and Halogens. *Journal of Chemical Theory and Computation* **2020**, *16*, 4192–4202, DOI: 10.1021/acs.jctc.0c00121, PMID: 32543858.

[9] von Lilienfeld, O. A.; Burke, K. Retrospective on a decade of machine learning for chemical discovery. *Nature Communications* **2020**, *11*, DOI: 10.1038/s41467-020-18556-9.

[10] Dral, P. O. Quantum Chemistry in the Age of Machine Learning. *The Journal of Physical Chemistry Letters* **2020**, *11*, 2336–2347, DOI: 10.1021/acs.jpclett.9b03664.

[11] Qiao, Z.; Welborn, M.; Anandkumar, A.; Manby, F. R.; Miller, T. F. OrbNet: Deep learning for quantum chemistry using symmetry-adapted atomic-orbital features. *The Journal of Chemical Physics* **2020**, *153*, 124111, DOI: 10.1063/5.0021955.

[12] Sinitskiy, A. V.; Pande, V. S. Deep Neural Network Computes Electron Densities and Energies of a Large Set of Organic Molecules Faster than Density Functional Theory (DFT). 2018.

[13] Sinitskiy, A. V.; Pande, V. S. Physical machine learning outperforms "human learning" in Quantum Chemistry. 2020.

[14] Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. SchNet – A deep learning architecture for molecules and materials. *The Journal of Chemical Physics* **2018**, *148*, 241722, DOI: 10.1063/1.5019779.

[15] Weininger, D. SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules. *Journal of Chemical Information and Computer Sciences* **1988**, *28*, 31–36, DOI: 10.1021/CI00057A005.

[16] Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *Journal of Chemical Information and Computer Sciences* **1989**, *29*, 97–101, DOI: 10.1021/ci00062a008.

[17] Weininger, D. Smiles. 3. Depict. Graphical Depiction of Chemical Structures. *Journal of Chemical Information and Computer Sciences* **1990**, *30*, 237–243, DOI: 10.1021/CI00067A005.

[18] Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *Journal of Cheminformatics 2015 7:1* **2015**, *7*, 1–34, DOI: 10.1186/S13321-015-0068-4.

[19] Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Central Science* **2018**, *4*, 268–276, DOI: 10.1021/acscentsci.7b00572, PMID: 29532027.

[20] Kusner, M. J.; Paige, B.; Hernández-Lobato, J. M. Grammar Variational Autoencoder. 2017.

[21] O'Boyle, N.; Dalke, A. DeepSMILES: An Adaptation of SMILES for Use in Machine-Learning of Chemical Structures. **2018**, DOI: 10.26434/CHEMRXIV.7097960.V1.

[22] Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 058301, DOI: 10.1103/PhysRevLett.108.058301.

[23] Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; von Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A. Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space. *The Journal of Physical Chemistry Letters* **2015**, *6*, 2326–2331, DOI: 10.1021/acs.jpclett.5b00831.

[24] Huang, B.; von Lilienfeld, O. A. Communication: Understanding molecular representations in machine learning: The role of uniqueness and target similarity. *The Journal of Chemical Physics* **2016**, *145*, 161102, DOI: 10.1063/1.4964627.

[25] Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling* **2010**, *50*, 742–754, DOI: 10.1021/CI100050T.

[26] Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular graph convolutions: moving beyond fingerprints. *Journal of Computer-Aided Molecular Design 2016 30:8* **2016**, *30*, 595–608, DOI: 10.1007/S10822-016-9938-8.

[27] Smith, J. S.; Nebgen, B. T.; Zubatyuk, R.; Lubbers, N.; Devereux, C.; Barros, K.; Tretiak, S.; Isayev, O.; Roitberg, A. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. **2019**, DOI: 10.26434/chemrxiv.6744440.v2.

[28] Faber, F. A.; Christensen, A. S.; Huang, B.; von Lilienfeld, O. A. Alchemical and structural distribution based representation for universal quantum machine learning. *The Journal of Chemical Physics* **2018**, *148*, 241717, DOI: 10.1063/1.5020710.

[29] Christensen, A. S.; Bratholm, L. A.; Faber, F. A.; von Lilienfeld, O. A. FCHL revisited: Faster and more accurate quantum machine learning. *The Journal of Chemical Physics* **2020**, *152*, 044107, DOI: 10.1063/1.5126701.

[30] Christensen, A. S.; Sirumalla, S. K.; Qiao, Z.; O'Connor, M. B.; Smith, D. G. A.; Ding, F.; Bygrave, P. J.; Anandkumar, A.; Welborn, M.; Manby, F. R.; Miller, T. F. OrbNet Denali: A machine learning potential for biological and organic chemistry with semi-empirical cost and DFT accuracy. *The Journal of Chemical Physics* **2021**, *155*, 204103, DOI: 10.1063/5.0061990.

[31] Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864–2875, DOI: 10.1021/ci300415d, PMID: 23088335.

[32] Blum, L. C.; Reymond, J.-L. 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13. *J. Am. Chem. Soc.* **2009**, *131*, 8732.

[33] Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data* **2014**, *1*.

[34] Wu, Z.; Ramsundar, B.; Feinberg, E.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* **2018**, *9*, 513–530, DOI: 10.1039/C7SC02664A.

[35] Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1, A data set of 20 million calculated off-equilibrium conformations for organic molecules. *Scientific Data* **2017**, *4*, 170193, DOI: 10.1038/sdata.2017.193.

[36] Laghuvarapu, S.; Pathak, Y.; Priyakumar, U. D. BAND NN: A Deep Learning Framework for Energy Prediction and Geometry Optimization of Organic Small Molecules. *Journal of Computational Chemistry* **2020**, *41*, 790–799, DOI: 10.1002/jcc.26128.

[37] Riniker, S.; Landrum, G. A. Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *Journal of Chemical Information and Modeling* **2015**, *55*, 2562–2574, DOI: 10.1021/acs.jcim.5b00654.

[38] Grimme, S. *Reviews in Computational Chemistry*; John Wiley & Sons Inc., 2004; pp 153–218, DOI: 10.1002/0471678856.ch3.

[39] Lodewyk, M. W.; Siebert, M. R.; Tantillo, D. J. Computational Prediction of1H and13C Chemical Shifts: A Useful Tool for Natural Product Mechanistic, and Synthetic Organic Chemistry. *Chemical Reviews* **2011**, *112*, 1839–1862, DOI: 10.1021/cr200106v.

[40] Jackson, N. E.; Savoie, B. M.; Kohlstedt, K. L.; Marks, T. J.; Chen, L. X.; Ratner, M. A. Structural and Conformational Dispersion in the Rational Design of Conjugated Polymers. *Macromolecules* **2014**, *47*, 987–992, DOI: 10.1021/ma4023923.

[41] Hawkins, P. C. D. Conformation Generation: The State of the Art. *Journal of Chemical Information and Modeling* **2017**, *57*, 1747–1756, DOI: 10.1021/acs.jcim.7b00221.

[42] Kanal, I. Y.; Keith, J. A.; Hutchison, G. R. A sobering assessment of small-molecule force field methods for low energy conformer predictions. *International Journal of Quantum Chemistry* **2017**, *118*, e25512, DOI: 10.1002/qua.25512.

[43] Habgood, M.; James, T.; Heifetz, A. *Quantum Mechanics in Drug Discovery*; Springer US, 2020; pp 207–229, DOI: 10.1007/978-1-0716-0282-$9_1$4.

[44] Sharapa, D. I.; Genaev, A.; Cavallo, L.; Minenkov, Y. A Robust and Cost-Efficient Scheme for Accurate Conformational Energies of Organic Molecules. *ChemPhysChem* **2018**, DOI: 10.1002/cphc.201801063.

[45] Kesharwani, M. K.; Karton, A.; Martin, J. M. L. Benchmark ab Initio Conformational Energies for the Proteinogenic Amino Acids through Explicitly Correlated Methods. Assessment of Density Functional Methods. *Journal of Chemical Theory and Computation* **2015**, *12*, 444–454, DOI: 10.1021/acs.jctc.5b01066.

[46] Řezáč, J.; Bím, D.; Gutten, O.; Rulíšek, L. Toward Accurate Conformational Energies of Smaller Peptides and Medium-Sized Macrocycles: MPCONF196 Benchmark Energy Data Set. *Journal of Chemical Theory and Computation* **2018**, *14*, 1254–1266, DOI: 10.1021/acs.jctc.7b01074.

[47] Prasad, V. K.; de-la Roza, A. O.; DiLabio, G. A. PEPCONF a diverse data set of peptide conformational energies. *Scientific Data* **2019**, *6*, DOI: 10.1038/sdata.2018.310.

[48] Kang, Y. K.; Park, H. S. Exploring conformational preferences of alanine tetrapeptide by CCSD(T) MP2, and dispersion-corrected DFT methods. *Chemical Physics Letters* **2018**, *702*, 69–75, DOI: 10.1016/j.cplett.2018.05.006.

[49] Yuan, Y.; Mills, M. J. L.; Popelier, P. L. A.; Jensen, F. Comprehensive Analysis of Energy Minima of the 20 Natural Amino Acids. *The Journal of Physical Chemistry A* **2014**, *118*, 7876–7891, DOI: 10.1021/jp503460m.

[50] Rai, B. K.; Sresht, V.; Yang, Q.; Unwalla, R.; Tu, M.; Mathiowetz, A. M.; Bakken, G. A. Comprehensive Assessment of Torsional Strain in Crystal Structures of Small Molecules and Protein–Ligand Complexes using ab Initio Calculations. *Journal of Chemical Information and Modeling* **2019**, *59*, 4195–4208, DOI: 10.1021/acs.jcim.9b00373.

[51] Foloppe, N.; Chen, I.-J. Energy windows for computed compound conformers: covering artefacts or truly large reorganization energies? *Future Medicinal Chemistry* **2019**, *11*, 97–118, DOI: 10.4155/fmc-2018-0400.

[52] Johansson, M. P.; Olsen, J. Torsional Barriers and Equilibrium Angle of Biphenyl: Reconciling Theory with Experiment. *Journal of Chemical Theory and Computation* **2008**, *4*, 1460–1471, DOI: 10.1021/ct800182e.

[53] Jackson, N. E.; Savoie, B. M.; Kohlstedt, K. L.; de la Cruz, M. O.; Schatz, G. C.; Chen, L. X.; Ratner, M. A. Controlling Conformations of Conjugated Polymers and Small Molecules: The Role of Nonbonding Interactions. *Journal of the American Chemical Society* **2013**, *135*, 10475–10483, DOI: 10.1021/ja403667s.

[54] Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Rassolov, V.; Pople, J. A. Gaussian-3 (G3) theory for molecules containing first and second-row atoms. *The Journal of Chemical Physics* **1998**, *109*, 7764–7776, DOI: 10.1063/1.477422.

[55] Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. Gaussian-4 theory. *The Journal of Chemical Physics* **2007**, *126*, 084108, DOI: 10.1063/1.2436888.

[56] Martin, J. M. L.; de Oliveira, G. Towards standard methods for benchmark quality ab initio thermochemistry—W1 and W2 theory. *The Journal of Chemical Physics* **1999**, *111*, 1843–1856, DOI: 10.1063/1.479454.

[57] Parthiban, S.; Martin, J. M. L. Assessment of W1 and W2 theories for the computation of electron affinities ionization potentials, heats of formation, and proton affinities. *The Journal of Chemical Physics* **2001**, *114*, 6014–6029, DOI: 10.1063/1.1356014.

[58] Karton, A.; Rabinovich, E.; Martin, J. M. L.; Ruscic, B. W4 theory for computational thermochemistry: In pursuit of confident sub-kJ/mol predictions. *The Journal of Chemical Physics* **2006**, *125*, 144108, DOI: 10.1063/1.2348881.

[59] Ghahremanpour, M. M.; van Maaren, P. J.; Ditz, J. C.; Lindh, R.; van der Spoel, D. Large-scale calculations of gas phase thermochemistry: Enthalpy of formation standard entropy, and heat capacity. *The Journal of Chemical Physics* **2016**, *145*, 114305, DOI: 10.1063/1.4962627.

[60] Hawkins, P.; Skillman, A.; Warren, G.; Ellingson, B.; Stahl, M. Conformer generation with OMEGA: algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database. *J Chem Inf Model* **2010**, *50*, 572–84.

[61] Juárez-Jiménez, J.; Barril, X.; Orozco, M.; Pouplana, R.; Luque, F. J. Assessing the Suitability of the Multilevel Strategy for the Conformational Analysis of Small Ligands. *The Journal of Physical Chemistry B* **2014**, *119*, 1164–1172, DOI: 10.1021/jp506779y.

[62] O'Boyle, N.; Banck, M.; James, C.; Morley, C.; Vandermeersch, T.; Hutchison, G. Open Babel: An open chemical toolbox. *J Cheminform* **2011**, *3*, 33.

[63] Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *Journal of Computational Chemistry* **1996**, *17*, 490–519.

[64] Halgren, T. A. Merck molecular force field. II. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions. *Journal of Computational Chemistry* **1996**, *17*, 520–552.

[65] Halgren, T. A. Merck molecular force field. III. Molecular geometries and vibrational frequencies for MMFF94. *Journal of Computational Chemistry* **1996**, *17*, 553–586.

[66] Halgren, T. A.; Nachbar, R. B. Merck molecular force field. IV. conformational energies and geometries for MMFF94. *Journal of Computational Chemistry* **1996**, *17*, 587–615.

[67] Halgren, T. A. Merck molecular force field. V. Extension of MMFF94 using experimental data, additional computational data, and empirical rules. *Journal of Computational Chemistry* **1996**, *17*, 616–641.

[68] Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A.; Skiff, W. M. UFF a full periodic table force field for molecular mechanics and molecular dynamics simulations. *Journal of the American Chemical Society* **1992**, *114*, 10024–10035, DOI: 10.1021/ja00051a040.

[69] Casewit, C. J.; Colwell, K. S.; Rappe, A. K. Application of a universal force field to organic molecules. *Journal of the American Chemical Society* **1992**, *114*, 10035–10046, DOI: 10.1021/ja00051a041.

[70] Stewart, J. J. P. Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters. *Journal of Molecular Modeling* **2012**, *19*, 1–32, DOI: 10.1007/s00894-012-1667-x.

[71] grimme-lab/xtb. GitHub, `https://github.com/grimme-lab/xtb`, Accessed on Thu, October 03, 2019.

[72] Pracht, P.; Caldeweyher, E.; Ehlert, S.; Grimme, S. A Robust Non-Self-Consistent Tight-Binding Quantum Chemistry Method for large Molecules. **2019**, DOI: 10.26434/chemrxiv.8326202.v1.

[73] Grimme, S.; Bannwarth, C.; Shushkov, P. A Robust and Accurate Tight-Binding Quantum Chemical Method for Structures Vibrational Frequencies, and Noncovalent Interactions of Large Molecular Systems Parametrized for All spd-Block Elements (Z = 1–86). *Journal of Chemical Theory and Computation* **2017**, *13*, 1989–2009, DOI: 10.1021/acs.jctc.7b00118.

[74] Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB - an Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. **2018**, DOI: 10.26434/chemrxiv.7246238.v2.

[75] Neese, F. The ORCA program system. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2011**, *2*, 73–78, DOI: 10.1002/wcms.81.

[76] Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *Journal of Computational Chemistry* **2011**, *32*, 1456–1465, DOI: 10.1002/jcc.21759.

[77] Becke, A. D.; Johnson, E. R. A density-functional model of the dispersion interaction. *The Journal of Chemical Physics* **2005**, *123*, 154101, DOI: 10.1063/1.2065267.

[78] Johnson, E. R.; Becke, A. D. A post-Hartree–Fock model of intermolecular interactions. *The Journal of Chemical Physics* **2005**, *123*, 024101, DOI: 10.1063/1.1949201.

[79] Johnson, E. R.; Becke, A. D. A post-Hartree-Fock model of intermolecular interactions: Inclusion of higher-order corrections. *The Journal of Chemical Physics* **2006**, *124*, 174104, DOI: 10.1063/1.2190220.

[80] Chai, J.-D.; Head-Gordon, M. Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections. *Physical Chemistry Chemical Physics* **2008**, *10*, 6615, DOI: 10.1039/b810189b.

[81] Kossmann, S.; Neese, F. Efficient Structure Optimization with Second-Order Many-Body Perturbation Theory: The RIJCOSX-MP2 Method. *Journal of Chemical Theory and Computation* **2010**, *6*, 2325–2338, DOI: 10.1021/ct100199k.

[82] Liakos, D. G.; Neese, F. Is It Possible To Obtain Coupled Cluster Quality Energies at near Density Functional Theory Cost? Domain-Based Local Pair Natural Orbital Coupled Cluster vs Modern Density Functional Theory. *Journal of Chemical Theory and Computation* **2015**, *11*, 4054–4063, DOI: 10.1021/acs.jctc.5b00359.

[83] Guo, Y.; Riplinger, C.; Becker, U.; Liakos, D. G.; Minenkov, Y.; Cavallo, L.; Neese, F. Communication: An improved linear scaling perturbative triples correction for the domain based local pair-natural orbital based singles and doubles coupled cluster method [DLPNO-CCSD(T)]. *The Journal of Chemical Physics* **2018**, *148*, 011101, DOI: 10.1063/1.5011798.

[84] Dunning, T. H. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *The Journal of Chemical Physics* **1989**, *90*, 1007–1023, DOI: 10.1063/1.456153.

[85] Kendall, R. A.; Dunning, T. H.; Harrison, R. J. Electron affinities of the first-row atoms revisited. Systematic basis sets and wave functions. *The Journal of Chemical Physics* **1992**, *96*, 6796–6806, DOI: 10.1063/1.462569.

[86] O'boyle, N. M.; Tenderholt, A. L.; Langner, K. M. cclib: A library for package-independent computational chemistry algorithms. *Journal of Computational Chemistry* **2008**, *29*, 839–845, DOI: 10.1002/jcc.20823.

[87] O'Boyle, N. M.; Morley, C.; Hutchison, G. R. Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit. *Chemistry Central Journal* **2008**, *2*, 5, DOI: 10.1186/1752-153X-2-5.

[88] Folmsbee, D.; Upadhyay, S.; Dumi, A.; Hiener, D.; Mulvey, D. chemreps/chemreps: Molecular Machine Learning Representations. 2019; `https://doi.org/10.5281/zenodo.3333856`.

[89] Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

[90] McKinney, W. Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference. 2010; pp 51 – 56.

[91] van der Walt, S.; Colbert, S. C.; Varoquaux, G. The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science & Engineering* **2011**, *13*, 22–30, DOI: 10.1109/mcse.2011.37.

[92] Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods* **2020**, *17*, 261–272, DOI: 10.1038/s41592-019-0686-2.

[93] Rego, N.; Koes, D. 3Dmol.js: molecular visualization with WebGL. *Bioinformatics* **2014**, *31*, 1322–1324, DOI: 10.1093/bioinformatics/btu829.

[94] Inc., P. T.

[95] Ebejer, J.; Morris, G.; Deane, C. Freely available conformer generation methods: how good are they? *J Chem Inf Model* **2012**, *52*, 1146–58.

[96] Hartshorn, M.; Verdonk, M.; Chessari, G.; Brewerton, S.; Mooij, W.; Mortenson, P.; Murray, C. Diverse, high-quality test set for the validation of protein-ligand docking performance. *J Med Chem* **2007**, *50*, 726–41.

[97] Weininger, D. SMILES a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Modeling* **1988**, *28*, 31–36, DOI: 10.1021/ci00057a005.

[98] Paulechka, E.; Kazakov, A. Efficient DLPNO–CCSD(T)-Based Estimation of Formation Enthalpies for C- H-, O-, and N-Containing Closed-Shell Compounds Validated Against Critically Evaluated Experimental Data. *The Journal of Physical Chemistry A* **2017**, *121*, 4379–4387, DOI: 10.1021/acs.jpca.7b03195.

[99] Liakos, D. G.; Guo, Y.; Neese, F. Comprehensive Benchmark Results for the Domain Based Local Pair Natural Orbital Coupled Cluster Method (DLPNO-CCSD(T)) for Closed- and Open-Shell Systems. *The Journal of Physical Chemistry A* **2019**, DOI: 10.1021/acs.jpca.9b05734.

[100] Vo, M. N.; Call, M.; Kowall, C.; Johnson, J. K. Method for Predicting Dipole Moments of Complex Molecules for Use in Thermophysical Property Estimation. *Industrial & Engineering Chemistry Research* **2019**, DOI: 10.1021/acs.iecr.9b03699.

[101] Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *Journal of Computational Chemistry* **2004**, *25*, 1157–1174, DOI: 10.1002/jcc.20035.

[102] Grimme, S.; Brandenburg, J. G.; Bannwarth, C.; Hansen, A. Consistent structures and interactions by density functional theory with small atomic orbital basis sets. *The Journal of Chemical Physics* **2015**, *143*, 054107, DOI: 10.1063/1.4927476.

[103] Brandenburg, J. G.; Bannwarth, C.; Hansen, A.; Grimme, S. B97-3c: A revised low-cost variant of the B97-D density functional method. *The Journal of Chemical Physics* **2018**, *148*, 064104, DOI: 10.1063/1.5012601.

[104] Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Physical Review B* **1988**, *37*, 785–789, DOI: 10.1103/physrevb.37.785.

[105] Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Physical Review A* **1988**, *38*, 3098–3100, DOI: 10.1103/physreva.38.3098.

[106] Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *The Journal of Physical Chemistry* **1994**, *98*, 11623–11627, DOI: 10.1021/j100096a001.

[107] Vosko, S. H.; Wilk, L.; Nusair, M. Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis. *Canadian Journal of Physics* **1980**, *58*, 1200–1211, DOI: 10.1139/p80-159.

[108] Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple [Phys. Rev. Lett. 77 3865 (1996)]. *Physical Review Letters* **1997**, *78*, 1396–1396, DOI: 10.1103/physrevlett.78.1396.

[109] Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Physical Review Letters* **1996**, *77*, 3865–3868, DOI: 10.1103/physrevlett.77.3865.

[110] Weigend, F.; Ahlrichs, R. Balanced basis sets of split valence triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Physical Chemistry Chemical Physics* **2005**, *7*, 3297, DOI: 10.1039/b508541a.

[111] Weigend, F. Accurate Coulomb-fitting basis sets for H to Rn. *Physical Chemistry Chemical Physics* **2006**, *8*, 1057, DOI: 10.1039/b515623h.

[112] Martin, J. M. L. What Can We Learn about Dispersion from the Conformer Surface ofn-Pentane? *The Journal of Physical Chemistry A* **2013**, *117*, 3118–3132, DOI: 10.1021/jp401429u.

[113] Gruzman, D.; Karton, A.; Martin, J. M. L. Performance of Ab Initio and Density Functional Methods for Conformational Equilibria ofCnH2n+2Alkane Isomers (n= 4-8)†. *The Journal of Physical Chemistry A* **2009**, *113*, 11974–11983, DOI: 10.1021/jp903640h.

[114] Caldeweyher, E.; Bannwarth, C.; Grimme, S. Extension of the D3 dispersion coefficient model. *The Journal of Chemical Physics* **2017**, *147*, 034112, DOI: 10.1063/1.4993215.

[115] Caldeweyher, E.; Ehlert, S.; Hansen, A.; Neugebauer, H.; Spicher, S.; Bannwarth, C.; Grimme, S. A generally applicable atomic-charge dependent London dispersion correction. *The Journal of Chemical Physics* **2019**, *150*, 154122, DOI: 10.1063/1.5090222.

[116] Grimme, S. Density functional theory with London dispersion corrections. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2011**, *1*, 211–228, DOI: 10.1002/wcms.30.

[117] Johnson, E. R.; Mackie, I. D.; DiLabio, G. A. Dispersion interactions in density-functional theory. *Journal of Physical Organic Chemistry* **2009**, *22*, 1127–1135, DOI: 10.1002/poc.1606.

[118] Witte, J.; Mardirossian, N.; Neaton, J. B.; Head-Gordon, M. Assessing DFT-D3 Damping Functions Across Widely Used Density Functionals: Can We Do Better? *Journal of Chemical Theory and Computation* **2017**, *13*, 2043–2052, DOI: 10.1021/acs.jctc.7b00176.

[119] Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. *Journal of the American Chemical Society* **1985**, *107*, 3902–3909, DOI: 10.1021/ja00299a024.

[120] Stewart, J. J. P. Optimization of parameters for semiempirical methods I. Method. *Journal of Computational Chemistry* **1989**, *10*, 209–220, DOI: 10.1002/jcc.540100208.

[121] Stewart, J. J. P. Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements. *Journal of Molecular Modeling* **2007**, *13*, 1173–1213, DOI: 10.1007/s00894-007-0233-4.

[122] Hansen, K.; Montavon, G.; Biegler, F.; Fazli, S.; Rupp, M.; Scheffler, M.; von Lilienfeld, O. A.; Tkatchenko, A.; Müller, K.-R. Assessment and Validation of Machine

Learning Methods for Predicting Molecular Atomization Energies. *Journal of Chemical Theory and Computation* **2013**, *9*, 3404–3419, DOI: 10.1021/ct400195d, PMID: 26584096.

[123] Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data* **2014**, *1*, DOI: 10.1038/sdata.2014.22.

[124] dftd4/dftd4. https://github.com/dftd4/dftd4, `https://github.com/dftd4/dftd4`, Accessed on Tue, May 12, 2020.

[125] Laghuvarapu, S.; Pathak, Y.; Priyakumar, U. D. BAND NN: A Deep Learning Framework for Energy Prediction and Geometry Optimization of Organic Small Molecules. *Journal of Computational Chemistry* **2019**, *41*, 790–799, DOI: 10.1002/jcc.26128.

[126] Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv* **2015**, *abs/1502.03167*.

[127] Klambauer, G.; Unterthiner, T.; Mayr, A.; Hochreiter, S. In *Advances in Neural Information Processing Systems 30*; Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc., 2017; pp 971–980.

[128] Wahl, J.; Freyss, J.; von Korff, M.; Sander, T. Accuracy evaluation and addition of improved dihedral parameters for the MMFF94s. *Journal of Cheminformatics* **2019**, *11*, DOI: 10.1186/s13321-019-0371-6.

[129] van der Spoel, D.; Ghahremanpour, M. M.; Lemkul, J. A. Small Molecule Thermochemistry: A Tool for Empirical Force Field Development. *The Journal of Physical Chemistry A* **2018**, *122*, 8982–8988, DOI: 10.1021/acs.jpca.8b09867.

[130] Roos, K.; Wu, C.; Damm, W.; Reboul, M.; Stevenson, J. M.; Lu, C.; Dahlgren, M. K.; Mondal, S.; Chen, W.; Wang, L.; Abel, R.; Friesner, R. A.; Harder, E. D. OPLS3e: Extending Force Field Coverage for Drug-Like Small Molecules. *Journal of Chemical Theory and Computation* **2019**, *15*, 1863–1874, DOI: 10.1021/acs.jctc.8b01026.

[131] Harder, E.; Damm, W.; Maple, J.; Wu, C.; Reboul, M.; Xiang, J. Y.; Wang, L.; Lupyan, D.; Dahlgren, M. K.; Knight, J. L.; Kaus, J. W.; Cerutti, D. S.; Krilov, G.; Jorgensen, W. L.; Abel, R.; Friesner, R. A. OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. *Journal of Chemical Theory and Computation* **2015**, *12*, 281–296, DOI: 10.1021/acs.jctc.5b00864.

[132] Lin, F.-Y.; MacKerell, A. D. In *Biomolecular Simulations: Methods and Protocols*; Bonomi, M., Camilloni, C., Eds.; Springer New York: New York, NY, 2019; pp 21–54, DOI: 10.1007/978-1-4939-9608-7_2.

[133] Inakollu, V. S.; Geerke, D. P.; Rowley, C. N.; Yu, H. Polarisable force fields: what do they add in biomolecular simulations? *Current Opinion in Structural Biology* **2020**, *61*, 182–190, DOI: 10.1016/j.sbi.2019.12.012.

[134] Warshel, A.; Kato, M.; Pisliakov, A. V. Polarizable Force Fields: History Test Cases, and Prospects. *Journal of Chemical Theory and Computation* **2007**, *3*, 2034–2045, DOI: 10.1021/ct700127w.

[135] Jing, Z.; Liu, C.; Cheng, S. Y.; Qi, R.; Walker, B. D.; Piquemal, J.-P.; Ren, P. Polarizable Force Fields for Biomolecular Simulations: Recent Advances and Applications. *Annual Review of Biophysics* **2019**, *48*, 371–394, DOI: 10.1146/annurev-biophys-070317-033349.

[136] Zhang, C.; Lu, C.; Jing, Z.; Wu, C.; Piquemal, J.-P.; Ponder, J. W.; Ren, P. AMOEBA Polarizable Atomic Multipole Force Field for Nucleic Acids. *Journal of Chemical Theory and Computation* **2018**, *14*, 2084–2108, DOI: 10.1021/acs.jctc.7b01169.

[137] Rackers, J. A.; Wang, Q.; Liu, C.; Piquemal, J.-P.; Ren, P.; Ponder, J. W. An optimized charge penetration model for use with the AMOEBA force field. *Physical Chemistry Chemical Physics* **2017**, *19*, 276–291, DOI: 10.1039/c6cp06017j.

[138] Ponder, J. W.; Wu, C.; Ren, P.; Pande, V. S.; Chodera, J. D.; Schnieders, M. J.; Haque, I.; Mobley, D. L.; Lambrecht, D. S.; DiStasio, R. A.; Head-Gordon, M.; Clark, G. N. I.; Johnson, M. E.; Head-Gordon, T. Current Status of the AMOEBA Polarizable Force Field. *The Journal of Physical Chemistry B* **2010**, *114*, 2549–2564, DOI: 10.1021/jp910674d.

[139] Liu, C.; Piquemal, J.-P.; Ren, P. AMOEBA+ Classical Potential for Modeling Molecular Interactions. *Journal of Chemical Theory and Computation* **2019**, *15*, 4122–4139, DOI: 10.1021/acs.jctc.9b00261.

[140] The Open Force Field 1.0 small molecule force field, our first optimized force field (codename Parsley). https://openforcefield.org/news/introducing-openforcefield-1.0/, `//openforcefield.org/news/introducing-openforcefield-1.0/`, Accessed on Fri, February 14, 2020.

[141] Beauchamp, K. A.; Behr, J. M.; Rustenburg, A. S.; Bayly, C. I.; Kroenlein, K.; Chodera, J. D. Toward Automated Benchmarking of Atomistic Force Fields: Neat Liquid Densities and Static Dielectric Constants from the ThermoML Data Archive. *The Journal of Physical Chemistry B* **2015**, *119*, 12912–12920, DOI: 10.1021/acs.jpcb.5b06703.

[142] Zanette, C.; Bannan, C. C.; Bayly, C. I.; Fass, J.; Gilson, M. K.; Shirts, M. R.; Chodera, J. D.; Mobley, D. L. Toward Learned Chemical Perception of Force Field Typing Rules. *Journal of Chemical Theory and Computation* **2018**, *15*, 402–423, DOI: 10.1021/acs.jctc.8b00821.

[143] Waldher, B.; Kuta, J.; Chen, S.; Henson, N.; Clark, A. E. ForceFit: A code to fit classical force fields to quantum mechanical potential energy surfaces. *Journal of Computational Chemistry* **2010**, NA–NA, DOI: 10.1002/jcc.21523.

[144] Zahariev, F.; Silva, N. D.; Gordon, M. S.; Windus, T. L.; Dick-Perez, M. ParFit: A Python-Based Object-Oriented Program for Fitting Molecular Mechanics Parameters to ab Initio Data. *Journal of Chemical Information and Modeling* **2017**, *57*, 391–396, DOI: 10.1021/acs.jcim.6b00654.

[145] Grimme, S. A General Quantum Mechanically Derived Force Field (QMDFF) for Molecules and Condensed Phase Simulations. *Journal of Chemical Theory and Computation* **2014**, *10*, 4497–4514, DOI: 10.1021/ct500573f.

[146] Basdogan, Y.; Maldonado, A. M.; Keith, J. A. Advances and challenges in modeling solvated reaction mechanisms for renewable fuels and chemicals. *WIREs Computational Molecular Science* **2019**, *10*, DOI: 10.1002/wcms.1446.

[147] Basdogan, Y.; Keith, J. A. A paramedic treatment for modeling explicitly solvated chemical reaction mechanisms. *Chemical Science* **2018**, *9*, 5341–5346, DOI: 10.1039/c8sc01424h.

[148] Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864–2875, DOI: 10.1021/ci300415d, PMID: 23088335.

[149] Folmsbee, D.; Hutchison, G. Assessing conformer energies using electronic structure and machine learning methods. *International Journal of Quantum Chemistry n/a*, e26381, DOI: 10.1002/qua.26381.

[150] DuBay, K. H.; Hall, M. L.; Hughes, T. F.; Wu, C.; Reichman, D. R.; Friesner, R. A. Accurate Force Field Development for Modeling Conjugated Polymers. *Journal of Chemical Theory and Computation* **2012**, *8*, 4556–4569, DOI: 10.1021/ct300175w.

[151] Wildman, J.; Repiščák, P.; Paterson, M. J.; Galbraith, I. General Force-Field Parametrization Scheme for Molecular Dynamics Simulations of Conjugated Materials in Solution. *Journal of Chemical Theory and Computation* **2016**, *12*, 3813–3824, DOI: 10.1021/acs.jctc.5b01195.

[152] Kanal, I. Y.; Keith, J. A.; Hutchison, G. R. A sobering assessment of small-molecule force field methods for low energy conformer predictions. *International Journal of Quantum Chemistry* **2017**, *118*, e25512, DOI: 10.1002/qua.25512.

[153] Chai, J.-D.; Head-Gordon, M. Systematic optimization of long-range corrected hybrid density functionals. *The Journal of Chemical Physics* **2008**, *128*, 084106, DOI: 10.1063/1.2834918.

[154] Neese, F. The ORCA program system. *WIREs Computational Molecular Science* **2012**, *2*, 73–78, DOI: 10.1002/wcms.81.

[155] Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *Journal of Computational Chemistry* **2004**, *25*, 1157–1174, DOI: 10.1002/jcc.20035.

[156] O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *Journal of Cheminformatics* **2011**, *3*, 33, DOI: 10.1186/1758-2946-3-33.

[157] Christensen, A.; Faber, F.; Huang, B.; Bratholm, L.; Tkatchenko, A.; Müller, K.; von Lilienfeld, O. QML: A Python Toolkit for Quantum Machine Learning. 2017; `https://github.com/qmlcode/qml`.

[158] Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling* **2010**, *50*, 742–754, DOI: 10.1021/ci100050t, PMID: 20426451.

[159] Wójcikowski, M.; Zielenkiewicz, P.; Siedlecki, P. Open Drug Discovery Toolkit (ODDT): a new open-source player in the drug discovery field. *Journal of Cheminformatics* **2015**, *7*, 26, DOI: 10.1186/s13321-015-0078-2.

[160] Ragoza, M.; Hochuli, J.; Idrobo, E.; Sunseri, J.; Koes, D. R. Protein–ligand scoring with convolutional neural networks. *Journal of chemical information and modeling* **2017**, *57*, 942–957.

[161] Jiménez, J.; Skalic, M.; Martinez-Rosell, G.; De Fabritiis, G. K deep: Protein–ligand absolute binding affinity prediction via 3d-convolutional neural networks. *Journal of chemical information and modeling* **2018**, *58*, 287–296.

[162] Sunseri, J.; Koes, D. R. libmolgrid: Graphics Processing Unit Accelerated Molecular Gridding for Deep Learning Applications. *Journal of Chemical Information and Modeling* **2020**, *60*, 1079–1084.

[163] Smith, J. S.; Zubatyuk, R.; Nebgen, B.; Lubbers, N.; Barros, K.; Roitberg, A. E.; Isayev, O.; Tretiak, S. The ANI-1ccx and ANI-1x data sets, coupled-cluster and density functional theory properties for molecules. *Scientific data* **2020**, *7*, 1–10.

[164] Nakata, M.; Shimazaki, T. PubChemQC Project: A Large-Scale First-Principles Electronic Structure Database for Data-Driven Chemistry. *Journal of Chemical Information and Modeling* **2017**, *57*, 1300–1308, DOI: 10.1021/acs.jcim.7b00083, PMID: 28481528.

[165] Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *Journal of Chemical Theory and Computation* **2019**, *15*, 1652–1671, DOI: 10.1021/acs.jctc.8b01176, PMID: 30741547.

[166] Nakata, M.; Shimazaki, T. PubChemQC Project: A Large-Scale First-Principles Electronic Structure Database for Data-Driven Chemistry. *Journal of Chemical Information and Modeling* **2017**, *57*, 1300–1308, DOI: 10.1021/acs.jcim.7b00083, PMID: 28481528.

[167] Gražulis, S.; Merkys, A.; Vaitkus, A.; Chateigner, D.; Lutterotti, L.; Moeck, P.; Quiros, M.; Downs, R. T.; Kaminsky, W.; Bail, A. L. In *Materials Informatics: Methods, Tools and Applications*; Isayev, O., Tropsha, A., Curtarolo, S., Eds.; Wiley, 2019; Chapter 1, pp 1–39, DOI: 10.1002/9783527802265.ch1.

[168] Sterling, T.; Irwin, J. J. ZINC 15 – Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55*, 2324–2337, DOI: 10.1021/acs.jcim.5b00559.

[169] Friedrich, N.-O.; Meyder, A.; de Bruyn Kops, C.; Sommer, K.; Flachsenberg, F.; Rarey, M.; Kirchmair, J. High-Quality Dataset of Protein-Bound Ligand Conformations and Its Application to Benchmarking Conformer Ensemble Generators. *Journal of Chemical Information and Modeling* **2017**, *57*, 529–539, DOI: 10.1021/acs.jcim.6b00613.

[170] Wang, S.; Witek, J.; Landrum, G. A.; Riniker, S. Improving Conformer Generation for Small Rings and Macrocycles Based on Distance Geometry and Experimental Torsional-Angle Preferences. *J. Chem. Inf. Model.* **2020**, *60*, 2044–2058.

[171] Landrum, G. RDKit: Open-Source Cheminformatics. Available at `http://www.rdkit.org`, 2020; `http://www.rdkit.org`.

[172] Folmsbee, D.; Hutchison, G. Assessing conformer energies using electronic structure and machine learning methods. *International Journal of Quantum Chemistry* **2021**, *121*, e26381, DOI: 10.1002/qua.26381.

[173] Curtarolo, S.; Hart, G. L. W.; Nardelli, M. B.; Mingo, N.; Sanvito, S.; Levy, O. The high-throughput highway to computational materials design. *Nature Materials* **2013**, *12*, 191–201, DOI: 10.1038/nmat3568.

[174] Schleder, G. R.; Padilha, A. C. M.; Acosta, C. M.; Costa, M.; Fazzio, A. From DFT to machine learning: recent approaches to materials science–a review. *Journal of Physics: Materials* **2019**, *2*, 032001, DOI: 10.1088/2515-7639/ab084b.

[175] Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science* **2018**, *361*, 360–365, DOI: 10.1126/science.aat2663.

[176] Schwalbe-Koda, D.; Gómez-Bombarelli, R. In *Machine Learning Meets Quantum Physics*; Schütt, K. T., Chmiela, S., von Lilienfeld, O. A., Tkatchenko, A., Tsuda, K., Müller, K.-R., Eds.; Springer International Publishing: Cham, 2020; pp 445–467, DOI: 10.1007/978-3-030-40245-$7_2$1.

[177] Kirkpatrick, S.; Gelatt, C. D.; Vecchi, M. P. Optimization by Simulated Annealing. **1983**, *220*, 671–680, DOI: 10.1126/science.220.4598.671.

[178] Kushner, H. J. A New Method of Locating the Maximum Point of an Arbitrary Multipeak Curve in the Presence of Noise. *Journal of Basic Engineering* **1964**, *86*, 97–106, DOI: 10.1115/1.3653121.

[179] Močkus, J. *Optimization Techniques IFIP Technical Conference*; Springer Berlin Heidelberg, 1975; pp 400–404, DOI: 10.1007/978-3-662-38527-$2_5$5.

[180] Holland, J. H., et al. *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*; MIT press, 1992.

[181] Forrest, S. Genetic algorithms: principles of natural selection applied to computation. *Science* **1993**, *261*, 872–878.

[182] Krenn, M.; Häse, F.; Nigam, A.; Friederich, P.; Aspuru-Guzik, A. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Machine Learning: Science and Technology* **2020**, *1*, 045024, DOI: 10.1088/2632-2153/aba947.

[183] Jensen, J. H. Graph-based Genetic Algorithm and Generative Model/Monte Carlo Tree Search for the Exploration of Chemical Space. **2019**, DOI: 10.26434/chemrxiv.7240751.v2.

[184] Henault, E. S.; Rasmussen, M. H.; Jensen, J. H. Chemical space exploration: how genetic algorithms find the needle in the haystack. *PeerJ Physical Chemistry* **2020**, DOI: I10.7717/peerj-pchem.11.

[185] Grimme, S.; Bannwarth, C.; Shushkov, P. A Robust and Accurate Tight-Binding Quantum Chemical Method for Structures, Vibrational Frequencies, and Noncovalent Interactions of Large Molecular Systems Parametrized for All spd-Block Elements (Z = 1-86). *J Chem Theory Comput* **2017**, *13*, 1989–2009.

[186] Hiener, D.; Folmsbee, D.; Langkamp, L.; Hutchison, G. Evaluating Fast Methods for Static Polarizabilities on Extended Conjugated Oligomers. *ChemRxiv* **2021**, DOI: 10.26434/chemrxiv-2021-pv54s-v2.

[187] Hiener, D.; Hutchison, G. Pareto Optimization of Oligomer Polarizability and Dipole Moment using a Genetic Algorithm. **2021**, DOI: 10.26434/chemrxiv.14043782.v1.

[188] Steinmann, C.; Jensen, J. H. Using a genetic algorithm to find molecules with good docking scores. *PeerJ Physical Chemistry* **2021**, *3*, e18, DOI: 10.7717/peerj-pchem.18.

[189] Esra Bozkurt, R. H. N. J. B., Marta A. S. Perez; Rothlisberger, U. Genetic Algorithm Based Design and Experimental Characterization of a Highly Thermostable Metalloprotein. *J. Am. Chem. Soc.* **2018**, *140*, 4517–4521, DOI: 10.1021/jacs.7b10660.

[190] Elizabeth Brunk, P. A. U. R., Marta A. S. Perez Genetic-Algorithm-Based Optimization of a Peptidic Scaffold for Sequestration and Hydration of CO2. *ChemPhysChem* **2016**, *17*, 3831–3835, DOI: 10.1002/cphc.201601034.

[191] Daniel J. Kozuch, F. H. S.; Debenedetti, P. G. Genetic Algorithm Approach for the Optimization of Protein Antifreeze Activity Using Molecular Simulations. *J. Chem. Theory Comput.* **2020**, *16*, 7866–7873, DOI: 10.1021/acs.jctc.0c00773.

[192] Yang, H.; Wong, M. W. Automatic Conformational Search of Transition States for Catalytic Reactions Using Genetic Algorithm. *J. Phys. Chem. A* **2019**, *123*, 10303–10314, DOI: 10.1021/acs.jpca.9b09543.

[193] Gui-Fang Shao, T.-D. L. L.-Y. X. Y.-H. W., Na-Na Tu Structural studies of Au–Pd bimetallic nanoparticles by a genetic algorithm method. *Physica E: Low-dimensional Systems and Nanostructures* **2015**, *70*, 11–20, DOI: 10.1016/j.physe.2015.02.008.

[194] Tun-Dong Liu, G.-F. S. N.-N. T. J.-P. T. Y.-H. W., Liang-You Xu Structural optimization of Pt–Pd–Rh trimetallic nanoparticles using improved genetic algorithm. *Journal of Alloys and Compounds* **2015**, *663*, 466–4730, DOI: 10.1016/j.jallcom.2015.12.146.

[195] Guifang Shao, J. T. J. Z. T. L.-Y. W., Yali Shangguan An improved genetic algorithm for structural optimization of Au–Ag bimetallic nanoparticles. *Applied Soft Computing* **2018**, *73*, 39–49, DOI: 10.1016/j.asoc.2018.08.019.

[196] Lazauskas, T.; Sokol, A. A.; Woodley, S. M. An efficient genetic algorithm for structure prediction at the nanoscale. *Nanoscale* **2017**, *9*, 3850–3864, DOI: 10.1039/C6NR09072A.

[197] Curtis, F.; Li, X.; Rose, T.; Vázquez-Mayagoitia, Á.; Bhattacharya, S.; Ghiringhelli, L. M.; Marom, N. GAtor: A First-Principles Genetic Algorithm for Molecular Crystal Structure Prediction. *Journal of Chemical Theory and Computation* **2018**, *14*, 2246–2264, DOI: 10.1021/acs.jctc.7b01152.

[198] Gustafson, J. A.; Wilmer, C. E. Intelligent Selection of Metal–Organic Framework Arrays for Methane Sensing via Genetic Algorithms. *ACS Sensors* **2019**, *4*, 1586–1593, DOI: 10.1021/acssensors.9b00268.

[199] Day, B. A.; Wilmer, C. E. Genetic Algorithm Design of MOF-based Gas Sensor Arrays for CO2-in-Air Sensing. *Sensors* **2020**, *20*, 924, DOI: 10.3390/s20030924.

[200] Jensen, J. H.; Ree, N.; Koerstz, M.; Mikkelsen, K. V. Virtual screening of norbornadiene-based molecular solar thermal energy storage systems using a genetic algorithm. **2021**, DOI: 10.33774/chemrxiv-2021-zd39r.

[201] Koerstz, M.; Christensen, A. S.; Mikkelsen, K. V.; Nielsen, M. B.; Jensen, J. H. High throughput virtual screening of 230 billion molecular solar heat battery candidates. *PeerJ Physical Chemistry* **2021**, *3*, e16, DOI: 10.7717/peerj-pchem.16.

[202] Kanal, I. Y.; Owens, S. G.; Bechtel, J. S.; Hutchison, G. R. Efficient Computational Screening of Organic Polymer Photovoltaics. *J. Phys. Chem. Lett.* **2013**, *4*, 1613–1623, DOI: 10.1021/jz400215j.

[203] Probst, D.; Reymond, J.-L. A probabilistic molecular fingerprint for big data settings. *Journal of Cheminformatics 2018 10:1* **2018**, *10*, 1–12, DOI: 10.1186/S13321-018-0321-8.

[204] Gutman, I.; Vidović, D.; Popović, L. Graph representation of organic molecules Cayley's plerograms vs. his kenograms. *Journal of the Chemical Society, Faraday Transactions* **1998**, *94*, 857–860, DOI: 10.1039/A708076J.

[205] Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. *Journal of Chemical Information and Modeling* **2019**, *59*, 3370–3388, DOI: 10.1021/ACS.JCIM.9B00237.

[206] Stokes, J. M.; Yang, K.; Swanson, K.; Jin, W.; Cubillos-Ruiz, A.; Donghia, N. M.; MacNair, C. R.; French, S.; Carfrae, L. A.; Bloom-Ackermann, Z.; Tran, V. M.; Chiappino-Pepe, A.; Badran, A. H.; Andrews, I. W.; Chory, E. J.; Church, G. M.; Brown, E. D.; Jaakkola, T. S.; Barzilay, R.; Collins, J. J. A Deep Learning Approach to Antibiotic Discovery. *Cell* **2020**, *180*, 688–702.e13, DOI: 10.1016/J.CELL.2020.01.021.

[207] Lin, T.-S.; Coley, C. W.; Mochigase, H.; Beech, H. K.; Wang, W.; Wang, Z.; Woods, E.; Craig, S. L.; Johnson, J. A.; Kalow, J. A.; Jensen, K. F.; Olsen, B. D. BigSMILES: A Structurally-Based Line Notation for Describing Macromolecules. *ACS Central Science* **2019**, *5*, 1523–1531, DOI: 10.1021/ACSCENTSCI.9B00476.

[208] Folmsbee, D. L.; Koes, D. R.; Hutchison, G. R. Evaluation of Thermochemical Machine Learning for Potential Energy Curves and Geometry Optimization. *The Journal of Physical Chemistry A* **2021**, *125*, 1987–1993, DOI: 10.1021/acs.jpca.0c10147, PMID: 33630611.

[209] Lu, Y.; Anand, S.; Shirley, W.; Gedeck, P.; Kelley, B. P.; Skolnik, S.; Rodde, S.; Nguyen, M.; Lindvall, M.; Jia, W. Prediction of pKa Using Machine Learning Methods with Rooted Topological Torsion Fingerprints: Application to Aliphatic Amines. *Journal of Chemical Information and Modeling* **2019**, *59*, 4706–4719, DOI: 10.1021/ACS.JCIM.9B00498.

[210] Cui, Q.; Lu, S.; Ni, B.; Zeng, X.; Tan, Y.; Chen, Y. D.; Zhao, H. Improved Prediction of Aqueous Solubility of Novel Compounds by Going Deeper With Deep Learning. *Frontiers in Oncology* **2020**, *0*, 121, DOI: 10.3389/FONC.2020.00121.

[211] Xiong, J.; Li, Z.; Wang, G.; Fu, Z.; Zhong, F.; Xu, T.; Liu, X.; Huang, Z.; Liu, X.; Chen, K.; Jiang, H.; Zheng, M. Multi-instance learning of graph neural networks for aqueous pKa prediction. *Bioinformatics* **2021**, 0–0, DOI: 10.1093/BIOINFORMATICS/BTAB714.

[212] Kawai, H.; Nakagawa, Y. O. Predicting excited states from ground state wavefunction by supervised quantum machine learning. *Machine Learning: Science and Technology* **2020**, *1*, 045027, DOI: 10.1088/2632-2153/ABA183.

[213] Bukov, M.; Schmitt, M.; Dupont, M. Learning the ground state of a non-stoquastic quantum Hamiltonian in a rugged neural network landscape. *SciPost Physics* **2021**, *10*, 147, DOI: 10.21468/SCIPOSTPHYS.10.6.147.

[214] Himmetoglu, B. Tree based machine learning framework for predicting ground state energies of molecules. *The Journal of Chemical Physics* **2016**, *145*, 134101, DOI: 10.1063/1.4964093.

[215] Westermayr, J.; Marquetand, P. Machine Learning for Electronically Excited States of Molecules. *Chemical Reviews* **2020**, *121*, 9873–9926, DOI: 10.1021/ACS.CHEMREV.0C00749.

[216] Kiyohara, S.; Tsubaki, M.; Mizoguchi, T. Learning excited states from ground states by using an artificial neural network. *npj Computational Materials 2020 6:1* **2020**, *6*, 1–6, DOI: 10.1038/s41524-020-0336-3.

[217] Babaei, M.; Azar, Y. T.; Sadeghi, A. Locality meets machine learning: Excited and ground-state energy surfaces of large systems at the cost of small ones. *Physical Review B* **2020**, *101*, 115132, DOI: 10.1103/PhysRevB.101.115132.

[218] Francoeur, P. G.; Koes, D. R. SolTranNet–A Machine Learning Tool for Fast Aqueous Solubility Prediction. *Journal of Chemical Information and Modeling* **2021**, *61*, 2530–2536, DOI: 10.1021/ACS.JCIM.1C00331.

[219] Lovrić, M.; Pavlović, K.; Žuvela, P.; Spataru, A.; Lučić, B.; Kern, R.; Wong, M. W. Machine learning in prediction of intrinsic aqueous solubility of drug-like compounds: Generalization, complexity, or predictive ability? *Journal of Chemometrics* **2021**, *35*, e3349, DOI: 10.1002/CEM.3349.

[220] Abarbanel, O. D.; Hutchison, G. R. Machine learning to accelerate screening for Marcus reorganization energies. *Journal of Chemical Physics* **2021**, *155*, 54106, DOI: 10.1063/5.0059682.

[221] Guo, T.; Wu, L.; Li, T. Machine Learning Accelerated, High Throughput, Multi-Objective Optimization of Multiprincipal Element Alloys. *Small* **2021**, *17*, 2102972, DOI: 10.1002/smll.202102972, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/smll.202102972.

[222] Torkamanian-Afshar, M.; Nematzadeh, S.; Tabarzad, M.; Najafi, A.; Lanjanian, H.; Masoudi-Nejad, A. In silico design of novel aptamers utilizing a hybrid method of machine learning and genetic algorithm. *Molecular Diversity* **2021**, *25*, 1395–1407, DOI: 10.1007/s11030-021-10192-9.

[223] Yoshida, M.; Hinkley, T.; Tsuda, S.; Abul-Haija, Y. M.; McBurney, R. T.; Kulikov, V.; Mathieson, J. S.; Galiñanes Reyes, S.; Castro, M. D.; Cronin, L. Using Evolutionary Algorithms and Machine Learning to Explore Sequence Space for the Discovery of Antimicrobial Peptides. *Chem* **2018**, *4*, 533–543, DOI: 10.1016/j.chempr.2018.01.005.

[224] Kwon, Y.; Kang, S.; Choi, Y.-S.; Kim, I. Evolutionary design of molecules based on deep learning and a genetic algorithm. *Scientific Reports* **2021**, *11*, 17304, DOI: 10.1038/s41598-021-96812-8, Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Cheminformatics;Chemistry;Drug discovery;Materials chemistry;Materials science;Mathematics and computing;Optics and photonics;Organic chemistry Subject_term_id: cheminformatics;chemistry;drug-discovery;materials-chemistry;materials-science;mathematics-and-computing;optics-and-photonics;organic-chemistry.

[225] Janet, J. P.; Chan, L.; Kulik, H. J. Accelerating Chemical Discovery with Machine Learning: Simulated Evolution of Spin Crossover Complexes with an Artificial Neural Network. *The Journal of Physical Chemistry Letters* **2018**, *9*, 1064–1071, DOI: 10.1021/acs.jpclett.8b00170, Publisher: American Chemical Society.

[226] Browning, N. J.; Ramakrishnan, R.; von Lilienfeld, O. A.; Roethlisberger, U. Genetic Optimization of Training Sets for Improved Machine Learning Models of Molecular Properties. *The Journal of Physical Chemistry Letters* **2017**, *8*, 1351–1359, DOI: 10.1021/acs.jpclett.7b00038, Publisher: American Chemical Society.