Assessing Conformer Energies using Electronic Structure and Machine Learning Methods

Dakota Folmsbee[1] and Geoffrey Hutchison[1]

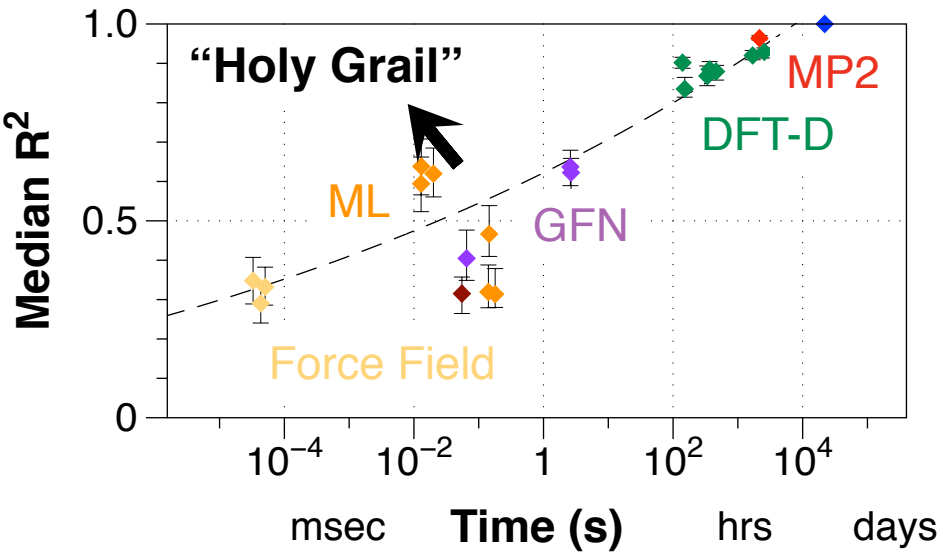
[1]Department of Chemistry, University of Pittsburgh

June 12, 2020

Abstract

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.

Keywords: conformers, thermochemistry, machine learning, density functional, semiempirical, DFTB, coupled-cluster



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.[1],[2],[3] In principle, for each rotatable bond, the number of possible minima increases exponentially. Consequently, most conformer sampling methods[4] use classical small-molecule force fields to evaluate energies because of their fast performance, despite potentially poor correlation with quantum mechanical methods.[5]

Multiple efforts have evaluated the success of wavefunction and density functional first-principles methods to compare the energetics of different conformers.[6],[7],[8],[9],[10],[11],[12] 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.[13],[14]

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.[15] Other works have documented the need for accurate treatment of non-covalent interactions to model conformers in π-conjugated oligomers.[16]

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,[17] G4,[18] and W1[19],[20] to W4[21] 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.[22] As mentioned above, efforts for conformer sampling have often focused on classical force fields or multi-level approaches using semiempirical methods.[4],[23],[24]

In our previous paper,[5] 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.

Computational Methods

Calculations were performed using Open Babel version 3.0[25] for all force field calculations (MMFF94[26],[27],[28],[29],[30] and UFF[31],[32]) `OpenMOPAC` for PM7,[33] `xtb` version 6.2[34] for GFN0[35] GFN1[36] and GFN2 calculations,[37] and `Orca` 4.0.1[38] for all density functional and *ab initio* calculations, unless otherwise indicated. For density functional methods, the D3(BJ)[39],[40],[41],[42] dispersion correction scheme was used as indicated, except for ωB97X-D3[43] which uses a similar approach. For *ab initio* methods,

`Orca` 4.0.1 was used for MP2[44] and DLPNO-CCSD(T)[45],[46] with "TightPNO" using the cc-pVTZ basis set.[47],[48] Energies are read from all output files using the `cclib`[49] version 1.6.2, and `pybel` version 3.0.[50]

Machine learning methods included "bag-of-features" representations and ANI-1x[51], ANI-1ccx[52], and ANI-2x[53] models. The Bag-of-Features representations chosen were Bag of Bonds[54](BOB), Bond Angle Torsion [55](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`.[56] 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`[57] 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`[58] data frames, using `numpy`[59] and `scipy`[60] functions for analysis. 3DMol.js was used for interactive molecular visualization of conformers.[61] Plotly was used for interactive plots.[62]

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.

Test Set Selection

As in our previous work,[5] a dataset consisting of experimental crystal structures of 700 small molecules capable of multiple conformer geometries was provided to us by Ebejer[63] and were derived from the work of Hawkins et al.[23] along with ligands from the Astex Diverse Set.[64] These compounds have been repeatedly used to evaluate the quality of conformer generation.[23],[63] 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).[23] A list of Simplified Molecular Input Line Entry Specification (SMILES)[65] 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.

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.[5]

Using this set of DFT-optimized minima, in this work, single point atomization energies were computed using the DLPNO-CCSD(T)[45],[46] method using the cc-pVTZ basis set.(Dunning 1989, Kendall 1992) 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.[66],[67],[45] 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.

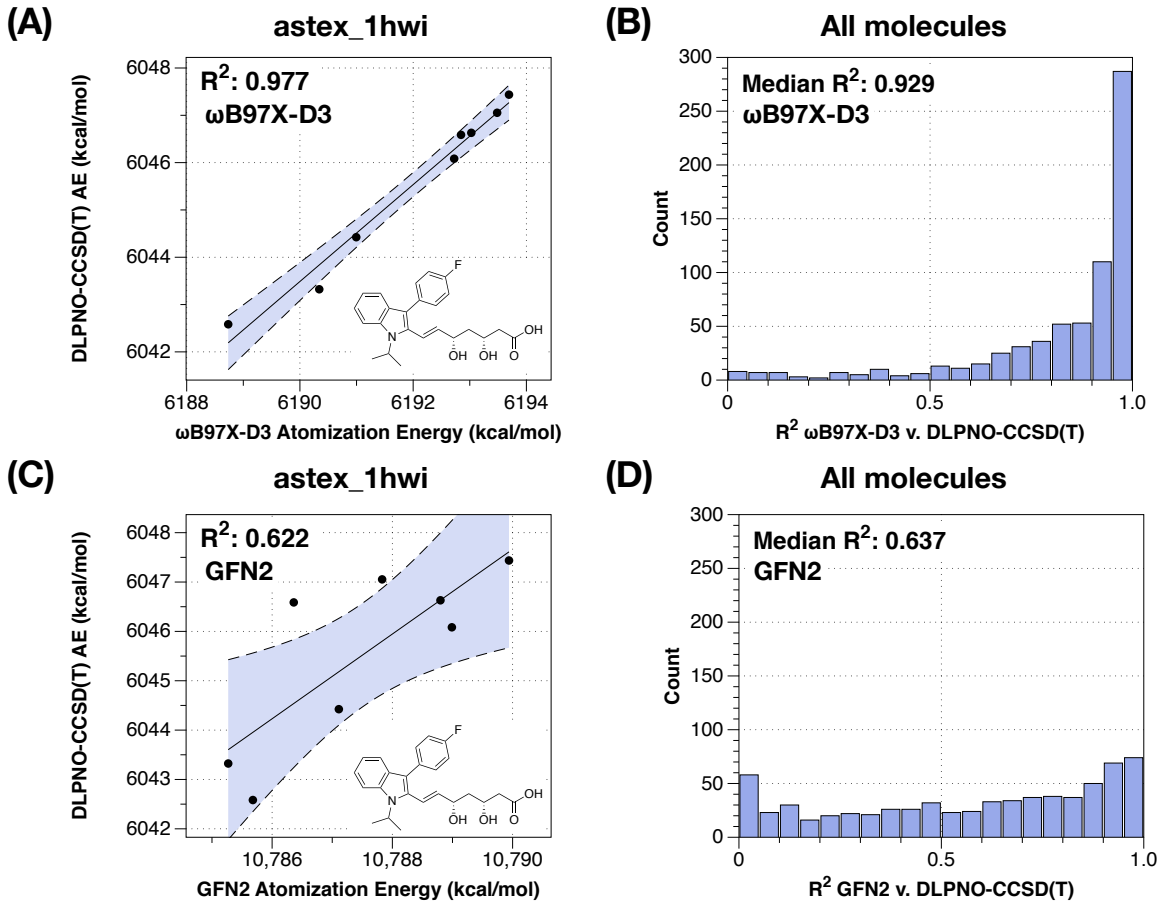


Figure 1: Example analysis of ωB97X-D3 and GFN2 methods, starting with (A) correlation between ωB97X-D3 and DLPNO-CCSD(T) energies for a single molecule, (B) histogram of R^2 correlations across all molecules, (C) correlation between GFN2 and DLPNO-CCSD(T) energies, and (D) corresponding histogram of R^2 correlations across all molecules.

As illustrated in Figure 1, each method is correlated with DLPNO-CCSD(T) / cc-pVTZ energies for each molecule (e.g., `astex_1hwi` in Figure 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, the median value was compiled as illustrated in Figure 1, to represent the overall quality of a given method.

Since the metrics are unlikely to reflect normal distributions (e.g., Figure 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 ± 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,[7] 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,[2],[1] reactions, and even understanding the effects of dipole moments on solvent viscosity.[68] Consequently, deriving accurate relative energies of molecular conformers is a crucial task, as discussed below.

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,[69],[27],[28],[29],[30] UFF,[31] GAFF[70]
- **Semiempirical wave function:** PM7[33]
- **Density functional tight binding**: GFN0,[35] GFN1,[71] GFN2[37]
- **Low-cost density functional approximation**s: PBEh-3c,[72] B97-3c[73]
- **Dispersion-corrected density functionals**: B3LYP,[74],[75],[76],[77] PBE[78],[79], ωB97X-D[43] with dispersion correction (using def2-TZVP basis set[80],[81])
- **Møller-Plesset RI-MP2**[44] with the cc-pVTZ basis set[47],[48]

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.

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±0.064 to 0.920±0.025 for B3LYP-D3BJ and from 0.835±0.025 to 0.885±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

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

Table 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.

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.

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.[82],[83] Density functional methods are known to incorrectly account for dispersion interactions, which has led to a variety of empirical corrections.[39],[40],[41],[42],[84],[85],[86],[87] 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±0.012 to 0.706±0.050 without the D3BJ dispersion correction.

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]

Table 2: Effect of dispersion correction for DFT methods. Values in brackets indicate 95% confidence intervals from bootstrap sampling.

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[88] or the new D4 method[85],[84] will hopefully improve their accuracy further.

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, hybrid density functional methods such as B3LYP-D3BJ require significant single-computational time for single-point energies of medium-sized organic molecules (median 26±0.3 minutes) compared to GGA methods such as PBE or approximate density functional tight binding methods such

as GFN1 / GFN2 (median 2.6±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).

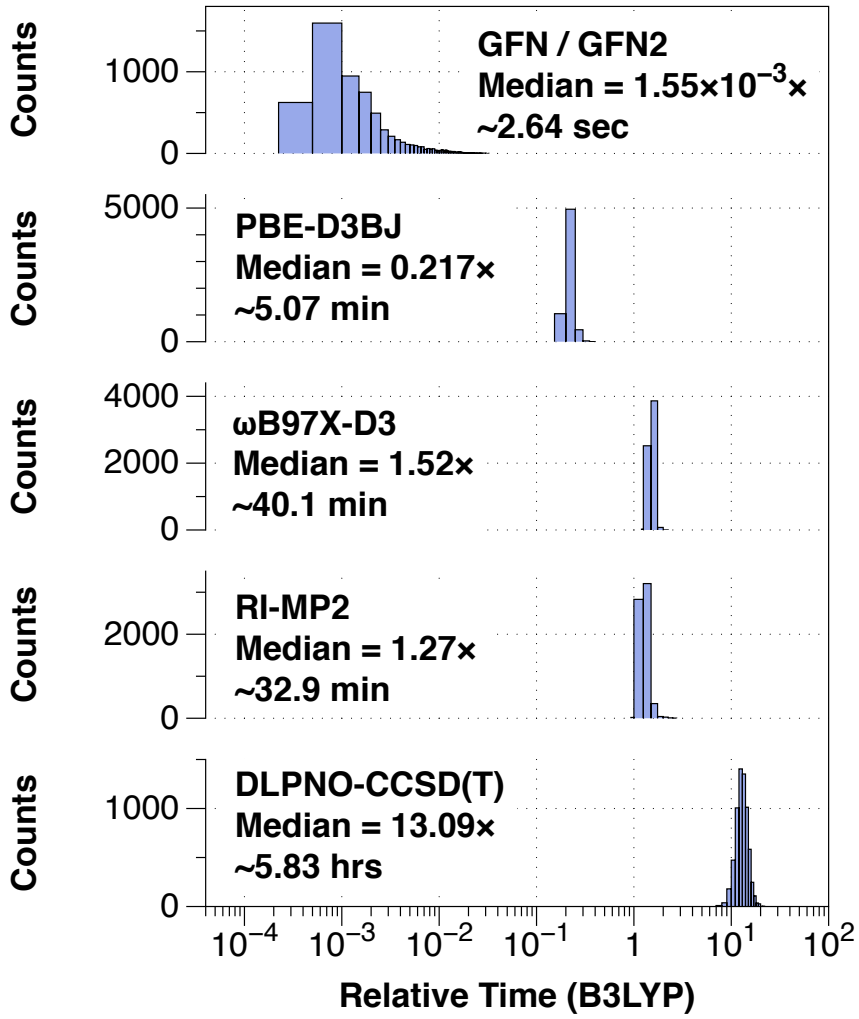


Figure 2: Histograms of relative timings for key methods considered, normalized to B3LYP-D3BJ single points on the same molecule, using ORCA 4.0.1. Median relative times and median wall clock times for single-core runs are included for reference.

Consequently, an important consideration is also the typical trade-off in computational chemistry between thermochemical accuracy and computational time. Since traditional 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,[89] PM3,[90] PM6,[91] and PM7[33] 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.

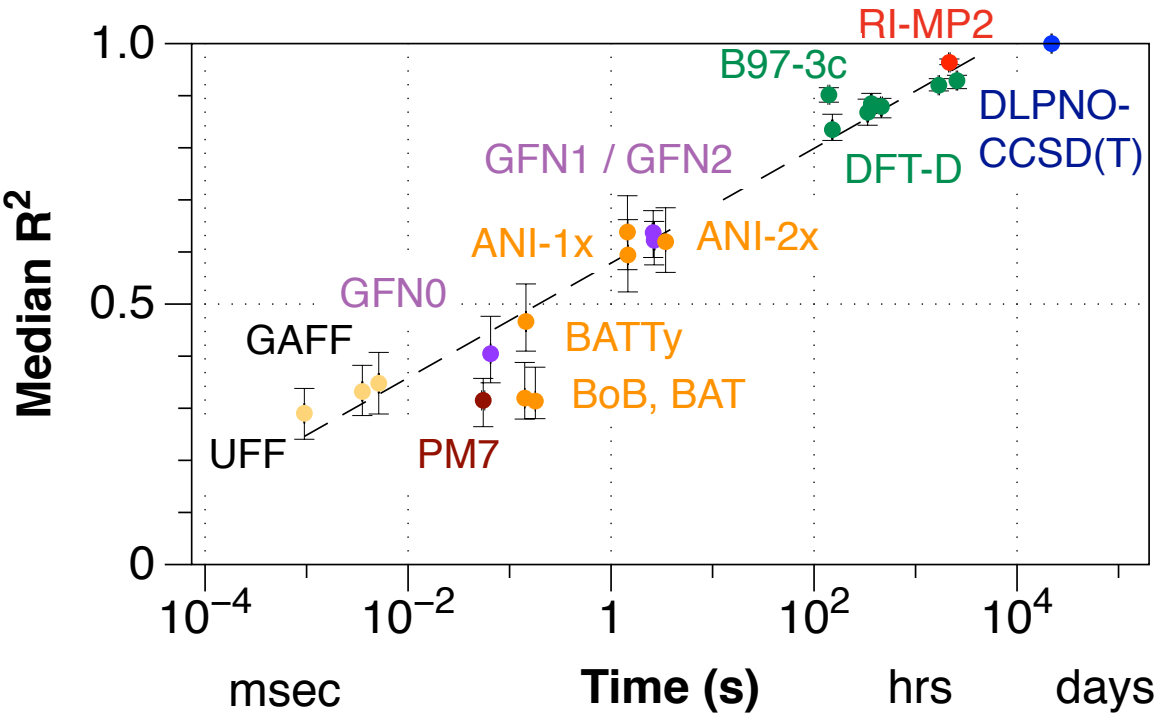


Figure 3: Comparison of single-core computational time required for energy evaluation (in log scale) to median R^2 found when compared to DLPNO-CCSD(T) energies. Error bars indicate 95% confidence intervals of time and medain R^2 from bootstrap sampling. Dashed line indicates approximate "best current 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 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.

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.[92],[93],[94] Typically, these ML methods use deep neural networks (DNN) and have been trained to density functional calculations, particularly hybrid B3LYP or ωB97X atomization energies[95][96] although recent efforts have included training on coupled-cluster quality data as well.[52]

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 graphics 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 1 and Figure 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 1 and Figure 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.

In principal, it is possible to perform *post hoc* addition of a D3 dispersion correction to both ANI-1x and ANI-2x. Table 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[85],[84] can be calculated using the DFTD4 program,[97] 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.

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]

Table 3: Comparison of post hoc dispersion correction for ANI machine learning methods. Values in brackets indicate 95% confidence intervals from bootstrap sampling.

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.[98]

ML commonly employs techniques to normalize the data, improving the model's training.[99],[100] 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.

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]

Table 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.

Discussion

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.[7] Certainly, one might imagine that when considering multiple geometries with only small differences in energies, random errors are magnified. Figure 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 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.

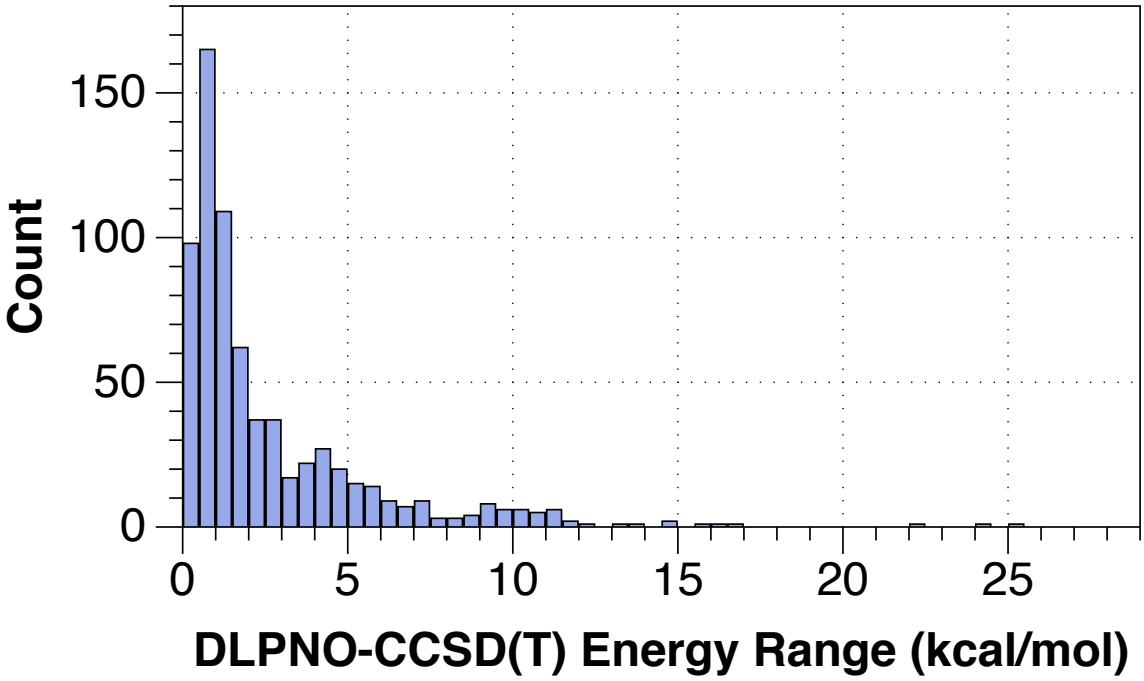


Figure 4: Histogram of relative DLPNO-CCSD(T) energy ranges across multiple conformers.

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 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 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).

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 small energy windows to other geometries. One recent example comes from understanding the effects of dipole moments on solvent viscosity.[68] Finding

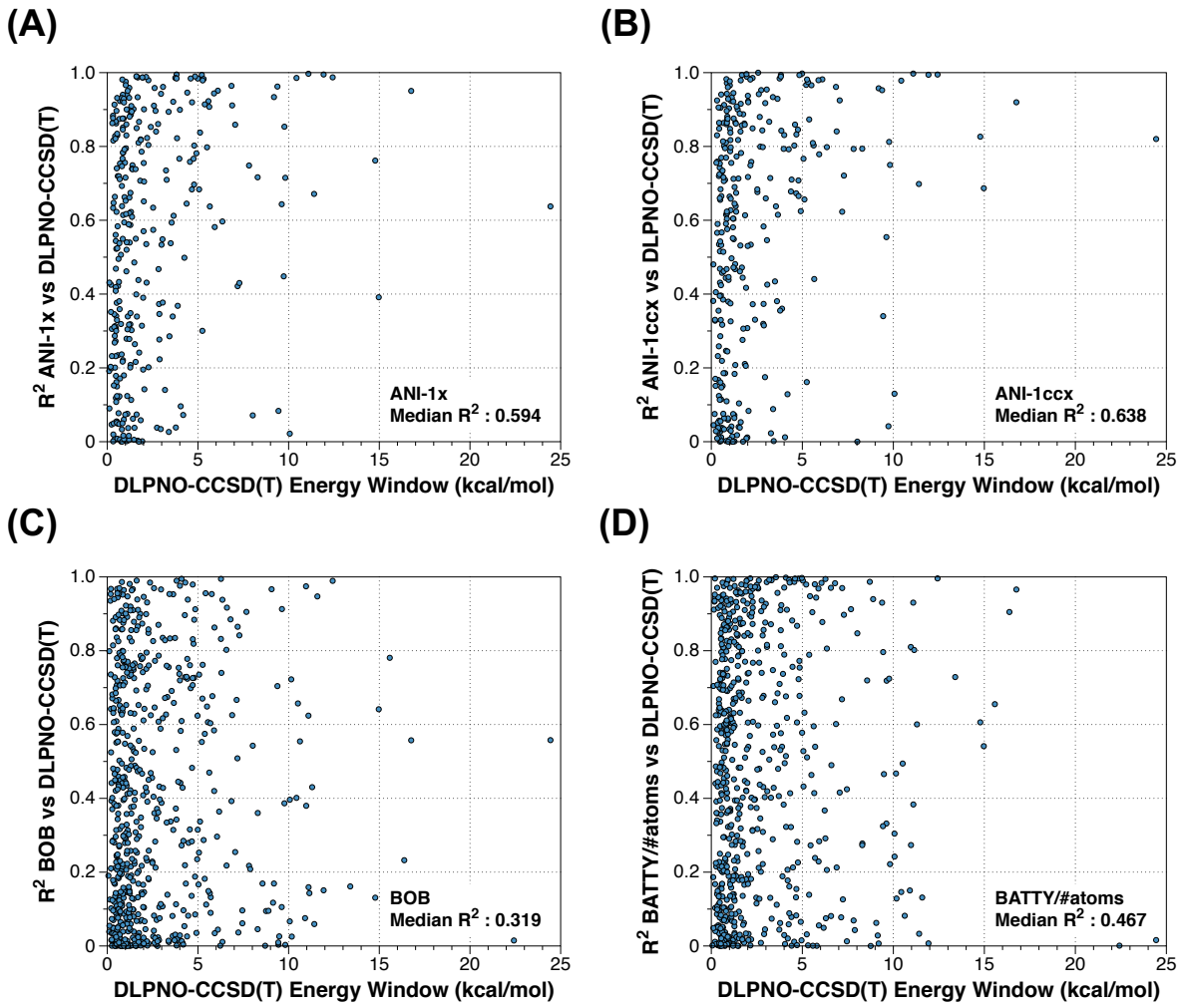


Figure 5: Examples of the relation of energy windows to R^2 for the ML methods (A) ANI-1x, (B) ANI-1ccx, (C) BOB, and (D) BATTY/# atoms.

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 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 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

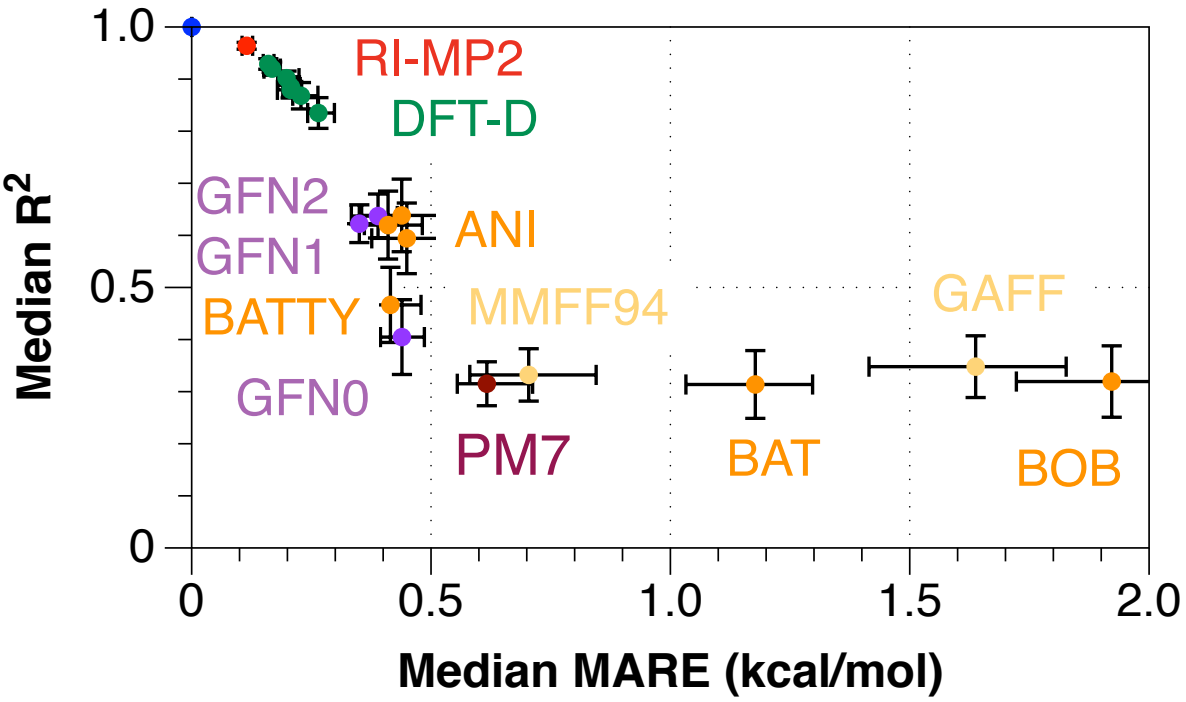


Figure 6: Correlation between mean absolute relative energies (MARE) and median R^2 correlation. Since the R^2 metric minimizes systematic errors, the high degree of correlation between the two metrics indicate most methods exhibit relatively random / non-systematic errors. Error bars indicate 95% confidence intervals from bootstrap sampling.

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.[68]

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 5 indicates the median sequential times from Table 1 and median time 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.

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

Table 5: Comparison of single-core median sequential time to median batch time (in seconds), and relative speedups for batch evaluation.

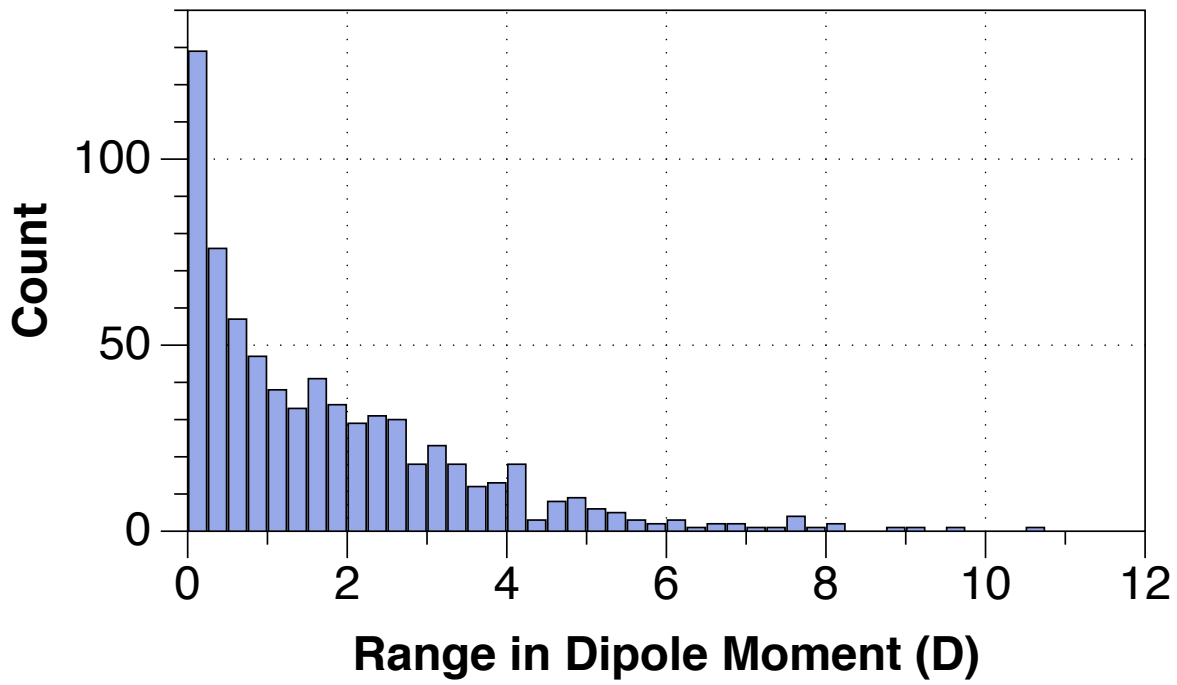


Figure 7: Histogram of the range of B3LYP-computed dipole moments in Debye across the conformers considered in this work. While most molecules show only small differences in polarity across conformers, many have over 3-4 Debye ranges.

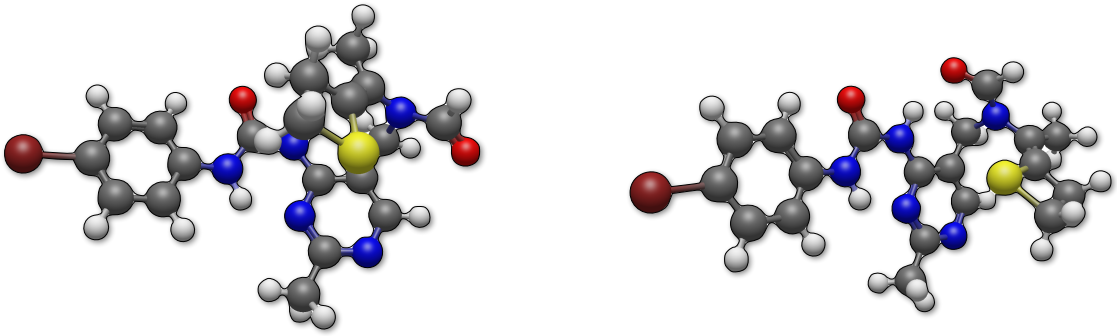


Figure 8: Example of conformational diversity in dipole moment in the molecule omegacsd_CNBPCT reflecting anti-parallel carbonyl (*left* - rmsd45) or parallel carbonyl groups (*right* - rmsd92), with B3LYP-D3BJ def2-TZVP computed dipole moments ranging from 1.41D to 9.78D, respectively. The two geometries 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.

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,[52] 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,[101],[102],[103],[104] inclusion of polarizable atomic charges,[105],[106],[107],[108],[109],[110],[111],[112] novel force fields from experimental data, density functional and other quantum methods,[113],[114],[115],[116],[117],[118] and continued development of approximate semiempirical quantum methods.[37]

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.[15] Moreover, much chemistry is performed in solution, thus work on understanding conformer energies in solvation is also critical.[119],[120]

Acknowledgments

GRH and DLF acknowledge the National Science Foundation (CHE-1800435) for support and the University of Pittsburgh Center for Research Computing through the computational resources provided. The authors thank Olexandr Isayev and Justin Smith for access to the ANI-2x model.

Supporting Information

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

References

[1]S. Grimme, in *Reviews in Computational Chemistry*, John Wiley & Sons Inc., **2004**, pp. 153–218.

[2]M. W. Lodewyk, M. R. Siebert, D. J. Tantillo, *Chemical Reviews* **2011**, *112*, 1839–1862.

[3]N. E. Jackson, B. M. Savoie, K. L. Kohlstedt, T. J. Marks, L. X. Chen, M. A. Ratner, *Macromolecules* **2014**, *47*, 987–992.

[4]P. C. D. Hawkins, *Journal of Chemical Information and Modeling* **2017**, *57*, 1747–1756.

[5]I. Y. Kanal, J. A. Keith, G. R. Hutchison, *International Journal of Quantum Chemistry* **2017**, *118*, e25512.

[6]M. Habgood, T. James, A. Heifetz, *Conformational Searching with Quantum Mechanics*, Springer US, **2020**.

[7]D. I. Sharapa, A. Genaev, L. Cavallo, Y. Minenkov, *ChemPhysChem* **2018**, DOI 10.1002/cphc.201801063.

[8]M. K. Kesharwani, A. Karton, J. M. L. Martin, *Journal of Chemical Theory and Computation* **2015**, *12*, 444–454.

[9]J. Řezáč, D. Bím, O. Gutten, L. Rulíšek, *Journal of Chemical Theory and Computation* **2018**, *14*, 1254–1266.

[10]V. K. Prasad, A. Otero-de-la-Roza, G. A. DiLabio, *Scientific Data* **2019**, *6*, DOI 10.1038/sdata.2018.310.

[11]Y. K. Kang, H. S. Park, *Chemical Physics Letters* **2018**, *702*, 69–75.

[12]Y. Yuan, M. J. L. Mills, P. L. A. Popelier, F. Jensen, *The Journal of Physical Chemistry A* **2014**, *118*, 7876–7891.

[13]B. K. Rai, V. Sresht, Q. Yang, R. Unwalla, M. Tu, A. M. Mathiowetz, G. A. Bakken, *Journal of Chemical Information and Modeling* **2019**, *59*, 4195–4208.

[14]N. Foloppe, I.-J. Chen, *Future Medicinal Chemistry* **2019**, *11*, 97–118.

[15]M. P. Johansson, J. Olsen, *Journal of Chemical Theory and Computation* **2008**, *4*, 1460–1471.

[16]N. E. Jackson, B. M. Savoie, K. L. Kohlstedt, M. O. de la Cruz, G. C. Schatz, L. X. Chen, M. A. Ratner, *Journal of the American Chemical Society* **2013**, *135*, 10475–10483.

[17]L. A. Curtiss, K. Raghavachari, P. C. Redfern, V. Rassolov, J. A. Pople, *The Journal of Chemical Physics* **1998**, *109*, 7764–7776.

[18]L. A. Curtiss, P. C. Redfern, K. Raghavachari, *The Journal of Chemical Physics* **2007**, *126*, 084108.

[19]J. M. L. Martin, G. de Oliveira, *The Journal of Chemical Physics* **1999**, *111*, 1843–1856.

[20]S. Parthiban, J. M. L. Martin, *The Journal of Chemical Physics* **2001**, *114*, 6014–6029.

[21]A. Karton, E. Rabinovich, J. M. L. Martin, B. Ruscic, *The Journal of Chemical Physics* **2006**, *125*, 144108.

[22]M. M. Ghahremanpour, P. J. van Maaren, J. C. Ditz, R. Lindh, D. van der Spoel, *The Journal of Chemical Physics* **2016**, *145*, 114305.

[23]P. C. Hawkins, A. G. Skillman, G. L. Warren, B. A. Ellingson, M. T. Stahl, *J Chem Inf Model* **2010**, *50*, 572–84.

[24]J. Juárez-Jiménez, X. Barril, M. Orozco, R. Pouplana, F. J. Luque, *The Journal of Physical Chemistry B* **2014**, *119*, 1164–1172.

[25]N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, G. R. Hutchison, *J Cheminform* **2011**, *3*, 33.

[26]T. A. Halgren, *Journal of Computational Chemistry* **1996**, *17*, 490–519.

[27]T. A. Halgren, *Journal of Computational Chemistry* **1996**, *17*, 520–552.

[28]T. A. Halgren, *Journal of Computational Chemistry* **1996**, *17*, 553–586.

[29]T. A. Halgren, R. B. Nachbar, *Journal of Computational Chemistry* **1996**, *17*, 587–615.

[30]T. A. Halgren, *Journal of Computational Chemistry* **1996**, *17*, 616–641.

[31]A. K. Rappe, C. J. Casewit, K. S. Colwell, W. A. Goddard, W. M. Skiff, *Journal of the American Chemical Society* **1992**, *114*, 10024–10035.

[32]C. J. Casewit, K. S. Colwell, A. K. Rappe, *Journal of the American Chemical Society* **1992**, *114*, 10035–10046.

[33]J. J. P. Stewart, *Journal of Molecular Modeling* **2012**, *19*, 1–32.

[34].

[35]P. Pracht, E. Caldeweyher, S. Ehlert, S. Grimme, **2019**, DOI 10.26434/chemrxiv.8326202.v1.

[36]S. Grimme, C. Bannwarth, P. Shushkov, *J Chem Theory Comput* **2017**, *13*, 1989–2009.

[37]C. Bannwarth, S. Ehlert, S. Grimme, **2018**, DOI 10.26434/chemrxiv.7246238.v2.

[38]F. Neese, *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2011**, *2*, 73–78.

[39]S. Grimme, S. Ehrlich, L. Goerigk, *Journal of Computational Chemistry* **2011**, *32*, 1456–1465.

[40]A. D. Becke, E. R. Johnson, *The Journal of Chemical Physics* **2005**, *123*, 154101.

[41]E. R. Johnson, A. D. Becke, *The Journal of Chemical Physics* **2005**, *123*, 024101.

[42]E. R. Johnson, A. D. Becke, *The Journal of Chemical Physics* **2006**, *124*, 174104.

[43]J.-D. Chai, M. Head-Gordon, *Physical Chemistry Chemical Physics* **2008**, *10*, 6615.

[44]S. Kossmann, F. Neese, *Journal of Chemical Theory and Computation* **2010**, *6*, 2325–2338.

[45]D. G. Liakos, F. Neese, *Journal of Chemical Theory and Computation* **2015**, *11*, 4054–4063.

[46]Y. Guo, C. Riplinger, U. Becker, D. G. Liakos, Y. Minenkov, L. Cavallo, F. Neese, *The Journal of Chemical Physics* **2018**, *148*, 011101.

[47]T. H. Dunning, *The Journal of Chemical Physics* **1989**, *90*, 1007–1023.

[48]R. A. Kendall, T. H. Dunning, R. J. Harrison, *The Journal of Chemical Physics* **1992**, *96*, 6796–6806.

[49]N. M. O'boyle, A. L. Tenderholt, K. M. Langner, *Journal of Computational Chemistry* **2008**, *29*, 839–845.

[50]N. M. O'Boyle, C. Morley, G. R. Hutchison, *Chemistry Central Journal* **2008**, *2*, DOI 10.1186/1752-153x-2-5.

[51]J. S. Smith, B. Nebgen, N. Lubbers, O. Isayev, A. E. Roitberg, *The Journal of Chemical Physics* **2018**, *148*, 241733.

[52]J. S. Smith, B. T. Nebgen, R. Zubatyuk, N. Lubbers, C. Devereux, K. Barros, S. Tretiak, O. Isayev, A. Roitberg, **2019**, DOI 10.26434/chemrxiv.6744440.v2.

[53]C. Devereux, J. Smith, K. Davis, K. Barros, R. Zubatyuk, O. Isayev, A. Roitberg, *ChemRxiv* **2020**, DOI 10.26434/chemrxiv.11819268.v1.

[54]K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. von Lilienfeld, K.-R. Müller, A. Tkatchenko, *The Journal of Physical Chemistry Letters* **2015**, *6*, 2326–2331.

[55]B. Huang, O. A. von Lilienfeld, *The Journal of Chemical Physics* **2016**, *145*, 161102.

[56]D. Folmsbee, S. Upadhyay, A. Dumi, D. Hiener, D. Mulvey, **2019**, DOI 10.5281/zenodo.3333856.

[57]F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

[58]W. McKinney, in *Proceedings of the 9th Python in Science Conference* (Eds.: S. van der Walt, J. Millman), **2010**, pp. 51–56.

[59]S. van der Walt, S. C. Colbert, G. Varoquaux, *Computing in Science & Engineering* **2011**, *13*, 22–30.

[60]P. Virtanen, and Ralf Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, et al., *Nature Methods* **2020**, *17*, 261–272.

[61]N. Rego, D. Koes, *Bioinformatics* **2014**, *31*, 1322–1324.

[62]P. T. Inc., **2015**.

[63]J. P. Ebejer, G. M. Morris, C. M. Deane, *J Chem Inf Model* **2012**, *52*, 1146–58.

[64]M. J. Hartshorn, M. L. Verdonk, G. Chessari, S. C. Brewerton, W. T. Mooij, P. N. Mortenson, C. W. Murray, *J Med Chem* **2007**, *50*, 726–41.

[65]D. Weininger, *Journal of Chemical Information and Modeling* **1988**, *28*, 31–36.

[66]E. Paulechka, A. Kazakov, *The Journal of Physical Chemistry A* **2017**, *121*, 4379–4387.

[67]D. G. Liakos, Y. Guo, F. Neese, *The Journal of Physical Chemistry A* **2019**, DOI 10.1021/acs.jpca.9b05734.

[68]M. N. Vo, M. Call, C. Kowall, J. K. Johnson, *Industrial & Engineering Chemistry Research* **2019**, DOI 10.1021/acs.iecr.9b03699.

[69]T. A. Halgren, *Journal of Computational Chemistry* **1996**, *17*, 490–519.

[70]J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, D. A. Case, *Journal of Computational Chemistry* **2004**, *25*, 1157–1174.

[71]S. Grimme, C. Bannwarth, P. Shushkov, *Journal of Chemical Theory and Computation* **2017**, *13*, 1989–2009.

[72]S. Grimme, J. G. Brandenburg, C. Bannwarth, A. Hansen, *The Journal of Chemical Physics* **2015**, *143*, 054107.

[73]J. G. Brandenburg, C. Bannwarth, A. Hansen, S. Grimme, *The Journal of Chemical Physics* **2018**, *148*, 064104.

[74]C. Lee, W. Yang, R. G. Parr, *Physical Review B* **1988**, *37*, 785–789.

[75]A. D. Becke, *Physical Review A* **1988**, *38*, 3098–3100.

[76]P. J. Stephens, F. J. Devlin, C. F. Chabalowski, M. J. Frisch, *The Journal of Physical Chemistry* **1994**, *98*, 11623–11627.

[77]S. H. Vosko, L. Wilk, M. Nusair, *Canadian Journal of Physics* **1980**, *58*, 1200–1211.

[78]J. P. Perdew, K. Burke, M. Ernzerhof, *Physical Review Letters* **1997**, *78*, 1396–1396.

[79]J. P. Perdew, K. Burke, M. Ernzerhof, *Physical Review Letters* **1996**, *77*, 3865–3868.

[80]F. Weigend, R. Ahlrichs, *Physical Chemistry Chemical Physics* **2005**, *7*, 3297.

[81]F. Weigend, *Physical Chemistry Chemical Physics* **2006**, *8*, 1057.

[82]J. M. L. Martin, *The Journal of Physical Chemistry A* **2013**, *117*, 3118–3132.

[83]D. Gruzman, A. Karton, J. M. L. Martin, *The Journal of Physical Chemistry A* **2009**, *113*, 11974–11983.

[84]E. Caldeweyher, C. Bannwarth, S. Grimme, *The Journal of Chemical Physics* **2017**, *147*, 034112.

[85]E. Caldeweyher, S. Ehlert, A. Hansen, H. Neugebauer, S. Spicher, C. Bannwarth, S. Grimme, *The Journal of Chemical Physics* **2019**, *150*, 154122.

[86]S. Grimme, *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2011**, *1*, 211–228.

[87]E. R. Johnson, I. D. Mackie, G. A. DiLabio, *Journal of Physical Organic Chemistry* **2009**, *22*, 1127–1135.

[88]J. Witte, N. Mardirossian, J. B. Neaton, M. Head-Gordon, *Journal of Chemical Theory and Computation* **2017**, *13*, 2043–2052.

[89]M. J. S. Dewar, E. G. Zoebisch, E. F. Healy, J. J. P. Stewart, *Journal of the American Chemical Society* **1985**, *107*, 3902–3909.

[90]J. J. P. Stewart, *Journal of Computational Chemistry* **1989**, *10*, 209–220.

[91]J. J. P. Stewart, *Journal of Molecular Modeling* **2007**, *13*, 1173–1213.

[92]M. Rupp, A. Tkatchenko, K.-R. Müller, O. A. von Lilienfeld, *Physical Review Letters* **2012**, *108*, DOI 10.1103/physrevlett.108.058301.

[93]K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. von Lilienfeld, A. Tkatchenko, K.-R. Müller, *Journal of Chemical Theory and Computation* **2013**, *9*, 3404–3419.

[94]F. A. Faber, L. Hutchison, B. Huang, J. Gilmer, S. S. Schoenholz, G. E. Dahl, O. Vinyals, S. Kearnes, P. F. Riley, O. A. von Lilienfeld, *Journal of Chemical Theory and Computation* **2017**, *13*, 5255–5264.

[95]R. Ramakrishnan, P. O. Dral, M. Rupp, O. A. von Lilienfeld, *Scientific Data* **2014**, *1*, DOI 10.1038/sdata.2014.22.

[96]J. S. Smith, O. Isayev, A. E. Roitberg, *Chemical Science* **2017**, *8*, 3192–3203.

[97].

[98]S. Laghuvarapu, Y. Pathak, U. D. Priyakumar, *Journal of Computational Chemistry* **2019**, *41*, 790–799.

[99]S. Ioffe, C. Szegedy, *arXiv* **2015**, *abs/1502.03167*.

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

[101]J. Wahl, J. Freyss, M. von Korff, T. Sander, *Journal of Cheminformatics* **2019**, *11*, DOI 10.1186/s13321-019-0371-6.

[102]D. van der Spoel, M. M. Ghahremanpour, J. A. Lemkul, *The Journal of Physical Chemistry A* **2018**, *122*, 8982–8988.

[103]K. Roos, C. Wu, W. Damm, M. Reboul, J. M. Stevenson, C. Lu, M. K. Dahlgren, S. Mondal, W. Chen, L. Wang, et al., *Journal of Chemical Theory and Computation* **2019**, *15*, 1863–1874.

[104]E. Harder, W. Damm, J. Maple, C. Wu, M. Reboul, J. Y. Xiang, L. Wang, D. Lupyan, M. K. Dahlgren, J. L. Knight, et al., *Journal of Chemical Theory and Computation* **2015**, *12*, 281–296.

[105]F.-Y. Lin, A. D. MacKerell, in *Biomolecular Simulations: Methods and Protocols* (Eds.: M. Bonomi, C. Camilloni), Springer New York, New York, NY, **2019**, pp. 21–54.

[106]V. S. S. Inakollu, D. P. Geerke, C. N. Rowley, H. Yu, *Current Opinion in Structural Biology* **2020**, *61*, 182–190.

[107]A. Warshel, M. Kato, A. V. Pisliakov, *Journal of Chemical Theory and Computation* **2007**, *3*, 2034–2045.

[108]Z. Jing, C. Liu, S. Y. Cheng, R. Qi, B. D. Walker, J.-P. Piquemal, P. Ren, *Annual Review of Biophysics* **2019**, *48*, 371–394.

[109]C. Zhang, C. Lu, Z. Jing, C. Wu, J.-P. Piquemal, J. W. Ponder, P. Ren, *Journal of Chemical Theory and Computation* **2018**, *14*, 2084–2108.

[110]J. A. Rackers, Q. Wang, C. Liu, J.-P. Piquemal, P. Ren, J. W. Ponder, *Physical Chemistry Chemical Physics* **2017**, *19*, 276–291.

[111]J. W. Ponder, C. Wu, P. Ren, V. S. Pande, J. D. Chodera, M. J. Schnieders, I. Haque, D. L. Mobley, D. S. Lambrecht, R. A. DiStasio, et al., *The Journal of Physical Chemistry B* **2010**, *114*, 2549–2564.

[112]C. Liu, J.-P. Piquemal, P. Ren, *Journal of Chemical Theory and Computation* **2019**, *15*, 4122–4139.

[113].

[114]K. A. Beauchamp, J. M. Behr, A. S. Rustenburg, C. I. Bayly, K. Kroenlein, J. D. Chodera, *The Journal of Physical Chemistry B* **2015**, *119*, 12912–12920.

[115]C. Zanette, C. C. Bannan, C. I. Bayly, J. Fass, M. K. Gilson, M. R. Shirts, J. D. Chodera, D. L. Mobley, *Journal of Chemical Theory and Computation* **2018**, *15*, 402–423.

[116]B. Waldher, J. Kuta, S. Chen, N. Henson, A. E. Clark, *Journal of Computational Chemistry* **2010**, NA–NA.

[117]F. Zahariev, N. D. Silva, M. S. Gordon, T. L. Windus, M. Dick-Perez, *Journal of Chemical Information and Modeling* **2017**, *57*, 391–396.

[118]S. Grimme, *Journal of Chemical Theory and Computation* **2014**, *10*, 4497–4514.

[119]Y. Basdogan, A. M. Maldonado, J. A. Keith, *WIREs Computational Molecular Science* **2019**, *10*, DOI 10.1002/wcms.1446.

[120]Y. Basdogan, J. A. Keith, *Chemical Science* **2018**, *9*, 5341–5346.