Benchmarking Semiempirical QM Methods for Calculating the Dipole Moment of Organic Molecules

Ademola Soyemi [a] *and Tibor Szilvási* [a]*

[a] Department of Chemical and Biological Engineering, University of Alabama, Tuscaloosa, Alabama 35487, United States.

* Email: tibor.szilvasi@ua.edu

ABSTRACT

The dipole moment is a simple descriptor of the charge distribution and polarity and is important for understanding and predicting various molecular properties. Semiempirical (SE) methods offer a cost-effective way to calculate dipole moment that can be used in high-throughput screening applications although the accuracy of the methods is still in question. Therefore, we have evaluated AM1, GFN0-xTB, GFN1-xTB, GFN2-xTB, PM3, PM6, PM7, B97-3c, HF-3c, and PBEh-3c SE methods, which cover a variety of SE approximations, to directly assess the performance of SE methods in predicting molecular dipole moments and their directions using 7211 organic molecules contained in the QM7b database. We find that B97-3c and PBEh-3c perform best against coupled cluster reference values yielding dipole moments with a mean absolute error (MAE) of 0.10 D and 0.11 D, respectively, which is similar to the MAE of DFT methods (~0.1 D)

reported earlier. Analysis of the atomic composition shows that all SE methods show good performance for hydrocarbons for which the spread of error was within 1 D of the reference data, while the worst performances are for sulfur-containing compounds for which only B97-3c and PBEh-3c show acceptable performance. We also evaluate the effect of SE optimized geometry instead of the benchmark DFT geometry that shows a dramatic drop in the performance of AM1 and PM3 for which the range of error tripled. Based on our overall findings, we highlight that there is an optimal compromise between accuracy and computational cost using GFN2-xTB (MAE: 0.25 D) that is three orders of magnitude faster than B97-3c and PBEh-3c. Thus, we recommend using GFN2-xTB for cost-efficient calculation of the dipole moment of organic molecules containing C, H, O, and N atoms whereas for sulfur containing organic molecules, we suggest PBEh-3c.

1. INTRODUCTION

The dipole moment of molecules is a simple measure of the electron density and polarity and as such plays an important role in understanding intermolecular and intramolecular interactions relevant for predicting bulk and molecular properties. For example, strong dipole-dipole interactions can rationalize strong intermolecular interactions that lead to higher melting and boiling points.[1] The dipole moment can also be used as a descriptor for designing solvents.[2-6] Molecules with large dipole moments readily polarize other molecules and promote dipole-dipole or dipole-induced dipole interactions that help miscibility and solubility. In addition, the dipole moment is also critical to understanding photon absorption induced transitions in rotational and vibrational spectroscopy to name a few applications. Thus, with the wide range of applicability of the dipole moment, it is imperative to understand how accurately dipole moment can be calculated with different computational methods.

Wave function theory, specifically coupled cluster (CC) theory, is the gold standard for obtaining reliable dipole moment data, similar to reaction energies, building on the systematically improvable nature of the underlying theory. As in energy calculations, calculating dipole moment using CC theory with an acceptably large basis set is a computationally costly endeavor and limited only to small molecules. Therefore, wave function theory and CC calculations mainly serve as benchmarks in calculating the dipole moment of molecules.

Density Functional Theory (DFT) is currently the most popular method for calculating the electronic structure of molecules because of its favorable cost-to-performance ratio which allows for reasonably accurate calculations of medium sized (50-200 atoms) molecules. To provide guidance on which DFs with what basis sets are the most accurate, CC calculations have been often employed as benchmarks[7-9] however only a few studies used large enough dataset to provide statistically relevant conclusions. Notably, Hait and Head-Gordon have recently calculated dipole moments for a set of 152 small molecules at the CCSD(T)[10] level of theory and used the results to test the performance of 88 popular DFs from different rungs on Jacob's ladder. In their study, they found that double-hybrid functionals (e.g. ωB97X-2)[11] gave the best performance, yielding dipole moments within 3.6-4.5% regularized root-mean-square (RMS) error compared to the reference CCSD(T) which was comparable to the approximately 4% regularized RMS error produced by CCSD at a greater computational cost. Hybrid functionals (e.g ωB97X-V)[12] also provided good performance with regularized root-mean-square errors in the 5-6% range.[13] Building on Hait and Head-Gordon's work, Zapata and McKemmish evaluated the performance of 38 basis sets of single- up to triple-zeta quality paired with 9 different DFT methods.[14] In their study, they found that in agreement with Hait and Head-Gordon's work, hybrid functionals performed best. However, they also observed that the calculated dipole moment was more sensitive to the basis set

size than the DFT method of choice. For example, the regularized RMS produced by the hybrid functional ωB97X-V when paired with double-ζ-quality basis sets varied between a range of about 5% for aug-pc-1[15-17] and 38% for 6-31G,[18] while for triple- ζ-quality basis sets the regularized RMS varied between approximately 5% for pVTZ[19] and 36% for 6-311+G.[20, 21] Because the best performing double- and triple-ζ-quality basis sets yielded similar performance, the authors recommended that the best compromise between accuracy and computational efficiency is achieved when the augmented double-ζ-quality basis set aug-cc-pVDZ[22] is paired with hybrid functionals (e.g. ωB97X-V). Similar to Zapata and Mckemmish's work, Hickey and Rowley also studied the basis set dependence of dipole moment prediction using DFT methods.[23] In their work, the performance of 7 DFT methods paired with 5 basis sets of double- and triple-ζ-quality was studied for a dataset of 46 molecules containing elements C, Si, N, O, S, F, and Cl in which they found that hybrid functionals produced the best performance compared to experimental dipole moment data. Hickey and Rowley showed that hybrid functionals such as B3LYP,[24-26] and PBE0[27] when paired with the aug-cc-pVTZ[22, 28] basis set yielded comparable predictions to CCSD/aug-cc-pVTZ compared to experimental dipole moment data with mean absolute errors of 0.09, 0.09, and 0.07 D respectively. Thus, the study further supports the usage of hybrid functionals with augmented double- and triple-ζ-quality basis sets for the prediction of molecular dipole moments although the dataset is too small to provide statistically relevant conclusions.

Semiempirical (SE) methods constitutes the third category whereby the SE methods are still based on ab initio formalism but neglect terms and make use of parameterization to decrease computational cost by orders of magnitude while also sacrificing some accuracy at the same time.[29, 30] Semiempirical methods based on Hartree-Fock formalism have been used since the 1970s and since then a number of methods have been introduced including AM1,[31] PM3,[32] PM6,[33]

and PM7.[34] In this time, there have been several studies which benchmarked the performance of SE methods in predicting the dipole moments of specific sets of molecules,[35-39] however these studies focused on a narrow range of molecules for which experimental dipole moment data was available. For other molecules, these early studies used ab-initio reference data with very small basis set,[39, 40] due to computational cost reasons, which are now known to be inadequate for benchmark purposes.[41-43] For example, Anisimov et al. evaluated the performance of the AM1 and PM3 SE methods against dipole moment data calculated at the MP2/6-31+Gx (where 'x' indicates that the d-orbital exponent was set to 0.2) level of theory for a set of 20 natural amino acids.[40] Thus, the accuracy of SE methods in predicting dipole moments cannot be accurately accessed based on these studies alone.

Recently, there has been a renewed interest in SE methods to find faster alternatives to accurate but costly DFT calculations with large basis sets for medium-sized and large (~500 atoms) molecular systems. This effort has led to the development of the so-called '3c' composite methods, such as HF-3c[44], PBEh-3c[45], and B97-3c[46], which provide relatively accurate results by introducing three physically motivated atom pair-wise correction terms for dispersion interactions, basis set superposition error, and short-ranged basis set incompleteness effects.[47, 48] These methods have been developed for obtaining accurate and affordable geometries and relative energies, but accurate description of the charge density is challenging due to their rather compact orbital basis set expansions. Due to inaccuracies in describing the charge density, there can be errors in predicting the dipole moment of molecules. To evaluate the performance of the 3c methods in predicting dipole moments, Caldeweyher and Brandenburg compared the performance of HF-3c, PBEh-3c, and B97-3c against experimental dipole moment data for 43 molecules in Hickey and Rowley's benchmark.[48] The authors showed that PBEh-3c and B97-3c perform comparable to

DFT methods studied by Hickey and Rowley yielding mean absolute deviations of 0.11 and 0.09 D, respectively, against methods such as PBE, PBE0, and B3LYP/aug-cc-pVTZ[22] which all yielded mean absolute deviations of 0.09 D while HF-3c performed worse with a mean absolute deviation of 0.21 D. These results are very promising and suggest composite methods are a good alternative to DFT for dipole moment calculations, however, as noted previously as well, 43 molecules are not a statistically relevant dataset.

As the '3c' SE methods are still computational costly for number of problems, for example conformer search, the GFNn-xTB semiempirical methods were designed as special purpose tools to provide affordable geometries, vibrational frequencies, and noncovalent interactions for large systems.[47] Bannwarth et al. compared the performance of the GFN1-xTB,[49] GFN2-xTB,[50] and PM6 against Hait and Head-Gordons's CCSD(T) dipole moment benchmark data to estimate the reliability of these SE methods for dipole moment calculations.[50] The authors showed that GFN1-xTB, GFN2-xTB, and PM6 provide predictions with a mean absolute deviation of 0.69, 0.45, and 0.52 D, respectively. We however note that the small molecules Hait and Head-Gordon introduced in their benchmark dataset is not representative for the chemical systems SE methods are typically used for. The dataset contains a significant number of unstable inorganic species that pose a challenge for most electronic structure methods, while SE methods are generally applied to treat stable closed shell organic molecules, which can explain their weak performance in this benchmark study. To illustrate, we note that the mean absolute deviation of GFN1-xTB, and GFN2-xTB is higher for open shell species (0.74 and 0.49 D) compared to closed shell species (0.67 and 0.44 D). Therefore, there is a need for a comprehensive evaluation of semiempirical methods on closed shell organic molecules to understand their performance in calculating the dipole moment.

In this work, we present the performance of a set of ten SE methods namely AM1, PM3, PM6, PM7, GFN0-xTB, GFN1-xTB, GFN2-xTB, HF-3c, B97-3c, and PBEh-3c, which represent a wide range of SE approximations, against high level CC dipole moment data composed by Yang et al.[51] The dipole moment calculations are performed across the 7211 organic molecules contained in the QM7b dataset,[41-43] which provides a comprehensive library of organic molecules to obtain statistically relevant data and answer our questions on the reliability of these methods for calculating dipole moment for practical applications, e.g., organic solvent design. Comparing these results to results in Bannwarth et al.,[50] we highlight that the performance of GFN2-xTB is significantly better for organic molecules studied here with a mean absolute error of 0.27 D compared to their 0.45 D and is comparable to the performance of HF-3c. Thus, we suggest GFN2-xTB as a method of choice from SE methods given its optimal cost/performance ratio to calculate the dipole moment of organic molecules except for sulfur containing compounds for which PBEh-3c is suggested.

2. COMPUTATIONAL METHODS

A set of nine SE methods representing a wide range of semiempirical methods have been evaluated in this study namely; (i) Neglect of Diatomic Differential Overlap (NDDO) methods AM1,[31] PM3,[32] PM6,[33] and PM7[34] which employ the valence-only minimal basis set and are based on the NDDO approximation where all three, and four center two-electron integrals are completely neglected; (ii) Grimme's extended tight-binding methods GFN0-xTB, [52] GFN1-xTB,[49] GFN2-xTB[50] which are derived from a perturbation expansion of the electron density similar to the density functional tight binding model. And finally (iii) composite methods HF-3c,[44] PBEh-3c[45], and B97-3c,[46] in which three corrections (hence '3c' in their names), namely the D3 scheme to incorporate London dispersion, a geometrical counterpoise correction to handle intra- and

intermolecular basis set superposition error, and a short-range term to correct basis set deficiencies are applied to the HF,[53] PBE,[54] and B97-D[55, 56] density functionals and coupled with small Gaussian atomic orbital basis sets.

The performance of the tested methods is evaluated by comparing results against high-level CC data. In our analysis, we will use the following metrics and abbreviations: mean absolute error (MAE), maximum absolute deviation (MAD), standard deviation (SD), mean percent error (mean %error), full width at half maximum (FWHM), and regularized relative root-mean-square error (RMSE). We also define all metrics used in our analysis below:

The error between calculated and benchmark dipole moment values is defined as:

$$\mu' = \mu_{CCSD} - \mu \qquad (1)$$

Where μ_{CCSD} is the reference dipole moment calculated at the CCSD/d-aug-cc-pVDZ level of theory, and μ is the dipole moment calculated with the SE method.

The MAE is defined as:

$$\frac{\Sigma|\mu\prime|}{N} \qquad (2)$$

The MAD is defined as:

$$Max(|\mu'|) \qquad (3)$$

The SD is defined as:

$$\sqrt{\frac{\Sigma(\mu' - \mu'_{mean})^2}{N}} \qquad (4)$$

Where μ'_{mean} is the mean error between the dipole moment calculated by the same SE method and the reference dipole moment which we define as:

$$\mu'_{mean} = \frac{\Sigma\mu\prime}{N} \qquad (5)$$

The mean %error is defined as follows:

$$\frac{\Sigma(\frac{|\mu'|}{\mu_{CCSD}})}{N} \times 100\% \qquad (6)$$

The FWHM is defined as:

$$2.355 \times \sigma \qquad (7)$$

where σ is the SD.

As in Hait and Head-Gordon's work,[13] the RMSE is defined to be:

$$\frac{\mu - \mu_{CCSD}}{\max(\mu_{CCSD}, 1D)} \times 100\% \quad (8)$$

where a value of 1D is used for regularization.

The range is defined as:

$$\mu'_{max} - \mu'_{min} \quad (9)$$

where μ'_{max} and μ'_{min} are the maximum and minimum dipole moment calculated by the same SE method, respectively.

To compare each SE method across all metrics, we have defined a condensed metric score. For the calculations associated with the benchmark DFT optimized geometries, we normalize the errors in each metric by the maximum value of that metric for GFN1-xTB which provided the highest errors in every metric for all methods excluding GFN0-xTB. We did not pick GFN0-xTB as base line because its large errors would have made very accurate methods hardly distinguishable from each other based on the condensed metric score. Then we average each normalized metric which yields a single value between 0 and 1 for all methods except GFN0-xTB. To obtain a similar condensed metric score for calculations using SE optimized geometries, we normalize the errors in each metric by the GFN1-xTB value of that metric obtained for DFT optimized geometries. Thus, condensed metric scores obtained for SE optimized geometries can be higher than 1 for all methods if there was a significant drop in the performance compared to the results obtained with DFT optimized geometries. In general, a higher condensed metric score indicates worse performance.

We considered the 7211 organic molecules which contain C, H, N, O, S, and Cl in the QM7b database.[41-43] All molecular geometries for all species were obtained online via Materials Cloud in the sdf format (accessed on 04/28/2021).[51] Table 1 details the number of the organic molecules by elemental composition in the QM7b dataset. In our analysis, we will consider subcategories formed based on elemental composition to obtain more detailed information on the performance of the studied semiempirical methods. Subcategories are defined to include statistically relevant number of molecules (>300 molecules) to give meaningful information. Thus, the defined subcategories are based on CH, CHN, CHO, CHON, and S-containing compositions. The S-containing subgroup, which will be subsequently referred to as 'S-X' includes CHNS, CHNOS, CHOS, CHS, CHNSCl, and CHSCl compositions as shown in Table 1. We highlight that each of our subcategories contains more molecules than most of the previous dipole moment benchmark studies.

Table 1. Number of organic compounds based on elemental composition in the QM7b dataset.

Composition	Molecule count
CHON	2580
CHN	1928
CHO	1867
CH	498
CHNS	134
CHNOS	78
CHOS	73
CHNCl	15
CHS	14
CHNOCl	9
CHNSCl	5

CHOCl	4
CHSCl	4
CHCl	1
CN	1
Total	7211

Benchmark values were obtained from Yang et al.[51] For high-level couple-cluster, the authors computed the dipole moments for all 7211 molecules employing LR-CCSD with the doubly augmented d-aug-cc-pVDZ basis set.[57] In Hickey and Rowley's and Hait and Head-Gordon's works, CCSD/aug-cc-pVDZ has been shown to provide accurate dipole moment predictions with a mean absolute error of 0.08 D compared to experiment,[23] and approximately 4% regularized RMS error compared to CCSD(T).[13] These findings indicated to us that LR-CCSD/d-aug-cc-pVTZ used in the QM7b dataset can serve as a reliable reference for benchmarking semiempirical methods given the similar and somewhat larger basis set used in the previous benchmarks and that we expected the mean absolute error of the SE methods to be an order of magnitude larger compared to CC results, of which the latter assumption was indeed proved in our study.

Molecular geometries in the sdf format were converted to the xyz format using the OpenBabel software (Version 2.3.1).[58, 59] All AM1, PM3, PM6, and PM7 single point and geometry optimization calculations were carried out using the Gaussian software program (Version 16, Revision C.01) using default settings.[60] All GFN0-xTB, GFN1-xTB and GFN2-xTB single point and geometry optimization calculations were carried out using the xTB software package and default settings (Version 6.4).[47, 61] All HF-3c, B97-3c, and PBEh-3c single point and geometry optimization calculations were carried out using the Orca software program and default settings (Version 4.2.1).[62, 63]

3. RESULTS AND DISCUSSION

We have calculated the dipole moment given by AM1, PM3, PM6, PM7, GFN0-xTB, GFN1-xTB, GFN2-xTB, HF-3c, PBEh-3c, and B97-3c SE methods using the benchmark DFT optimized geometries of the QM7b dataset as well as the SE optimized geometries. The error metrics defined in the Computational Methods section were used to evaluate the performance of each SE method in predicting dipole moments for both DFT and SE optimized geometries. For clarity, we will discuss the results using the benchmark DFT optimized structures first then we will analyze the results related to the SE optimized structures in a separate section.

3.1. PERFORMANCE OF SE METHODS USING BENCHMARK DFT OPTIMIZED STRUCTURES

Generally, all SE methods studied herein were able to give reasonable dipole moments based on the R^2 value (Figure 1) for each method compared against benchmark CCSD. For the benchmark DFT optimized geometries, the SE methods studied showed strong prediction ability with R^2 values that ranged from 0.93 for the worst performing method GFN1-xTB to 0.99 for the best performing method PBEh-3c.

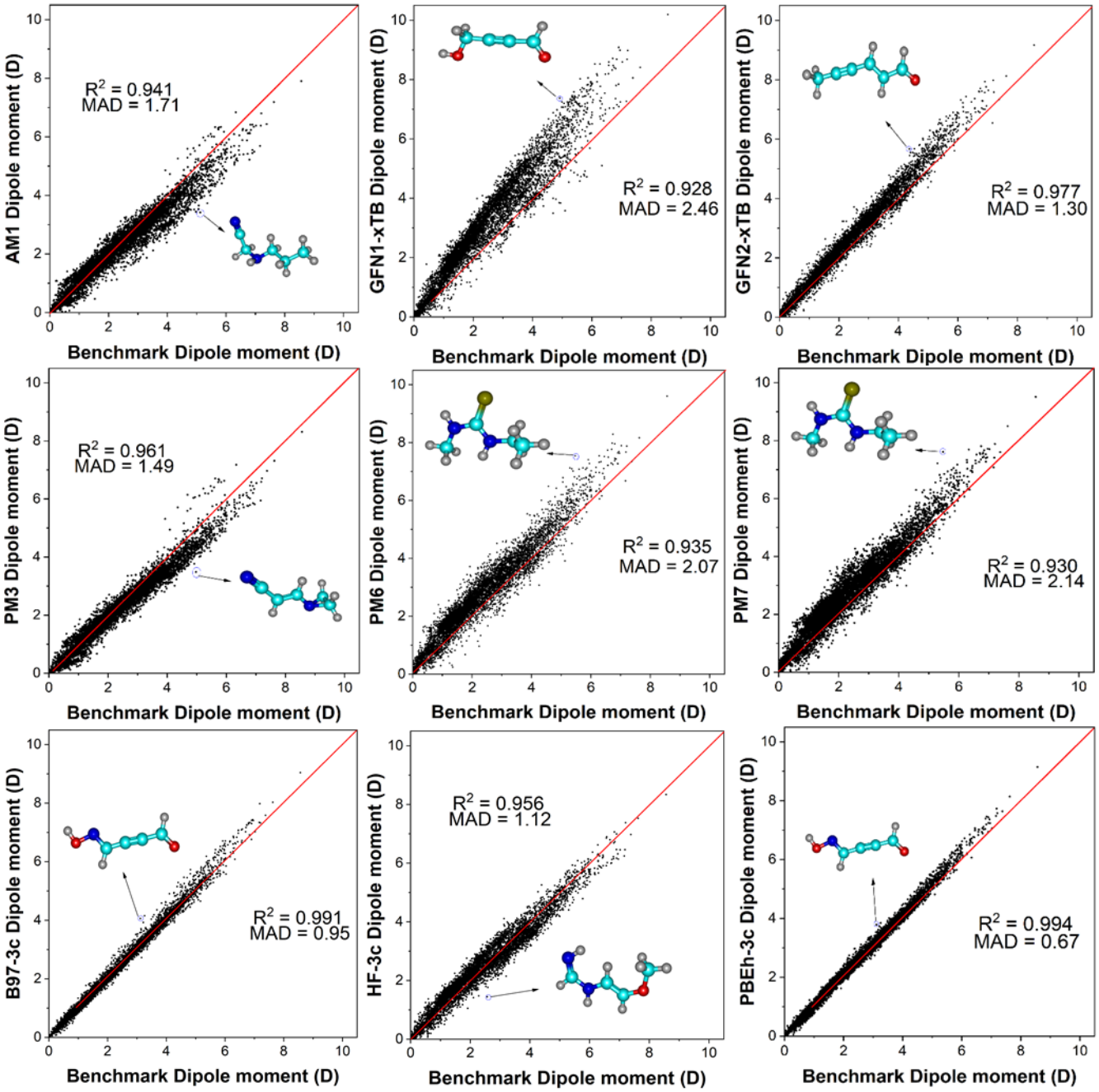


Figure 1. Parity plots comparing dipole moment predicted for the SE methods studied (see Supporting information for GFN0-xTB) using DFT optimized geometries to benchmark dipole moment data. The red diagonal line is to help the reader see the ideal correlation. The blue circle and black arrow point to the molecule with the largest deviation from its CCSD dipole moment.

Figure 2a shows a radial plot where all methods evaluated in this study are compared based on multiple error metrics. The different errors for each SE method are also given in Table S1. From Figure 2a we can immediately observe that GFN1-xTB gave the worst performance across each metric with MAE, Mean %error, MAD, SD, and RMSE of 0.70 D, 37%, 2.46 D, 0.49 D, and 38%, respectively. Next, PM6 and PM7 give nearly identical performance across each metric with MAE, Mean %error, MAD, SD, and RMSE of 0.46 D, 32%, 2.07 D, 0.36 D, and 31% for PM6 and 0.43 D, 31%, 2.14 D, 0.37 D, and 32% for PM7. However, PM6 and PM7 showed improvement compared to GFN1-xTB in each metric especially in terms of the MAE where there is more than a 0.2 D difference. AM1, PM3, GFN2-xTB, and HF-3c form the next group with similar levels of performance across the MAE, Mean %error, and RMSE metrics. AM1 and GFN2-xTB have a MAE of 0.27 D while PM3 and HF-3c show a MAE of 0.25 D. For the Mean %error, AM1 and GFN2-xTB produced Mean %errors of 16% while PM3 and HF-3c give Mean %errors of 15% and 16%, respectively. For the SD, GFN2-xTB performed best compared to AM1, PM3, and HF-3c with SD of 0.24, 0.36, 0.29, and 0.29 D, respectively. This indicates that GFN2-xTB has the smallest spread of error among these four methods. HF-3c significantly outperforms AM1 and PM3 in terms of MAD, with MADs of 1.12, 1.71, and 1.49 D, respectively, but is still comparable to the MAD of GFN2-xTB (1.30 D). Compared to the performance of all other methods studied here, B97-3c and PBEh-3c showed superior performances in every metric. B97-3c and PBEh-3c showed similar performance in MAE, RMSE, and SD metrics, with MAE of 0.10 and 0.11 D, RMSE of 6 and 7%, and SD of 0.13, and 0.12 D, respectively. PBEh-3c however significantly outperforms B97-3c in Mean %error and MAD, with Mean %errors of 2 and 6%, and MADs of 0.67 and 0.95 D, respectively.

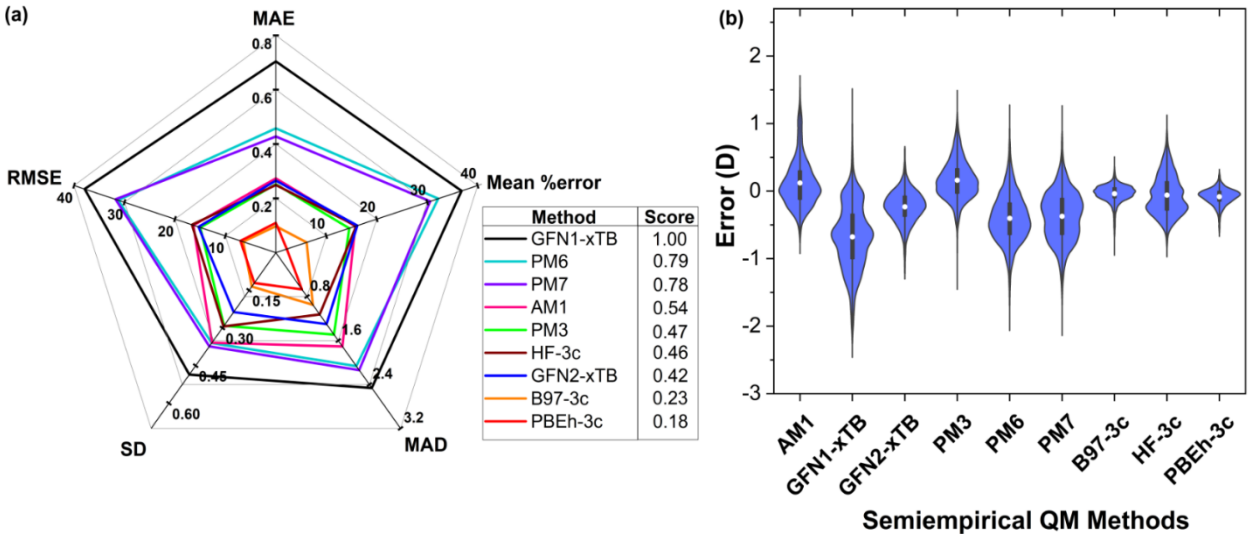


Figure 2. (a) Radial plot showing the performance of each SE method based on different error metrics and (b) Violin plots for each SE method showing the kernel smooth probability density distribution of the errors using DFT optimized geometries. The score is a normalized average of each metric and serves as a proxy for the area under each polygon. See definitions of all metrics in the Computational Methods section.

Based on our condensed error metric score (Figure 2a), GFN1-xTB was the worst performing method for the DFT optimized geometries with a score of 1.00. Additionally, we observed that the older AM1 and PM3 methods with scores 0.54 and 0.47, respectively, outperformed the newer PM6 and PM7 methods which gave a similar level of performance with scores 0.79 and 0.78, respectively. This poor performance for organic molecules is not surprising given that PM6 and PM7 are parametrized with reference data sets that were expanded to cover more of the periodic table compared to AM1 and PM3.[34] Interestingly, GFN2-xTB and PM3 with scores of 0.45 and 0.47, respectively, performed at a similar level to HF-3c which had a score of 0.46 while also being more computationally costly. B97-3c and PBEh-3c gave the best performance of all methods studied with scores of 0.23 and 0.18, respectively, which was not surprising given their sounder theoretical background compared to other semiempirical methods.

In order to give further insights into the performance of each SE method, we have explored the probability density distribution of errors of each SE method which gives the probabilities that the error given by a SE method will fall at a certain value. In order to compare each method, we have presented the probability density distribution of each method studied here as violin plots in Figure 2b. In Figure 2b, we can again immediately observe the poor performance of GFN1-xTB given the spread of the errors based on the range of 3.98 D and the presence of two peaks (i.e. a bimodal probability density distribution) at 0 D and -0.65 D, respectively, which indicated that GFN1-xTB behaved differently for certain group of molecules. The position of the mean error of GFN1-xTB (-0.68 D) relative to 0 D also indicated that GFN1-xTB systematically overpredicted the dipole moment compared to the CCSD reference data. For PM6 and PM7, there is also a wide spread of errors with ranges of 3.34 and 3.40 D, respectively. The position of the mean errors for PM6 (-0.40 D) and PM7 (-0.37 D) relative to 0 D also indicated that these methods also systematically overpredict the dipole moment. Additionally, the presence of two peaks for PM7 at -0.61 and -0.19 D was attributed to behaving differently for two groups of compounds. For GFN2-xTB and AM1, the spread of the error is significantly improved compared to the other the previously mentioned SE methods with ranges of 1.96 and 2.63 D, respectively, in addition to producing only single peaks at -0.23 and 0 D, respectively. The position of the mean error of GFN2-xTB (-0.23 D) indicated GFN2-xTB generally overpredicts the dipole moment, while the mean error of AM1 (0.12 D) shows AM1 generally underpredicted the dipole moment compared to the CCSD reference data. For the '3c' methods, PBEh-3c provided the smallest spread in error with a range of 0.99 D followed by B97-3c with a range of 1.46 D, while HF-3c had a range of 2.10 D, which was worse than the range of 1.96 D given by GFN2-xTB. Additionally, B97-3c and PBEh-3c showed only single peaks at 0 and -0.05 D, respectively, while HF-3c had two peaks at 0 and -0.24

D. The ‘3c’ methods only slightly overpredict the dipole moment with mean errors of -0.06, -0.04, and -0.08 D for HF-3c, B97-3c, and PBEh-3c, respectively.

3.1.1. PERFORMANCE OF SE METHODS FOR DIFFERENT ATOMIC COMPOSITIONS

Due to the multiple peaks that appeared in the violin plots, we were interested in understanding the performance of each SE method for molecules with different atomic composition that formed five subcategories (see Computational Methods section). Figure 3 shows the probability distribution of errors for each SE method for each subcategory. For all methods, the most consistent performance was observed for molecules composed of C and H only (i.e. hydrocarbons), for which there was low spread of error with ranges of 1.02, 1.14, 0.75, 0.97, 0.72, 0.37, and 0.49 D for AM1, PM3, GFN1-xTB, GFN2-xTB, HF-3c, PBEh-3c, and B97-3c, respectively. Furthermore, for the CH category there was always a single peak except for PM6 and PM7 which notably overpredict the dipole moment of hydrocarbons with mean errors of -0.24 D and -0.18 D, respectively, and have a relatively wider spread of error with ranges of 1.73 and 1.41 D, respectively, in addition to having no well-defined peak. The good performance of most of the SE methods is not surprising since hydrocarbons are apolar and have very low dipole moments such that SE methods will have little problem providing accurate absolute dipole moment predictions.

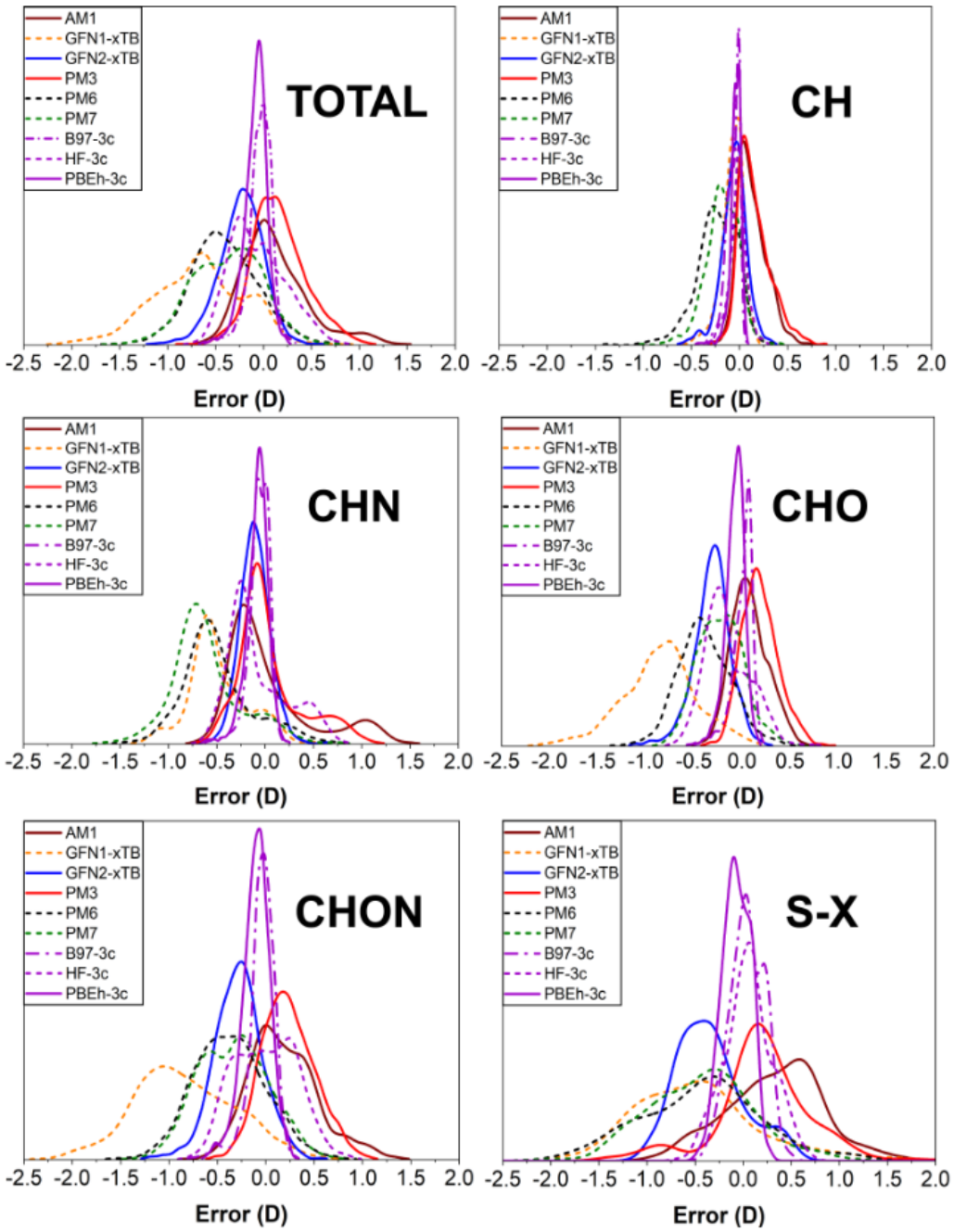


Figure 3. Probability density distribution of errors for each SE method using benchmark DFT optimized geometries for the CH, CHO, CHN, CHON, and S-X compositional subcategories. The vertical axis of each plot is the probability density of the error. For comparison, the probability density distribution of errors for the total dataset is also disclosed.

In the CHN subcategory, B97-3c, PBEh-3c, and GFN2-xTB gave the best performances and produced probability density distributions with single peaks at -0.07 D, -0.08 D, and -0.11 D, respectively, and relatively low spread of error with ranges of 1.04 D, 0.92 D, and 1.50 D, respectively. AM1, PM3, PM6, PM7 in addition to GFN1-xTB and HF-3c however had poorer performances. GFN1-xTB, PM6, and PM7 systematically overpredict the dipole moment of molecules in this subcategory with mean errors of -0.49 D, -0.49 D, and -0.61 D, respectively, in addition to having wide spread of error with ranges of 2.34 D, 2.57 D, and 2.69 D, respectively. Despite AM1, PM3 and HF-3c having insignificant systematic error given mean errors of 0.04 D, 0.04 D, and -0.07 D, respectively, these methods gave probability density distributions with a second peak in addition to having wide spreads of error with ranges of 2.38 D, 2.31 D, and 1.82 D respectively.

In the CHO subcategory, majority of the SE methods performed well with AM1, B97-3c, and PBEh-3c performing best with essentially no systematic error given their mean errors of 0.08 D, 0.00 D, -0.07 D, respectively, and narrow spread error with ranges of 1.31 D, 1.13 D, and 0.89 D, respectively. GFN2-xTB, HF-3c, PM6, and PM7 provided comparable performances in which these methods systematically overpredicted the dipole moment of molecules in this subcategory with mean errors of -0.31 D, -0.15 D, -0.39 D, and -0.22 D, respectively, while PM3 slightly overpredicted the dipole moment with a mean error of 0.19 D. Additionally, GFN2-xTB , HF-3c, and PM3 yielded similar spread of error with ranges of 1.61 D, 1.55 D, and 1.38 D, respectively, while PM6 and PM7 had a wider spread of error with ranges of 2.21 D and 1.89 D, respectively. GFN1-xTB significantly overpredicted the dipole moments with a mean error of -0.86 D while producing a wide range of error 2.74 D.

In the CHON subcategory, B97-3c and PBEh-3c again were the best performing SE methods, providing minimal systematic error given by their mean errors of -0.05 and -0.1 D, respectively, as well as the relatively narrow ranges of 1.27 D and 0.99 D, respectively. GFN2-xTB and PM3 yielded similar levels of performance with ranges of 1.90 D and 1.87 D, respectively, with GFN2-xTB overpredicting the dipole moments having a mean error of -0.29 D and PM3 underpredicting the dipole moments with a mean error of 0.23 D. AM1's performance was comparable to that of PM3, with AM1 also underpredicting the dipole moments (mean error of 0.19 D), however AM1 had a significantly larger spread of error with a range of 2.46 D. Similar to its performance in the CHN subcategory, HF-3c on average had no systematic error, however HF-3c gave a relatively wide spread of error with its range of 2.10 D and had a probability density distribution with two peaks. In parallel to the CHN subcategory, PM6 and PM7 had similar performance. Although PM6 and PM7 both overpredicted the dipole moments with mean errors of -0.37 D and -0.33 D, respectively, and had similar spread of error with ranges of 2.58 D and 2.45 D, respectively, PM7 however showed a probability density distribution with two peaks. GFN1-xTB was the worst performing SE method, significantly overpredicting the dipole moment with a mean error of -0.83 D and the widest spread of error for the CHO subcategory with a range of 3.32 D.

For the S-X subcategory, only PBEh-3c, B97-3c and to a lesser extent HF-3c performed acceptably. PBEh-3c, B97-3c, and HF-3c had practically no systematic error with mean errors of -0.07 D, 0.04 D, and 0.05 D, respectively, however PBEh-3c had a much narrower spread of error with a range of 0.77 D compared to the range of 1.36 D for HF-3c. AM1 and PM3 yielded similar levels of performance for the S-X subcategory, with comparable ranges of 2.63 D, and 2.70 D, respectively, while GFN2-xTB yielded a narrower range of 1.65 D. GFN2-xTB slightly overpredicted the dipole moment with a mean error of -0.34 D, while AM1 and PM3

underpredicted the dipole moment with mean errors of 0.33 D and 0.14 D, respectively. As we saw in other subcategories, PM6 and PM7 had nearly matching performances, with both methods overpredicting the dipole moment of molecules in the S-X subcategory with mean errors -0.44 D and -0.43 D, respectively, and ranges of 3.34 D and 3.40 D, respectively. GFN1-xTB yielded a slightly worse performance compared to PM6 and PM7, overpredicting the dipole moment with a mean error of -0.50 D but produced a comparable spread of error with a range of 3.48 D.

3.1.2. EFFECT OF ATOMIC COMPOSITION ON OVERALL SE METHOD PERFORMANCE

To investigate further the clustering of errors which resulted in multiple peaks for GFN1-xTB, HF-3c, and PM7 in Figure 2b, we present the subcategory data organized by SE method in Figure 4.

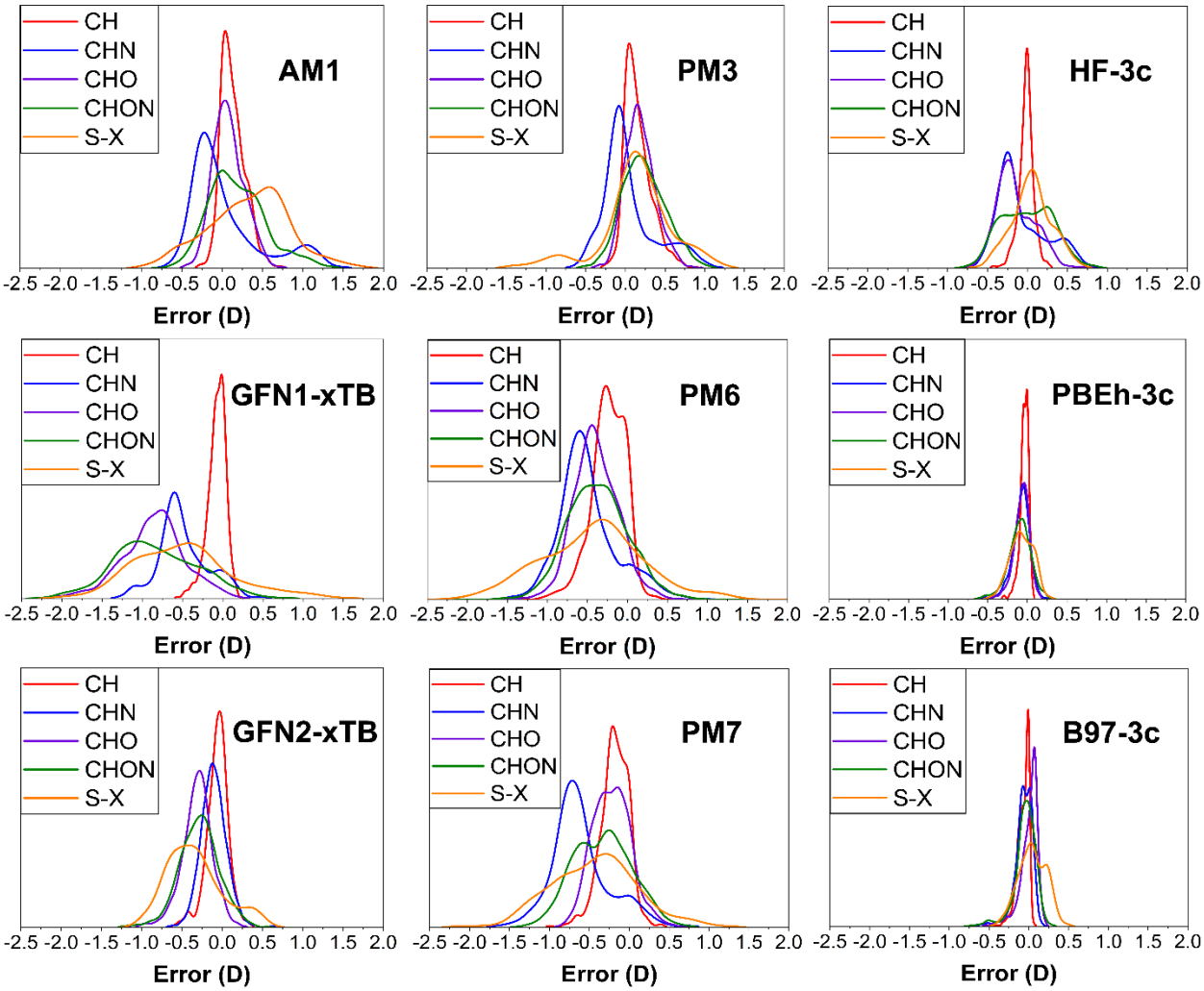


Figure 4. Probability density distribution of subcategory errors for each SE method using benchmark DFT optimized geometries. The vertical axis of each plot is the probability density of the error. For visualization purposes, categories are not normalized with its population. For plots with normalized subcategory data relative to the total dataset see the SI (Figure S17).

For GFN1-xTB, the performance was particularly poor due to the varied positions of the probability density distributions of the subcategories reflected in the mean errors of -0.08, 0.,49, -0.86, -0.83, and -0.50 D for the CH, CHN, CHO, CHON, and S-X subcategories, respectively. In addition to the varied positions of the probability density distributions, GFN1-xTB considerably overpredicted the dipole moment of every category except the CH subcategory based on the mean errors. By comparing Figure 2b and Figure 4 based on the mean errors of each subcategory, the peak of the probability density distribution in Figure 2b located at 0 D can be attributed to the CH subcategory, while the second peak located at -0.6 D can be attributed to the CHN, CHO, and CHON subcategories.

For HF-3c and PM7, the CHON subcategory contributed significantly to their poor performance. For HF-3c, there was a very wide spread of error in the CHON subcategory given by a SD of 0.32 D. In addition to the wide spread of error, two different peaks were present in Figure 5 positioned at -0.31 D and 0.25 D, respectively. Given that the CHN subcategory had a probability density distribution positioned at -0.24 D, we ascribed the -0.24 D peak observed in the probability density distribution for HF-3c in Figure 2b to the CHN and CHON subcategories. Analogously, the second smaller peak located at 0.00 D in the probability density distribution for HF-3c in Figure 2b was attributed to the CH (mean error: -0.02 D), CHO (mean error: -0.15 D), S-X (mean error: 0.05 D) subcategories as well as the second peak of the CHON subcategory which was located at 0.25 D. Similarly, for PM7, there was a wide spread of error in the CHON subcategory reflected in a SD of 0.37 D. In addition to the wide spread of error, two different peaks were present positioned at -0.26 D and -0.56 D, respectively. Also, given that the CHN subcategory had a probability density distribution positioned at -0.71 D, we ascribed the -0.66 D peak observed in the probability density distribution for HF-3c in Figure 2b to the CHN and CHON subcategories. While the second peak

at -0.19 D for the probability density distribution for PM7 in Figure 2b was explained by the CH (mean error: -0.18 D), CHO (mean error: -0.22 D) subcategories as well as the second peak of the CHON subcategory which was located at -0.26 D.

Among methods with a single peak in Figure 2b, PBEh-3c is particularly encouraging because the probability density distributions for all subcategories had similar mean errors (CH: -0.04 D, CHN: -0.08 D, CHO: -0.07, CHON: -0.1, and S-X: -0.07 D) and spread of error (CH SD: 0.05 D,CHN SD: 0.11 D,CHO SD: 0.10 D, CHON SD: 0.14 D, and S-X SD: 0.14 D) which resulted in only an overall mean error of -0.08 D and overall SD of 0.12 D. Additionally, all categories had comparable ranges for PBEh-3c which provided the lowest overall range (0.99 D) for all SE methods suggesting there was no drop in the performance for different compositions of organic molecules. A similar trend was observed for B97-3c where all subcategories had similar mean errors (CH: -0.07 D, CHN: -0.07 D, CHO: 0 D, CHON: -0.05 D, and S-X: 0.04 D) and spread of error (CH SD: 0.07 D, CHN SD: 0.12 D, CHO SD: 0.12 D, CHON SD: 0.15 D, and S-X SD: 0.17 D) which resulted in an overall mean error of -0.04 D and SD of 0.13 D. Furthermore, all categories had comparable ranges for B97-3c which provided the second lowest overall range (1.46 D) for all SE methods implying that there was no drop in the performance for different compositions of organic molecules.

For GFN2-xTB, we observed spread between the positions of the probability density distributions of the of the subcategories (CH: -0.06 D, CHN: -0.11 D, CHO: -0.31 D, CHON: -0.29 D, and S-X: -0.34 D), however, the CHO and CHON subcategories which together accounted for 62% of the dataset had essentially the same mean error (CHO: -0.31 D, and CHON: -0.29 D) and similar spread of error (CHO SD: 0.20 D, and CHON SD: 0.25 D. Thus, because of its lower population relative to the CHO and CHON subcategories, the CHN subcategory which had a

relatively close mean error of -0.11 D and comparable spread of error (CHN SD: 0.16 D) did not result in the appearance of a second peak.

Similar to GFN2-xTB, AM1 also exhibited variation between the mean errors of the subcategories in which the CHO and CHN subcategories gave mean errors of 0.08 D and 0.04 D, respectively. The CH and CHON subcategories formed the next group with mean errors of 0.13 D and 0.19 D, respectively, while the S-X subcategory had a mean error of 0.33 D. However, the positions of each subcategory distribution were close enough to prevent the appearance of a second peak.

For PM3, the dipole moment of the CH, CHO, CHON and S-X subcategories were significantly underpredicted given the mean errors of each subcategory (CH: 0.16 D, CHN: 0.04 D, CHO: 0.19 D, CHON: 0.23 D, and S-X: 0.14 D). In addition, the performance of PM3 was worse for the CHON (SD: 0.28 D) and S-X (SD: 0.50 D) subcategories compared to the CH (SD: 0.17 D) and CHO (SD: 0.19 D) subcategories. Furthermore, although the CHN subcategory had an acceptable mean error, the performance of PM3 in terms of the spread of error was poorer for the CHN subcategory (SD: 0.34 D) compared to the other subcategories.

For PM6, all subcategories except for the CH and S-X subcategories had relatively similar mean errors (CH: -0.24 D, CHN: -0.49 D, CHO: -0.39 D, CHON: -0.37 D, and S-X: -0.44 D) and spread of error (CH: 0.22 D, CHN: 0.37 D, CHO: 0.28 D, CHON: 0.37 D, and S-X: 0.62 D). While there is a significant drop in performance for the S-X subcategory, the poor performance in the S-X subcategory is mitigated by the relatively small population of the S-X subcategory and as such its effect is minimal in the overall spread of error for PM6.

3.2. PERFORMANCE OF SE METHODS IN PREDICTING THE DIPOLE MOMENT DIRECTION

Due to the nature of the dipole moment as a vector quantity, we were also interested in evaluating whether the SE methods studied here could reproduce the direction of the dipole moment and not just its magnitude. This is important because it is possible for a correct magnitude to be predicted without preserving the direction of the dipole. Therefore, to quantify the ability of the SE methods in reproducing the direction of the dipole moment, we collected the components of the dipole moment for all 7211 molecules and calculated the angle between the dipole moment vector predicted by the SE method and the CCSD benchmark data provided in the QM7b dataset using the dot product of these two vectors.

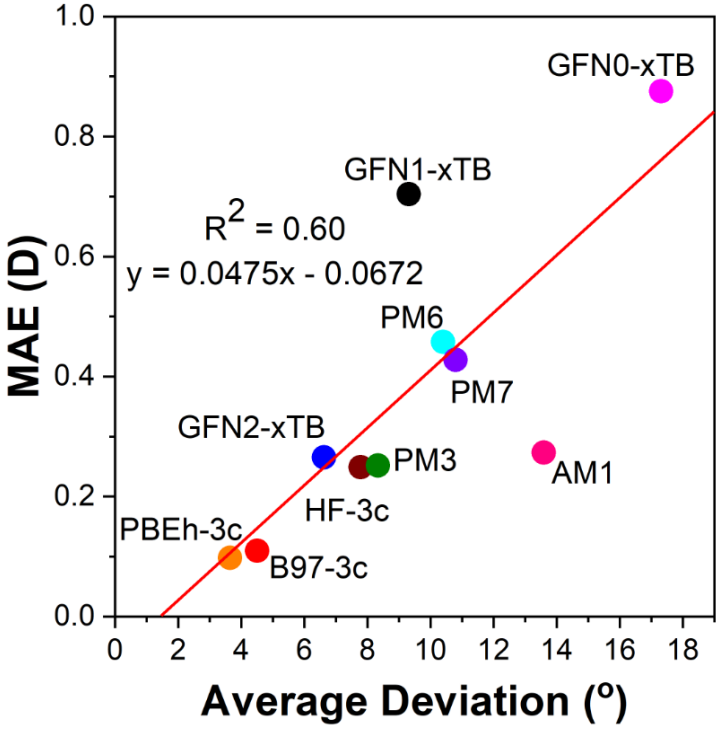


Figure 5. Correlation between accuracy in dipole moment magnitude in terms of MAE and accuracy in dipole moment direction in terms of the average deviation. The red line shows the fitted equation.

In Figure 5, we compare the average angle between the benchmark dipole moment vector and SE dipole moment vector to the MAE given by each method. In this figure, we see that B97-3c and PBEh-3c perform the best in terms of reproducing the direction of the dipole moment with average deviations of 3.6 and 4.5 degrees, respectively. GFN2-xTB, HF-3c, and PM3 form the next group of methods with average deviations of 6.6, 7.8, and 8.3 degrees respectively. PM6 and PM7 show nearly identical performance in predicting the dipole moment direction with average deviations of 10.4 and 10.8 degrees, respectively, while GFN0-xTB was the worst performing method with an average deviation of 17.3 degrees. Despite the accuracy of AM1 in predicting the magnitude of the dipole moment, it showed a poor performance in reproducing the direction of the dipole moment with an average deviation of 13.6 degrees. Conversely, GFN1-xTB showed a good ability to accurately predict the direction of the dipole moment with an average deviation of 9.3 degrees despite its poor performance in predicting the magnitude of the dipole moment. Overall, the accuracy of the SE method in predicting the magnitude of the dipole moment generally trended with its ability to preserve the direction of the dipole moment as shown by the fitted correlation.

3.3. PERFORMANCE OF SE METHODS USING SE OPTIMIZED STRUCTURES

To further evaluate the reliability of the SE methods studied here, we also evaluated the performance of the SE methods using the SE optimized geometry instead of the DFT geometry provided with the QM7b dataset. We did this to evaluate the performance of SE methods under the practical scenario where the DFT optimized structure is not available and SE method is used to perform the geometry optimization to obtain dipole moment.

Here, we will focus on differences between the performance of each SE method using the benchmark DFT optimized geometries and the SE optimized geometries. Using SE optimized geometries, R^2 values ranged from 0.85 for the worst performing method PM6 to 0.99 for the best

performing method PBEh-3c. Parity plots as well as the R^2 values for each SE method using SE optimized geometries are also given in the Supporting Information. In the section below, we will give method-wise comparisons of performances using DFT and SE optimized geometries from the worse performing to the best performing methods (comparing Figures 2 and 6) and also discuss the effect of the subcategories on the overall performance of the SE methods (Figure 7).

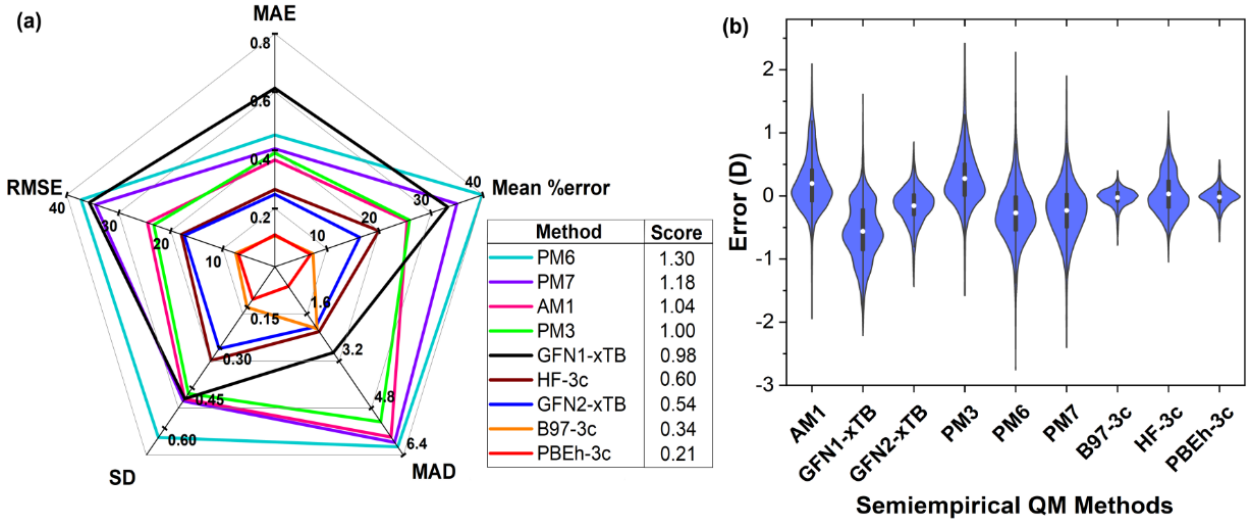


Figure 6. (a) Radial plot showing the performance of each SE method based on different error metrics and (b) Violin plots for each SE method showing the kernel smooth probability density distribution of the errors using SE optimized geometries. The score is a normalized average of each metric and serves as a proxy for the area under each polygon. See definitions of all metrics in the Computational Methods section

Unlike for DFT optimized geometries, the performances of PM6 and PM7 are more distinct from each other as shown in Figure 2a. Using the SE optimized geometries, PM6, performed significantly worse in every category except the MAE which resulted in PM6 becoming the worst performing SE method with a condensed metric score of 1.30 (Figure 6a) while PM7 performed better compared to PM6 with a condensed metric score of 1.18 in contrast to scores of 0.79 and 0.78 for PM6 and PM7, respectively, using DFT optimized geometries. We attributed this drop in

the performance for both PM6 and PM7 to structural differences between DFT and SE optimized geometries which we observed for a large number of molecules in the dataset and resulted in changes in dipole moment of more than 1 D. However, both PM6 and PM7 overpredicted the dipole moment to a lesser degree using SE optimized geometries given by the mean errors of -0.27 D and -0.23 D, respectively, compared to -0.40 D and -0.37 D using the benchmark DFT geometries. Despite the improvement in mean error, the spread of error for SE optimized geometries given by PM6 and PM7 (Figure 6b) were much wider with larger SD (PM6: 0.63 D, and PM7: 0.50 D) and FWHM (PM6: 1.49 D, and PM7: 1.17 D) compared to the SD (PM6: 0.36 D, and PM7: 0.37 D) and FWHM (PM6: 0.85 D, and PM7: 0.88 D) obtained using DFT optimized geometries. Additionally in Figure 6b, although PM7 gave a single peak in its probability density distribution using SE optimized geometries, the range of errors for PM6 and PM7 increased significantly (PM6 range: 9.51 D, and PM7 range: 8.94 D) compared to the range of errors given by PM6 (3.34 D) and PM7 (3.40 D) using DFT optimized geometries.

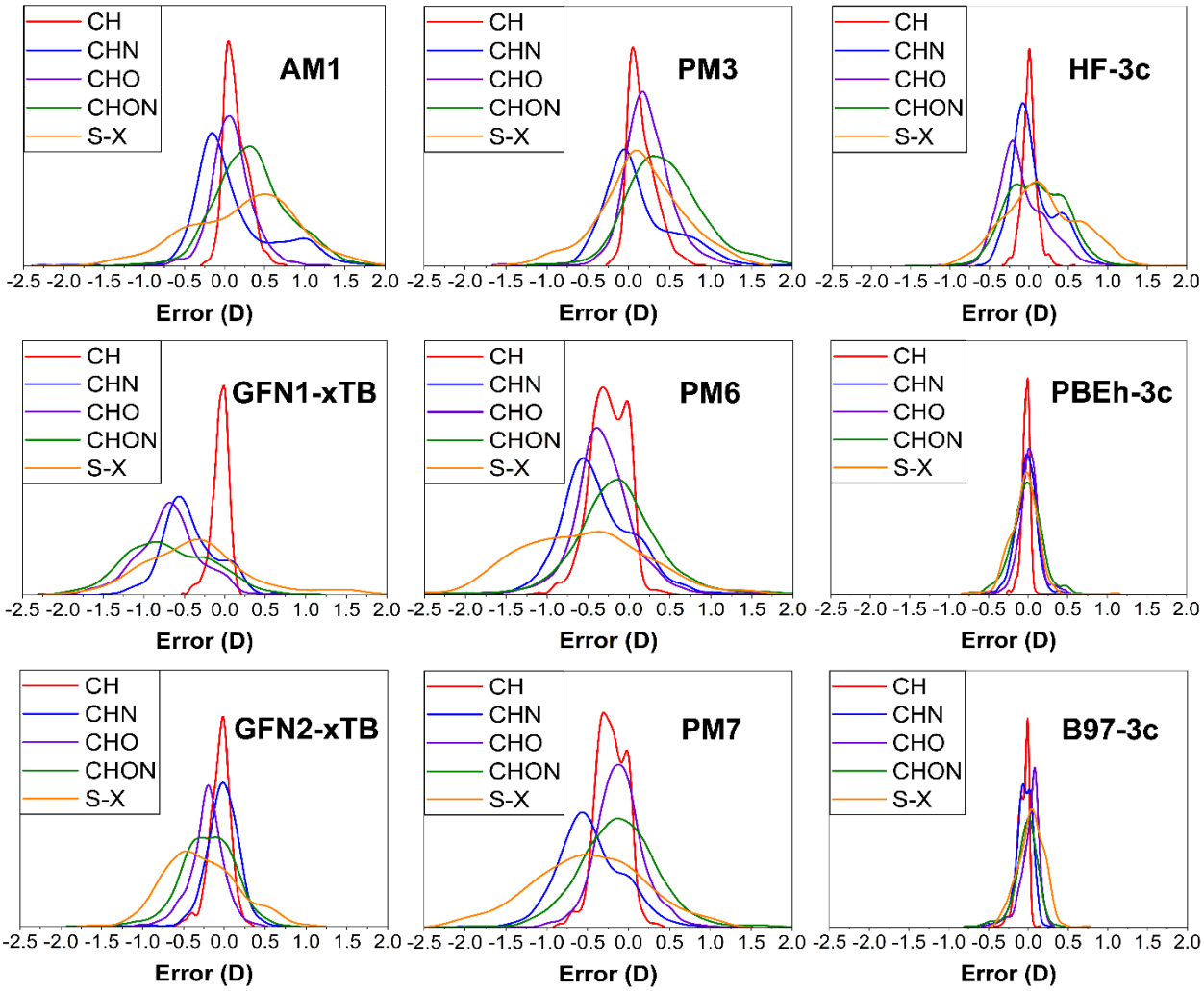


Figure 7. Probability density distribution of subcategory errors for each SE method using SE optimized geometries. The vertical axis of each plot is the probability density of the error. For visualization purposes, categories are not normalized with its population. For plots with normalized subcategory data relative to the total dataset see the SI (Figure S24).

For PM6, the dipole moment for each subcategory was generally more accurate given the lower mean errors for all subcategories except the CH subcategory using the SE optimized geometries (CH: -0.25 D, CHN: -0.38 D, CHO: -0.32 D, CHON: -0.12 D, and S-X: -0.57 D) compared to the DFT optimized geometries (CH: -0.24 D, CHN: -0.49 D, CHO: -0.39 D, CHON: -0.37 D, and S-X: -0.44 D). However, the spread of error for all subcategories except the CH subcategory was

significantly wider using SE optimized geometries (CH: 0.22 D, CHN: 0.50 D, CHO: 0.43 D, CHON: 0.60 D, and S-X: 0.75 D) compared to the much narrower spreads of error obtained using DFT optimized geometries (CH: 0.22 D, CHN: 0.37 D, CHO: 0.28 D, CHON: 0.37 D, and S-X: 0.62 D).

For PM7, we noted improvements in the dipole moment prediction using SE optimized geometries for the CHN, CHO, and CHON subcategories characterized by the lower mean errors of -0.61 D, -0.11 D, and -0.12 D, respectively, compared to -0.61 D, -0.22 D, and -0.33 D, respectively, obtained using DFT optimized geometries. However, the spread of error for these subcategories (CHN: 0.47 D, CHO: 0.35 D, CHON: 0.56 D) increased compared to the spread of error using DFT optimized geometries (CHN: 0.37 D, CHO: 0.24 D, CHON: 0.37 D). For the CH and S-X subcategories, the performance of PM7 dropped marginally shown in mean errors of -0.21 D and -0.49 D, respectively, compared to -0.18 D and -0.43 D using DFT optimized geometries. In addition, the spread of error for the CH (0.18 D) was unchanged while the spread of error for the S-X subcategory (0.67 D) worsened compared to the spread of error obtained using DFT optimized geometries (CH: 0.22 D, S-X: 0.54 D).

Interestingly, GFN1-xTB improved slightly in every metric presented in Figure 2a except the MAD which increased significantly from 2.46 D using DFT optimized geometries to 2.91 D using SE optimized geometries, and the SD which was unchanged. The most significant improvement was in the MAE (0.61 D) compared to 0.70 D for benchmark DFT optimized geometries, while the Mean %error and RMSE only improved by slightly by 4% and 2%, respectively. The improvements in the MAE, Mean %error, and RMSE however resulted in only a marginal improvement in the condensed metric score (0.98) compared to the condensed metric score obtained for the benchmark DFT geometries (1.00). Also, for SE optimized geometries, GFN1-

xTB overpredicted the dipole moment to a lesser degree compared to benchmark DFT optimized geometries shown in the improvement of the mean error from -0.68 D to -0.56 D. Furthermore, based on the probability density distributions for GFN1-xTB in Figure 6b, the range of errors was markedly wider using SE optimized geometries (4.99 D) compared to the range of errors given using benchmark DFT optimized geometries (3.98 D). Additionally, there was an improvement in the dipole moment prediction for every subcategory given the improvement in the mean error using SE optimized geometries (CH: -0.06 D, CHN: -0.44 D, CHO: -0.69 D, CHON: -0.69 D, S-X: -0.34 D) compared to DFT optimized geometries (CH: -0.08 D, CHN: -0.49 D, CHO: -0.86 D, CHON: -0.83 D, S-X: -0.50 D). However, the varied positions of the probability density distributions of the subcategories still led to the appearance of multiple peaks in Figure 6b.

For AM1, there was a dramatic drop in the performance in every metric given in Figure 2a, particularly the Mean %error, RMSE, and MAD which increased by 9%, 9% and 4 D, respectively. The poor performance in every metric resulted in the significant increase in the condensed metric score using SE optimized geometries (1.04) compared to the condensed metric score obtained using the benchmark DFT geometries (0.54). Additionally, AM1 more markedly underpredicted the dipole moment characterized by a mean error of 0.20 D obtained using the SE optimized structures compared to 0.12 D obtained using the benchmark DFT optimized geometries. The spread of error associated with AM1 also increased sharply for SE optimized geometries denoted by the SD (0.49) and FWHM (1.16) compared to the SD (0.36) and FWHM (0.84) obtained for DFT optimized geometries. Moreover, based on the probability density distributions for AM1 in Figure 6b, the range of errors was dramatically wider using SE optimized geometries (8.87 D) compared to the range of errors given by the benchmark DFT optimized geometries (2.63 D). To understand the dramatic drop in the performance for AM1, we manually compared some SE

optimized geometries to benchmark DFT optimized geometries. Based on visual inspection, we attributed the poor performance to differences in the SE and DFT optimized geometries. To illustrate, for molecule 324, 26, 32, and 55 (See Figure 8 and Figures S12, S13, and S14) in the dataset there was significant conformational change involving the rotation of an amine group which resulted in a change of 2.90 D, 1.48 D, 1.45 D, and 0.93 D, respectively. In addition to conformational changes, we also observed less dramatic differences in the SE and DFT optimized structures such as pyramidization of amine groups (see Figure 8) which led to planar structures becoming non-planar as well as changes in bond lengths, bond angles, and dihedral angles. Additionally, the dipole moment for the CH, CHN, and CHON subcategories was underpredicted to greater extent for SE optimized geometries reflected in the mean errors (CH: 0.14 D, CHN: 0.10 D, CHO: 0.08 D, CHON: 0.36 D, S-X: 0.22 D) compared to mean errors obtained using DFT optimized geometries (CH: 0.13 D, CHN: 0.04 D, CHO: 0.08 D, CHON: 0.19 D, and S-X: 0.33 D). For the S-X category, we noted an improvement in the dipole moment prediction characterized by the lower mean of 0.22 D compared to 0.33 D obtained using DFT optimized geometries, while the mean error for the CHO subcategory (0.08 D) was unchanged.

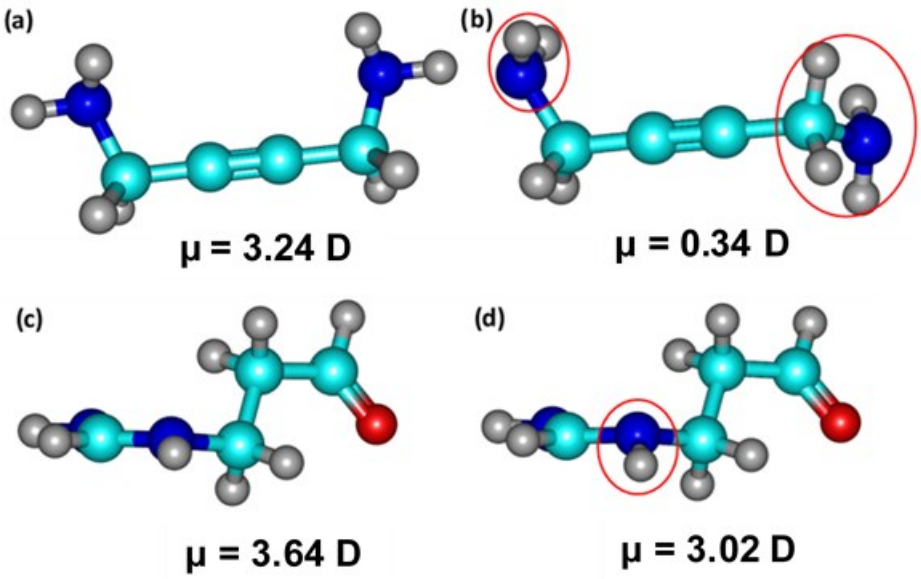


Figure 8. Molecule 324 (a and c) and 117 (b and d) of the QM7b dataset. (a and c) DFT optimized geometry and (b and d) AM1 optimized geometry. Red circle indicates location of conformation change after geometry optimization using AM1. Color code: C - cyan, H - gray, N – blue, O - red.

Similar to AM1, PM3 performed worse in every metric given in Figure 2a, particularly the Mean %error, RMSE, and MAD which increased by 10%, 8%, and 3.8 D, respectively. Due to the poor performance in each metric, the condensed metric score also rose to 1.00 using SE optimized geometries compared to 0.47 that was obtained using DFT optimized geometries. Additionally, for SE optimized geometries, PM3 underpredicted the dipole moment considerably characterized by a mean error of 0.28 D obtained for the SE optimized compared to 0.16 D for the benchmark DFT optimized geometries. The spread of error given for PM3 also increased sharply using SE optimized geometries denoted by the SD (0.47 D) and FWHM (1.11 D) compared to the SD (0.29 D) and FWHM (0.69 D) obtained using DFT optimized geometries. Moreover, based on the probability density distributions for PM3 in Figure 6b, the range of errors was dramatically wider using SE optimized geometries (8.38 D) compared to the range of errors given for benchmark DFT optimized geometries (2.95 D). Much like for AM1, we attributed the dramatic drop in the

performance of PM3 to structural differences between SE and DFT optimized geometries. For example, for molecules 26 and 324 (See Figure 9 and Figure S22) in the dataset there was conformational change involving amine groups, for molecule 12 there was a rotation of a methoxy group (See Figure S23), and for molecule 250 we observed pyramidization of the amine group (Figure 9) which resulted in changes in dipole moment of 2.01 D, 1.55 D, 1.33 D, and 1.45 D, respectively.

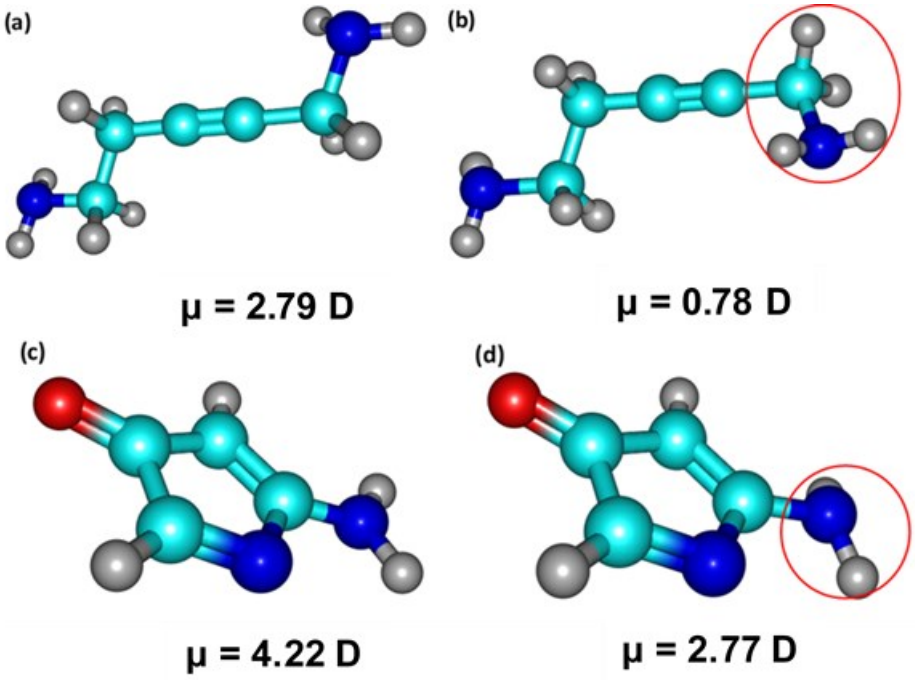


Figure 9. Molecule 26 (a and c) and 250 (b and d) of the QM7b dataset. (a and c) DFT optimized geometry and (b and d) PM3 optimized geometry. Red circle indicates location of conformation change after geometry optimization using PM3. Color code: C - cyan, H - gray, N – blue, O - red.

For PM3, the dipole moment in the CHN, CHO, and CHON subcategories were underpredicted to a greater degree given the mean errors of each subcategory SE optimized geometries (CHN: 0.14 D, CHO: 0.22 D, CHON: 0.46 D) compared to DFT optimized geometries (CHN: 0.04 D, CHO: 0.19 D, CHON: 0.23 D). For the CH subcategory, the mean error was unchanged while for the S-X subcategory the mean error (0.15 D) was slightly higher compared to the mean error of 0.14 D

given using DFT optimized geometries. In addition, the spread of error for the CHN, CHO, CHON, and S-X subcategories (CH SD: 0.17 D, CHN SD: 0.46, CHO SD: 0.36 D, CHON SD: 0.53 D, S-X SD: 0.51 D) also worsened compared to the spread of error shown using DFT optimized geometries (CH SD: 0.17 D, CHN SD: 0.34, CHO SD: 0.19 D, CHON SD: 0.28 D, S-X SD: 0.45 D).

The performance of GFN2-xTB based on metrics in Figure 6a remained relatively unchanged except for the MAD which increased for SE optimized geometries (2.04 D) compared to 1.30 D given by GFN2-xTB for the benchmark DFT optimized geometries. The large change in the MAD was caused due to conformational change in molecule 1256 which involved the rotation of a $=CH_2$ group (See Figure 10) and resulted in a 1.78 D change in the dipole moment.

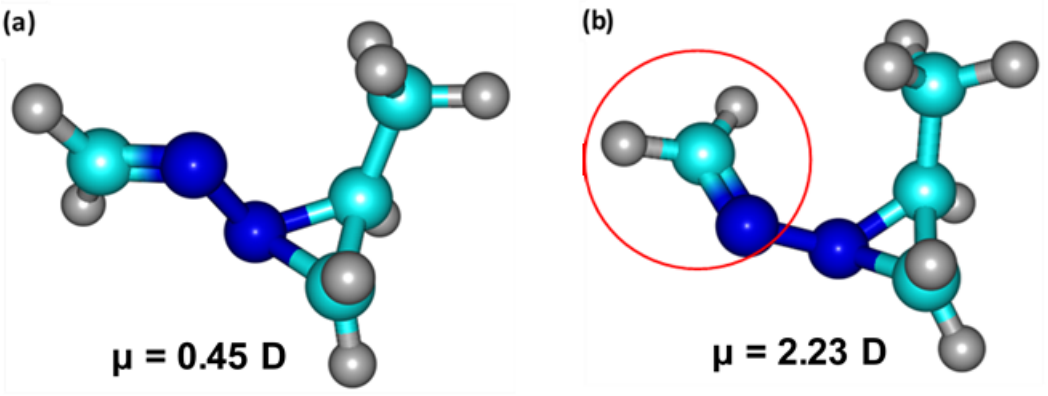


Figure 10. Molecule 1256 of the QM7b dataset. (a) DFT optimized geometry and (b) GFN2-xTB optimized geometry. Red circle indicates location of conformation change after geometry optimization using GFN2-xTB. Color code: C - cyan, H - gray, N – blue, O - red.

We note that the absence of large changes in the error metrics for GFN2-xTB reflected in the relatively small change to the condensed metric scores for benchmark DFT (0.42) and SE optimized geometries (0.54) indicated that GFN2-xTB can provide good dipole moment predictions regardless of the source of the input geometry. In addition, based on the mean error

GFN2-xTB also overpredicted the dipole moment using SE optimized geometries to a lesser degree with a mean error of -0.15 D than using the benchmark DFT optimized geometries which had a mean error of -0.23 D. However, GFN2-xTB gave a wider spread of error using the SE optimized geometries reflected in the SD (0.30) and FWHM (0.72) compared to spread of error observed using benchmark DFT optimized geometries which had a SD of 0.24 D and FWHM of 0.56. Moreover, based on the probability density distributions for GFN2-xTB in Figure 3, the range of errors was wider for SE optimized geometries (3.87 D) compared to the range of errors given for benchmark DFT optimized geometries (1.96 D) which we attributed to structural differences between SE and DFT optimized geometries which led to larger errors. For the subcategories, GFN2-xTB gave more accurate dipole moment prediction for the CHN, CHO, CHON, and S-X subcategories characterized by the smaller mean errors using SE optimized geometries (CHN: -0.03 D, CHO: -0.22 D, CHON: -0.20 D, S-X: -0.28 D) compared to using DFT optimized geometries (CHN: -0.11 D, CHO: -0.31 D, CHON: -0.29 D, S-X: -0.34 D). However, GFN2-xTB provided slightly worse predictions for the S-X subcategory given a mean error of -0.28 D compared to -0.24 D obtained using DFT optimized geometries while the performance in the CH subcategory (-0.06 D) was unchanged. Despite the improvements in the mean error, the spread of error increased for all subcategories except for the CH subcategories which was reflected in the larger SD using SE optimized geometries (CH: 0.13 D, CHN: 0.25 D, CHO: 0.23 D, CHON: 0.35 D, S-X: 0.45 D) compared to the narrower spread of error using DFT optimized geometries (CH: 0.14 D, CHN: 0.16 D, CHO: 0.20 D, CHON: 0.25 D, S-X: 0.35 D).

For HF-3c the only changes in performance across the metrics presented in Figure 1 were the Mean %error which increased by 4% and the MAD which increased by 1.1 D, however because of these the condensed metric score for HF-3c rose to 0.60 from the 0.46 obtained using DFT

optimized geometries. Additionally, HF-3c went from overpredicting the dipole moment with a mean error of -0.06 D for DFT optimized geometries to slightly underpredicting the dipole moment for SE optimized geometries reflected in a mean error of 0.03 D.

In Figure 6b, we observed the range of error for HF-3c increased to 4.26 D from the 2.10 D obtained for DFT optimized geometries. We attributed the dramatic drop in the performance of HF-3c to structural differences between SE and DFT optimized geometries. However, upon manual inspection unlike AM1 and PM3 there were fewer instances where the difference in the HF-3c dipole moment using DFT optimized geometries and SE optimized geometries exceeded 1 D. For example, for molecules 26 and 324 (See Figure 11) in the dataset there was conformational change involving amine groups which resulted in changes in dipole moment of 2.54 D, and 1.71 D, respectively.

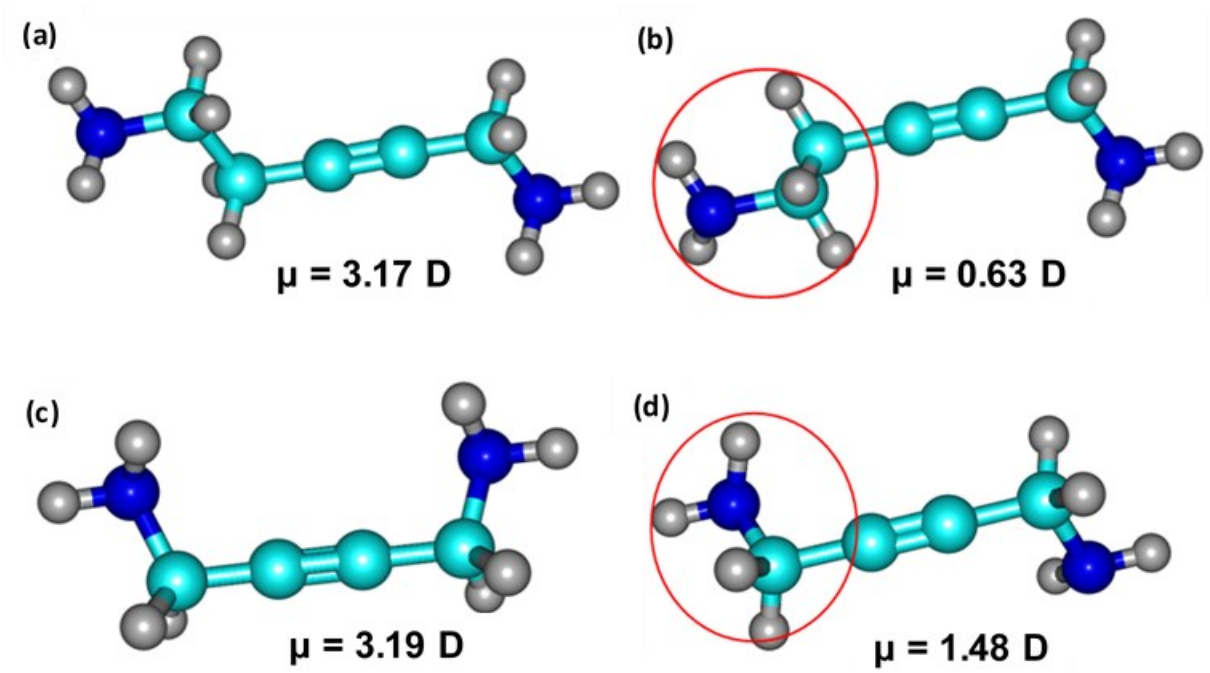


Figure 11. Molecule 26 (a and c) and 324 (b and d) of the QM7b dataset. (a and c) DFT optimized geometry and (b and d) HF-3c optimized geometry. Red circle indicates location of conformation

change after geometry optimization using HF-3c. Color code: C - cyan, H - gray, N – blue, O - red.

In terms of the subcategories, the only notable changes in performance of HF-3c were observed in the CHO and S-X subcategories. For the CHO subcategory, HF-3c still overpredicted the dipole moment but to a much smaller degree characterized by the much smaller mean error of -0.08 D compared to the mean error of -0.15 D given using DFT optimized geometries. In contrast to the improvement in mean error, the spread of error for the CHO subcategories was also markedly narrower using SE optimized geometries (CHO SD: 0.31 D) compared to the spread of error given using DFT optimized geometries (CHO SD: 0.24 D). While for the S-X subcategory, HF-3c more significantly underpredicted the dipole moment given by the larger mean error of 0.17 D compared to 0.05 D provided using DFT optimized geometries.

For B97-3c the only notable drop in the performance based on metrics in Figure 6a was in the MAD which went from 0.95 D for DFT optimized geometries to 2.10 D for SE optimized geometries, however this increase in the MAD still resulted in an increase of the condensed metric score for B97-3c which rose from 0.23 to 0.34. Furthermore, there was no significant change in the spread of error for B97-3c based on the SD of 0.13 D using DFT optimized geometries and 0.15 D using SE optimized geometries. Additionally, there was no considerable shift in the mean error of its probability density distribution in Figure 6b, however the range increased to 3.13 D compared to 1.46 D range for DFT optimized geometries which was attributed to differences in the SE optimized and DFT optimized geometries. However, for B97-3c, there were only four instances where the difference between B97-3c dipole moment using DFT optimized geometries and SE optimized geometries was greater than 1 D. For the subcategories, there were no notable changes in the performance of B97-3c which further indicates the reliability of this method.

For PBEh-3c, the only change in performance across the metrics presented in Figure 6a was the Mean %error which increased by approximately 7% and did not affect the condensed metric score significantly given a score of 0.21 compared to 0.18 using DFT optimized geometries. However, the spread of error PBEh-3c was larger for SE optimized geometries reflected in the FWHM of 0.41 D compared to 0.28 D obtained for DFT optimized geometries. Similarly, there was a larger range of error in the probability density distribution for SE optimized geometries (3.11 D) compared to DFT optimized geometries (0.99 D) shown in Figure 6b which was attributed to the overall increase in error as a result of structural changes between DFT and SE optimized geometries. However, the position of the mean improved marginally from -0.08 D for DFT optimized geometries to -0.02 D for SE optimized geometries. For the subcategories, there were no notable changes in the performance of PBEh-3c which further indicates the reliability of this method.

3.4. COMPUTATIONAL COST ANALYSIS

For studies considering a large number of target molecules, computational cost is also a major consideration before choosing a method. Therefore, to give an estimation of the timings of each method, we randomly chose 100 molecules from the dataset and ran single point calculations under comparable circumstances (1 core, 500 MB of memory on the same compute node) for each SE method using the DFT optimized geometries of the chosen molecules obtained from the QM7b database. To quantify the relative speeds of the SE methods, we averaged the timings for each method. In Figure 12, we compare the performance of all ten SE methods evaluated in this work in terms of approximate timings and error.

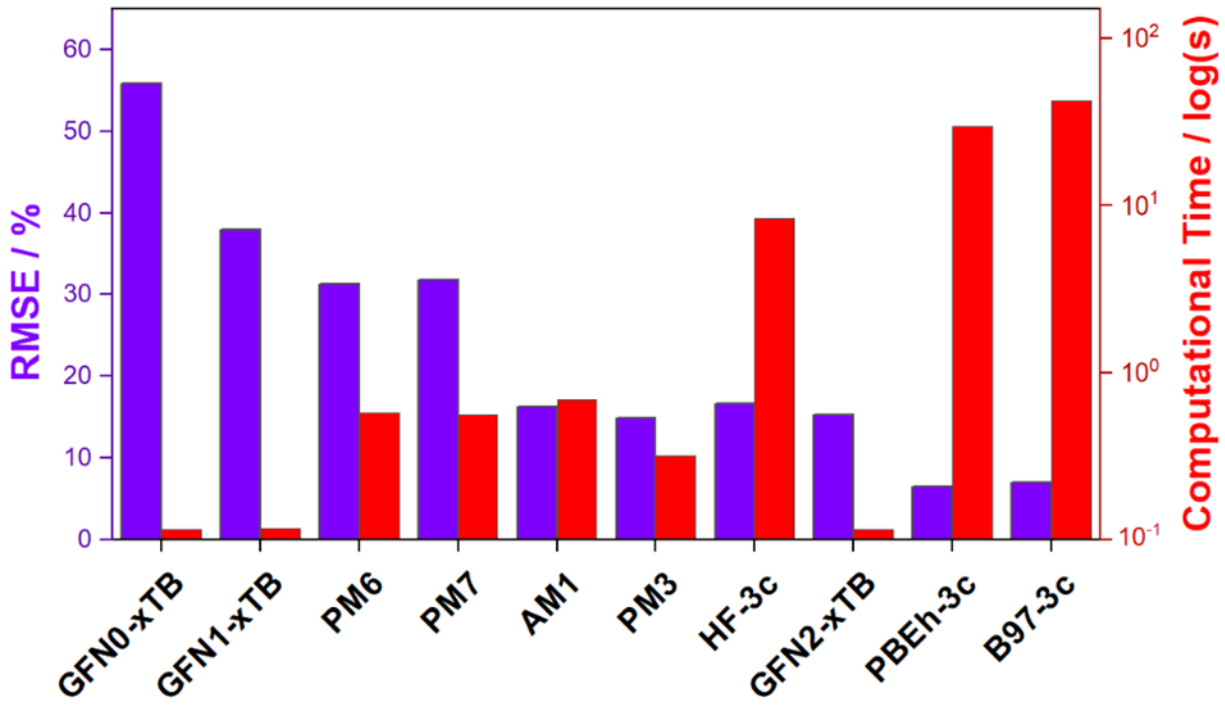


Figure 12. Performance of all studied SE methods in terms of the RMSE% error and computational cost. Purple bars represent the RMSE% error while red bars represent the computational time on the log scale. Computational times are estimated based on single point calculations using a single core. Results were obtained by averaging over 100 randomly chosen molecules from the QM7b database.

It clear from Figure 12 that the 3c methods were much more time-consuming compared to the other SE methods given their sounder theoretical form. The NNDO methods form the next group of SE methods with respect to computational time with AM1, PM3, PM6, and PM7 being two orders of magnitude faster than the best performing method PBEh-3c. The GFNn-xTB methods were the fastest methods among SE methods studied here given that they were three orders of magnitude faster than the best performing method PBEh-3c. Thus, given the good performance of GFN2-xTB compared to PBEh-3c, GFN2-xTB represents an acceptable compromise between accuracy and computational cost for the calculation of the dipole moment of organic molecules

except for sulfur containing compounds for which only B97-3c and PBEh-3c provided accurate results.

4. CONCLUSIONS

In this work, we have evaluated the performance of a set of nine SE methods namely AM1, GFN0-xTB, GFN1-xTB, GFN2-xTB, PM3, PM6, PM7, B97-3c, HF-3c, and PBEh-3c, in predicting dipole moments against high-level couple cluster dipole moment data of 7211 organic molecules contained in the QM7b dataset[41-43]. To understand how the composition of molecules in the QM7b dataset affects the performance of the SE methods, we defined subgroups CH, CHN, CHO, CHON, and S-X (i.e. sulfur containing molecules such as CHNS, CHNOS, CHOS, CHS, CHNSCl, and CHSCl) and calculated dipole moments of molecules in each subcategory against reference dipole moment data. To give insights into the ability of SE methods to predict the direction of the dipole moment, we also compared the direction predicted by each SE method against reference data. So as to provide further insights into the capability of SE methods to calculate dipole moment, we also evaluated the performance of the SE methods upon using SE geometries of the molecules instead of the DFT optimized benchmark geometry of the QM7b dataset. Finally, we give an assessment of the relative computational costs of each SE method to help the users in deciding on a method considering the computational costs as well as accuracy of each method.

Our results showed that B97-3c and PBEh-3c provided the most accurate dipole moment predictions using the benchmark DFT optimized geometries contained in the QM7b dataset with mean absolute errors of 0.10 and 0.11 D, respectively. Meanwhile, the performance of HF-3c was comparable to the performance of AM1, PM3, and GFN2-xTB with mean absolute errors of 0.25, 0.27, 0.25, and 0.27 D, respectively. Finally, PM6, PM7, and GFN1-xTB gave the weakest performances with mean absolute errors of 0.46, 0.43, and 0.70 D, respectively. For all methods,

the most consistent performance was observed for molecules composed of C and H only (i.e. hydrocarbons), for which there was low spread of error within 1 D for all methods. For the CHN, CHO, and CHON subcategories, B97-3c and PBEh-3c were the best performing methods again with ranges of error of 1 D. GFN2-xTB gave comparably performance to HF-3c in the CHN, CHO, and CHON subcategories, showing similar range of error of 1.50 (GFN2-xTB) and 1.82 D (HF-3c) for the CHN, 1.61 (GFN2-xTB) and 1.55 D (HF-3c) for the CHO, and 1.90 (GFN2-xTB) and 2.10 D (HF-3c) for the CHON subcategories, respectively. For the S-X subcategory, again B97-3c and PBEh-3c provided the best performance given their narrow ranges of error of 0.95 and 0.77 D, respectively, while GFN2-xTB followed closely with a range of 1.69 D. Results using semiempirical optimized geometries showed that the performance of all semiempirical methods dropped especially in the performance of AM1 and PM3 for which the range of error rose to 8.87 D and 8.38 D using semiempirical geometries from 2.63 D and 2.95 D, respectively, using DFT optimized geometries making AM1 and PM3 impractical for dipole moment calculations.

Based on our results, PBEh-3c was the best performing semiempirical method in absolute terms (MAE: 0.11 D, MAD: 0.67 D), which is comparable to DFT at the B3LYP/d-aug-cc-pVDZ level of theory (MAE: 0.09 D, MAD: 0.88 D)[51]. However, the computational cost of PBEh-3c, which is similar to DFT using small basis set sets, might make it impractical for large datasets such as QM7b. Our results showed however that there is an optimal balance between accuracy and computational cost by using GFN2-xTB (MAE: 0.25 D, MAD: 1.30 D) which is three of magnitude faster than PBEh-3c. Thus, we suggest GFN2-xTB as a method of choice among semiempirical methods because of its optimal cost/performance ratio in calculating the dipole moment of organic molecules except in the S-X subcategory where PBEh-3c is more preferred.

ASSOCIATED CONTENT

The data underlying this paper are available in the Supporting Information. The Supporting Information is available free of charge at SI link

Dipole moment data for each SE method for the full dataset, and subcategories using benchmark DFT optimized, and SE optimized geometries (XLSX)

Summary table of errors associated with each SE method using benchmark DFT optimized, and SE optimized geometries, parity plots showing correlation to benchmark data for each SE method using benchmark DFT geometries, and SE optimized geometries, supplementary discussion on the gaussian normal distributions showing distributions of errors for each SE method, supplementary discussion on performance of GFN0-xTB in dipole moment prediction, probability density distribution plot of subcategory errors for each SE method normalized by population using DFT optimized geometries, supplementary discussion on the performance of SE methods in the subcategories using SE optimized geometries, summary tables of errors in each subcategory associated with each SE method using SE optimized geometries, supplementary figures showing conformational change for a set of molecules using AM1, and PM3, and probability density distribution plot of subcategory errors for each SE method normalized by population using SE optimized geometries (PDF)

AUTHOR INFORMATION

Corresponding Author

*Tibor Szilvási – Department of Chemical and Biological Engineering, University of Alabama, Tuscaloosa, Alabama 35487, United States; Email: tszilvasi@ua.edu

Author

Ademola Soyemi - Department of Chemical and Biological Engineering, University of Alabama, Tuscaloosa, Alabama 35487, United States

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGEMENTS

A.S and T.S would like to acknowledge the financial support of the National Science Foundation (NSF) under grant number EFMA-2029387. Any opinions, findings, conclusions, and/or recommendations expressed in this material are those of the authors(s) and do not necessarily reflect the views of the NSF. A.S and T.S would also like to thank the University of Alabama and the Office of Information Technology for providing high-performance computing resources and support that has contributed to these research results. This work was also made possible in part by a grant of high-performance computing resources and technical support from the Alabama Supercomputer Authority.

REFERENCES

1. Moldoveanu, S. C. D., V. , Intermolecular Interactions. In *Essentials in Modern HPLC Separations*, Elsevier, **2013**.
2. Leal-Duaso, A.; Pérez, P.; Mayoral, J. A.; Pires, E.; García, J. I., Glycerol as a source of designer solvents: physicochemical properties of low melting mixtures containing glycerol ethers and ammonium salts. *Phys. Chem. Chem. Phys.* **2017,** *19*, 28302-28312.
3. Leal-Duaso, A.; Pérez, P.; Mayoral, J. A.; García, J. I.; Pires, E., Glycerol-Derived Solvents: Synthesis and Properties of Symmetric Glyceryl Diethers. *ACS Sustainable Chem. Eng.* **2019,** *7*, 13004-13014.
4. Qian, S.; Liu, X.; Emel'yanenko, V. N.; Sikorski, P.; Kammakakam, I.; Flowers, B. S.; Jones, T. A.; Turner, C. H.; Verevkin, S. P.; Bara, J. E., Synthesis and Properties of 1,2,3-

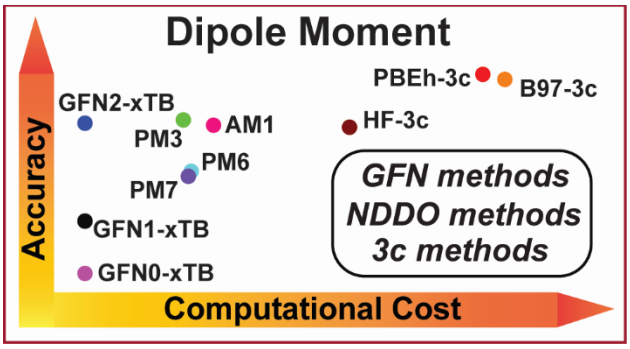
Triethoxypropane: A Glycerol-Derived Green Solvent Candidate. *Ind. Eng. Chem. Res.* **2020,** *59*, 20190-20200.
5. Flowers, B. S.; Mittenthal, M. S.; Jenkins, A. H.; Wallace, D. A.; Whitley, J. W.; Dennis, G. P.; Wang, M.; Turner, C. H.; Emel'yanenko, V. N.; Verevkin, S. P.; Bara, J. E., 1,2,3-Trimethoxypropane: A Glycerol-Derived Physical Solvent for CO2 Absorption. *ACS Sustainable Chem. Eng.* **2017,** *5*, 911-921.
6. Jessop, P. G., Searching for green solvents. *Green Chem.* **2011,** *13*, 1391-1398.
7. Grotjahn, R.; Lauter, G. J.; Haasler, M.; Kaupp, M., Evaluation of Local Hybrid Functionals for Electric Properties: Dipole Moments and Static and Dynamic Polarizabilities. *J. Phys. Chem. A* **2020,** *124*, 8346-8358.
8. Karne, A. S.; Vaval, N.; Pal, S.; Vásquez-Pérez, J. M.; Köster, A. M.; Calaminici, P., Systematic comparison of DFT and CCSD dipole moments, polarizabilities and hyperpolarizabilities. *Chem. Phys. Lett.* **2015,** *635*, 168-173.
9. Verma, P.; Truhlar, D. G., Can Kohn–Sham density functional theory predict accurate charge distributions for both single-reference and multi-reference molecules? *Phys. Chem. Chem. Phys.* **2017,** *19*, 12898-12912.
10. Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M., A fifth-order perturbation comparison of electron correlation theories. *Chem. Phys. Lett.* **1989,** *157*, 479-483.
11. Chai, J.-D.; Head-Gordon, M., Long-range corrected double-hybrid density functionals. *J. Chem. Phys.* **2009,** *131*, 174105.
12. Mardirossian, N.; Head-Gordon, M., ωB97X-V: A 10-parameter, range-separated hybrid, generalized gradient approximation density functional with nonlocal correlation, designed by a survival-of-the-fittest strategy. *Phys. Chem. Chem. Phys.* **2014,** *16*, 9904-9924.
13. Hait, D.; Head-Gordon, M., How Accurate Is Density Functional Theory at Predicting Dipole Moments? An Assessment Using a New Database of 200 Benchmark Values. *J. Chem. Theory Comp.* **2018,** *14*, 1969-1981.
14. Zapata, J. C.; McKemmish, L. K., Computation of Dipole Moments: A Recommendation on the Choice of the Basis Set and the Level of Theory. *J. Phys. Chem. A* **2020,** *124*, 7538-7548.
15. Jensen, F., Polarization consistent basis sets: Principles. *J. Chem. Phys.* **2001,** *115*, 9113-9125.
16. Jensen, F., Polarization consistent basis sets. II. Estimating the Kohn-Sham basis set limit. *J. Chem. Phys.* **2002,** *116*, 7372-7379.
17. Jensen, F., Polarization consistent basis sets. III. The importance of diffuse functions. *J. Chem. Phys.* **2002,** *117*, 9234-9240.
18. Hehre, W. J. D., R.;Pople, J. A., Self-Consistent Molecular Orbital Methods. XII. Further Extensions of Gaussian-Type Basis Sets for Use in Molecular Orbital Studies of Organic Molecules. *J. Chem. Phys.* **1972,** *56*, 2257-2261.
19. Sadlej, A. J., Medium-size polarized basis sets for high-level correlated calculations of molecular electric properties. *Collect. Czech. Chem. Commun.* **1988,** *53*, 1995-2016.
20. Krishnan, R. B., J. S.;Seeger, R.;Pople, J. A., Self-consistent molecular orbital methods. XX. A basis set for correlated wave functions. *J. Chem. Phys.* **1980,** *72*, 650-654.
21. Clark, T. C., Jayaraman;Spitznagel, Günther W.;Schleyer, Paul Von Ragué, Efficient diffuse function-augmented basis sets for anion calculations. III. The 3-21+G basis set for first-row elements, Li-F. *J. Comput. Chem.* **1983,** *4*, 294-301.
22. Kendall, R. A., Dunning Jr., T.H. and Harrison, R.J., Electron affinities of the first-row atoms revisited. Systematic basis sets and wave functions. *J. Chem. Phys.* **1992,** *96*, 6796-6806.

23. Hickey, A. L.; Rowley, C. N., Benchmarking Quantum Chemical Methods for the Calculation of Molecular Dipole Moments and Polarizabilities. *J. Phys. Chem. A* **2014,** *118*, 3678-3687.
24. Lee, C.; Yang, W.; Parr, R. G., Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **1988,** *37*, 785-789.
25. Vosko, S. H.; Wilk, L.; Nusair, M., Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis. *Can. J. Phys.* **1980,** *58*, 1200-1211.
26. Becke, A. D., Density-functional thermochemistry. III. The role of exact exchange. *J . Chem. Phys.* **1993,** *98*, 5648-5652.
27. Adamo, C.; Barone, V., Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phy.* **1999,** *110*, 6158-6170.
28. Dunning, T. H., Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *J. Chem. Phys.* **1989,** *90*, 1007-1023.
29. Thiel, W., Semiempirical quantum–chemical methods. *WIREs Comput. Mol. Sci.* **2014,** *4*, 145-157.
30. Christensen, A. S.; Kubař, T.; Cui, Q.; Elstner, M., Semiempirical Quantum Mechanical Methods for Noncovalent Interactions for Chemical and Biochemical Applications. *Chem. Rev.* **2016,** *116*, 5301-5337.
31. 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. *J. Am. Chem. Soc.* **1985,** *107*, 3902-3909.
32. Stewart, J. J. P., Optimization of parameters for semiempirical methods I. Method. *J. Comput. Chem.* **1989,** *10* (2), 209-220.
33. Stewart, J. J. P., Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements. *J. Mol. Model.* **2007,** *13*, 1173-1213.
34. Stewart, J. J. P., Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters. *J. Mol. Model.* **2013,** *19*, 1-32.
35. Pankratov, A. N., Semiempirical quantum chemical methods: testing of thermodynamic and molecular properties of cyclic non-aromatic hydrocarbons and unsaturated heterocycles. *J. Mol. Struct.: THEOCHEM* **1998,** *453*, 7-15.
36. Kurunczi, L.; Ilia, G., The MNDO, AM1 and PM3 semiempirical methods in dipole moment prediction for alkyl phosphonic acid dialkylesters. *Ann. West Univ. Timis., Ser. Chem.* **2003,** *12*, 167-174.
37. Pankratov, A. N.; Shchavlev, A. E., Semiempirical quantum chemical PM3 computations and evaluations of redox potentials, basicities, and dipole moments of the diphenylamine series as analytical reagents. *Can. J. Chem.* **1999,** *77*, 2053-2058.
38. Pankratov, A.; Shchavlev, A. E., Semiempirical quantum chemical methods: testing of physicochemical properties of acyclic and aromatic compounds. *J. Mol. Struct.: THEOCHEM* **1997,** *392*, 137-140.
39. Yatsenko, A.; Paseshnichenko, K., On the performance of semiempirical MO theory for dipole moments of dye molecules. *J. Mol. Model.* **2001,** *7*, 384-391.
40. Anisimov, V. M.; Anikin, N.; Bugaenko, V.; Bobrikov, V.; Andreyev, A., Accuracy assessment of semiempirical molecular electrostatic potential of proteins. *Theor. Chem. Acc.* **2003,** *109*, 213-219.

41. 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*, 8732-8733.
42. 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.
43. Montavon, G.; Rupp, M.; Gobre, V.; Vazquez-Mayagoitia, A.; Hansen, K.; Tkatchenko, A.; Müller, K.-R.; Anatole von Lilienfeld, O., Machine learning of molecular electronic properties in chemical compound space. *New J. Phys.* **2013,** *15*, 095003.
44. Sure, R.; Grimme, S., Corrected small basis set Hartree-Fock method for large systems. *J. Comput. Chem.* **2013,** *34* (19), 1672-1685.
45. Grimme, S.; Brandenburg, J. G.; Bannwarth, C.; Hansen, A., Consistent structures and interactions by density functional theory with small atomic orbital basis sets. *J. Chem. Phys.* **2015,** *143*, 054107.
46. Brandenburg, J. G.; Bannwarth, C.; Hansen, A.; Grimme, S., B97-3c: A revised low-cost variant of the B97-D density functional method. *J. Chem. Phys.* **2018,** *148* (6), 064104.
47. Bannwarth, C.; Caldeweyher, E.; Ehlert, S.; Hansen, A.; Pracht, P.; Seibert, J.; Spicher, S.; Grimme, S., Extended tight-binding quantum chemistry methods. *WIREs Comput. Mol. Sci.* **2021,** *11*, e1493.
48. Caldeweyher, E.; Brandenburg, J. G., Simplified DFT methods for consistent structures and energies of large systems. *J. Phys.: Condensed Matter* **2018,** *30*, 213001.
49. 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.
50. 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. *J. Chem. Theory Comput.* **2019,** *15*, 1652-1671.
51. Yang, Y.; Lao, K. U.; Wilkins, D. M.; Grisafi, A.; Ceriotti, M.; DiStasio, R. A., Quantum mechanical static dipole polarizabilities in the QM7b and AlphaML showcase databases. *Sci. Data* **2019,** *6* (1), 152.
52. Pracht, P. C., E.; Ehlert, S.; Grimme, S., A Robust Non-Self-Consistent Tight-Binding Quantum Chemistry Method for Large Molecules. *ChemRxiv,* June 27, **2019,** ver. 1. DOI: 10.26434/chemrxiv.8326202.v1 (accessed 2022-01-11)
53. Roothaan, C. C. J., New Developments in Molecular Orbital Theory. *Rev. Mod. Phys.* **1951,** *23*, 69-89.
54. Perdew, J. P.; Burke, K.; Ernzerhof, M., Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996,** *77*, 3865-3868.
55. Grimme, S., Semiempirical GGA-type density functional constructed with a long-range dispersion correction. *J. Comput. Chem.* **2006,** *27*, 1787-1799.
56. Becke, A. D., Density-functional thermochemistry. V. Systematic optimization of exchange-correlation functionals. *J. Chem. Phys.* **1997,** *107*, 8554-8560.
57. Woon, D. E.; Dunning, T. H., Gaussian basis sets for use in correlated molecular calculations. IV. Calculation of static electrical response properties. *J. Chem. Phys.* **1994,** *100*, 2975-2988.

58. O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R., Open Babel: An open chemical toolbox. *J. Cheminformatics* **2011,** *3* (1), 33.
59. *The Open Babel Package*, 2.3.1; 2011.
60. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H., et al. *Gaussian 16 Rev. C.01*, Wallingford, CT, 2016.
61. *Semiempirical extended tight-binding program xtb*, 6.4.
62. Neese, F., The ORCA program system. *WIREs Comput. Mol. Sci.* **2012,** *2*, 73-78.
63. Neese, F., Software update: the ORCA program system, version 4.0. *WIREs Comput. Mol. Sci.* **2018,** *8*, e1327.

TOC GRAPHIC



Supporting Information

Benchmarking Semiempirical QM Methods for Calculating the Dipole Moment of Organic Molecules

Ademola Soyemi and Tibor Szilvási[*]

[*]Corresponding author. Email: tibor.szilvasi@ua.edu

Department of Chemical and Biological Engineering, The University of Alabama, Tuscaloosa, AL 35487, United States

Table of Contents

1. **Summary of full dataset errors for each SE method**

Table S1. Errors associated with each SE method using benchmark DFT optimized geometries.

SE method	MAE (D)	Mean error (D)	Mean %error (%)	Max abs error (D)	SD (D)	RMSE (%)	FWHM (D)	Range (D)
AM1	0.27	0.12	15.75	1.71	0.36	16.35	0.84	2.63
GFN0-xTB	0.88	-0.75	48.57	8.60	1.02	55.89	2.40	11.17
GFN1-xTB	0.70	-0.68	36.83	2.46	0.49	37.91	1.14	3.98
GFN2-xTB	0.27	-0.23	16.16	1.30	0.24	15.34	0.56	1.96
PM3	0.25	0.16	14.55	1.49	0.29	14.98	0.69	2.95
PM6	0.46	-0.40	32.19	2.07	0.36	31.25	0.85	3.34
PM7	0.43	-0.37	30.60	2.14	0.37	31.78	0.88	3.40
B97-3c	0.10	-0.04	6.10	0.95	0.13	6.49	0.32	1.46
HF-3c	0.25	-0.06	15.95	1.12	0.29	16.63	0.69	2.10
PBEh-3c	0.11	-0.08	2.17	0.67	0.12	7.02	0.28	0.99

Table S2. Errors associated with each SE method using SE optimized geometries.

SE method	MAE (D)	Mean error (D)	Mean %error (%)	Max abs error (D)	SD (D)	RMSE (%)	FWHM (D)	Range (D)
AM1	0.37	0.20	25.42	5.79	0.49	24.49	1.16	8.87
GFN0-xTB	1.13	-0.95	60.91	15.20	1.51	78.03	3.56	18.69
GFN1-xTB	0.61	-0.56	33.24	2.91	0.49	35.68	1.15	4.99
GFN2-xTB	0.25	-0.15	16.32	2.04	0.30	17.42	0.72	3.87
PM3	0.39	0.28	25.90	5.26	0.47	23.28	1.11	8.38
PM6	0.45	-0.27	39.89	6.11	0.63	37.24	1.49	9.51
PM7	0.41	-0.23	34.99	5.97	0.50	34.52	1.17	8.94
B97-3c	0.11	-0.03	7.31	2.10	0.15	7.51	0.36	3.13
HF-3c	0.27	0.03	20.01	2.20	0.35	17.97	0.82	4.26
PBEh-3c	0.12	-0.02	8.55	1.62	0.17	9.57	0.41	3.11

Table S3. Performance of each SE method in predicting the dipole moment direction using DFT optimized geometries.

SE method	Average Deviation (D)
AM1	13.6
GFN0-xTB	17.3
GFN1-xTB	9.3
GFN2-xTB	6.6
PM3	8.3
PM6	10.4
PM7	10.8
B97-3c	3.6
HF-3c	7.8
PBEh-3c	4.5

2. Parity plots showing predictive power of each SE method

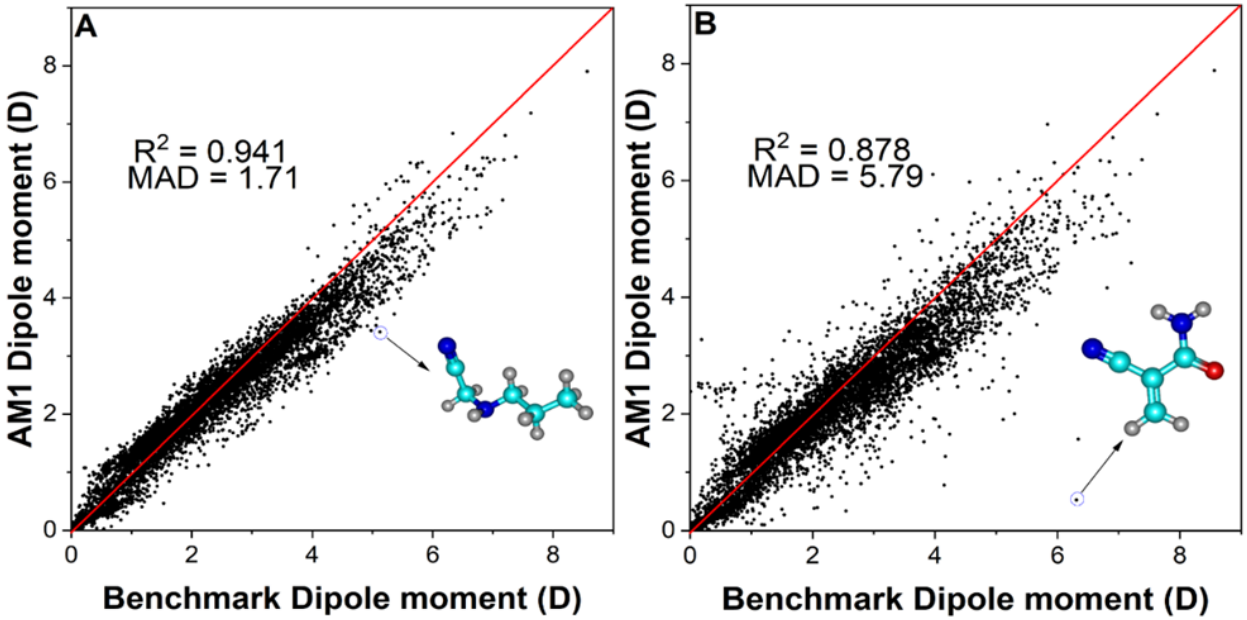


Figure S1. Parity plot comparing dipole moment predicted by AM1 using (a) benchmark DFT optimized geometries and (b) SE optimized geometries. The red diagonal line is to help the reader see the ideal correlation.

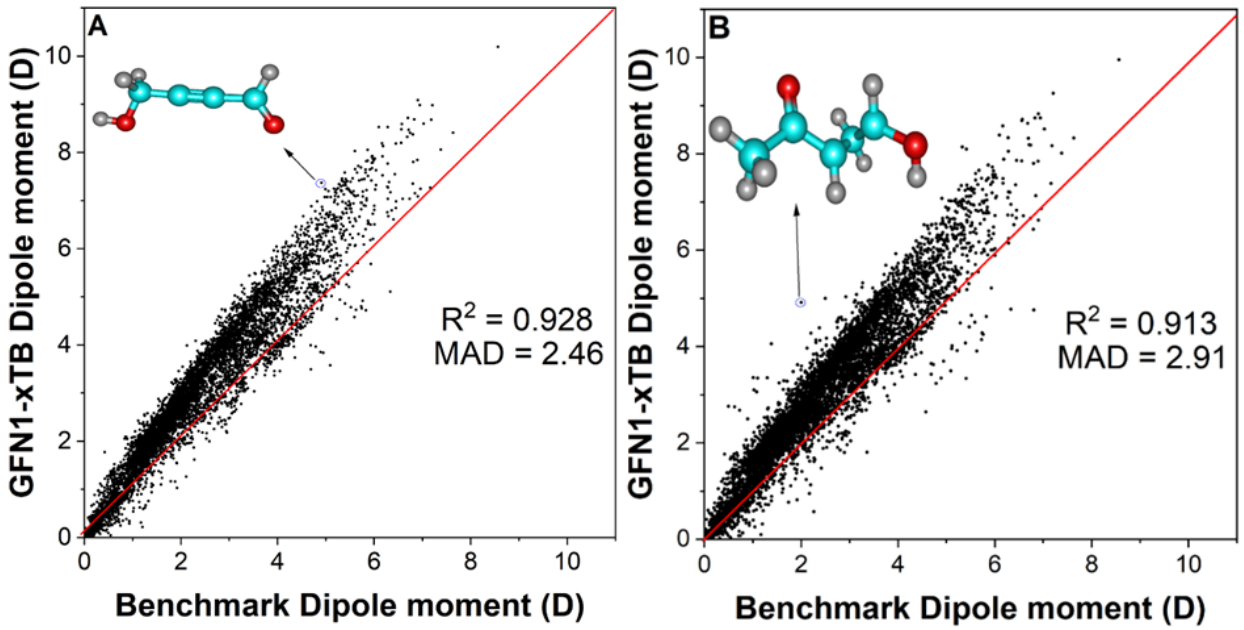


Figure S2. Parity plot comparing dipole moment predicted by GFN1-xTB for (a) DFT optimized geometries and (b) SE optimized geometries to benchmark dipole moment data. The red diagonal line is to help the reader see the ideal correlation.

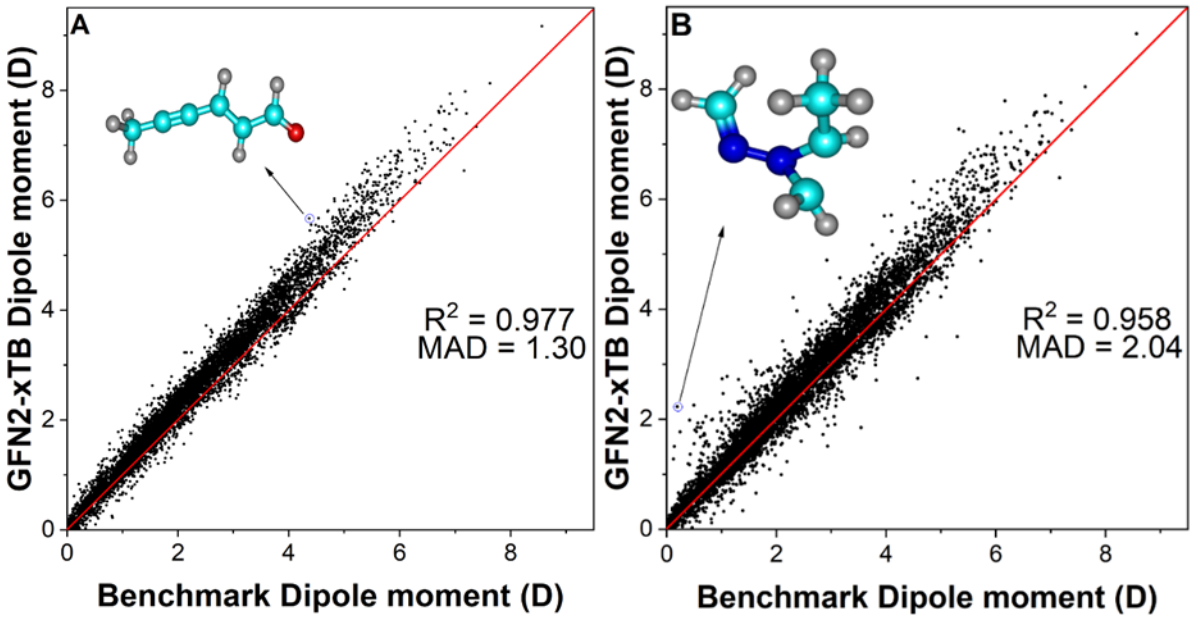


Figure S3. Parity plot comparing dipole moment predicted by GFN2-xTB for (a) DFT optimized geometries and (b) SE optimized geometries to benchmark dipole moment data. The red diagonal line is to help the reader see the ideal correlation.

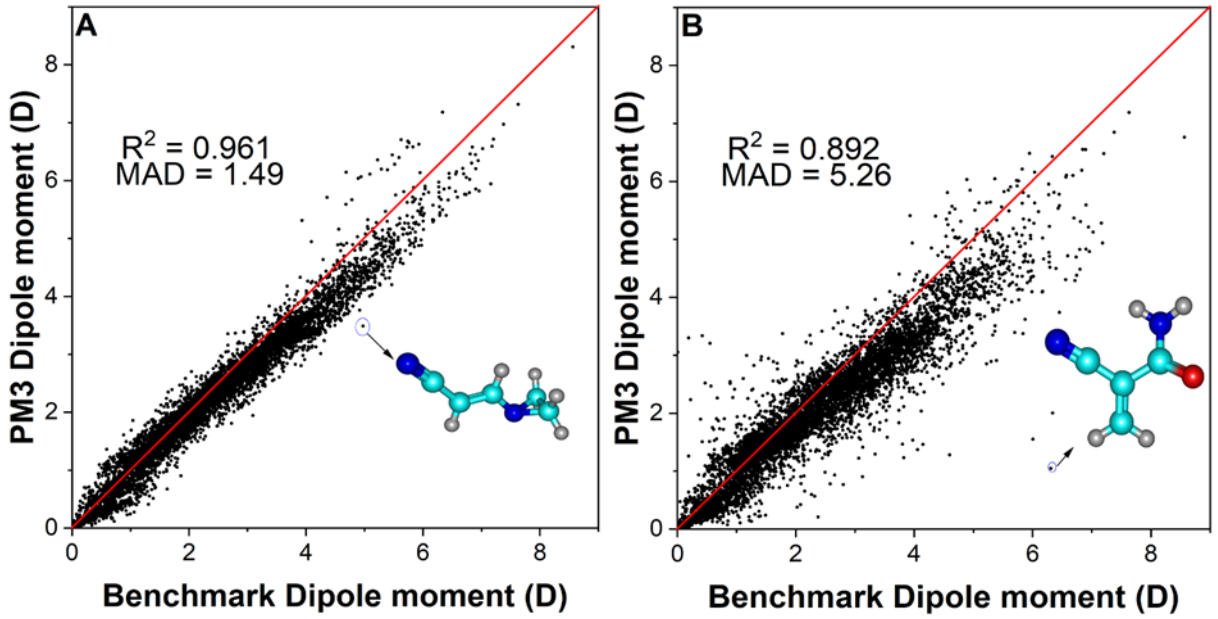


Figure S4. Parity plot comparing dipole moment predicted by PM3 for (a) DFT optimized geometries and (b) SE optimized geometries to benchmark dipole moment data. The red diagonal line is to help the reader see the ideal correlation.

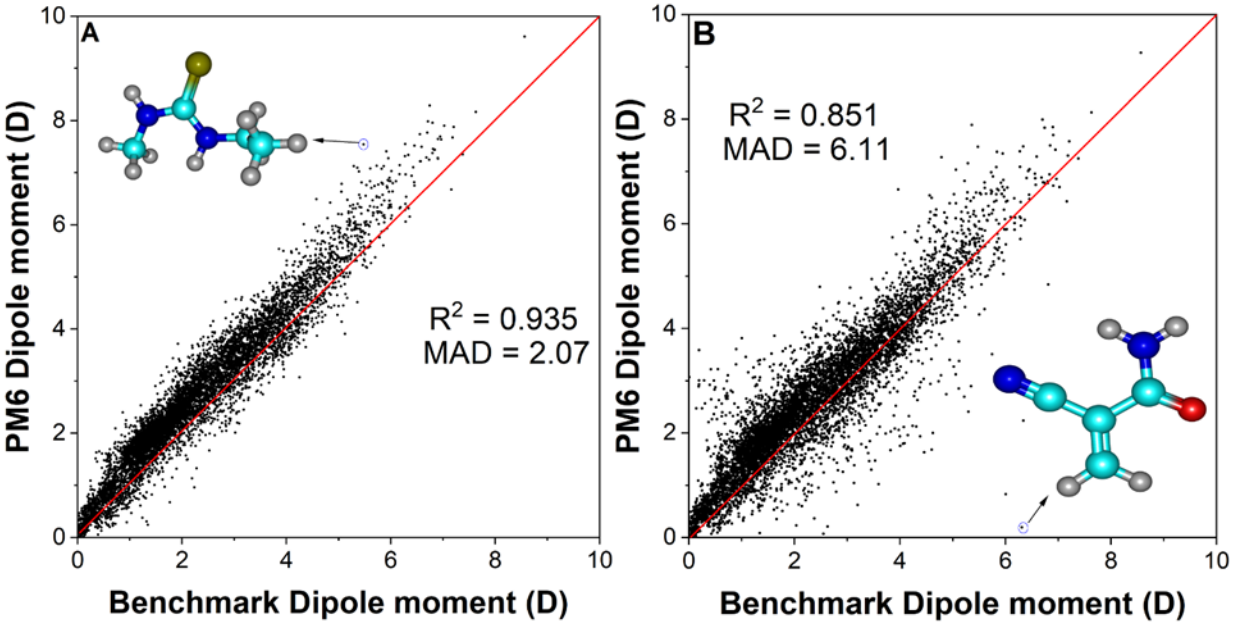


Figure S5. Parity plot comparing dipole moment predicted by PM6 for (a) DFT optimized geometries and (b) SE optimized geometries to benchmark dipole moment data. The red diagonal line is to help the reader see the ideal correlation.

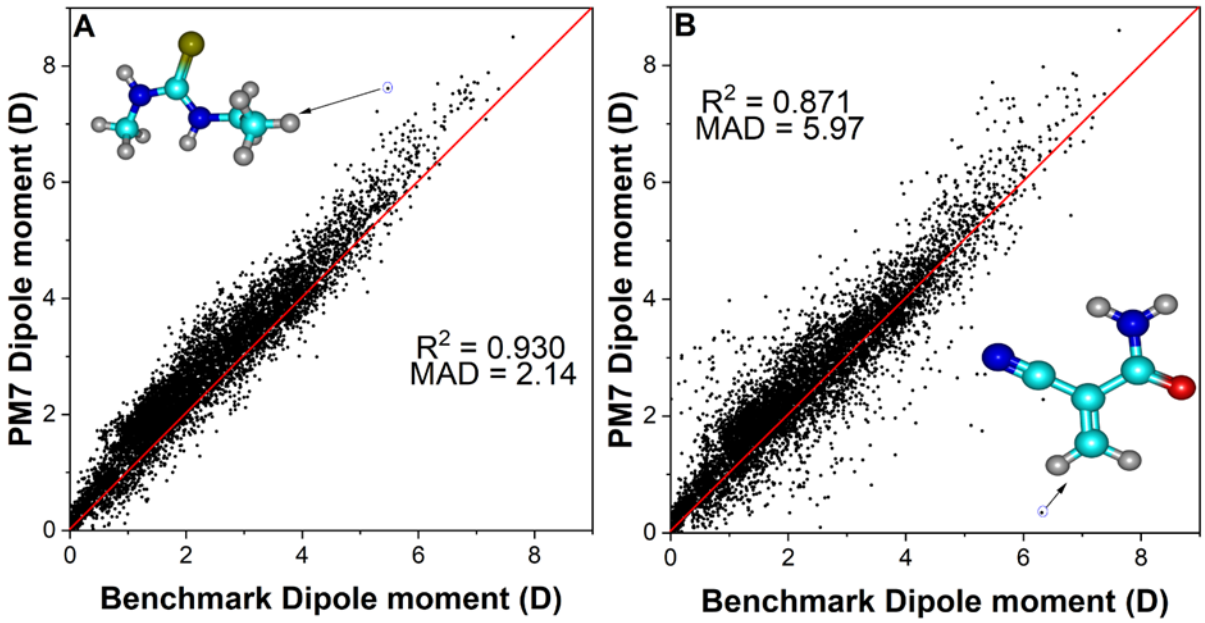


Figure S6. Parity plot comparing dipole moment predicted by PM7 for (a) DFT optimized geometries and (b) SE optimized geometries to benchmark dipole moment data. The red diagonal line is to help the reader see the ideal correlation.

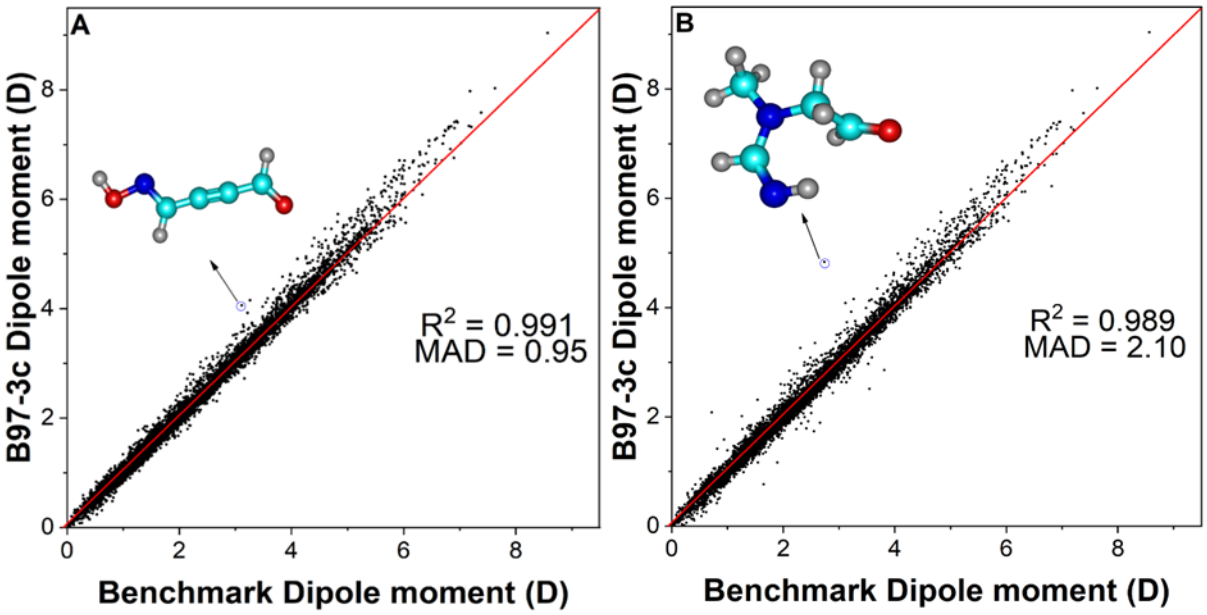


Figure S7. Parity plot comparing dipole moment predicted by B97-3c for (a) DFT optimized geometries and (b) SE optimized geometries to benchmark dipole moment data. The red diagonal line is to help the reader see the ideal correlation.

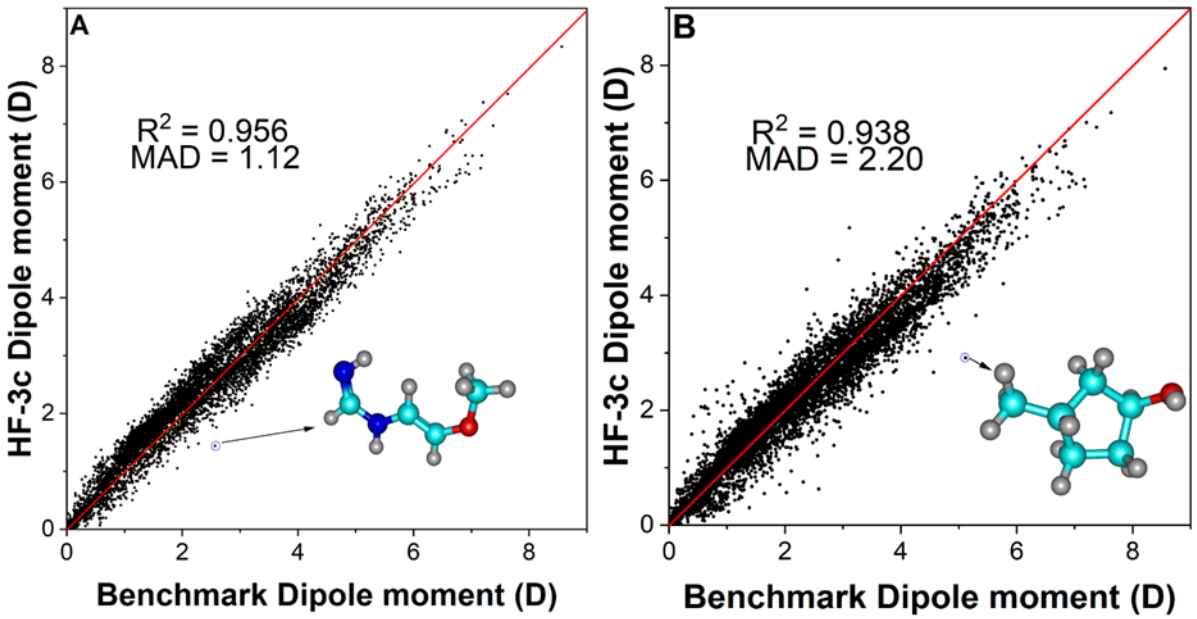


Figure S8. Parity plot comparing dipole moment predicted by HF-3c for (a) DFT optimized geometries and (b) SE optimized geometries to benchmark dipole moment data. The red diagonal line is to help the reader see the ideal correlation.

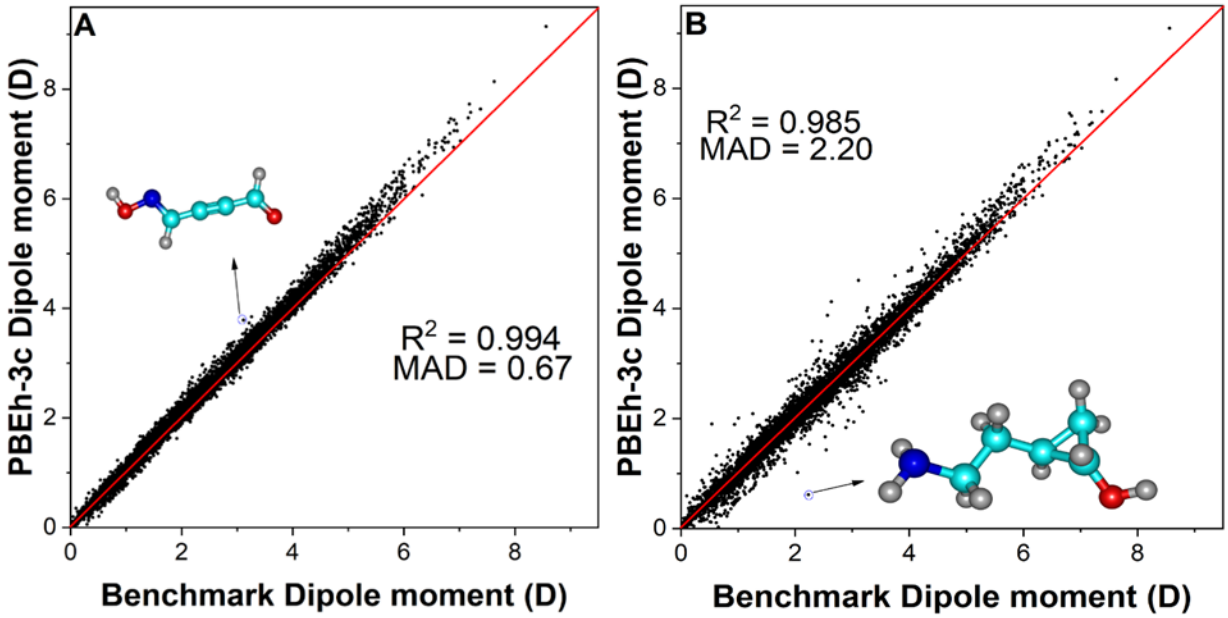
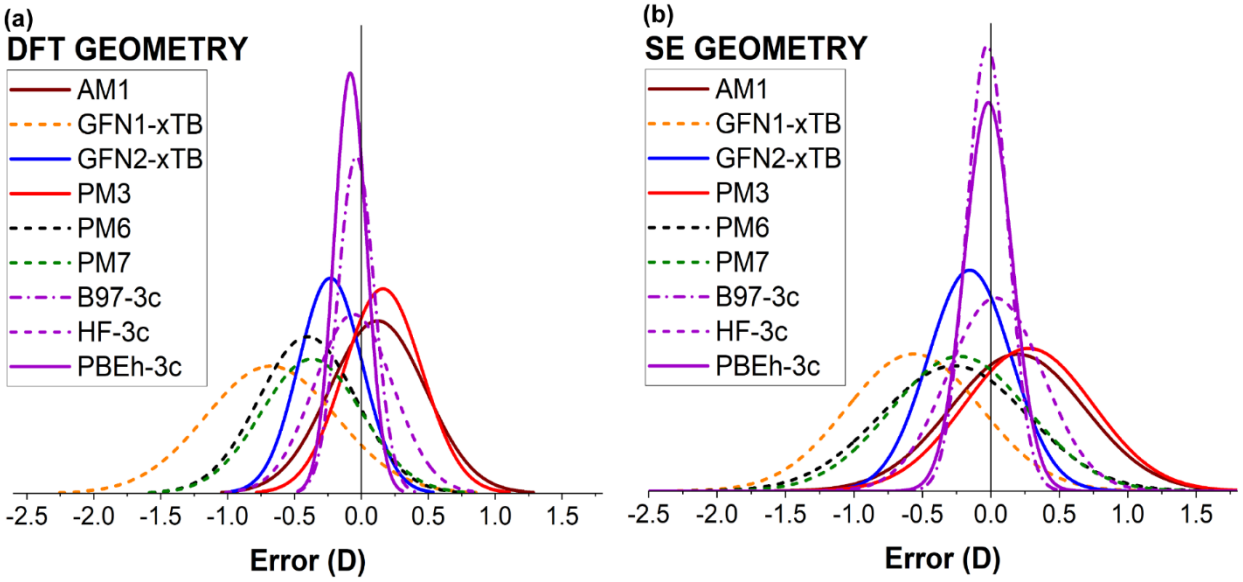


Figure S9. Parity plot comparing dipole moment predicted by PBEh-3c for (a) DFT optimized geometries and (b) SE optimized geometries to benchmark dipole moment data. The red diagonal line is to help the reader see the ideal correlation.

3. Supplementary discussion: Gaussian normal distributions showing distribution of errors for each SE method.

Based on the general features observed in Figure 2 and Table S1, we sought to understand the systematic errors of each method as well as the spread of errors within each method. Figure 2a shows the error distribution plots for DFT optimized geometries subdivided for the '3c' methods (purple lines), and other semiempirical methods GFN1-xTB, GFN2-xTB, AM1, PM3, PM6, and PM7 (all have different colors). Color choice was done to guide the eye of the reader and to differentiate the '3c' methods from the other semiempirical methods. For PBEh-3c and B97-3c, we observed mean errors of -0.08 and -0.04 D which taken together with the low FWHM of 0.28 D and 0.32 D, respectively, indicated that these methods have almost no systematic error in predicting dipole moments compared to the CCSD reference data. For HF-3c, mean error of -0.06 D and FWHM of 0.69 D were comparable to the error distribution of PM3 and GFN2-xTB which had FWHM of 0.69 and 0.56 D, and mean errors of 0.16 and -0.23 D, respectively. For other semiempirical methods, GFN2-xTB, PM3, and to a slightly lesser extent AM1 performed best, having mean errors of -0.23, 0.16, and 0.12 D, respectively. Moreover, GFN2-xTB, PM3, and AM1 yielded a lower spread in error characterized by FWHM of 0.56, 0.69, and 0.84 D, respectively, compared to GFN1-xTB, PM6, and PM7. Meanwhile, GFN1-xTB, PM6, and PM7



were the worst performing methods reflected in mean errors of -0.68, -0.40, and -0.37 D, and FWHM of 1.14, 0.85, and 0.88 D, respectively.

Figure S10. Gaussian normal distribution plots for each SE method showing the spread of error using (a) benchmark DFT optimized geometries and (b) SE optimized geometries.

4. Supplementary discussion: Performance of GFN0-xTB in dipole moment prediction

We have calculated the dipole moment given by GFN0-xTB using the benchmark DFT optimized geometries of the QM7b dataset as well as the GFN0-xTB optimized geometries. The error metrics defined in the Computational Methods section were used to evaluate the performance of GFN0-xTB in predicting dipole moments for both DFT and GFN0-xTB optimized geometries. We will discuss the results for GFN0-xTB using the benchmark DFT optimized structures first then we analyze the results related to the SE optimized structures.

4.1. Performance of GFN0-xTB using benchmark DFT optimized geometries

Figure SX shows parity plots comparing the predictive power of GFN0-xTB and the second worst method, GFN1-xTB, compared to benchmark CCSD data. Overall, GFN0-xTB provides poor dipole moment predictions indicated by the R^2 value of 0.78 and MAD of 8.60 D compared to the next worst method GFN1-xTB which had a R^2 value of 0.93 and a MAD of 2.46 D.

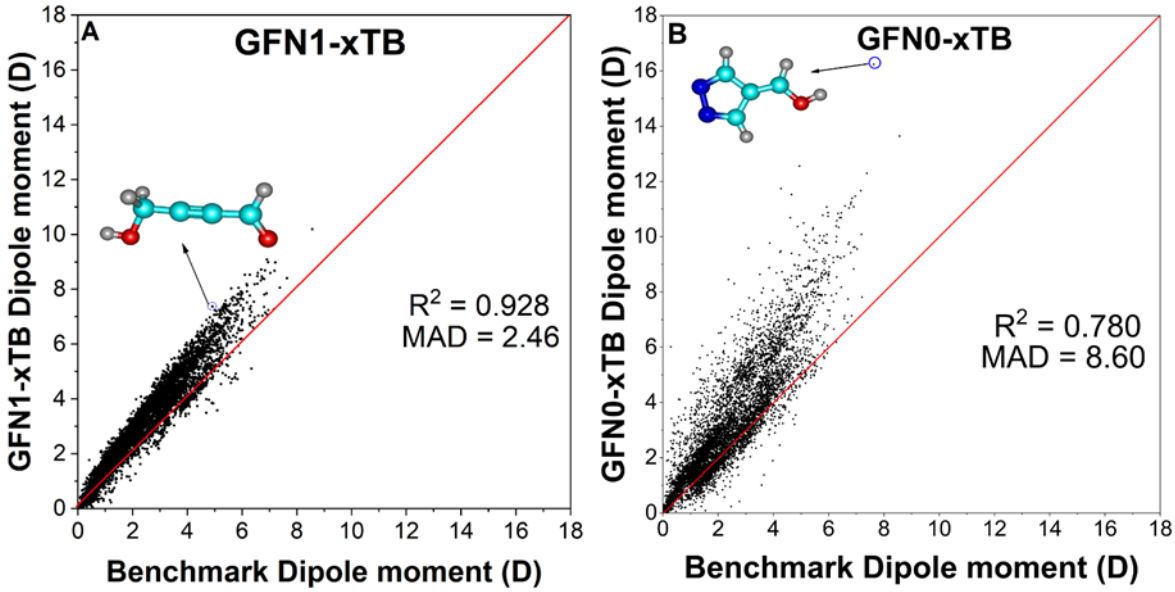


Figure S11. Parity plots comparing dipole moment predicted by (a) GFN1-xTB and (b) GFN0-xTB using DFT optimized geometries to benchmark dipole moment data. The red diagonal line is to help the reader see the ideal correlation. The blue circle and black arrow point to the molecule with the largest deviation from its CCSD dipole moment.

Figure S12a shows a radial plot where all GFN0-xTB is compared to the nine other methods evaluated in this study based on multiple error metrics. From Figure S12a we immediately observe that GFN0-xTB gave the worst performance across each metric with MAE, Mean %error, MAD, SD, and RMSE of 0.88 D, 49%, 8.60 D, 1.02 D, and 56%, respectively. Based on our condensed error metric score (Figure S12a), GFN0-xTB was almost two times worse overall (1.93) than the second worst GFN1-xTB. This poor performance of GFN0-xTB is not surprising given GFN0-xTB solves the electron density in a non-self-consistent manner. [1]

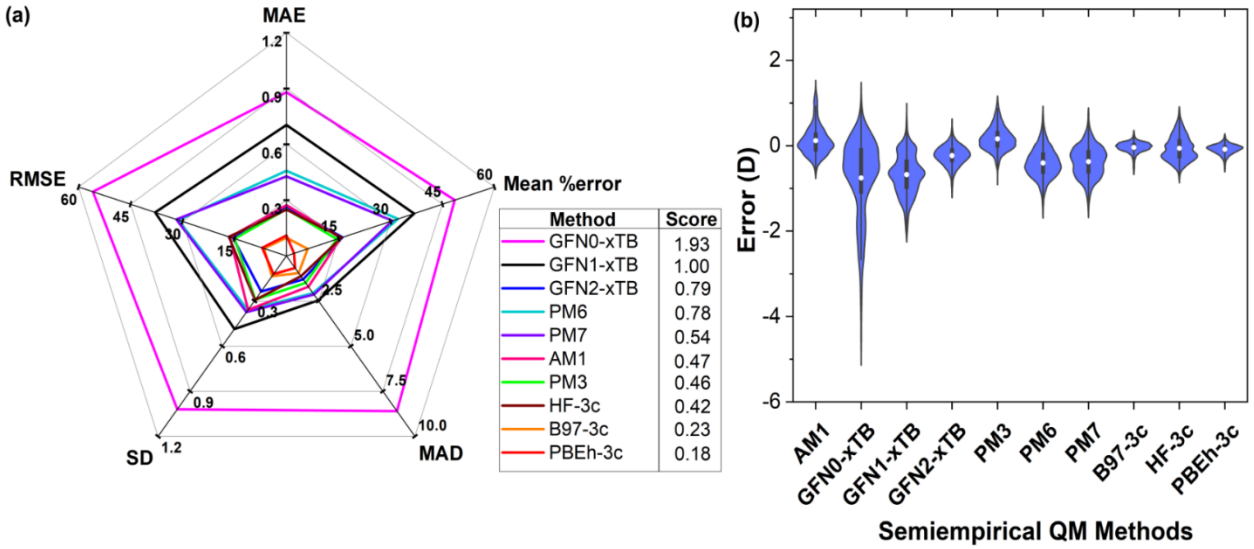


Figure S12. (a) Radial plot showing performance of GFN0-xTB compared to the nine other methods based on different error metrics. (b) Violin plots showing performance of GFN0-xTB relative to the nine other SE methods using DFT optimized geometries.

Figure S12b shows the probability density distribution of GFN0-xTB compared to all other methods evaluated in this study. Again, we see the poor performance of GFN0-xTB given the very wide spread of error based on the range of 11.17 D and having no well-defined peak. Additionally, based on the position of the distribution relative to 0 D, GFN0-xTB generally overpredicts the dipole moment of the organic molecules in the QM7b dataset.

4.2. Performance of GFN0-xTB for different atomic compositions using benchmark DFT optimized geometries

In order to understand the performance of GFN0-xTB for the different atomic compositions that formed five subcategories (see Computational Methods section), we present probability density distributions of error for GFN0-xTB for each subcategory (See Table S4-S8 for raw numbers).

In the CH subcategory, the performance of GFN0-xTB closely mirrored the performance of PM6, which gave the poorest performance for the CH subcategory among the other nine methods, in that GFN0-xTB notably overpredicts the dipole moment of hydrocarbons with a mean error of -0.28 D (PM6 mean error: -0.24 D) and an even wider range of error of 2.55 D (PM6 range: 1.73 D).

For the CHN subcategory, GFN0-xTB shows significantly worse performance compared to PM7 which was the worst performing method among the other nine methods for the CHN subcategory. GFN0-xTB considerably underpredicts the dipole moment of molecules in this subcategory with

a mean error of -1.18 D compared to PM7's -0.61 D, in addition to having more than double the range of error of PM7 (2.69 D) with a range of 5.55 D.

In the CHO subcategory, GFN0-xTB showed its best performance in terms of the mean error, showing almost no systematic error with a mean error of -0.09 D which was comparable to AM1 (mean error: 0.08 D). Despite the low mean error however, GFN0-xTB gave a very large range of error with a range of 7.38 D which is almost three times the range of the next worst SE method GFN1-xTB (range: 2.74 D).

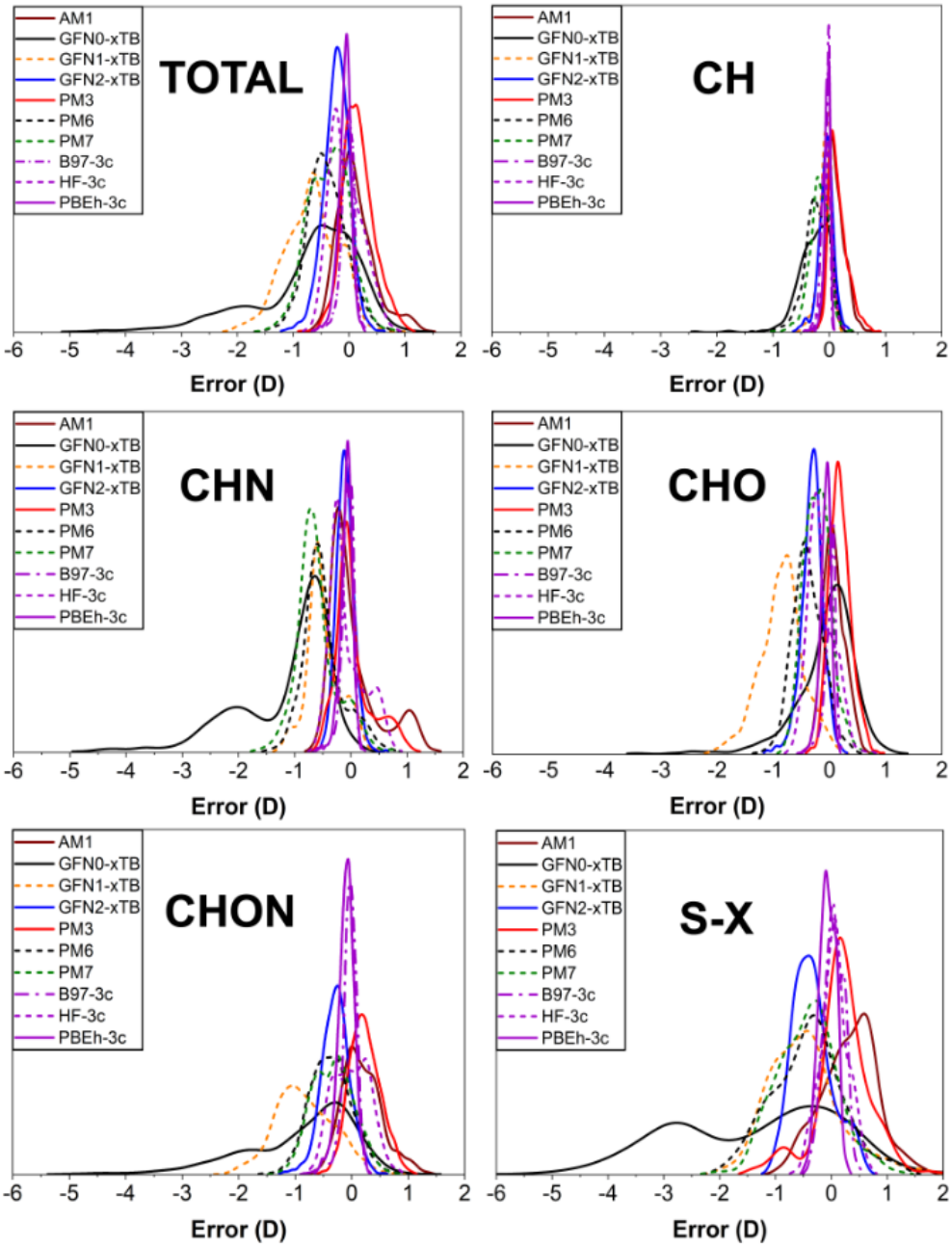


Figure S13. Kernel smooth probability density distribution of errors for GFN0-xTB using benchmark DFT optimized geometries for the CH, CHO, CHN, CHON, and S-X compositional subcategories. The vertical axis of each plot is the probability density of the error value. For comparison, the probability density distribution of errors for the total dataset is also disclosed.

In the CHON subcategory, GFN0-xTB underpredicted the dipole moment molecules with a mean error of -0.93 D which was similar to GFN1-xTB with a mean error of -0.83 D. However, GFN0-xTB showed a very wide range of error with a range of 11.17 D which was more than three times the range of the next worst SE method in the CHON subcategory GFN1-xTB (3.32 D).

For the S-X subcategory, GFN0-xTB again shows very poor performance given its mean error of -1.38 D indicating GFN0-xTB considerably underpredicted the dipole moment of molecules in

this subcategory with the next worst SE method GFN1-xTB having a mean error of -0.47 D. As we saw for other subcategories, GFN0-xTB had a very large spread of error shown in the range of 8.84 D.

4.3. Performance of GFN0-xTB using GFN0-xTB optimized geometries

We also evaluated the performance of GFN0-xTB using the GFN0-xTB optimized geometry instead of the DFT optimized geometry contained in the QM7b dataset. As previously discussed, we did this to evaluate the performance of GFN0-xTB under the practical scenario where DFT optimized geometries are unavailable and GFN0-xTB must be used to carry out a geometry optimization to obtain the dipole moment. Here, we will focus on the differences in performance using DFT optimized, and GFN0-xTB optimized geometries.

Figure S14 shows parity plots comparing the predictive power of GFN0-xTB and the second worst method, PM6, compared to benchmark CCSD data. Overall, GFN0-xTB provides very poor dipole moment predictions using GFN0-xTB optimized geometries indicated by the R^2 value of 0.63 and MAD of 15.20 D compared to R^2 value and MAD of PM6 (0.85, 6.11 D).

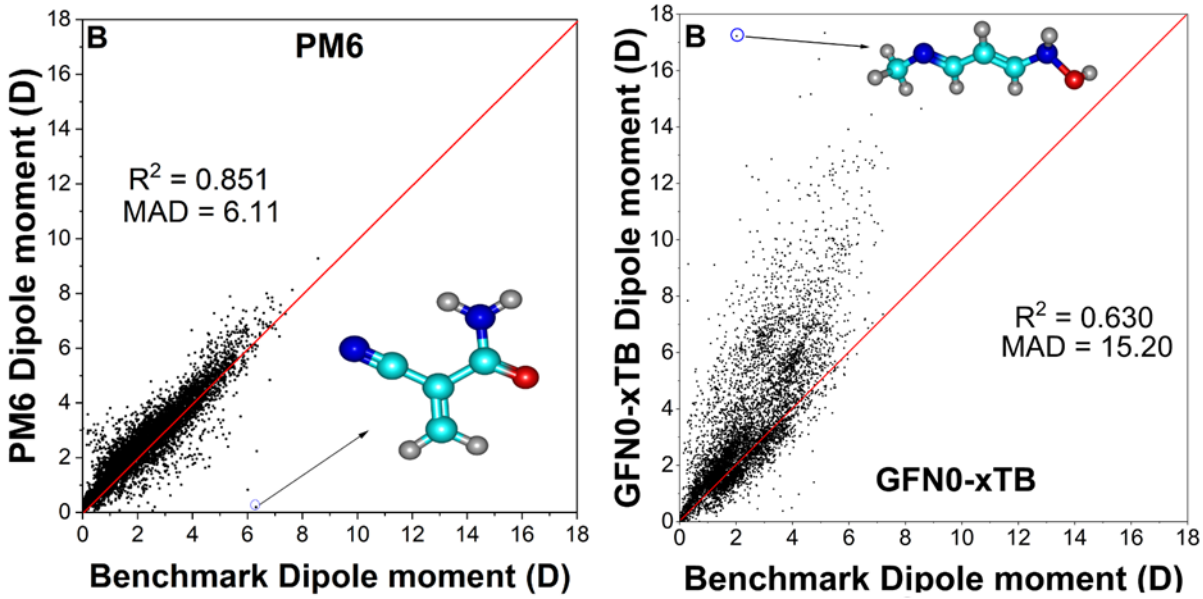


Figure S14. Parity plot comparing dipole moment predicted by (a) PM6 and (b) GFN0-xTB using SE optimized geometries to benchmark dipole moment data. The red diagonal line is to help the reader see the ideal correlation. The blue circle and black arrow point to the molecule with the largest deviation from its CCSD dipole moment.

Figure S15a shows a radial plot where all GFN0-xTB is compared to the nine other methods evaluated in this study based on multiple error metrics.

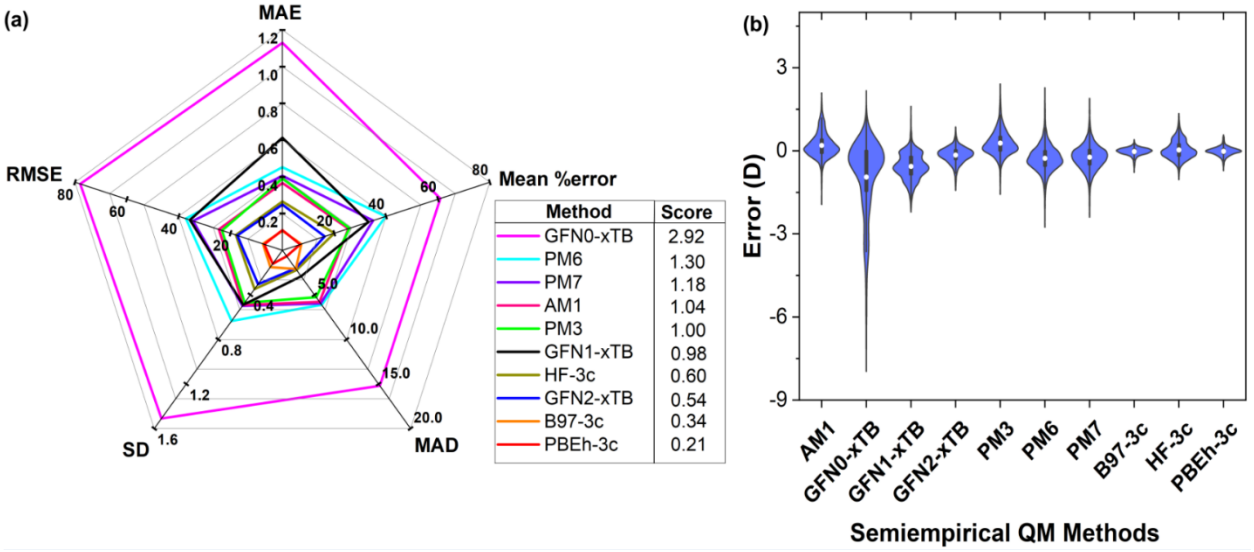


Figure S15. (a) Radial plot showing performance of GFN0-xTB compared to the nine other methods based on different error metrics. (b) Violin plots showing performance of GFN0-xTB relative to the nine other SE methods using DFT optimized geometries.

Again, we see GFN0-xTB's poor performance across each metric which resulted in the condensed metric score dramatically rising to 2.92 using GFN0-xTB optimized geometries compared to 1.93 that was obtained using DFT optimized geometries. Figure S15b shows the probability density distribution of GFN0-xTB compared to all other methods evaluated in this study. Using GFN0-xTB optimized geometries, the performance of GFN0-xTB decreased significantly indicated by the much larger range of 18.87 D. On the positive side, the probability density distribution has a well-defined peak.

4.4. Performance of GFN0-xTB for different atomic compositions using SE optimized geometries

For completeness, we also evaluated the performance of GFN0-xTB in the different subcategories using GFN0-xTB optimized geometries. Figure S16 shows the probability distribution of errors for each SE method for each subcategory. See Tables S3 to S4 for raw numbers.

In the CH subcategory, we saw some marginal improvements in the performance of GFN0-xTB shown in the mean error which decreased marginally to -0.26 D from -0.28 D obtained using DFT optimized geometries. Additionally, the range slightly decreased from 2.55 D using DFT optimized geometries to 2.41 D using GFN0-xTB optimized geometries.

For the CHN subcategory, GFN0-xTB shows significantly worse performance using GFN0-xTB optimized geometries. GFN0-xTB underpredicts the dipole moment of molecules in this subcategory more significantly with a mean error of -1.29 D compared to -1.18 D using DFT optimized geometries. In addition, the range of error is almost doubled using SE optimized geometries (9.18 D) compared to the range obtained using benchmark DFT geometries (5.55 D).

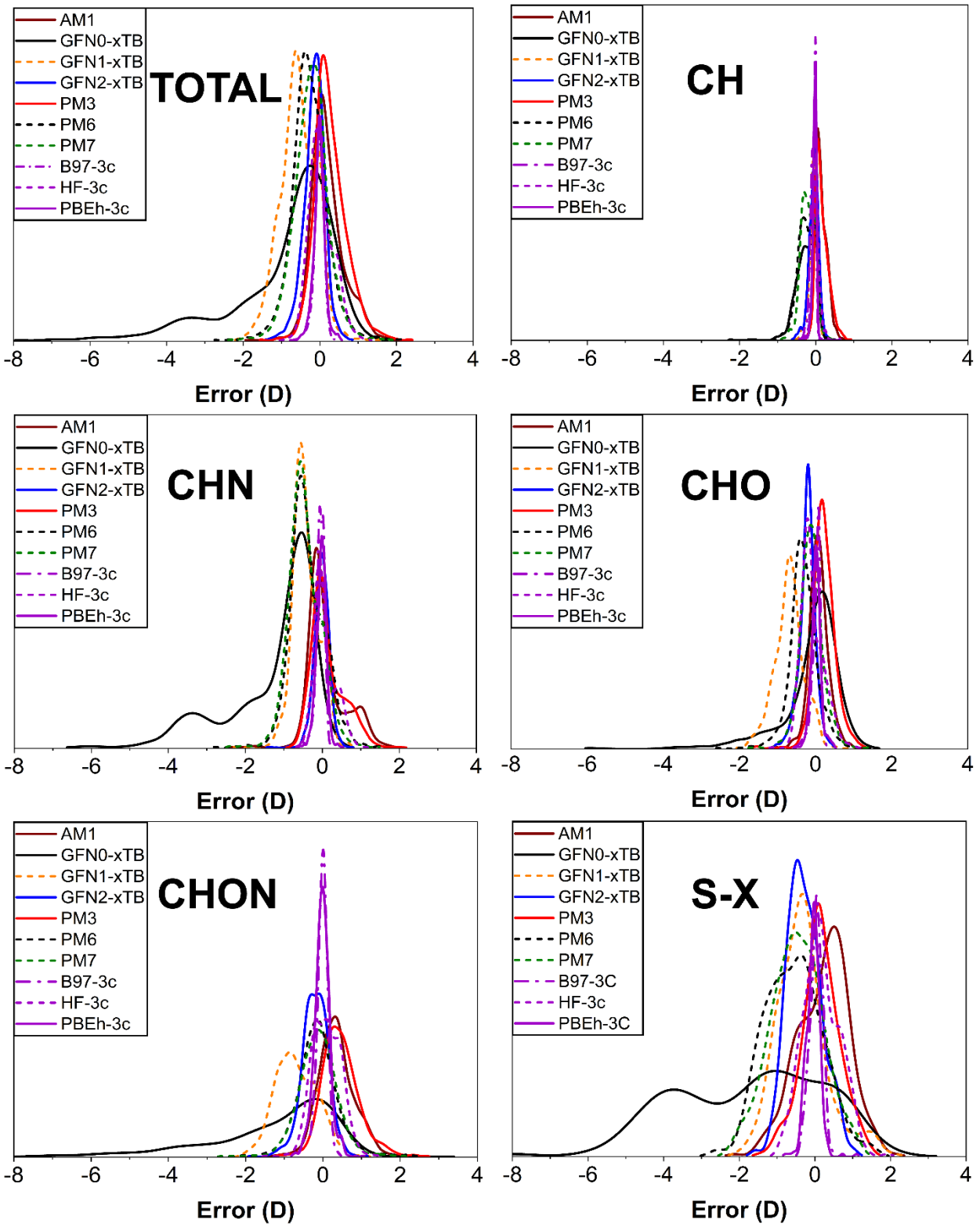


Figure S16. Kernel smooth probability density distribution of errors for GFN0-xTB using SE optimized geometries for the CH, CHO, CHN, CHON, and S-X compositional subcategories. The vertical axis of each plot is the probability density of the error value. For comparison, the probability density distribution of errors for the total dataset is also disclosed.

Similarly, in the CHO subcategory the performance of GFN0-xTB was somewhat worse using GFN0-xTB optimized geometries as the mean error increased slightly to -0.14 D compared to -0.09 D using benchmark DFT optimized geometries. Additionally, GFN0-xTB gave a very large range of error with a range of 10.18 D compared to the range of 7.38 D obtained using DFT optimized geometries.

In the CHON subcategory, GFN0-xTB more significantly underpredicted the dipole moment molecules with a mean error of -1.29 D compared to the mean error of -0.93 D obtained using DFT optimized geometries while there was again a dramatic increase in the range of error which rose to 18.69 D from 11.17 D.

Following the same trend for the other subcategories, the performance of GFN0-xTB dropped significantly in the S-X subcategory using GFN0-xTB optimized geometries compared to DFT optimized geometries. For this subcategory, GFN0-xTB considerably underpredicted the dipole moment shown in the mean error of -1.78 D compared to -1.38 D obtained using DFT optimized geometries while the range of error also increased to 11.65 D compared to 8.84 D obtained using DFT optimized geometries.

5. **Probability density distribution of subcategory errors for each SE method normalized by population using DFT optimized geometries.**

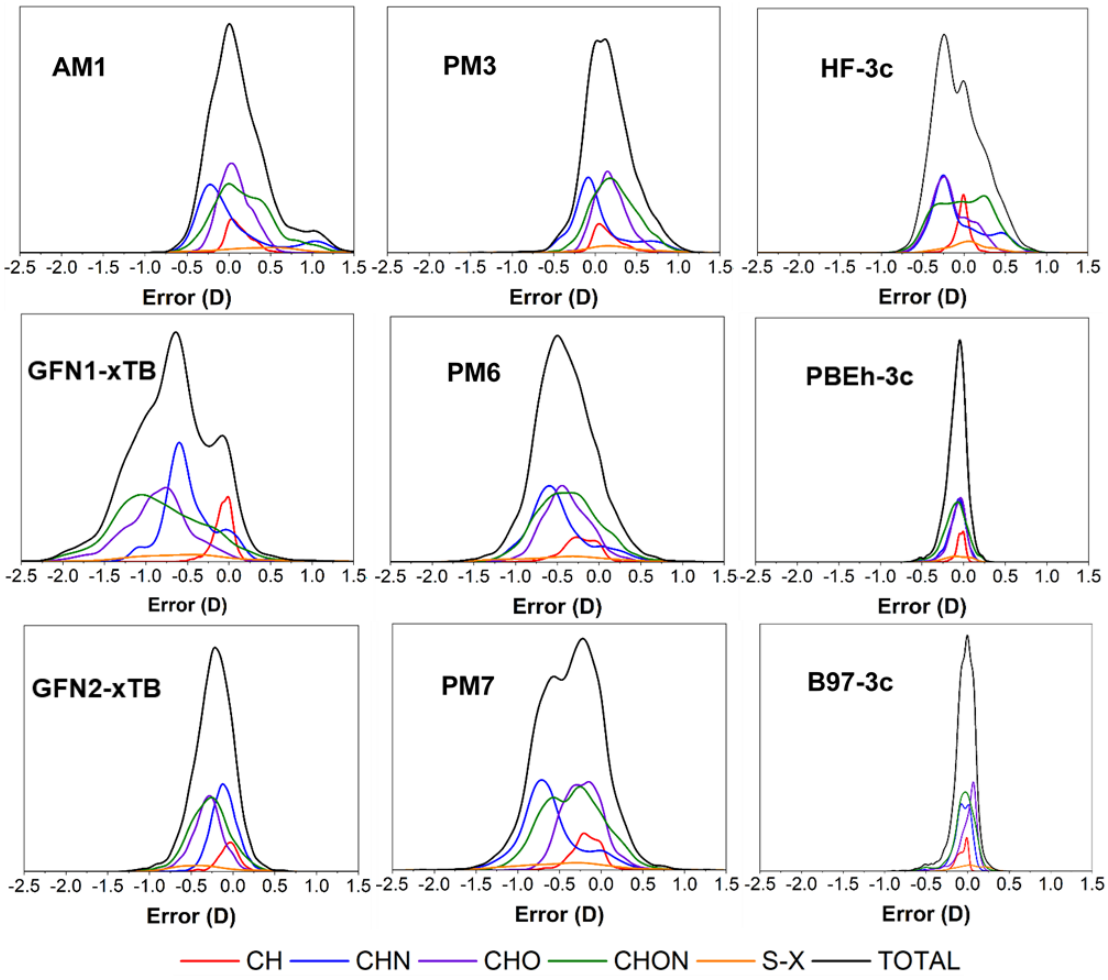


Figure S17. Kernel smooth probability density distribution of subcategory errors for each SE method using benchmark DFT optimized geometries. The vertical axis of each plot is the probability of obtaining a certain error value. Each subcategory has been normalized by its population.

6. Supplementary discussion: Performance of SE methods for different atomic compositions using SE optimized geometries

For completeness, we also evaluated the performance of the SE methods in the different subcategories using SE optimized geometries. Figure S18 shows the probability distribution of errors for each SE method for each subcategory. See Tables S3 to S4 for raw numbers.

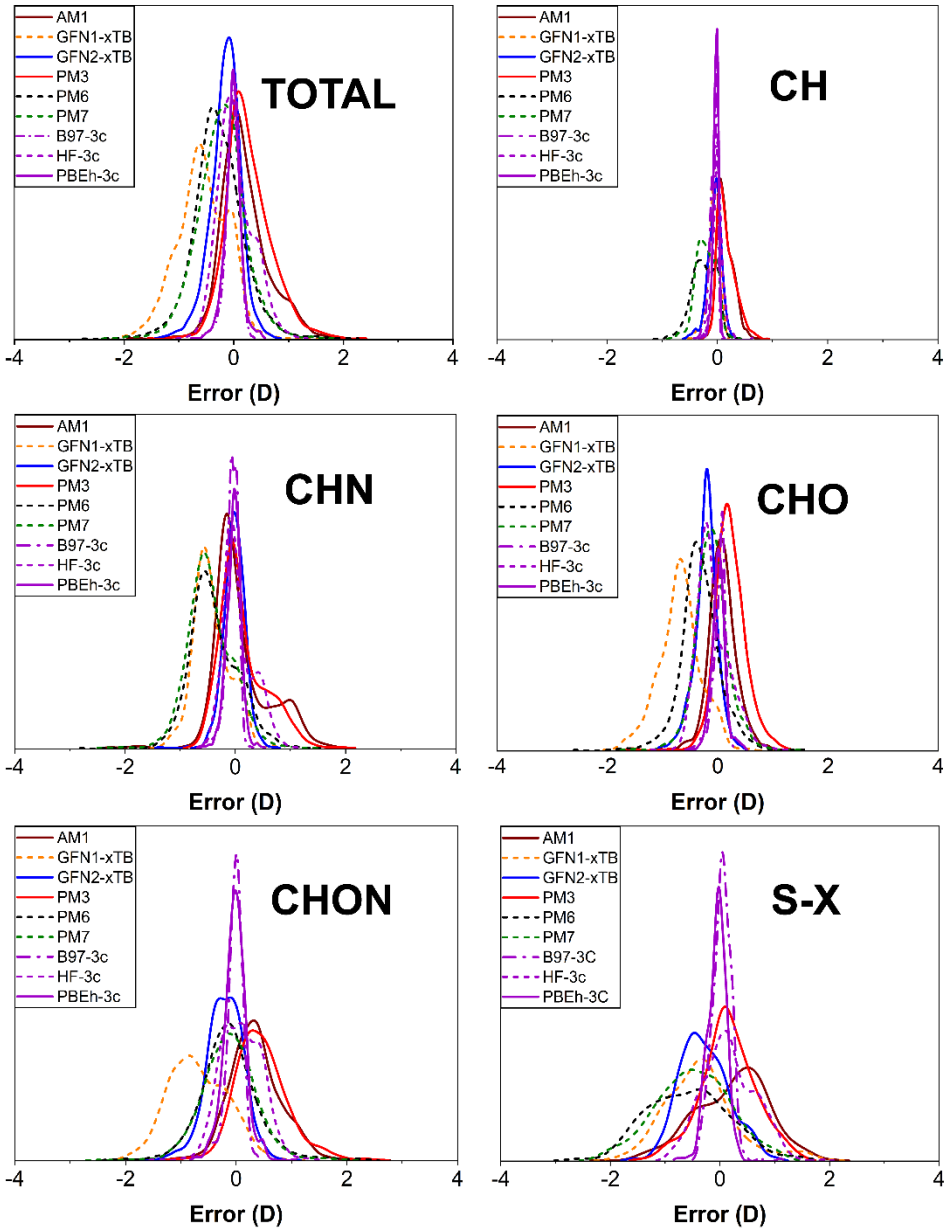


Figure S18. Kernel smooth probability density distribution of errors for each SE method using SE optimized geometries for the CH, CHO, CHN, CHON, and S-X compositional subcategories. The

vertical axis of each plot is the probability of obtaining a certain error value. For comparison, the probability density distribution of errors for the total dataset is also disclosed.

For the CH subcategory, there was no considerable change in the performance of the SE methods shown in the comparable mean errors (AM1: 0.14 D, GFN1-xTB: -0.06 D, GFN2-xTB: -0.06 D, PM3: 0.16 D, PM6: -0.25 D, PM7: -0.21 D, B97-3c: -0.07 D, HF-3c: 0.00 D, and PBEh-3C: -0.03 D) and ranges (AM1: 0.98 D, GFN1-xTB: 0.68 D, GFN2-xTB: 0.93 D, PM3: 1.12 D, PM6: 1.56 D, PM7: 1.23 D, B97-3c: 0.57 D, HF-3c: 0.89 D, and PBEh-3c: 0.88 D) using SE optimized geometries and the mean errors (AM1: 0.13 D, GFN1-xTB: -0.08 D, GFN2-xTB: -0.06 D, PM3: 0.16 D, PM6: -0.24 D, PM7: -0.18 D, B97-3c: -0.07 D, HF-3c: -0.02 D, and PBEh-3c: -0.04 D and ranges (AM1: 1.02 D, GFN1-xTB: 0.75 D, GFN2-xTB: 0.97 D, PM3: 1.14 D, PM6: 1.73 D, PM7: 1.41 D, B97-3c: 0.49 D, HF-3c: 0.72 D, and PBEh-3c: 0.37 D) using DFT optimized geometries. Consequently, just as we saw using DFT optimize geometries, the most consistent performance for all methods was shown in the CH subcategory.

In the CHN subcategory, there were improvements in the mean errors for GFN1-xTB, GFN2-xTB, PM6 and PM7 characterized by their smaller mean errors of -0.44 D, -0.03 D,-0.38 D, and -0.46 D, respectively, given using SE optimized geometries compared to -0.49 D, -0.11 D, -0.49 D, and -0.61 D, respectively, obtained using DFT optimized geometries. However for these methods, the spread of error increased dramatically shown in the ranges of 3.45 D, 3.27 D, 6.91 D, and 7.03 D, respectively, compared to much smaller ranges of 2.34 D, 1.50 D, 2.57 D, and 2.69 D, respectively, shown using DFT optimized geometries. For B97-3c, HF-3c, PBEh-3c, and PM3 there were no considerable changes in the mean errors using SE optimized geometries (B97-3c: -0.07 D, HF-3c: 0.00 D, PBEh-3c: -0.03 D, and PM3: 0.16 D) compared to using DFT optimized geometries (B97-3c: -0.07 D, HF-3c: -0.02 D, PBEh-3c: -0.04 D, and PM3: 0.16 D). However, we observed significant increase in the spread of error shown in the ranges of 2.18 D, 3.69 D, 1.87 D, and 5.67 D compared to 0.49 D, 1.82 D, 0.92 D, and 2.31 D, respectively, provided using DFT optimized geometries.

In the CHO subcategory, there were notable improvements in the mean errors for GFN1-xTB, GFN2-xTB, and PM7. This is characterized by their smaller mean errors of -0.69 D, -0.22 D, and -0.11 D, respectively, given using SE optimized geometries compared to -0.86 D, -0.31 D, and -0.22 D, respectively, obtained using DFT optimized geometries. However, for these methods, the spread of error increased dramatically shown in the ranges of 4.61 D, 2.27 D, and 5.39 D, respectively, compared to much smaller ranges of 2.74 D, 1.61 D, and 1.89 D, respectively, shown using DFT optimized geometries. In contrast, HF-3c overpredicted the dipole moment for molecules in the CHO subcategory to a greater extent given the increase in the mean error from -0.08 D to -0.15 D. In addition, there was also an increase in the spread of error for HF-3c shown as the range increased from 1.55 D using DFT optimized geometries to 2.76 D using SE optimized geometries, respectively. Meanwhile, for AM1, B97-3c, PBEh-3c, PM3, and PM6, the only changes in performance were an increased spread of error for these methods shown in the range (AM1: 5.14 D, B97-3c: 1.48 D, PBEh-3c: 2.39 D, PM3: 5.55 D, and PM6: 5.97 D) compared to those given using DFT optimized geometries (AM1: 1.31 D, B97-3c: 1.13 D, PBEh-3c: 0.89 D, PM3: 1.38 D, and PM6: 2.21 D).

For the CHON subcategory, GFN1-xTB, GFN2-xTB, PM6, and PM7 provided lower systematic error reflected in their lower mean errors using SE optimized geometries (GFN1-xTB: -0.55 D, GFN2-xTB: -0.20 D, PM6: -0.12 D, and PM7: -0.12 D) compared to their performance using DFT

optimized geometries (GFN1-xTB: -0.83 D, GFN2-xTB: -0.29 D, PM6: -0.37 D, and PM7: -0.33 D). Despite the improvement in the mean error, there were considerable increases in the spread of error (GFN1-xTB: 4.87 D, GFN2-xTB: 3.77 D, PM6: 8.91 D, and PM7: 8.72 D) compared to the narrower spread of error shown using DFT optimized geometries (GFN1-xTB: 3.32 D, GFN2-xTB: 1.90 D, PM6: 2.58 D, and PM7: 2.45 D). In contrast, AM1 and PM3 showed larger systematic error characterized by lower mean errors obtained using SE optimized geometries (AM1: 0.36 D, PM3: 0.46 D) compared to DFT optimized geometries (AM1: 0.19 D, PM3: 0.23 D). Additionally, there was a significant increase in the spread of error given by AM1 and PM3 using SE optimized geometries shown in their much wider ranges of 7.76 D and 8.18 D, respectively, compared to 2.46 D and 1.87 D, respectively, provided using DFT optimized geometries. We consequently attributed this drastic drop in performance to structural differences between the SE and DFT optimized geometries. Meanwhile, for the '3c' methods, the only notable changes were in the spread of error which increased for each method using SE optimized geometries (B97-3c: 3.13 D, HF-3c: 3.73 D, PBEh-3c: 3.02 D) relative to the spread of error shown using DFT optimized geometries (B97-3c: 1.27 D, HF-3c: 2.10 D, PBEh-3c: 0.99 D).

In the S-X subcategory, we observed small improvements in the range of error for PM7 (3.34 D) using SE optimized geometries compared to DFT optimized geometries (3.40 D) while PM3, and PM6 produced wider ranges of error using SE optimized geometries (PM3: 3.02 D, PM6: 4.27 D) compared to DFT optimized geometries (PM3: 2.70 D, PM6: 3.40 D). However, the mean error was essentially unchanged for PM3 (0.15 D) using SE optimized geometries while PM6 (-0.57 D) and PM7 (-0.49 D) overpredicted the dipole moment to a slightly greater degree using SE optimized geometries compared to DFT optimized geometries (PM3: 0.14 D, PM6: -0.26, and PM7: -0.43 D). For AM1 and GFN1-xTB, the mean error (AM1: 0.22 D, and GFN1-xTB: -0.34 D) also improved using SE optimized geometries compared to DFT optimized geometries (AM1: 0.33 D, and GFN1-xTB: 0.50 D). However, we observed an increase in the range of error given by GFN1-xTB (4.12 D) using SE optimized geometries compared to DFT optimized geometries (3.48 D) which we again attributed to structural differences between SE optimized geometries and DFT optimized geometries. For the '3c' methods and GFN2-xTB however, there were no notable changes in performance.

7. Summary of subcategory errors for each SE method

Table S4. Errors associated with each SE method for the CH subcategory using DFT and SE optimized geometries.

SE method	Mean error (D)		SD (D)		FWHM (D)		Range (D)	
	DFT geometry	SE geometry	DFT geometry	SE geometry	DFT geometry	SE geometry	DFT geometry	SE geometry
AM1	0.13	0.14	0.15	0.15	0.35	0.35	1.02	0.98
GFN0-xTB	-0.28	-0.26	0.28	0.27	0.66	0.63	2.55	2.41
GFN1-xTB	-0.08	-0.06	0.11	0.10	0.27	0.25	0.75	0.68

GFN2-xTB	-0.06	-0.06	0.14	0.13	0.32	0.31	0.97	0.93
PM3	0.16	0.16	0.17	0.17	0.40	0.40	1.14	1.12
PM6	-0.24	-0.25	0.22	0.22	0.51	0.52	1.73	1.56
PM7	-0.18	-0.21	0.18	0.18	0.42	0.43	1.41	1.23
B97-3c	-0.07	-0.07	0.07	0.07	0.17	0.17	0.49	0.57
HF-3c	-0.02	0.00	0.09	0.10	0.22	0.23	0.72	0.89
PBEh-3c	-0.04	-0.03	0.05	0.06	0.13	0.14	0.37	0.88

Table S5. Errors associated with each SE method for the CHN subcategory using DFT and SE optimized geometries.

SE method	Mean error (D)		SD (D)		FWHM (D)		Range (D)	
	DFT geometry	SE geometry	DFT geometry	SE geometry	DFT geometry	SE geometry	DFT geometry	SE geometry
AM1	0.04	0.10	0.45	0.55	1.07	1.31	2.38	7.16
GFN0-xTB	-1.18	-1.29	0.88	1.24	2.06	2.92	5.55	9.18
GFN1-xTB	-0.49	-0.44	0.31	0.35	0.74	0.84	2.34	3.45
GFN2-xTB	-0.11	-0.03	0.16	0.25	0.38	0.58	1.50	3.27
PM3	0.04	0.14	0.34	0.46	0.80	1.08	2.31	5.67
PM6	-0.49	-0.38	0.37	0.50	0.87	1.19	2.57	6.91
PM7	-0.61	-0.46	0.37	0.47	0.88	1.10	2.69	7.03
B97-3c	-0.07	-0.06	0.12	0.13	0.27	0.30	1.04	2.18
HF-3c	-0.07	0.06	0.31	0.30	0.74	0.71	1.82	3.69
PBEh-3c	-0.08	-0.03	0.11	0.14	0.27	0.34	0.92	1.88

Table S6. Errors associated with each SE method for the CHO subcategory using DFT and SE optimized geometries.

	Mean error (D)	SD (D)	FWHM (D)	Range (D)

SE method	DFT geometry	SE geometry	DFT geometry	SE geometry	DFT geometry	SE geometry	DFT geometry	SE geometry
AM1	0.08	0.08	0.19	0.31	0.45	0.72	1.31	5.14
GFN0-xTB	-0.09	-0.14	0.62	0.90	1.46	2.11	7.38	10.18
GFN1-xTB	-0.86	-0.69	0.39	0.41	0.92	0.96	2.74	4.61
GFN2-xTB	-0.31	-0.22	0.20	0.23	0.46	0.54	1.61	2.27
PM3	0.19	0.22	0.19	0.36	0.45	0.84	1.38	5.55
PM6	-0.39	-0.32	0.28	0.43	0.66	1.01	2.21	5.97
PM7	-0.22	-0.11	0.24	0.35	0.57	0.83	1.89	5.39
B97-3c	0.00	0.02	0.12	0.14	0.29	0.32	1.13	1.47
HF-3c	-0.15	-0.08	0.24	0.31	0.56	0.74	1.55	2.76
PBEh-3c	-0.07	0.01	0.10	0.14	0.23	0.34	0.89	2.39

Table S7. Errors associated with each SE method for the CHON subcategory using DFT and SE optimized geometries.

SE method	Mean error (D)		SD (D)		FWHM (D)		Range (D)	
	DFT geometry	SE geometry	DFT geometry	SE geometry	DFT geometry	SE geometry	DFT geometry	SE geometry
AM1	0.19	0.36	0.36	0.52	0.85	1.23	2.46	7.76
GFN0-xTB	-0.93	-1.29	1.08	1.82	2.55	4.28	11.17	18.69
GFN1-xTB	-0.83	-0.69	0.53	0.55	1.24	1.30	3.32	4.87
GFN2-xTB	-0.29	-0.20	0.25	0.35	0.60	0.83	1.90	3.77
PM3	0.23	0.46	0.28	0.53	0.65	1.24	1.87	8.18
PM6	-0.37	-0.12	0.37	0.60	0.88	1.41	2.58	8.91
PM7	-0.33	-0.12	0.37	0.56	0.86	1.31	2.45	8.72
B97-3c	-0.05	-0.03	0.15	0.17	0.35	0.41	1.27	3.13
HF-3c	0.00	0.09	0.32	0.39	0.76	0.91	2.10	3.73
PBEh-3c	-0.10	-0.02	0.14	0.22	0.34	0.51	0.99	3.02

Table S8. Errors associated with each SE method for the S-X subcategory using DFT and SE optimized geometries.

SE method	Mean error (D)		SD (D)		FWHM (D)		Range (D)	
	DFT geometry	SE geometry	DFT geometry	SE geometry	DFT geometry	SE geometry	DFT geometry	SE geometry
AM1	0.33	0.22	0.49	0.67	1.15	1.59	2.63	4.25
GFN0-xTB	-1.38	-1.78	1.50	2.01	3.54	4.75	8.84	11.65
GFN1-xTB	-0.50	-0.34	0.61	0.67	1.44	1.58	3.48	4.12
GFN2-xTB	-0.34	-0.28	0.35	0.45	0.82	1.06	1.69	2.66
PM3	0.14	0.15	0.45	0.51	1.06	1.20	2.70	3.02
PM6	-0.44	-0.57	0.62	0.75	1.47	1.76	3.34	4.27
PM7	-0.43	-0.49	0.54	0.67	1.26	1.59	3.40	3.33
B97-3c	0.04	0.03	0.17	0.17	0.41	0.39	0.95	1.16
HF-3c	0.05	0.17	0.24	0.47	0.58	1.12	1.36	2.82
PBEh-3c	-0.07	-0.05	0.14	0.18	0.34	0.42	0.77	1.88

8. Examples of SE optimized geometries compared to DFT optimized geometries

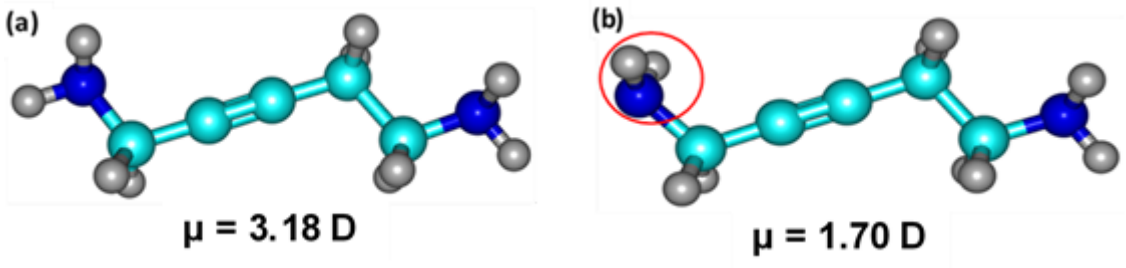


Figure S19. Molecule 26 of the QM7b dataset. (a) DFT optimized geometry and (b) AM1 optimized geometry. Red circle indicates location of conformation change after geometry optimization using AM1. Color code: C - cyan, H - gray, N – blue.

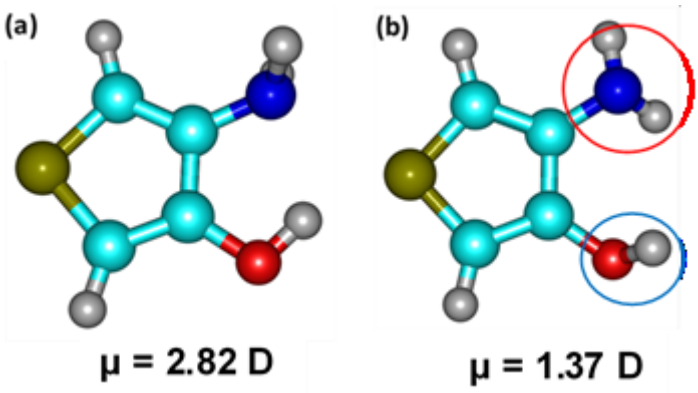


Figure S20. Molecule 32 of the QM7b dataset. (a) DFT optimized geometry and (b) AM1 optimized geometry. Red and blue circles indicate locations of conformation changes after geometry optimization using AM1. Color code: C - cyan, H - gray, N – blue, O – red, S – dark yellow.

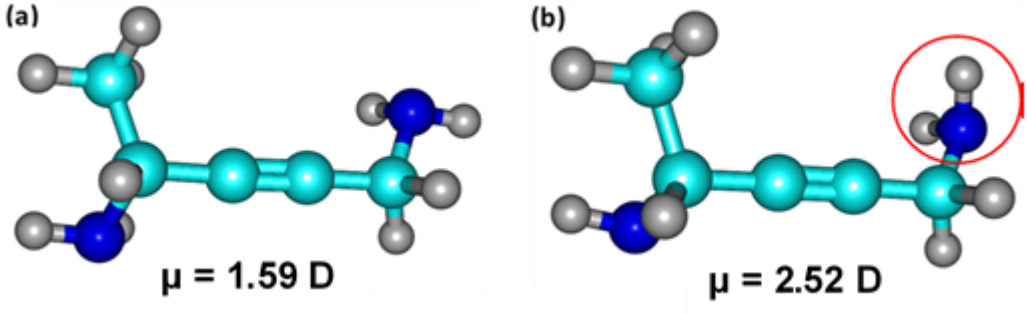


Figure S21. Molecule 55 of the QM7b dataset. (a) DFT optimized geometry and (b) AM1 optimized geometry. Red circle indicates location of conformation change after geometry optimization using AM1. Color code: C - cyan, H - gray, N – blue.

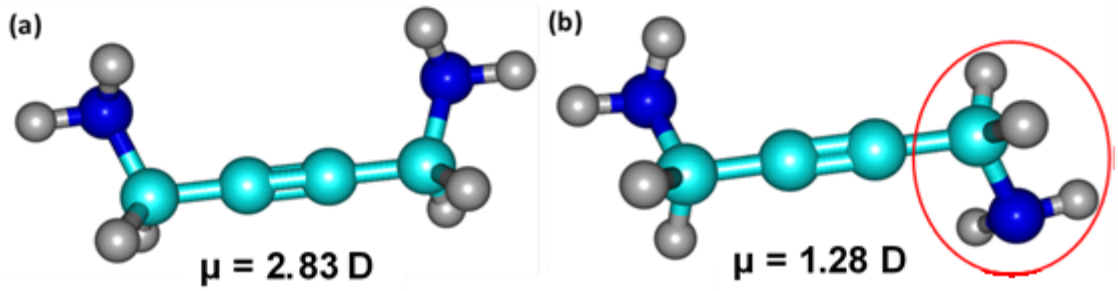


Figure S22. Molecule 324 of the QM7b dataset. (a) DFT optimized geometry and (b) PM3 optimized geometry. Red circle indicates location of conformation change after geometry optimization using PM3. Color code: C - cyan, H - gray, N – blue.

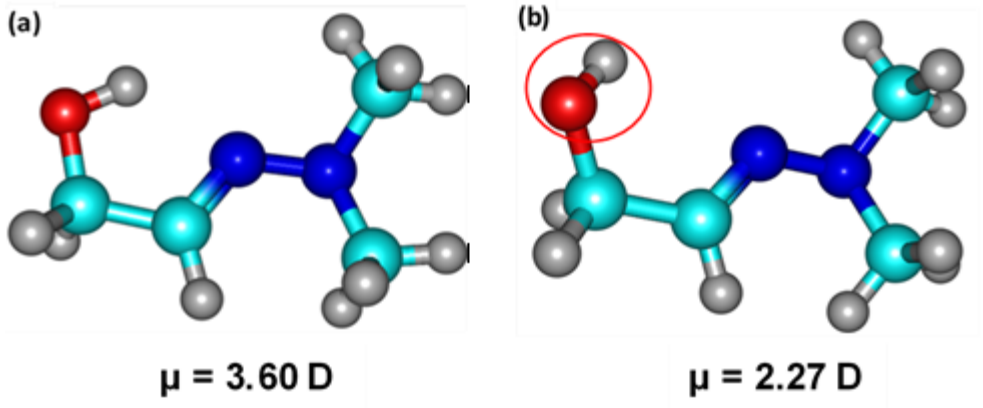


Figure S23. Molecule 12 for the QM7b dataset. (a) DFT optimized geometry and (b) PM3 optimized geometry. Red circle indicates location of conformation change after geometry optimization using PM3. Color code: C - cyan, H - gray, N – blue, O - red.

9. **Probability density distribution of subcategory errors for each SE method normalized by population using SE optimized geometries**

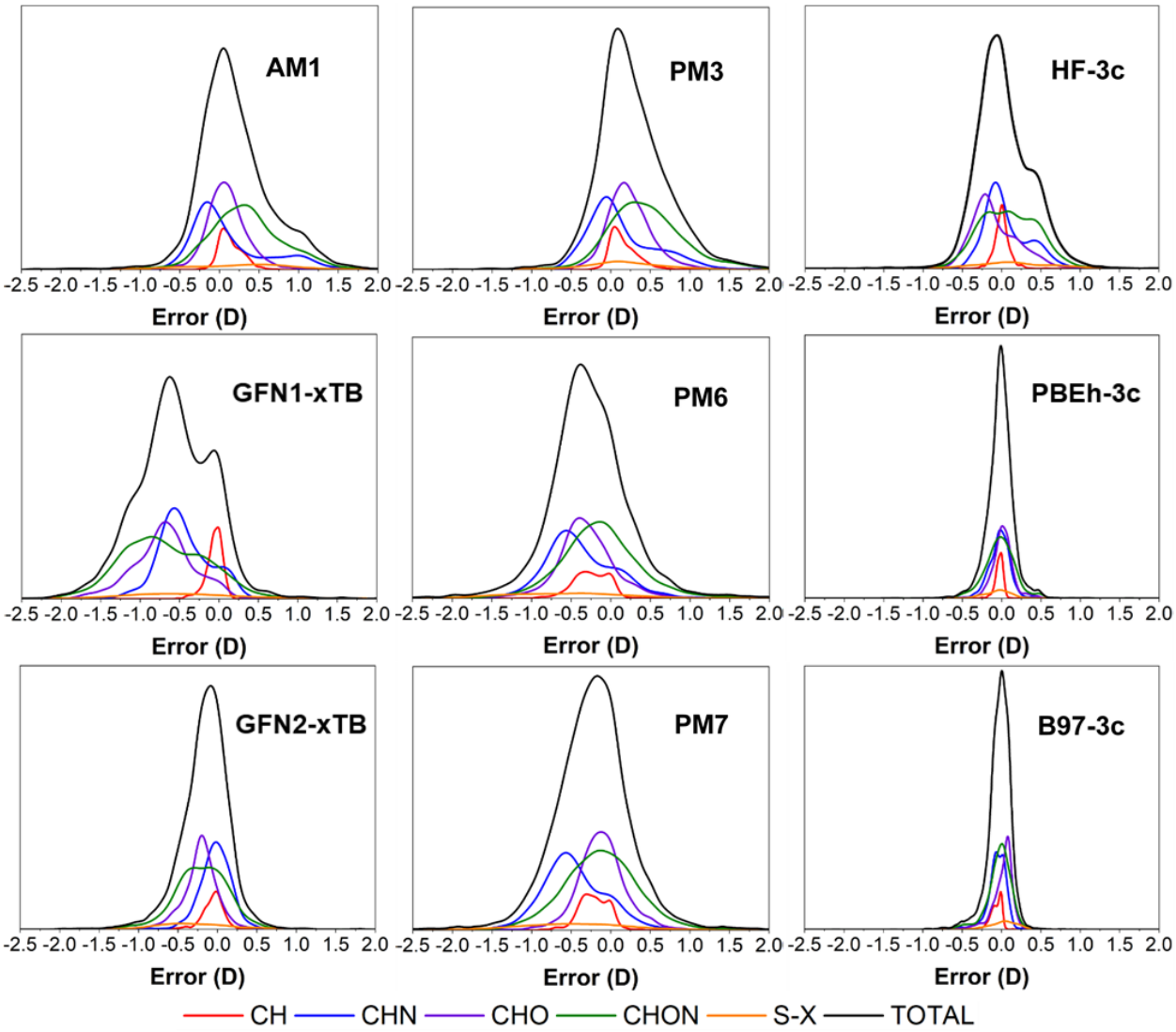


Figure S24. Kernel smooth probability density distribution of subcategory errors for each SE method using SE optimized geometries. The vertical axis of each plot is the probability of obtaining a certain error value. Each subcategory has been normalized by its population.

REFERENCES

1. Pracht, P. C., E.; Ehlert, S.; Grimme, S., A Robust Non-Self-Consistent Tight-Binding Quantum Chemistry Method for Large Molecules. *ChemRxiv,* June 27, **2019**, ver. 1. DOI: 10.26434/chemrxiv.8326202.v1 (accessed 2022-01-11)