Evaluating Fast Methods for Static Polarizabilities on Extended Conjugated Oligomers

Danielle C. Hiener, Dakota L. Folmsbee, Luke A. Langkamp, and Geoffrey R. Hutchison*

Department of Chemistry, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, United States

E-mail: geoffh@pitt.edu

Abstract

Given the importance of accurate polarizability calculations to many chemical applications, coupled with the need for efficiency when calculating the properties of sets of molecules or large oligomers, we present a benchmark study examining possible calculation methods for polarizable materials. We first investigate the accuracy of highly-efficient semi-empirical tight-binding method GFN2-xTB, comparing its polarizability calculations to ωB97XD results for a subset of PubChemQC and a compiled benchmark set of molecules spanning polarizabilities from approximately 3Å^3 to 600 Å^3, with a few compounds in the range of approximately 1200 Å^3-1400 Å^3. Although we find GFN2 to have large errors with polarizability calculations, on large oligomers it would appear a correction factor can remedy this. We also compare the accuracy of DFT polarizability calculations run using basis sets of varying size and level of augmentation, determining that a non-augmented basis set may be used for highly polarizable species in conjunction with a linear correction factor to achieve accuracy extremely close to that of aug-cc-pVTZ.

Introduction

Polarizability plays a key role in many chemical processes and phenomena, and its accurate calculation is therefore crucial to a variety of applications. Because of its fundamental role in explaining dispersion forces,[1] it is a key component of widely-used dispersion corrections for computational calculations.[2] The importance of accurate electrostatic interactions has led to the development and widespread use of polarizable force field models for studying systems such as biomolecules[3] and ionic liquids.[4] Computationally-derived Raman spectra also rely on the calculation of polarizability tensors.[5] Polarizability values are also necessary to calculate the values of more complex material properties such as refractive index[6] and dielectric constant,[7] often in the context of molecular screening.

Because of its wide utility, polarizability has been the topic of a number of recent computational benchmark studies. Hait and Head-Gordon provided a thorough examination of the performance of a large number of density functional theory (DFT) functionals at the complete basis set (CBS) limit for 132 small molecules.[8] Frediani et. al. used a subset of the Head-Gordon study's molecule set to test the veracity of that study's CBS limit claim using alternative multiwavelet bases in order to reduce potential error.[9] Sauer and co-workers used a benchmark set of 14 heteroaromatic molecules to assess the accuracy of various second-order methods for both static and frequency dependent polarizabilities.[10] Afzal and Hachmann tested various DFT methods to determine the best way to balance accuracy and efficiency for high-throughput polymer screening.[11]

While all of these studies provide valuable insight into the relative accuracy of various polarizability methods for different applications, none of them examine such methods for the high polarizability limit. As shown in our previous work using a genetic algorithm (GA) to search for high dielectric oligomers, there is a need for a polarizability method capable of calculating large polarizabilities (on the scale of 10^2 Å^3) while making efficient use of time and computational resources.[7] As a point of reference, all of the molecular species examined by the previously mentioned studies possess isotropic polarizabilities less than 40 Å^3. The Hachmann study suggests the viability of extrapolating polymer polarizabilities from oligomers for non-conjugated species, but notes that this proves untenable for species with conjugated backbones due to electron correlation effects in the π-system.[11] Wong and coworkers performed a polarizability benchmark on oligomers of polydiacetylene and polybutatriene which included some hexamer polarizabilities greater than 200 Å^3.[12] While providing useful data, including the effects of short-range exchange on polarizability and CCSD(T) calculations for these systems, the scope of that study was not intended to examine resource efficient polarizability methods for large oligomers. It is also worth noting that despite augmented basis sets being recommended for accurate polarizability calculations,[13] the basis sets used in the Wong study were not augmented due resource constraints. With the realm of computationally-generated novel materials continuing to grow, data is needed on resource-efficient methods and basis sets to find accurate polarizabilities of largely polarizable molecules.

In this work, we analyze the viability of using a semi-empirical method and smaller basis sets to make large polarizability calculations more tractable. We test the durability of popular semi-empirical tight-binding method GFN2-xTB (GFN2) on highly polarizable species; this is of special interest since atomic polarizabilities are central to the D4 dispersion correction used widely with DFT methods.[2,14] We also examine four basis sets of varying size and level of augmentation to determine whether smaller basis sets can be used to calculate large polarizabilities since they minimize issues regarding computation time and potential linear dependence.

Computational methods

Two primary data sets were analyzed in this benchmark. The first set is a randomly chosen subset of approximately 8,400 species from PubChemQC's approximately 3.2 million known small (molecular weights less than 500 a.u.) molecules.[15] The second "wide [polarizability] range" set is drawn from previous studies and designed to cover a very wide range of polarizabilities. Drawing from the pool of hexamer structures we had created with our GA, we constructed a set of 54 hexamers with GFN2 predicted polarizability values in the approximate range of 80-280 Å^3. In order to balance out our benchmark set, we also added 19 conjugated oligomers and small molecules with GFN2 predicted polarizability values in the "medium polarizability" range of 4-91 Å^3. Hexamer equilibrium geometries were found using preliminary force-field optimization using OpenBabel[16] with MMFF94[17–21] or UFF[22,23] followed by geometry optimization using GFN2.[14] Equilibrium geometries for the medium polarizability molecules were calculated with ORCA 4.0.0.2[24] using DFT with the B3LYP functional[25–28] and 6-31G(d) basis set.[29,30]

All polarizabilities reported in this study are isotropic, meaning they are the average of the diagonal elements of the polarizability tensor. For GFN2 calculations, polarizabilities were calculated using xTB which relies on the D4 method using atomic polarizabilities.[14] DFT calculations were performed in Gaussian 09[31] for the PubChemQC set and ORCA 4.0.0.2.[24]

For GFN2 comparison studies with both the PubChemQC and wide range sets, non-augmented basis sets were chosen with an emphasis on efficiency over absolute accuracy. This was done to estimate how well GFN2 compared generally to DFT, without facing potential resource and/or linear dependence issues likely to arise when using augmented basis sets with large molecules. For the comparison of basis set accuracy, aug-cc-pVTZ[32,33] was selected as the standard of comparison. This basis set was chosen as the standard because diffuse functions are necessary to describe the long-range electron behavior and electron correlation important for polarizability calculations, and it has been shown to perform better than both a similar non-augmented triple-zeta basis set and an augmented double-zeta basis set.[13] Additionally, Sauer and co-workers found using larger augmented basis sets do not yield substantial accuracy increases for polarizability calculations despite the increased time required.[10]

Results and discussion

Due to our interest in finding an efficient method for calculating molecular polarizabilities for novel molecular searches, we sought to test the accuracy of GFN2/D4. As an initial experiment, we calculated the polarizabilities for approximately 8,400 species from the PubChemQC dataset, using both GFN2 and DFT with the ωB97XD functional[34] and the cc-pVTZ basis set. Although this does not allow polarizabilites to be as accurate as when calculated with augmented basis sets, it allowed us to perform thousands of calculations quickly and gave us an initial baseline against which we could compare GFN2. As discussed below, such results can be scaled to augmented basis sets.

Due to the presence of a few outliers, robust linear regression was performed using SciKit learn's Huber regressor method[35,36] with default epsilon value of 1.35 to limit outlier effects. The y-intercept was also forced to zero, representing the physical reality that a completely non-polarizable molecule should be computed to have zero polarizability by any method. After performing Huber linear regression (Figure 1), two notable observations were apparent. While the trendline's slope was very close to one, the values calculated with GFN2 were often substantially lower than those calculated with DFT, with differences as great as over 100 Å^3 between the two methods. This error appears to be somewhat systematic, as species with lower polarizabilities generally have smaller differences in calculated values (Figure 1A) whereas those with high polarizabilities generally have larger differences in calculated values (Figure 1C). A substantial number of species' values appear as outliers from the regression line, suggesting a level of random error in GFN2's calculations.

Highly Polarizable Oligomers

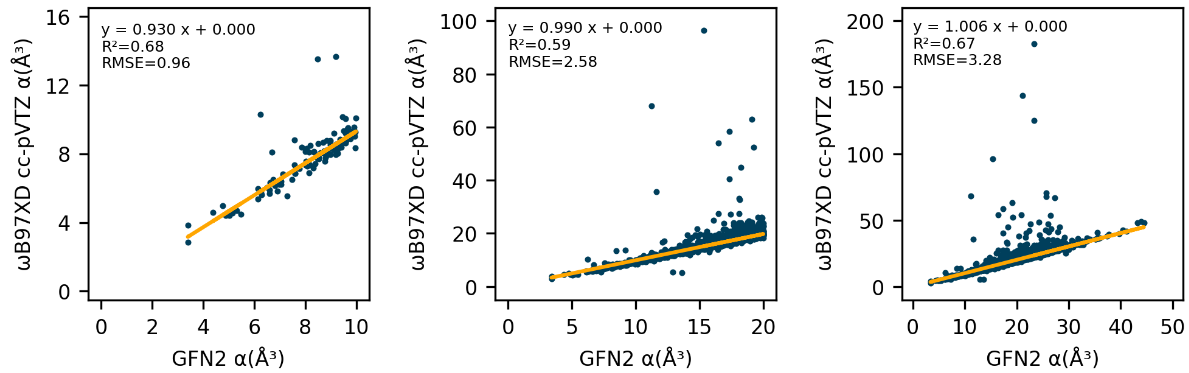

Figure 1: Comparison of PubChemQC polarizabilities calculated with GFN2 to those calcuated with DFT functional ωB97X.

In order to further explore the performance of GFN2 for large polarizability calculations, we pursued testing a smaller group of molecules with a wider range of polarizabilites. Because of its vast speed-up compared to ab initio methods, we previously used GFN2 to calculate the polarizabilities of novel hexamer structures generated by a genetic algorithm

(GA)-driven search for high-dielectric organic conjugated oligomers.[7] For that work, we chose to use GFN2 despite suspecting a degree of inaccuracy at high polarizabilities. The speed-up GFN2 provides over DFT methods was necessary to complete the thousands of hexamer polarizability calculations in a reasonable amount of computation time. We qualified our work by noting that while GFN2 appeared to vastly underestimate large polarizabilities, it appeared to do so in a systematic way, such that the order of the polarizabilities' magnitudes relative to one another was preserved (allowing us to accurately rank molecules by polarizability, as was needed for the GA).

To test the integrity of the GFN2 polarizability results for the 73 member "wide range" benchmark set, we ran single point DFT polarizability calculations using the ωB97X functional[34] and the cc-pVTZ basis set.[32,33,37–39] This functional was chosen because it allowed us to compare the DFT results for the "wide range" set to the previously computed results for the PubChemQC subset, since the former was performed in ORCA 4.0.0.2, which does not have an option for the exact dispersion-corrected functional used in the latter's calculations in Gaussian 09. This basis set was chosen because it was the largest basis set that we were able to reach convergence with for all species in the "wide range" set in a reasonable amount of computation time.

Performing Huber regression with a fixed intercept at the origin on the GFN2 and ωB97X polarizabilities from our "wide range" set (Figure 2), we observed trends similar to those seen in the PubChemQC study. We again observed smaller differences in polarizabilities less than 50 Å^3, with a mean absolute error of 4.72 Å^3 (Figure 2A), and increasingly larger differences as polarizabilities increased, with an MAE of 37.31 Å^3 for polarizabilities less than 100 Å^3(Figure 2B), and an MAE of 145.02Å^3 for the entire set (Figure 2C). We similarly observed greater variation in polarizability differences as their magnitudes grew, shown by the drastically smaller R^2 value for the full set as compared to the subset with values less than 50 Å^3. Three extreme outliers appeared (Figure 2C), in which the percent error of the GFN2 calculated value was in excess of 80%. These outliers were run with the same DFT method and basis set using Gaussian (Table S1), which confirmed the ORCA results and the presence of troubling random errors in GFN2's polarizability calculations. Examining their chemical structures (Figure 3), all notably include sulfur ring systems, suggesting the possibility that this particular motif is not well accounted for by GFN2.

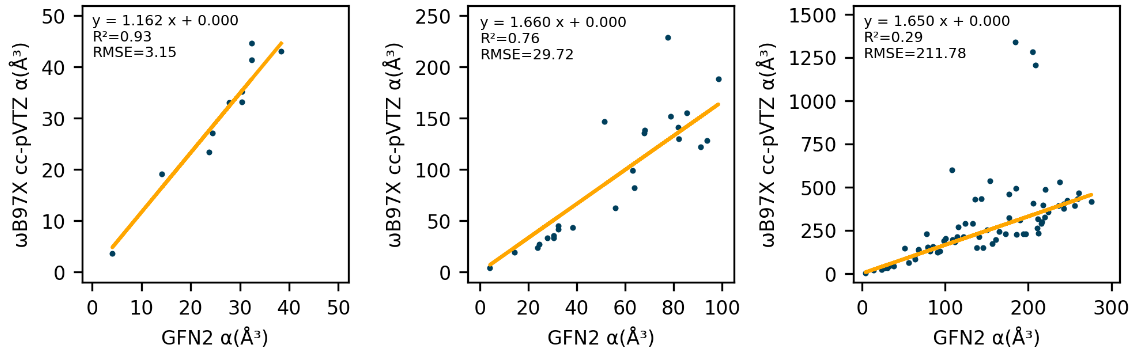

Figure 2: Using the Huber Regressor to perform linear regression robust to outliers (and forcing the y-intercept to 0), GFN2 shows some linear correlation with ωB97X cc-pVTZ for isotropic polarizability calculations.

Investigation of Potential GFN2 Improvement Strategies

As shown in the above assessment, GFN2 performs well when calculating relatively small isotropic polarizabilities, in the range <50Å^3 common for most small molecules. For species with larger polarizabilities, especially for long conjugated systems like the hexamers in the "wide range" set, GFN2 appears to systematically underestimate the polarizability. This derives from its use of an atom-additive polarizability model, neglecting the nonlocal polarizability-enhancing effects of electron delocalization. Our results suggest a correction is needed to allow GFN2 to more accurately predict larger polarizabilities.

We began by constructing an additive polarizability model as a baseline comparison for GFN2's performance. For this model, each atom in a molecule was assigned its GAFF atom type,[40] which was used to further assign it to an Alexandria polarizability type and then finally a corresponding atomic polarizability.[41] The molecular polarizability was calculated as the simple sum of these atomic polarizabilities. We found that the additive model performed

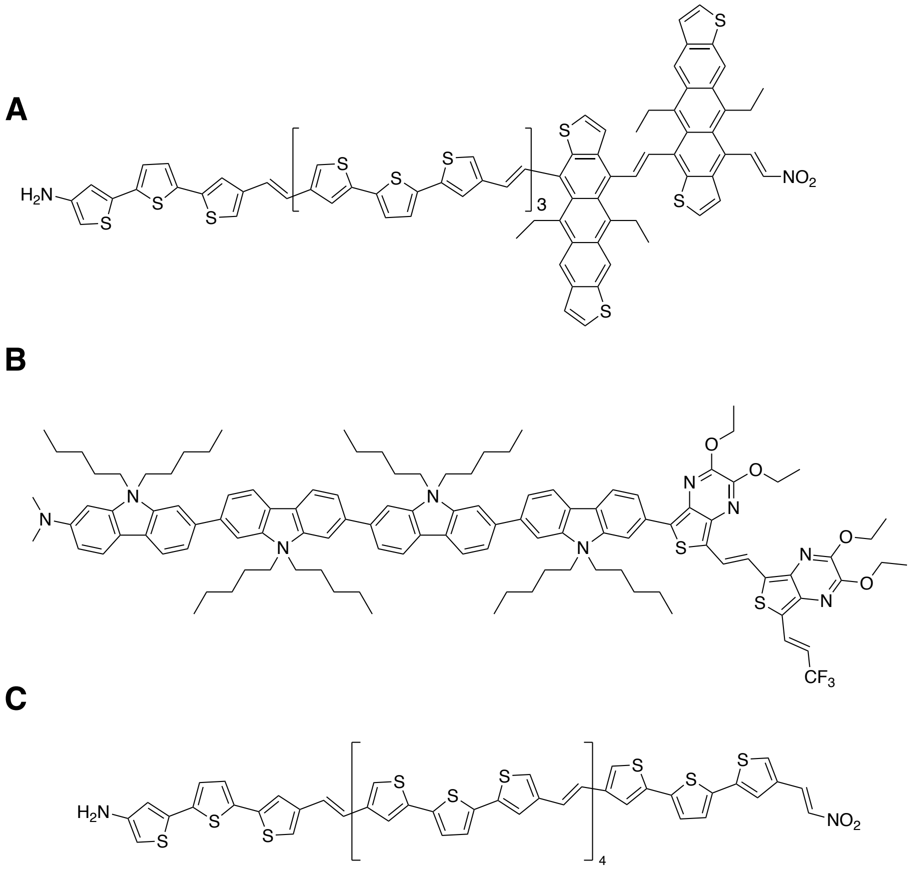

Figure 3: Chemical structures of the hexamers seen as the three major outliers in Figure 2C.

extremely similarly to GFN2 (Figure S1), providing computed polarizabilities with high correlation with GFN2/D4 (Figure S2).

Given GFN2's systematically increasing inaccuracy for large polarizability calculations, we considered additional chemical properties related to electron delocalization that could potentially correct the additive GFN2 polarizability. First, we examined the GFN2-computed HOMO-LUMO gaps for both the PubChemQC subset and the hexamers from the wide range polarizability set. Unfortunately, there was not a useful correlation between the calculated polarizability and molecular HOMO-LUMO gap (Figure 4A). We then used an empirical descriptor of the geometric size of the largest conjugated π-system.[42,43] While better correlated to GFN2 polarizability than HOMO-LUMO gap, this information was not enough to meaningfully correct large polarizabilities (Figure 4B).

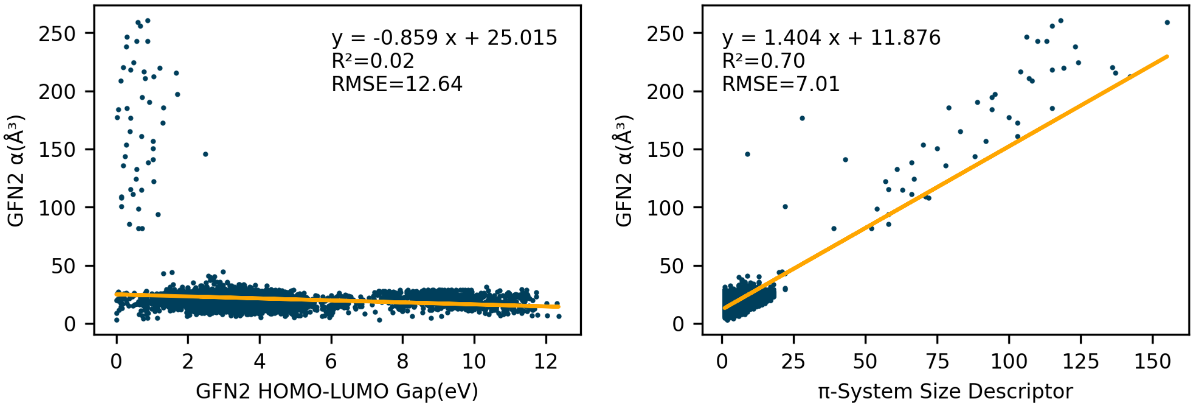

Figure 4: Linear regression demonstrates the lack of useful correlation between GFN2 calculated polarizabilities and both GFN2 calculated HOMO-LUMO gap (A) and a π-system size descriptor (B).

Plotting DFT polarizabilities against GFN2 polarizabilities for the PubChemQC subset and the "wide range" set, we examine the effects of using a polynomial fit. As an aside, because the PubChemQC subset and "wide range" sets were computed at different times using slightly different methods, with Gaussian with a dispersion correction and with ORCA without a dispersion correction, respectively, the functional has been labeled ωb97X(D) here to indicate that for part of the data set a dispersion correction was used. We do not believe the dispersion correction or program makes a meaningful difference in this case, as shown by the low MAE demonstrated for a sample of PubChemQC species in Table S2, and therefore the PubChemQC and "wide range" results may be grouped together and treated as one large dataset. We note that a quadratic fit provides a better correlation description than a linear fit, where the former has a MAE of 2.47 Å^3 compared to ωB97X(D), while the latter has an MAE of 7.94Å^3 compared to ωbB7X(D) (Figure 5). Although not as physically meaningful as an adjustment based on a related molecular property, we find a quadratic fit with zero intercept, and note the linear coefficient remains close to unity.

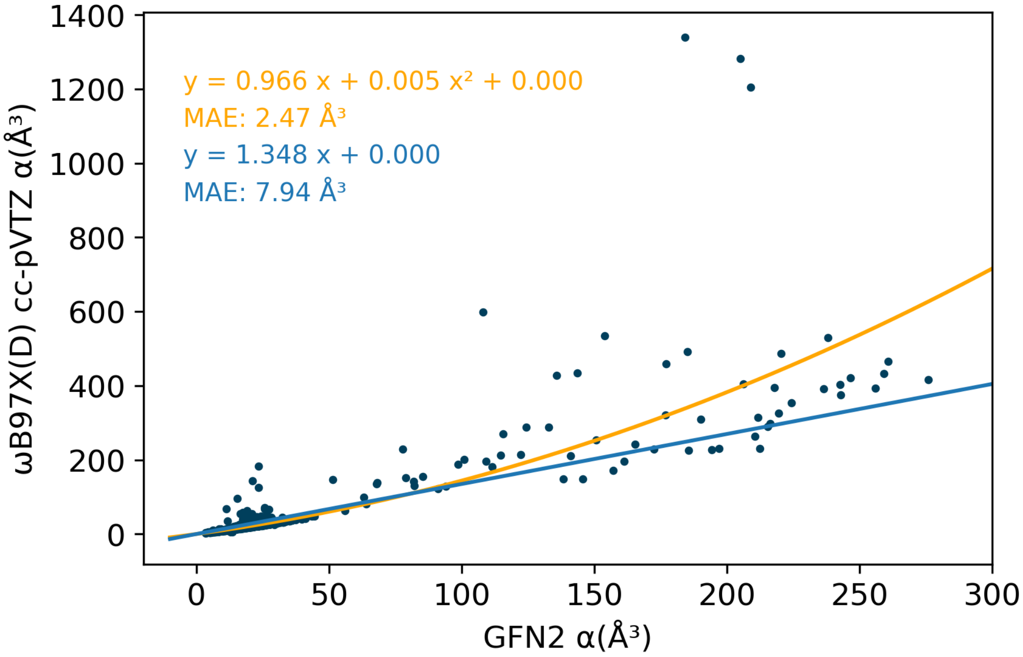

Figure 5: Linear and quadratic regression are performed on the combined PubChemQC subset and "wide range" set.

Basis Set Comparison

While an augmented basis set is ideal for polarizability calculations, running calculations with a large basis set for the larger species in the "wide range" set presented convergence issues. Both the large amount of computation time needed and the possibility of linear dependence concerns led us to choose a non-augmented triple-zeta set for our comparative DFT calculations above. Because we were interested in both the magnitude of the increase in accuracy provided by diffuse basis functions and the correlation between polarizabilities calculated with different basis sets, we ran the "wide range" set subset of low to medium polarizability species with four different basis sets for comparison. The sets we used were cc-pVDZ,[32,33,37–39] cc-pVTZ,[32,33,37–39] jun-cc-pVTZ,[44] and aug-cc-pVTZ,[32,33] the latter two providing increasing amounts of diffuse functions. Pairwise comparison of increasingly accurate basis sets (Figure 6) reveals incredibly linear correlations, with simple linear regression analysis showing slopes close to one and an R^2 value of 1.00 for all three comparisons. In summary, while differences in computed polarizabilities exist using larger and augmented basis sets, across a wide range of molecular polarizabilities, such effects appear small.

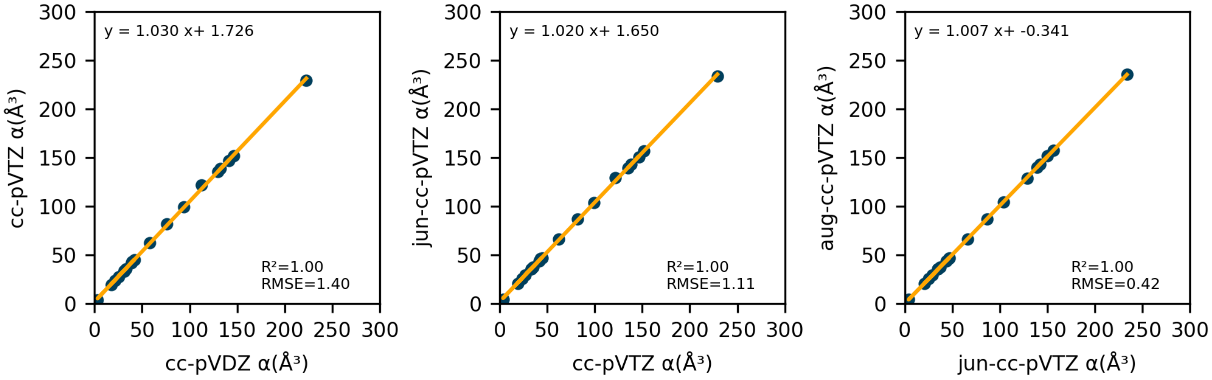

Figure 6: Linear regression is performed on isotropic polarizabilities calculated with systematically increasing basis set size for species less than 250 Å^3.

Comparing each smaller basis set to the largest set considered, aug-cc-pVTZ, we see similar results to the increasing pairwise comparison (Table 1). Again, simple linear regression analysis reveals slopes close to one and R^2 values of or nearly 1.00, even when comparing the largest basis set to a non-augmented double zeta basis set. The speed-ups are also worth noting, since even using a partially-augmented basis set (jun-cc-pVTZ) provides over a 2x speedup over the traditionally augmented set. Timings are shown in more detail in the box plot in Figure 7, where the range of calculation times for each basis set is shown to decrease substantially as the sets become smaller.

Table 1: Comparison of Smaller Basis Sets to aug-cc-pVTZ

Basis Set	Linear Regression Line	R^2	RMSE	Speed-up
cc-pVDZ	y = 1.059 x + 3.108	0.999	2.163	63.211
cc-pVTZ	y = 1.028 x + 1.312	1.000	0.755	5.941
jun-cc-pVTZ	y = 1.007 x - 0.341	1.000	0.424	2.213

The large speed-ups provided by non or partially augmented basis sets, combined with lower risk of linear dependence issues, make them better, albeit less accurate, choices for polarizability calculations for large systems. The linearity between systematically larger basis sets suggests that for species with large polarizabilities, the increase in accuracy of the magnitude of the polarizability is not substantial, and that a simple linear correlation coeffi-

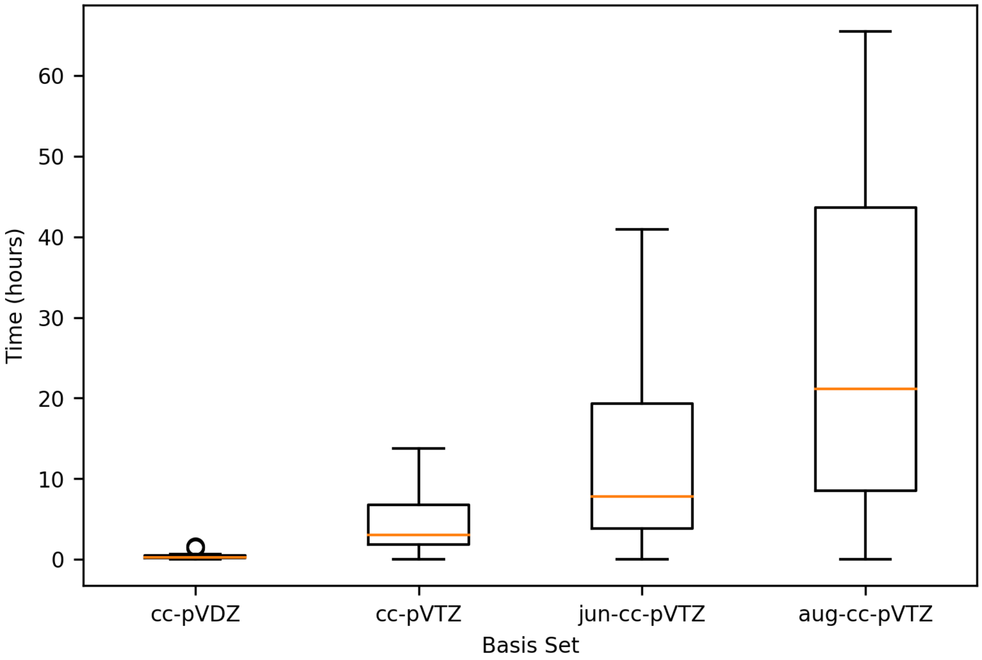

Figure 7: Mean CPU time and the range of CPU time distribution are shown for systematically larger basis set, displaying dramatic increases in both as basis sets become larger.

cient could be used to correct large polarizabilities found with smaller basis sets. Given the increased RMSE observed with cc-pVDZ, we suggest that for routine use on large molecules, non-augmented triple zeta basis sets, as used here, are an efficient balance of time and accuracy.

Conclusion

Based on our studies, GFN2 appears best parameterized for species with polarizabilities less than 50Å^3. In its current implementation GFN2 does not compare favorably to DFT-computed polarizabilities in highly polarizable oligomers. The method's underestimation of polarizability values, which systematically grows as polarizability increases, indicates that the application of a quadratic scaling correction factor could provide a relatively simple solution to drastically improve the accuracy of large polarizability calculations. The presence of three significant outliers in the GFN2 comparison data, all containing similar sulfur motifs, suggests the need to examine GFN2's parameterization for such chemical structures. Beyond increasing GFN2's usefulness for efficient polarizability calculations for large polarizability molecular screening applications, these improvements would notably also improve the accuracy of the D4 dispersion method for large π-conjugated species.

With regard to the accuracy and efficiency of basis sets, our study suggests that using smaller, even non-augmented basis sets to save time and resources is appropriate for large polarizability calculations. The substantial linear correlations seen between methods of varying size and levels of augmentation suggests that using a basis set such as cc-pVTZ with a linear scaling factor is appropriate for large polarizability molecules. By calculating polarizability in this manner for molecules with polarizabilities over 200Å^3, we believe accuracies near the level of those achieved with an aug-cc-pVTZ basis set are attainable at nearly a six-fold speed-up and without the convergence issues we faced when attempting to use this basis set to calculate polarizabilities in this range. Using a basis set with a linear correction factor opens up the possibility of calculating highly accurate polarizabilities for increasingly large molecules using conventional DFT methods. Additional work will need to be done to test the limits of the polarizability magnitudes that can be accurately calculated in this manner.

We hope that the results of this study aid work where the calculation of large polarizability values is crucial. While GFN2 is not currently fit to provide accurate calculations for large polarizability values, we hope that after some minor corrections it will be a viable method for such applications and improve the accuracy of future dispersion correction methods. Considering it's incredible efficiency, this would provide a valuable tool for future molecular screening studies of highly polarizable materials. Meanwhile, using lower-cost non-augmented basis sets with a correction factor vastly increases the number of potential molecular species for which highly accurate polarizabilities can be now obtained.

Acknowledgement

We acknowledge support from Department of Energy-Basic Energy Sciences Computational and Theoretical Chemistry (Award DE-SC0019335) and the University of Pittsburgh Center for Research Computing through the computational resources provided.

Supporting Information Available

Verification of outlier polarizabilities, analysis of additive and GFN2 / D4 polarizabilities, comparison between polarizabilities calculated with Gaussian and ORCA are included in the supporting information. This information is available free of charge via the Internet at `https://pubs.acs.org`

Full data files and analysis notebooks are available at `https://github.com/hutchisonlab/conjugated-polarizability`

References

(1) Hermann, J.; Jr., R. A. D.; Tkatchenko, A. First-Principles Models for van der Waals Interactions in Molecules and Materials: Concepts, Theory, and Applications. *Chemical Reviews* **2017**, *117*, 4714–4758.

(2) Caldeweyher, E.; Ehlert, S.; Hansen, A.; Neugebauer, H.; Spicher, S.; Bannwarth, C.; Grimme, S. A generally applicable atomic-charge dependent London dispersion correction. *Journal of Chemical Physics* **2019**, *150*, 154122.

(3) Jing, Z.; Liu, C.; Cheng, S. Y.; Qi, R.; Walker, B. D.; Piquemal, J.-P.; Ren, P. Polarizable Force Fields for Biomolecular Simulations: Recent Advances and Applications. *Annual Review of Biophysics* **2019**, *48*, 371–394.

(4) Bedrov, D.; Piquemal, J.-P.; Borodin, O.; Jr., A. D. M.; Roux, B.; Schröder, C. Molec-

ular Dynamics Simulations of Ionic Liquids and Electrolytes Using Polarizable Force Fields. *Chemical Reviews* **2019**, *119*, 7940–7995.

(5) Porezag, D.; Pederson, M. R. Infrared intensities and Raman-scattering activities within density-functional theory. *Physical Review B* **1996**, *54*, 7830–7836.

(6) Afzal, M. A. F.; Haghighatlari, M.; Ganesh, S. P.; Cheng, C.; Hachmann, J. Accelerated Discovery of High-Refractive-Index Polyimides via First-Principles Molecular Modeling, Virtual High-Throughput Screening, and Data Mining. *Journal of Physical Chemistry C* **2019**, *123*, 14610–14618.

(7) Hiener, D.; Hutchison, G. *Pareto Optimization of Oligomer Polarizability and Dipole Moment using a Genetic Algorithm*; 2021.

(8) Hait, D.; Head-Gordon, M. How accurate are static polarizability predictions from density functional theory? An assessment over 132 species at equilibrium geometry. *Physical Chemistry Chemical Physics* **2018**, *20*, 19800–19810.

(9) Brakestad, A.; Jensen, S. R.; Wind, P.; D'Alessandro, M.; Genovese, L.; Hopmann, K. H.; Frediani, L. Static Polarizabilities at the Basis Set Limit: A Benchmark of 124 Species. *Journal of Chemical Theory and Computation* **2020**, *16*, 4874–4882.

(10) Jørgensen, M. W.; Faber, R.; Ligabue, A.; Sauer, S. P. A. Benchmarking Correlated Methods for Frequency-Dependent Polarizabilities: Aromatic Molecules with the CC3, CCSD, CC2, SOPPA, SOPPA(CC2), and SOPPA(CCSD) Methods. *Journal of Chemical Theory and Computation* **2020**, *16*, 3006–3018.

(11) Afzal, M. A. F.; Hachmann, J. Benchmarking DFT approaches for the calculation of polarizability inputs for refractive index predictions in organic polymers. *Physical Chemistry Chemical Physics* **2019**, *21*, 4452–4460.

(12) Oviedo, M. B.; Ilawe, N. V.; Wong, B. M. Polarizabilities of π-Conjugated Chains Revisited: Improved Results from Broken-Symmetry Range-Separated DFT and New CCSD(T) Benchmarks. *Journal of Chemical Theory and Computation* **2016**, *12*, 3593–3602.

(13) Hickey, A. L.; Rowley, C. N. Benchmarking Quantum Chemical Methods for the Calculation of Molecular Dipole Moments and Polarizabilities. *Journal of Physical Chemistry A* **2014**, *118*, 3678–3687.

(14) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *Journal of Chemical Theory and Computation* **2019**, *15*, 1652–1671.

(15) Nakata, M.; Shimazaki, T. PubChemQC Project: A Large-Scale First-Principles Electronic Structure Database for Data-Driven Chemistry. *Journal of Chemical Information and Modeling* **2017**, *57*, 1300–1308.

(16) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *Journal of Cheminformatics* **2011**, *3*, 33.

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

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

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

(20) Halgren, T. A. Merck molecular force field. IV. conformational energies and geometries for MMFF94. *Journal of Computational Chemistry* **1996**, *17*, 587–615.

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

(22) Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; III, W. A. G.; Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *Journal of the American Chemical Society* **1992**, *114*, 10024–10035.

(23) Casewit, C. J.; Colwell, K. S.; Rappe, A. K. Application of a universal force field to organic molecules. *Journal of the American Chemical Society* **1992**, *114*, 10035–10046.

(24) Neese, F. The ORCA program system. *WIRES Computational Molecular Science* **2012**, *2*, 73–78.

(25) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Physical Review B* **1988**, *37*, 785–789.

(26) Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Physical Review A* **1988**, *38*, 3098–3100.

(27) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *Journal of Physical Chemistry* **1994**, *98*, 11623–11627.

(28) Vosko, S. H.; Wilk, L.; Nusair, M. Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis. *Canadian Journal of Physics* **1980**, *58*, 1200–1211.

(29) Petersson, G. A.; Bennett, A.; Tensfeldt, T. G.; Al-Laham, M. A.; Shirley, W. A. A

complete basis set model chemistry. I. The total energies of closed-shell atoms and hydrides of the first-row elements. *Journal of Chemical Physics* **1988**, *89*, 2193–2218.

(30) Petersson, G. A.; Al-Laham, M. A. A complete basis set model chemistry. II. Open-shell systems and the total energies of the first-row atoms. *Journal of Chemical Physics* **1991**, *94*, 6081–6090.

(31) 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 09. Gaussian, Inc.: Wallingford, CT, 2013.

(32) Kendall, R. A.; Jr., T. H. D. Electron affinities of the first-row atoms revisited. Systematic basis sets and wave functions. *Journal of Chemical Physics* **1992**, *96*, 6796–6806.

(33) Woon, D. E.; Jr., T. H. D. Gaussian basis sets for use in correlated molecular calculations. III. The atoms aluminum through argon. *Journal of Chemical Physics* **1993**, *98*, 1358–1371.

(34) Chai, J.-D.; Head-Gordon, M. Systematic optimization of long-range corrected hybrid density functionals. *Journal of Chemical Physics* **2008**, *128*, 084106.

(35) Huber, P. J.; Ronchetti, E. M. *Robust Statistics*, 2nd ed.; Wiley: Hoboken, New Jersey, 2009.

(36) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

(37) Jr., T. H. D. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *Journal of Chemical Physics* **1988**, *90*, 1007–1023.

(38) Peterson, K. A.; Woon, D. E.; Jr., T. H. D. Benchmark calculations with correlated molecular wave functions. IV. The classical barrier height of the H+H2→H2+H reaction. *Journal of Chemical Physics* **1994**, *100*, 7410–7415.

(39) K.Wilson, A.; Mourik, T.; Jr., T. H. Gaussian basis sets for use in correlated molecular calculations. VI. Sextuple zeta correlation consistent basis sets for boron through neon. *Journal of Molecular Structure: THEOCHEM* **1996**, *388*, 339–349.

(40) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *Journal of Computational Chemistry* **2004**, *25*, 1157–1174.

(41) Ghahremanpour, M. M.; van Maaren, P. J.; van der Spoel, D. The Alexandria library, a quantum-chemical database of molecular properties for force field development. *Scientific Data* **2018**, *5*, 180062.

(42) Abarbanel, O. D.; Hutchison, G. R. Machine learning to accelerate screening for Marcus reorganization energies. *The Journal of Chemical Physics* **2021**, *155*, 054106.

(43) Abarbanel, O. D.; Hutchison, G. R. Reorganization Energy. 2021; `https://github.com/hutchisonlab/ReorganizationEnergy`.

(44) Papajak, E.; Zheng, J.; Xu, X.; Leverentz, H. R.; Truhlar, D. G. Perspectives on Basis Sets Beautiful: Seasonal Plantings of Diffuse Basis Functions. *Journal of Chemical Theory and Computation* **2011**, *7*, 3027–3034.

Supplementary Information

Table S1: Outlier Polarizability Comparison

Hexamer ID	GFN2 (A^3)	ORCA (A^3)	Gaussian (A^3)
B:1073-1156-000011	205.010	1281.299	1278.897
C:520-515-111111	184.176	1339.674	1339.809

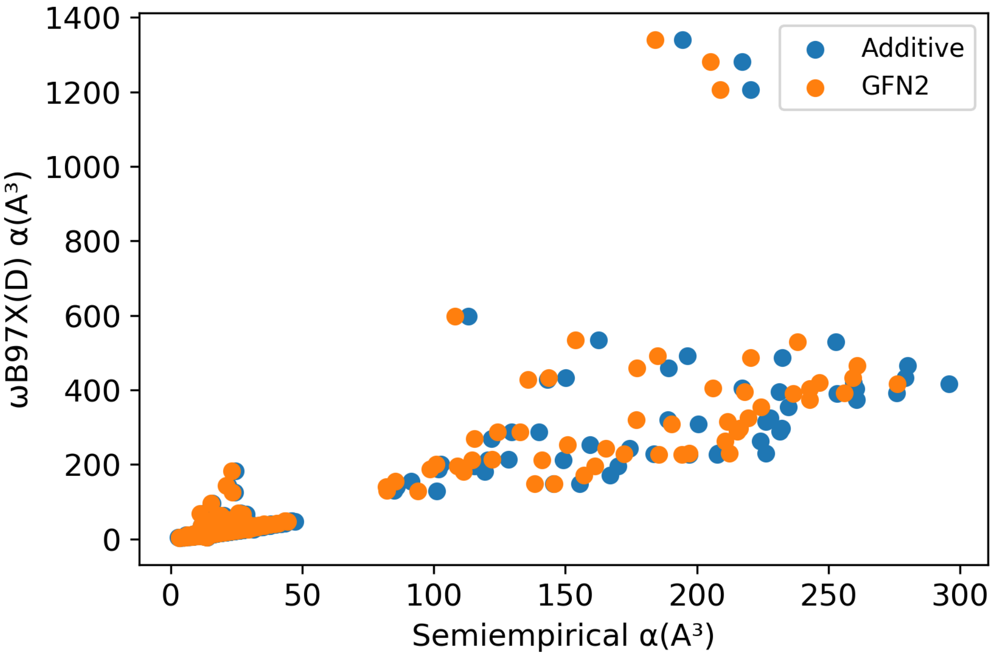

Figure S1: Comparison of additive model and GFN2 polarizabilities to DFT results.

Table S2: Polarizability Method Comparison for PubChemQC Sample

PubChemQC ID	Gaussian ωB97XD (Å^3)	ORCA ωB97X (Å^3)
10844636	15.8972868	15.78617472
19800353	31.9664682	31.60996769
21545508	23.31098235	23.09620301
65163876	26.31024675	26.07089537

The MAE comparing ORCA ωB97X to Gaussian ωB97XD for this sample of the PubChemQC subset is 0.184Å^3, which equates to approximately 0.7% error on the scale of this sample set.

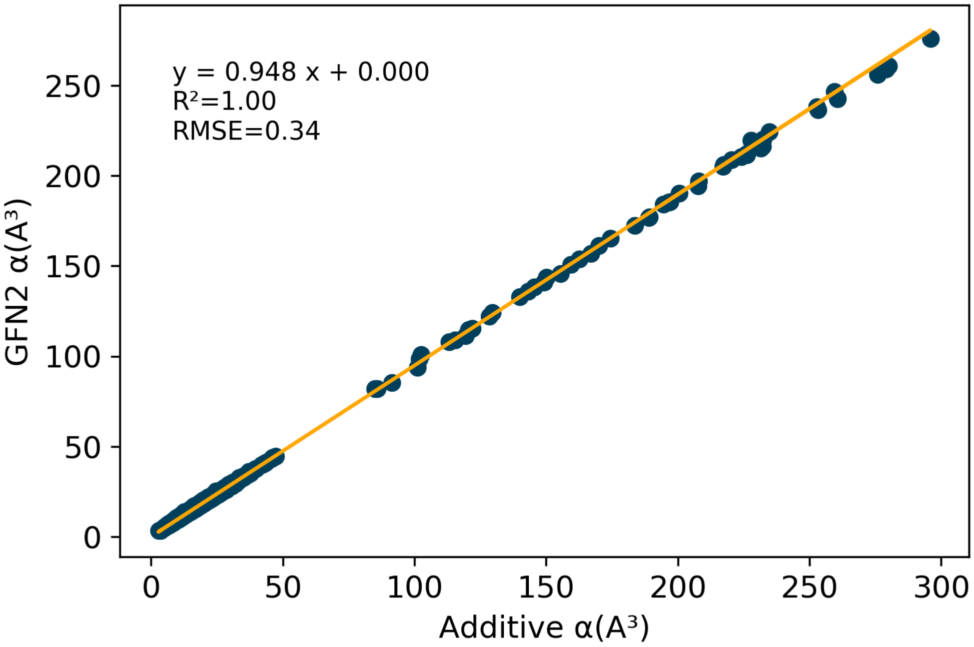

Figure S2: Strong correlation is shown between GFN2 and additive model polarizabilites.

Graphical TOC Entry