PCCP

Evaluating Fast Methods for Static Polarizabilities on Extended Conjugated Oligomers

Journal:	*Physical Chemistry Chemical Physics*
Manuscript ID	CP-ART-05-2022-002375.R1
Article Type:	Paper
Date Submitted by the Author:	19-Aug-2022
Complete List of Authors:	Hiener, Danielle; University of Pittsburgh, Department of Chemistry Folmsbee, Dakota; University of Pittsburgh, Department of Chemistry Langkamp, Luke; University of Pittsburgh, Department of Chemistry Hutchison, Geoffrey; University of Pittsburgh, Department of Chemistry

PCCP

ARTICLE TYPE

Cite this: DOI: 00.0000/xxxxxxxxxx

Evaluating Fast Methods for Static Polarizabilities on Extended Conjugated Oligomers[†]

Danielle C. Hiener,[a] Dakota L. Folmsbee,[a] Luke A. Langkamp,[a] and Geoffrey R. Hutchison[a,b]

Received Date
Accepted Date

DOI: 00.0000/xxxxxxxxxx

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 the additive model used in GFN2, a highly-efficient semi-empirical tight-binding method, and the D4 dispersion model, comparing its predicted additive polarizabilities 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 some compounds in the range of approximately 1200 Å^3-1400 Å^3. Although we find additive GFN2 polarizabilities, and thus D4, to have large errors with polarizability calculations on large conjugated oligomers, it would appear an empirical quadratic correction can largely 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 large, highly polarizable species in conjunction with a linear correction factor to achieve accuracy extremely close to that of aug-cc-pVTZ.

1 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[3,4] and ionic liquids.[5] Computationally-derived Raman spectra also rely on the calculation of polarizability tensors.[6] Polarizability values are also necessary to calculate the values of more complex material properties such as refractive index[7] and dielectric constant[8], 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.[9] 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.[10] 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.[11] Afzal and Hachmann tested various DFT methods to determine the best way to balance accuracy and efficiency for high-throughput non-conjugated polymer screening.[12]

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.[8] 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.[12] Wong and coworkers performed a polarizability benchmark on oligomers of polydiacetylene and polybutatriene which included some hexamer polarizabilities greater than 200 Å^3.[13] It is worth noting that despite augmented basis sets being recom-

[a] University of Pittsburgh, Department of Chemistry, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, United States
[b] University of Pittsburgh, Department of Chemical and Petroleum Engineering, University of Pittsburgh, Pittsburgh, Pennsylvania, 15261, United States

[†] Electronic Supplementary Information (ESI) available: Comparisons of outlier polarizabilities, comparison of simple additive model, comparison between Gaussian and Orca polarizabilities, and comparison of sTDA-XTB polarizabilities and timings.

mended for accurate polarizability calculations[14], the basis sets used in this study were not augmented due resource constraints. In another benchmark study from Wong, linear polarizabilities (and hyperpolarizabilities) were found for a range of streptocyanine oligomers.[15] By using CCSD(T)-F12 which enhances basis set convergence, they were able to give a basis set extrapolation of computed polarizabilities from a non-augmented triple zeta basis set. Both studies provide useful data, including the effects of short-range exchange on polarizability and CCSD(T) calculations in the former and an assessment of MP2 quality and the importance of potential lower-energy open-shell states in the latter, however the scope of neither study was intended to examine resource efficient polarizability methods for large oligomers. 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 the common additive polarizability model in the popular semi-empirical tight-binding method GFN2-xTB (GFN2) and the connected the D4 dispersion model, on both small organic molecules from PubChemQC[16] for which we expect GFN2/D4 to perform well, and highly polarizable conjugated species, for which we expect some degree of inaccuracy. Testing on a set of small organic molecules gives us a point of comparison which allows us to precisely identify GFN2/D4's strengths and limitations when it comes to polarizability calculations for larger extended systems. While it is understood that GFN2/D4 is empirically tuned to most accurately calculate the polarizabilities of relatively small molecules, testing this method on larger conjugated systems is of special interest since the same basic method used to compute GFN2 polarizabilities is also central to the D4 dispersion correction used widely with DFT methods on species of a wide variety of sizes and chemical structures.[2,17]

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. Diffuse functions are generally accepted as necessary to describe the long-range electron behavior and electron correlation important for polarizability calculations, demonstrated in by Rowley and co-workers' assessment that augmenting basis sets with diffuse functions leads to a substantial increase in accuracy, particularly for polarizability calculations.[14] Earlier works by Dykstra and Clementi found that while augmented basis functions were critical to accurate polarizabilities in small molecules, this effect generally diminishes as the molecule size increases, likely due to the the relatively great importance of intraatomic polarization in small systems.[18–20] The solutions proposed by these works generally involve finding the optimal number and type of diffuse functions necessary for accurate polarizability calculations on a given type/size of molecular system. We wanted extend this a step further, testing the correlation between a standard augmented basis set and several non-augmented and partially-augmented basis sets to determine if smaller basis sets could be used universally with some type of correction factor, saving compute time and the effort of selecting a particular basis set for each chemical system.

2 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.) organic molecules.[16] It was chosen to provide a strong basis of comparison as a set of molecules for which we expected additive GFN2/D4 to perform well. 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/D4 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/D4 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[21] with MMFF94[22–26] or UFF[27,28] followed by geometry optimization using GFN2.[17] Equilibrium geometries for the medium polarizability molecules were optimized with ORCA 4.0.0.2[29] using DFT with the B3LYP functional[30–33] and 6-31G(d) basis set.[34,35]

All polarizabilities reported in this study are isotropic, meaning they are the average of the diagonal elements of the polarizability tensor. We chose to focus on static, isotropic polarizabilities for this study due to their general applicability to a variety of theoretical and computational fields, including method development, dispersion correction, and polarizable electrostatics. For GFN2 calculations, polarizabilities were calculated using xTB which relies on the D4 method in which polarizabilities are calculated using a weighted sum of precomputed atomic polarizabilities.[2,17] DFT calculations were performed in Gaussian 09[36] for the PubChemQC set and ORCA 4.0.0.2[29] for the wide range set, both analytical derivatives and coupled-perturbed equations.

For GFN2/D4 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[37,38] was selected as the standard of comparison. As noted in the introduction, diffuse functions are generally considered necessary to describe the long-range electron behavior and electron correlation important for polarizability calculations. This particular basis set was chosen because Rowley and co-workers determined that aug-cc-pVTZ performed better than both a similar non-augmented triple-zeta basis set and an augmented double-zeta basis set.[14] 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.[11]

3 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 the additive GFN2/D4 model on common small molecules. As an initial experiment, we calculated the polarizabilities for approximately 8,400 species from the PubChemQC dataset, using both GFN2/D4 and DFT with the ωB97XD functional[39] 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/D4. 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[40,41] 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/D4 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/D4 calculations.

3.1 Highly Polarizable Oligomers

In order to further explore the performance of GFN2/D4 for large polarizability calculations, we pursued testing a smaller group of molecules with a wider range of polarizabilites. While we are aware that because of its minimal basis set approach GFN2 is best geared toward polarizability calculations for small molecules, we believe it is important to test its performance on a range of species, including conjugated oligomers. We are interested in the viability of GFN2/D4 additive polarizabilities for two primary reasons: first, to test them as resource-inexpensive albeit relatively low accuracy calculations that allow for bulk molecular screening. For example, because of its vast speed-up compared to *ab initio* methods, we previously used GFN2/D4 to calculate the polarizabilities of novel hexamer structures generated by a genetic algorithm (GA)-driven search for high-dielectric organic conjugated oligomers.[8] For that work, we chose to use GFN2/D4 despite suspecting a degree of inaccuracy at high polarizabilities. The speed-up GFN2/D4 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/D4 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). The second reason we believe it is both relevant and important to test GFN2's ability to calculate polarizabilities for a wide range of chemical species is that shares the additive polarizability model with the D4 disperson-correction method, which is widely used for a variety of species including those not limited to the types of small molecules for which GFN2 has been calibrated.

To test the integrity of the GFN2/D4 polarizability results for the 73 member "wide range" benchmark set, we ran single point DFT polarizability calculations using the ωB97X functional[39] and the cc-pVTZ basis set.[37,38,42–44] Although the accuracy of certain DFT methods, such as GGA functional LC-BLYP, has been noted to be somewhat poorer for polarizability than coupled cluster or MP2 method in the past[45], we believed that due to the large size of some of the molecules in our set DFT would scale better than either a coupled cluster or MP2 method. Additionally, in the polarizability benchmark by Hait and Head-Gordon[9], hybrid GGAs like ωB97X-D are noted as performing well for polarizability calculations with low RMS relative errors (RMSREs). For reference, that benchmark reported an RMSRE across its 132 member dataset of 5.18 for ωB97X-D, 11.12 for MP2, and 12.14 for BLYP. This particular hybrid GGA 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/D4 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/D4 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), they all contain sulfur ring systems, so it is possible that this particular motif can cause problems for polarizability calculations with GFN2/D4. We compared the chemical similarity of each of the outliers to the rest of the wide range set using Tanimoto coefficients (Tables S2, S3, and S4). Interestingly, outliers A and C were most similar to each other with a Tanimoto coefficient of 0.535, and then to a lesser degree to other hexamers containing the same terthiophene monomer unit. Given that the other hexamers containing this terthiophene monomer did not have similar polarizability calculation issues, it is not immediately clear why these two structures were outliers. Outlier B was not especially similar to other members of the

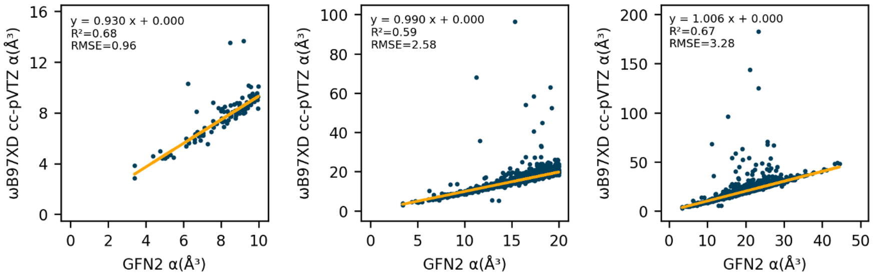

Fig. 1 Comparison of PubChemQC polarizabilities calculated with GFN2/D4 to those calcuated with DFT functional ωB97X.

wide range set, with the highest Tanimoto coefficient between it and another molecule being 0.324. More detailed investigation is needed to examine the exact cause of these structural outliers, but that is beyond the scope of this work.

3.2 Investigation of Potential GFN2/D4 Improvement Strategies

As shown in the above assessment, GFN2/D4 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/D4 appears to systematically underestimate the polarizability. We believed this derives from its use of an atom-additive polarizability model, neglecting the nonlocal polarizability-enhancing effects of electron delocalization. In order to verify that assumption, we calculated the polarizabilities for polyacetylene and polythiophene oligomers of increasing length using both GFN2 and DFT (Figures S7 and S8); this indeed showed that the GFN2 calculations do not follow the same quadratic trajectory with increasing number of monomer units as the DFT calculations do, but a rather more linear trajectory, indicating that GFN2 does not take into account the nonlocal effects of electron delocalization. Our results suggest a correction is needed to allow GFN2/D4 to more accurately predict larger polarizabilities.

We began by constructing an additive polarizability model as a baseline comparison for GFN2/D4 performance. For this model, each atom in a molecule was assigned its GAFF atom type[46], which was used to further assign it to an Alexandria polarizability type and then finally a corresponding atomic polarizability.[4] The molecular polarizability was calculated as the simple sum of these atomic polarizabilities. We found that the additive model performed extremely similarly to GFN2/D4 (Figure S1), providing computed polarizabilities with high correlation with GFN2/D4 (Figure S2).

Given GFN2/D4's systematically increasing inaccuracy for large polarizability calculations, we considered additional chemical properties related to electron delocalization that could potentially correct the additive GFN2/D4 polarizability.

Since the polarizability is connected to chemical hardness η in conceptual DFT,[47,48] which is defined by the HOMO-LUMO gaps, we first examined the GFN2-computed HOMO-LUMO gaps for both the PubChemQC subset and the hexamers from the wide range polarizability set. In principal, the highly polarizable π-conjugated species should have smaller HOMO-LUMO gaps. Unfortunately, there was not a useful correlation between the calculated polarizability and molecular HOMO-LUMO gap (Figure 4A), likely because GFN2 is not parameterized for HOMO or LUMO eigenvalues to connect with ionization potential or electron affinity.

We then used an empirical descriptor of the geometric size of the largest conjugated π-system.[49,50] While better correlated to GFN2/D4 polarizability than HOMO-LUMO gap, this information was not enough to meaningfully correct large polarizabilities (Figure 4B). Also, a further examination of the data for molecules with 25 π-systems or fewer showed almost no correlation (Figure S3).

Plotting DFT polarizabilities against GFN2/D4 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 that 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 S5, and therefore the PubChemQC and "wide range" results may be grouped together and treated as one large dataset. While in principle, dispersion could be scaled based on electron density and therefore could affect the final polarizability value, it is not part of the D4 implementation to do so and therefore does not make a difference.[2] 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 ωB7X(D) (Figure 5). We therefore conclude that although not as physically meaningful as an adjustment based on a related molecular property, we find a quadratic fit with zero intercept provides the best correction to GFN2/D4 for large polarizability calculations, and note the linear coefficient remains close to unity.

For comparison and the sake of completeness, we also tested the correlation between DFT and sTDA-xTB, a simplified time-dependent DFT procedure with a larger inherent basis set.[51] Although designed for orbital energies and electronic spectroscopy,

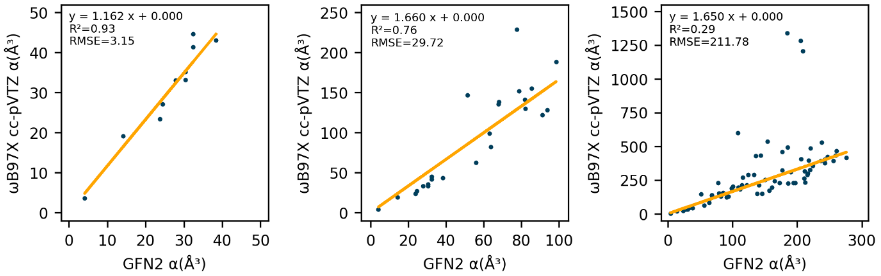

Fig. 2 Using the Huber Regressor to perform linear regression robust to outliers (and forcing the y-intercept to 0), GFN2/D4 shows some linear correlation with *ω*B97X cc-pVTZ for isotropic polarizability calculations.

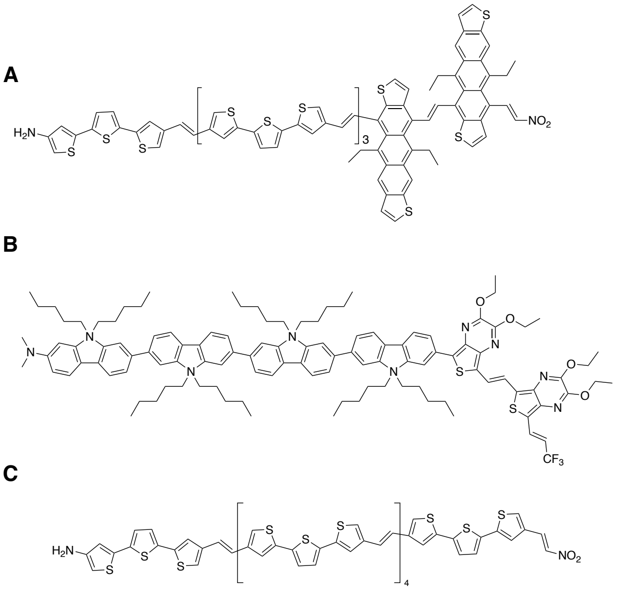

Fig. 3 Chemical structures of the hexamers seen as the three major outliers in Figure 2C.

this method has been used successfully for regarding electron density and polarizability calculations in specific cases.[52–55] Our attempt to use sTDA-xTB yielded worse results, with a linear fit MAE of 15.85Å(Figure S4). After this initial attempt, we performed additional sTDA-xTB calculations on the wide range data set to determine whether varying the energy threshold that controls the truncation of the CI space, which was set to 10 eV for the initial calculations, made a substantial difference. Figure S5 shows results comparing calculations for the wide range set with a 10 eV threshold to those with a 15 eV threshold. We also attempted to run sTDA-xTB with an energy threshold of 25 eV for the wide range set, but ran into time and memory limitation issues with the largest hexamers. Figure S6 shows comparison plots including only the 34 species from the wide range set which have successfully completed sTDA runs for all three energy thresholds. In both figures, performing linear regression with a forced zero intercept shows slight improvement in R^2 and RMSE, however it does not substantially change the results and does push the slope slightly farther away from one. Moreover, the results are still worse than the additive GFN2 / D4 model.

We also compared the results of sTDA using the 10 eV energy threshold for polyacetylene and polythiophene oligomers of increasing length to those of GFN2 and DFT with the *ω*B97X functional and the cc-pVTZ basis set corrected to aug-cc-pVTZ accuracy using our linear correction discussed later in this work from Figure S13. While the quadratic trajectories of the sTDA results seem to better match that of the DFT results as compared to the more linear trajectory of the GFN2 results (Figure /reffig:), upon closer inspection we noted that the quadratic regression line for DFT vs. GFN2 shows a better fit with much lower random error than the quadratic regression line for DFT vs. sTDA (Figure /reffig:). This is shown by the smaller residuals for the DFT vs. GFN2 regressions, indicating that GFN2 has much lower random error than sTDA. This means that GFN2 is reliably correctable for large conjugated systems, since the vast majority of its error is systematic. We believe this makes GFN2 a suitable and preferred method over sTDA for fast polarizability calculations on larger conjugated molecules, since the latter's errors are far more random in nature and therefore not systematically correctable.

Additionally, for the set of 34 species for which we have comparable data for the three energy thresholds tested, we note a meaningful difference in timings between energy thresholds (Figures S11 and S12). Because the median calculation time increases by orders of magnitude as the energy threshold increases, sTDA-xTB is not only relatively inaccurate for the conjugated oligomer systems we are interested in, but also impractical from a resource standpoint.

3.3 Basis Set Comparison

Augmented basis sets are regarded as ideal for accurate polarizability calculations[14], however 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

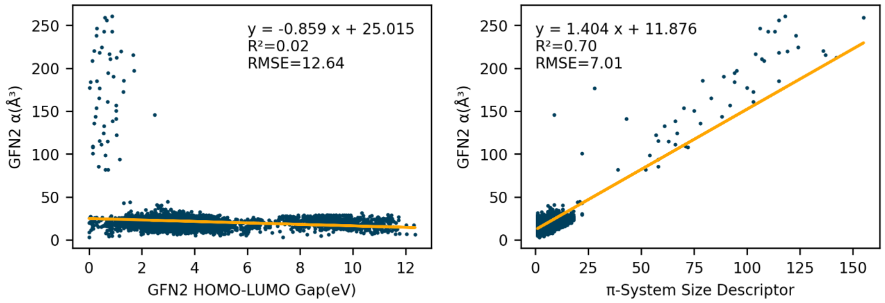

Fig. 4 Linear regression demonstrates the lack of useful correlation between GFN2/D4 calculated polarizabilities and both GFN2/D4 calculated HOMO-LUMO gap (A) and a π-system size descriptor (B).

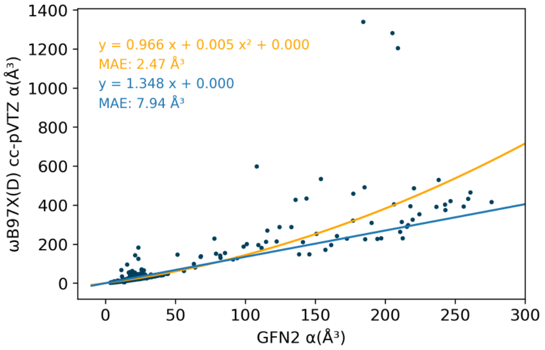

Fig. 5 Linear and quadratic regression are performed on the combined PubChemQC subset and "wide range" set.

low to medium polarizability species with four different basis sets for comparison. The sets we used were cc-pVDZ,[37,38,42–44] cc-pVTZ,[37,38,42–44] jun-cc-pVTZ,[56] and aug-cc-pVTZ,[37,38] 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 this figure, we did not choose to fix the intercept at zero in order to show the remarkably good correlation between increasingly augmented basis sets using unconstrained linear regression. Also, the y-intercepts found using this method were much closer to zero, and therefore adequately close to physical reality, than those found using linear regression without a fixed zero intercept in plots in the previous section. 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.

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.

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

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

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 conjugated 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 coefficient could be used to correct large polarizabilities found with smaller basis sets (Figure S13). 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.

4 Conclusion

Based on our studies, the additive GFN2/D4 model appears best parameterized for species with polarizabilities less than 50Å^3. In its current implementation, additive GFN2/D4 does not compare favorably to DFT-computed polarizabilities in highly polarizable oligomers. We note that in addition to limiting GFN2's usefulness in its present form as a polarizability method for larger and/or conjugated systems, this raises important concerns about the D4 model's accuracy for similar systems, since the methods share the additive polarizability model.

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/D4 comparison data, all containing similar sulfur motifs, suggests the need to examine GFN2/D4 additive parameterization for such chemical structures. Beyond increasing GFN2/D4 usefulness for efficient polarizability calculations for large polarizability molecular screening applications, these improvements would notably also improve the accuracy of

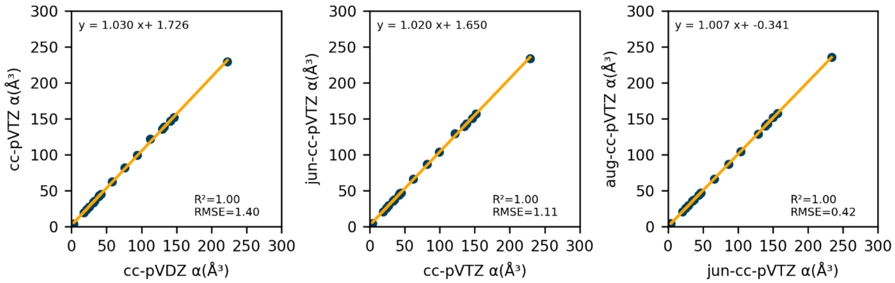

Fig. 6 Linear regression is performed on isotropic polarizabilities calculated with systematically increasing basis set size for species less than 250 Å^3.

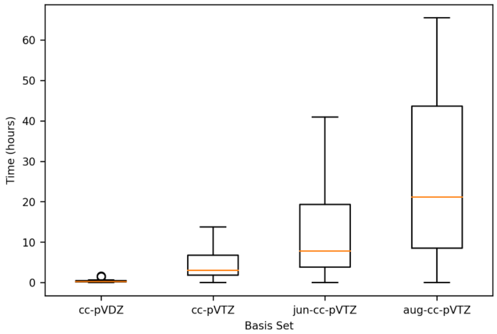

Fig. 7 Mean CPU time and the range of CPU time distribution are shown for systematically larger basis set, displaying dramatic increases in both as basis sets become larger.

the D4 dispersion method for large π-conjugated species. We also note that the additive model for GFN2/D4 remains more accurate than calculating polarizabilities using the related sTDA-xTB model.

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. We note that for some applications, such as dielectric device development, the frequency dependence of polarizability should be considered, but is outside the scope of this work. While GFN2/D4 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 the 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.

Author Contributions

We strongly encourage authors to include author contributions and recommend using CRediT for standardised contribution descriptions. Please refer to our general author guidelines for more information about authorship.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

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.

Notes and references

1 J. Hermann, R. A. D. Jr. and A. Tkatchenko, *Chemical Reviews*, 2017, **117**, 4714–4758.

2 E. Caldeweyher, S. Ehlert, A. Hansen, H. Neugebauer, S. Spicher, C. Bannwarth and S. Grimme, *Journal of Chemical Physics*, 2019, **150**, 154122.

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

4 M. M. Ghahremanpour, P. J. van Maaren, C. Caleman, G. R. Hutchison and D. van der Spoel, *Journal of Chemical Theory and Computation*, 2018, **14**, 5553–5566.

5 D. Bedrov, J.-P. Piquemal, O. Borodin, A. D. M. Jr., B. Roux and C. Schröder, *Chemical Reviews*, 2019, **119**, 7940–7995.

6 D. Porezag and M. R. Pederson, *Physical Review B*, 1996, **54**, 7830–7836.

7 M. A. F. Afzal, M. Haghighatlari, S. P. Ganesh, C. Cheng and J. Hachmann, *Journal of Physical Chemistry C*, 2019, **123**, 14610–14618.

8 D. C. Hiener and G. R. Hutchison, *The Journal of Physical Chemistry A*, 2022, **126**, 2750–2760.

9 D. Hait and M. Head-Gordon, *Physical Chemistry Chemical Physics*, 2018, **20**, 19800–19810.

10 A. Brakestad, S. R. Jensen, P. Wind, M. D'Alessandro, L. Genovese, K. H. Hopmann and L. Frediani, *Journal of Chemical Theory and Computation*, 2020, **16**, 4874–4882.

11 M. W. Jørgensen, R. Faber, A. Ligabue and S. P. A. Sauer, *Journal of Chemical Theory and Computation*, 2020, **16**, 3006–3018.

12 M. A. F. Afzal and J. Hachmann, *Physical Chemistry Chemical Physics*, 2019, **21**, 4452–4460.

13 M. B. Oviedo, N. V. Ilawe and B. M. Wong, *Journal of Chemical Theory and Computation*, 2016, **12**, 3593–3602.

14 A. L. Hickey and C. N. Rowley, *Journal of Physical Chemistry A*, 2014, **118**, 3678–3687.

15 L. Xu, A. Kumar and B. M. Wong, *Journal of Computational Chemistry*, 2018, **39**, 2350–2359.

16 M. Nakata and T. Shimazaki, *Journal of Chemical Information and Modeling*, 2017, **57**, 1300–1308.

17 C. Bannwarth, S. Ehlert and S. Grimme, *Journal of Chemical Theory and Computation*, 2019, **15**, 1652–1671.

18 S. Liu and C. E. Dykstra, *The Journal of Physical Chemistry*, 1987, **91**, 1749–1754.

19 T. Zhou and C. E. Dykstra, *The Journal of Physical Chemistry A*, 2000, **104**, 2204–2210.

20 G. J. B. Hurst, M. Dupuis and E. Clementi, *The Journal of Chemical Physics*, 1988, **89**, 385–395.

21 N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch and G. R. Hutchison, *Journal of Cheminformatics*, 2011, **3**, 33.

22 T. A. Halgren, *Journal of Computational Chemistry*, 1996, **17**, 490–519.

23 T. A. Halgren, *Journal of Computational Chemistry*, 1996, **17**, 520–552.

24 T. A. Halgren, *Journal of Computational Chemistry*, 1996, **17**, 553–586.

25 T. A. Halgren, *Journal of Computational Chemistry*, 1996, **17**, 587–615.

26 T. A. Halgren, *Journal of Computational Chemistry*, 1996, **17**, 616–641.

27 A. K. Rappe, C. J. Casewit, K. S. Colwell, W. A. G. III and W. M. Skiff, *Journal of the American Chemical Society*, 1992, **114**, 10024–10035.

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

29 F. Neese, *WIRES Computational Molecular Science*, 2012, **2**, 73–78.

30 C. Lee, W. Yang and R. G. Parr, *Physical Review B*, 1988, **37**, 785–789.

31 A. D. Becke, *Physical Review A*, 1988, **38**, 3098–3100.

32 P. J. Stephens, F. J. Devlin, C. F. Chabalowski and M. J. Frisch, *Journal of Physical Chemistry*, 1994, **98**, 11623–11627.

33 S. H. Vosko, L. Wilk and M. Nusair, *Canadian Journal of Physics*, 1980, **58**, 1200–1211.

34 G. A. Petersson, A. Bennett, T. G. Tensfeldt, M. A. Al-Laham and W. A. Shirley, *Journal of Chemical Physics*, 1988, **89**, 2193–2218.

35 G. A. Petersson and M. A. Al-Laham, *Journal of Chemical Physics*, 1991, **94**, 6081–6090.

36 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman and D. J. Fox, *Gaussian 09*, Gaussian, Inc., Wallingford, CT, 2013.

37 R. A. Kendall and T. H. D. Jr., *Journal of Chemical Physics*, 1992, **96**, 6796–6806.

38 D. E. Woon and T. H. D. Jr., *Journal of Chemical Physics*, 1993, **98**, 1358–1371.

39 J.-D. Chai and M. Head-Gordon, *Journal of Chemical Physics*, 2008, **128**, 084106.

40 P. J. Huber and E. M. Ronchetti, *Robust Statistics*, Wiley, Hoboken, New Jersey, 2nd edn, 2009.

41 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, *Journal of Machine Learning Research*, 2011, **12**, 2825–2830.

42 T. H. D. Jr., *Journal of Chemical Physics*, 1988, **90**, 1007–1023.

43 K. A. Peterson, D. E. Woon and T. H. D. Jr., *Journal of Chemical Physics*, 1994, **100**, 7410–7415.

44 A. K.Wilson, T. Mourik and T. H. Jr., *Journal of Molecular Structure: THEOCHEM*, 1996, **388**, 339–349.

45 B. Kirtman, S. Bonness, A. Ramirez-Solis, B. Champagne, H. Matsumoto and H. Sekino, *The Journal of Chemical Physics*, 2008, **128**, 114108.

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

47 P. Geerlings, F. D. Proft and W. Langenaeker, *Chemical Reviews*, 2003, **103**, 1793–1874.

48 R. G. Pearson, *Proceedings of the National Academy of Sciences*, 1986, **83**, 8440–8441.

49 O. D. Abarbanel and G. R. Hutchison, *The Journal of Chemical Physics*, 2021, **155**, 054106.

50 O. D. Abarbanel and G. R. Hutchison, *Reorganiza-*

tion Energy, 2021, `https://github.com/hutchisonlab/ReorganizationEnergy`.

51 S. Grimme and C. Bannwarth, *The Journal of Chemical Physics*, 2016, **145**, 054103.

52 M. de Wergifosse and S. Grimme, *The Journal of Chemical Physics*, 2018, **149**, 024108.

53 J. Seibert, B. Champagne, S. Grimme and M. de Wergifosse, *The Journal of Physical Chemistry B*, 2020, **124**, 2568–2578.

54 P. Beaujean, B. Champagne, S. Grimme and M. de Wergifosse, *The Journal of Physical Chemistry Letters*, 2021, **12**, 9684–9690.

55 L. Lescos, P. Beaujean, C. Tonnelé, P. Aurel, M. Blanchard-Desce, V. Rodriguez, M. de Wergifosse, B. Champagne, L. Muccioli and F. Castet, *Physical Chemistry Chemical Physics*, 2021, **23**, 23643–23654.

56 E. Papajak, J. Zheng, X. Xu, H. R. Leverentz and D. G. Truhlar, *Journal of Chemical Theory and Computation*, 2011, **7**, 3027–3034.