PCCP

PAPER

View Article Online
View Journal | View Issue

Cite this: *Phys. Chem. Chem. Phys.*, 2019, **21**, 4452

Benchmarking DFT approaches for the calculation of polarizability inputs for refractive index predictions in organic polymers†

Mohammad Atif Faiz Afzal *[a] and Johannes Hachmann *[abc]

Received 29th August 2018,
Accepted 24th January 2019

DOI: 10.1039/c8cp05492d

rsc.li/pccp

In a previous study, we introduced a new computational protocol to accurately predict the index of refraction (RI) of organic polymers using a combination of first-principles and data modeling. This protocol is based on the Lorentz–Lorenz equation and involves the calculation of static polarizabilities and number densities of oligomer sequences, which are extrapolated to the polymer limit. We chose to compute the polarizabilities within the density functional theory (DFT) framework using the PBE0/def2-TZVP-D3 model chemistry. While this *ad hoc* choice proved remarkably successful, it is also relatively expensive from a computational perspective. It represents the bottleneck step in the overall RI modeling protocol, thus limiting its utility for virtual high-throughput screening studies, in which efficiency is essential. For polymers that exhibit late-onset extensivity, the employed linear extrapolation scheme can require demanding calculations on long-oligomer sequences, thus becoming another bottleneck. In the work presented here, we benchmark DFT model chemistries to identify approaches that optimize the balance between accuracy and efficiency for this application domain. We compare results for conjugated and non-conjugated polymers, augment our original extrapolation approach with a non-linear option, analyze how the polarizability errors propagate into the RI predictions, and offer guidance for method selection.

I. Introduction

Organic materials with high index of refraction (RI) have gained considerable attention in recent years as they hold tremendous potential for applications in optic and optoelectronic devices.[1–5] The vast majority of carbon-based polymers has relatively low RI values (typically in the range of 1.3 to 1.5),[6,7] which has made the search for compounds with high and very high RIs (greater than 1.8) an active area of research.[8,9] The key to increasing the RI values of organic polymers is our ability to tailor their molecular structure.[6,9–11] However, the number of compounds that results from considering even only a modest selection of polymer building blocks is practically infinite. Experimental efforts alone are too time-, labor-, and resource-intensive to effectively survey the massive chemical space associated with this problem setting (and many others in the molecular sciences).

Computational high-throughput screening approaches have emerged as a way to rapidly characterize and assess large candidate pools, and to identify lead compounds for further in-detail investigations (see, *e.g.*, ref. 12–20). In the context of optical materials with large dielectric constants (and thus large RI values), the work by Ramprasad *et al.*[21–23] is particularly noteworthy. The foundation for any *in silico* screening study are suitable modeling protocols for the properties and compound classes of interest. For use in large-scale investigations, these protocols not only have to produce sufficiently accurate predictions, but they also have to be fast. A number of modeling approaches for the RI values of polymers have been introduced in the past,[21,22,24–29] each with distinct advantages and disadvantages in the areas of accuracy, reliability, robustness, cost, and range of applicability.

We recently introduced a new protocol[30] based on a synergistic combination of first-principles and data modeling. In this protocol, we calculate RI values n_r using the Lorentz–Lorenz equation with the number density N and polarizability α of a given candidate compound as input parameters. We obtain the former using the van der Waals volume and packing factor of the compound, and the latter directly from quantum chemistry. Specifically, we compute the van der Waals volumes using Slonimskii's method,[31] and for the packing fraction of the amorphous bulk polymer,

[a] *Department of Chemical and Biological Engineering, University at Buffalo, The State University of New York, Buffalo, NY 14260, USA. E-mail: m27@buffalo.edu, hachmann@buffalo.edu*
[b] *Computational and Data-Enabled Science and Engineering Graduate Program, University at Buffalo, The State University of New York, Buffalo, NY 14260, USA*
[c] *New York State Center of Excellence in Materials Informatics, Buffalo, NY 14203, USA*

† Electronic supplementary information (ESI) available: It provides details of all computational and experimental values displayed in the figures throughout this paper or that were used in the statistical analysis. We also give detailed definitions of all statistical metrics used in this work. See DOI: 10.1039/c8cp05492d

we introduced a support vector regression[32,33] (*i.e.*, machine learning) model. For the polarizabilities, we employ Kohn–Sham density functional theory (DFT)[34,35] using the PBE0 functional and def2-TZVP basis set along with D3 dispersion correction. As the protocol's target systems are quasi-infinite polymers, we obtain the asymptotic trends towards the polymer limit through a linear extrapolation scheme from a sequence of small-oligomer calculations. This scheme exploits the relatively short correlation length in many systems, in which it leads to an early onset of extensivity in the response properties. We tested the RI predictions of the protocol on 112 non-conjugated polymers and the results show very good agreement with the experimental values ($R^2 = 0.94$). The protocol is overall economical and suitable for high-throughput *in silico* studies. However, the polarizability calculations nonetheless stand out as the bottleneck that limits its efficiency. Another concern is that for conjugated systems with long-range correlation, the linear extrapolation scheme will require results of very long oligomers to reach extensivity, and thus become prohibitively costly.

In this paper, we present a benchmark study of several DFT model chemistries to identify approaches that deliver a more favorable balance of accuracy and efficiency for polarizabilities in the context of large-scale RI studies, and to understand the nature of prediction errors. We also revisit the protocol's extrapolation scheme, augmenting it with a non-linear option that uses shorter oligomer sequences below the extensivity threshold, and demonstrate its validity and performance. We provide an analysis of how the errors in the polarizability results propagate into the RI value predictions. In Section II, we introduce the benchmarking methodology and computational details of this study. Section III presents and discusses the results for the model chemistry performance analysis (Section IIIA), the improved extrapolation approach (Section IIIB), and the error propagation (Section IIIC). Our findings are summarized in Section IV.

II. Methods and computational details

As mentioned in Section I, the RI protocol introduced in ref. 30, employs the PBE0/def2-TZVP-D3 model chemistry to compute the static polarizabilities that serve as input for the Lorentz–Lorenz equation. The protocol calls for calculations on a sequence of small oligomers until a constant increase in the polarizability per added monomer unit is observed, which allows for a linear extrapolation to the polymer limit. For the non-conjugated polymers studied in ref. 30, extensivity was reached for very short oligomers (*i.e.*, $n \ll 10$ monomer units).

Computed polarizability values and the derived RI predictions generally depend on the employed quantum chemical approximation, and different model chemistries yield different results at different computational cost. In the DFT benchmark study at hand, we compare six reasonable functional choices from across Jacob's ladder,[36] covering generalized gradient approximation (BP86[37,38]), hybrid (B3LYP,[38–40] PBE0[41]), meta-hybrid (TPSSh[42]), highly-parametrized meta-hybrid (M06-2X[43,44]), and double-hybrid (B2PLYP[45]) functionals. Their formal cost scaling with system size is n^3 (generalized gradient approximation), n^4 (hybrid functionals), and n^5 (double-hybrid functionals), respectively. Each functional is tested with the atom-centered double-ζ def2-SVP and triple-ζ def2-TZVP basis sets (abbreviated for convenience as DZ and TZ, respectively) from the highly successful def2 basis set family by Ahlrichs *et al.*[46] The polarizability calculations were performed analytically *via* the solution of the coupled-perturbed self-consistent field equations in all-electron, closed-shell mode, and include Grimme's D3 dispersion correction[47] (we omit the D3 label for brevity). By default, we use single-point calculations on geometries optimized at the universal force field (UFF) level[48] *via* the OpenBabel code[49] as outlined in ref. 30. For the PA, PB, PE, and PT studies in Section IIIB, we utilize B3LYP/DZ optimized structures instead in order to obtain a clean results as UFF gives less reliable geometries for conjugated systems. We included these optimized geometries (*xyz* coordinates) in the ESI.† All DFT calculations were carried out using the ORCA 3.0.2 quantum chemistry program package[50] with default settings. The benchmark study involved about 5400 individual DFT calculations, which we performed using our automated virtual high-throughput screening code *ChemHTPS* 0.7.[51,52]

The RI predictions from the different model chemistries are compared with the experimentally known RI values of the same 112 non-conjugated polymers used in ref. 30. In addition, we perform an in-depth analysis of two prototypical examples for non-conjugated and conjugated polymers, *i.e.*, polyethylene (PE) and polyacetylene (PA), respectively. To study the transition from the conjugated to the non-conjugated regime, we further select polythiophene (PT) and poly(1,4-phenylene) (PB) as test cases, and break their conjugation by introducing non-planarity in the polymer chain (*i.e.*, by constraining consecutive rings as perpendicular to each other as shown in Fig. 1).

The following statistical measures are used in the error analyses: mean absolute error (MAE), mean absolute percentage error (MAPE), root mean squared error (RMSE), root mean squared percentage error (RMSPE), mean error (ME), mean percentage error (MPE), maximum absolute error (MaxAE), maximum absolute percentage error (MaxAPE), and difference of most extreme (*i.e.*, spread of largest positive and negative) errors (ΔMaxE). Aside from providing these direct measures, we also quantify the

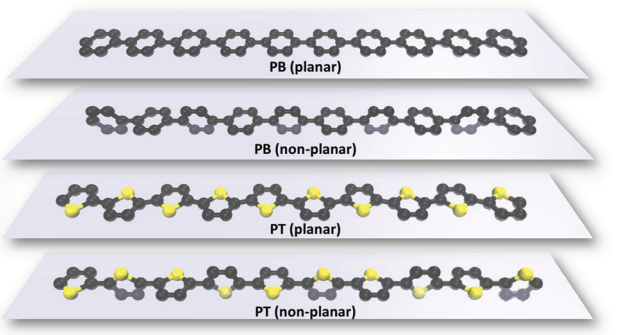

Fig. 1 Planar and non-planar structures of poly(1,4-phenylene) (PB) and polythiophene (PT).

extent of correlations between results of different methods by listing R^2, slope, and offset values of linear regressions.

III. Results and discussion

A. Model chemistry performance

Fig. 2 compares the RI predictions based on the different model chemistries with the experimentally known values of our 112-polymer data set, and Table 1 summarizes the corresponding error and correlation statistics. The analysis reveals that the PBE0/TZ and B3LYP/TZ results are favorable by most measures without a clear advantage for either one. This is a somewhat surprising finding considering the competition of more modern and more advanced functionals. Since PBE0/TZ was the method used in ref. 30, we choose it as the high-level reference throughout the following benchmarking. The RI prediction errors it yields are very reasonable with MAPE = 0.9%, RMSPE = 1.2%, and MaxAPE = 3.0%, and the results have very little directional bias as seen from MPE = −0.3%.

The basis set error between DZ and TZ is the most significant error contribution we observe, with the best DZ result being inferior to the worst one from TZ. Most RI values derived from TZ polarizabilities follow the experimental trends faithfully, with linear regression slopes close to 1, offsets close to 0, and only moderate spread (all $R^2 > 0.91$, except for BP86 with $R^2 = 0.86$). B2PLYP and BP86 show the largest MAE, MAPE, and RMSPE with B2PLYP systematically under- and BP86 systematically overpredicting (MPE = −1.7% and +1.8%, respectively). BP86 also shows the largest values for the worst-case metrics MaxAE, MaxAPE, and ΔMaxE. Despite its lower-order scaling, BP86/TZ is thus not a convincing alternative to the PBE0/TZ reference. The more expensive B2PLYP/TZ does not offer any benefits, despite featuring a higher-level functional.

Considering only the DZ results, this picture changes notably, with BP86 having the lowest MAE, MAPE, RMSE, RMSPE, ME, and MPE. B2PLYP is doing worst in most of these measures, while the remaining functionals show comparable performance without a distinct competitive edge. However, considering the correlation with the experimental data, B3LYP, PBE0, and TPSSh with R^2 values between 0.86 and 0.90 have an advantage over the other functionals (including BP86 with only 0.81). Given the considerable savings due to the smaller basis set, there is a case to be made for either BP86/DZ, B3PLY/DZ, PBE0/DZ, or TPSSh/DZ as low-cost alternatives to the PBE0/TZ reference. The errors for each of these methods are, however, 3–4 times larger. While BP86/DZ is the cheapest option with overall low errors, it does exhibit some of the most extreme failures and weakest correlation with respect to more accurate approaches. An interesting observation for the DZ results is that all functionals lead to underestimates of the RI predictions with negative MPE values between −2.8% and −5.1%. This bias is more pronounced for high-RI compounds, which is apparent from the linear regression slopes being significantly below 1.

In Fig. 3, we show the polarizability results underlying these RI predictions and compare them with the PBE0/TZ benchmark reference. An analysis of the errors and correlations is provided in Table 2. Despite the considerable differences in the employed

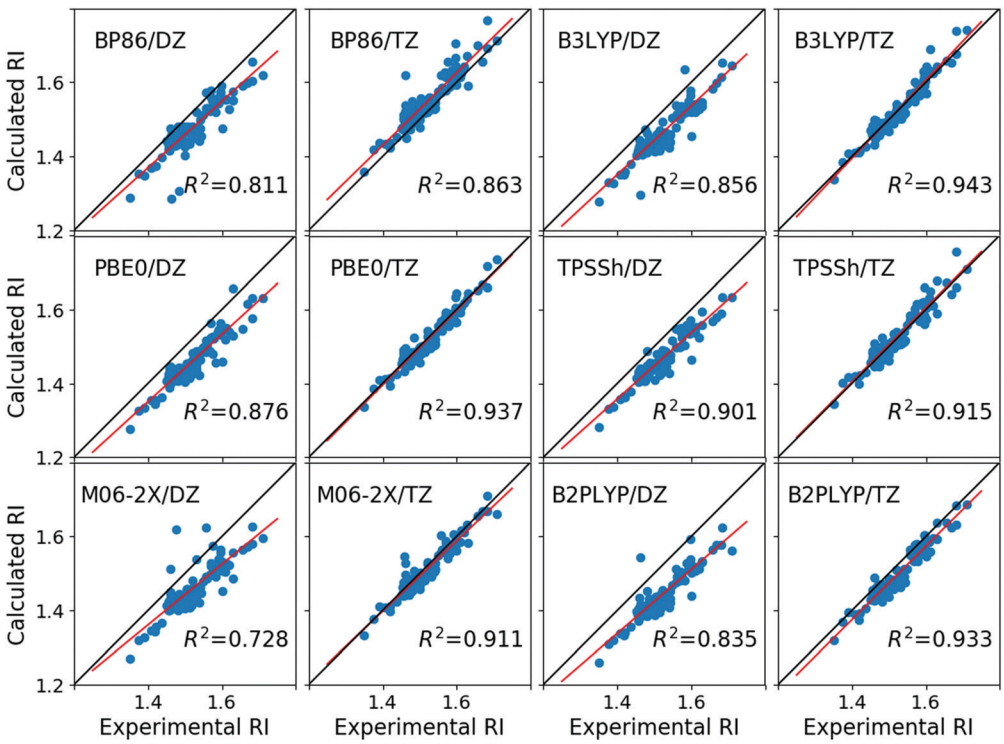

Fig. 2 Comparison of refractive index (RI) values calculated from polarizabilities of different model chemistries with experimental values of 112 non-conjugated polymers. Linear regressions (red lines) and their correlation coefficients (R^2) are provided in each case.

Table 1 Performance of different model chemistries for RI predictions. The error and correlation analysis compares the computational results to the experimentally known values of 112 non-conjugated polymers. The most favorable finding for each statistical measure is highlighted in bold and the best results within the DZ method spectrum in bold–italics

Functional	BP86		B3LYP		PBE0		TPSSh		M06-2X		B2PLYP	
Basis set	DZ	TZ	DZ	TZ	DZ	TZ	DZ	TZ	DZ	TZ	DZ	TZ
MAE	***0.043***	0.029	0.058	**0.013**	0.060	0.014	0.054	0.016	0.065	0.017	0.079	0.027
RMSE	***0.052***	0.038	0.082	0.031	0.063	**0.018**	0.064	**0.018**	0.058	0.021	0.070	0.022
ME	***−0.042***	0.027	−0.057	**0.001**	−0.060	−0.004	−0.054	0.005	−0.060	−0.008	−0.078	−0.026
MaxAE	0.177	0.159	0.166	0.077	0.142	**0.045**	*0.137*	0.073	0.149	0.088	0.161	0.068
deltaMaxE	0.195	0.197	0.218	0.111	0.170	**0.089**	0.144	0.115	0.290	0.137	0.242	0.096
MAPE	***2.8%***	1.9%	3.8%	0.8%	4.0%	0.9%	3.6%	1.1%	4.2%	1.1%	5.2%	1.8%
RMSPE	***3.4%***	2.5%	4.1%	**1.1%**	4.2%	1.2%	3.8%	1.4%	4.5%	1.4%	5.4%	2.1%
MPE	***−2.8%***	1.8%	−3.8%	**0.1%**	−3.9%	−0.3%	−3.5%	0.3%	−3.9%	−0.5%	−5.1%	−1.7%
MaxAPE	12.1%	10.9%	11.3%	4.8%	8.9%	**3.0%**	***8.6%***	4.3%	9.6%	6.1%	10.1%	4.4%
R^2	0.811	0.863	0.856	**0.943**	0.876	0.937	***0.901***	0.915	0.728	0.911	0.835	0.933
Slope	0.900	0.980	***0.930***	1.050	0.910	**1.000**	0.900	**1.000**	0.820	0.950	0.860	0.990
Offset	0.110	0.060	***0.050***	−0.080	0.070	−0.010	0.100	**0.000**	0.210	0.070	0.130	−0.020

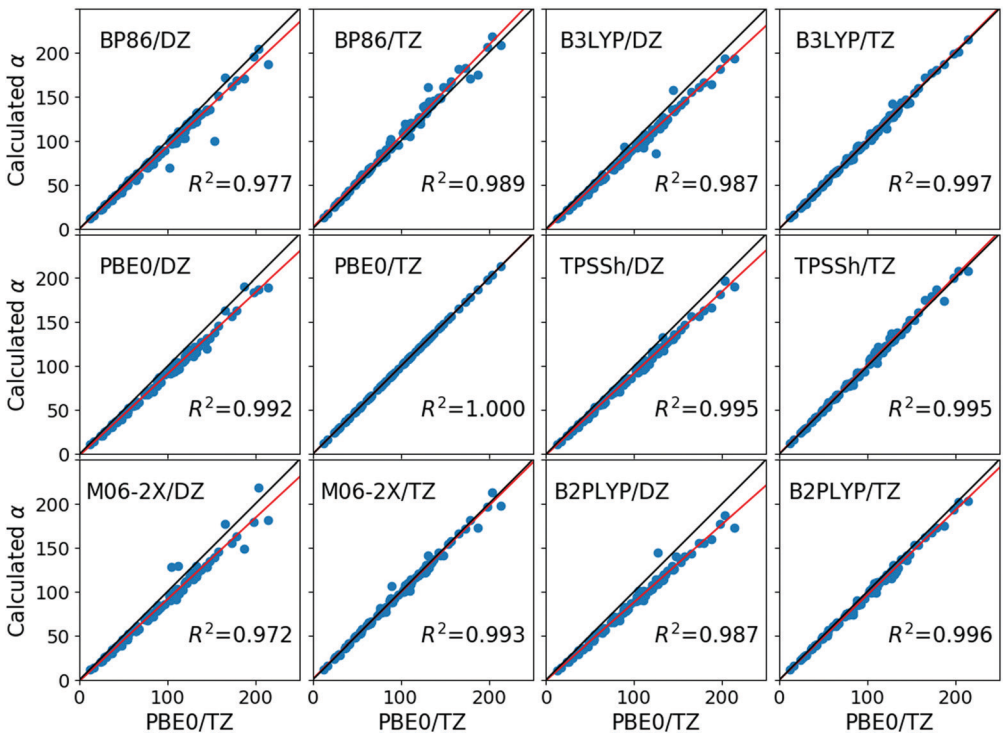

Fig. 3 Comparison of polarizability results α for 112 non-conjugate polymers from different model chemistries compared to PBE0/def2-TZVP-D3 (PBE0/TZ) benchmark reference results. Linear regressions (red lines) and their R^2 values are provided in each case.

electronic structure approximations, we can clearly see that the results of the different methods are very strongly correlated. With respect to PBE0/TZ, we find R^2 values of at least 0.97 (0.99 within the TZ approaches). However, the direct polarizability error metrics are many times larger than those for the RI predictions.

The DZ results stand out for all having large negative MPE values between −6.4% and −12.6%, slopes <1, and negative offsets, *i.e.*, they all systematically underestimate the polarizabilities, and the bias increases with increasing magnitude of the polarizability values. For TZ, the slope is in each case much closer to 1 and the offset closer to 0 (except for BP86 where TZ has a larger offset). The B3LYP/TZ results are closest to those of the reference, with TPSSh/TZ giving acceptable accuracy as well. Amongst the DZ results, BP86 again yielded the best results for several of the error metrics, however, it also exhibits some of the most extreme discrepancies, *i.e.*, with respect to MaxAE, MaxAPE, ΔMaxE, and R^2 it performs worst (together with M06-2X). B3LYP/DZ, PBE0/DZ, and TPSSh/DZ again yield very good results without the extreme error instances seen in BP86/DZ. B2PLYP and M06-2X do not offer any benefits in either basis.

In summary, we find that DZ approaches result in systematically lower polarizability values than the corresponding TZ models,

Table 2 Performance of different model chemistries for polarizability calculations. The error and correlation analysis compares the results of 112 non-conjugate polymers to those of the PBE0/TZ benchmark reference. The most favorable finding for each statistical measure is highlighted in bold and the best results within the DZ method spectrum in bold-italics

Functional	BP86		B3LYP		PBE0		TPSSh		M06-2X		B2PLYP	
Basis set	DZ	TZ	DZ	TZ	DZ	TZ	DZ	TZ	DZ	TZ	DZ	TZ
MAE	***6.087***	5.451	8.500	**1.539**	8.569	0.000	7.600	2.303	9.613	2.072	11.817	3.552
RMSE	8.914	6.905	13.166	4.549	9.816	0.000	9.661	**2.429**	***8.563***	3.429	11.126	3.631
ME	***−5.948***	4.855	−8.192	0.753	−8.510	0.000	−7.585	1.386	−8.381	**−0.646**	−11.511	−3.475
MaxAE	53.633	30.601	39.532	13.474	24.988	0.000	***23.220***	12.839	38.311	16.574	40.911	**12.447**
delta MaxE	60.449	42.629	52.706	21.072	28.291	0.000	***24.057***	23.107	62.492	32.505	58.001	**13.847**
MAPE	***6.5%***	5.6%	9.3%	**1.6%**	9.5%	0.0%	8.4%	2.2%	10.3%	2.0%	12.8%	3.8%
RMSPE	***81.1%***	64.1%	93.9%	**23.0%**	93.0%	0.0%	82.1%	30.9%	104.0%	32.3%	124.6%	42.8%
MPE	***−6.4%***	5.2%	−9.1%	0.8%	−9.5%	0.0%	−8.4%	1.4%	−9.4%	**−0.6%**	−12.6%	−3.7%
MaxAPE	35.0%	23.5%	31.6%	10.5%	17.3%	0.0%	***16.4%***	**9.2%**	23.2%	17.7%	21.3%	11.9%
R^2	0.977	0.989	0.987	0.997	**0.992**	1.000	0.995	0.995	0.972	0.993	0.987	0.996
Slope	***0.940***	1.036	0.928	1.006	**0.929**	1.000	0.930	1.011	0.928	0.983	0.887	0.964
Offset	***−0.276***	1.492	−1.420	0.183	−1.742	0.000	−0.952	0.338	−1.581	0.925	−0.800	**−0.083**

which we can rationalize based on the less flexible DZ expansion of the frontier orbitals that have to facilitate the electronic response. Between the functionals, BP86 yields the largest polarizabilities, which is consistent with the well-known over-delocalization of orbitals from generalized gradient approximation functionals. The order of the other functionals is (based on their MPEs): TPSSh > B3LYP > PBE0 > M06-2X > B2PLYP. For the hybrid functionals, this correlates directly with their increasing amount of exact exchange, *i.e.*, 10%, 20%, 25%, and 54%, respectively, which is known to lead to increasing orbital localization, thus damping the electronic response.[53] For the double-hybrid B2PLYP with 53% exact exchange, additional perturbation contributions play a role in further lowering the polarizability values it produces.

B. Extrapolation scheme

In Fig. 4, we show for the non-conjugated prototype polymer PE the incremental increase of the polarizability α with oligomer size, *i.e.*, with the number of monomer units n. We observe the rapid convergence of α/n to a constant value after only a few monomer units. This behavior represents the basis of the linear extrapolation scheme used to great effect in the original RI protocol. The initial decrease in α/n from monomer to trimer is due to finite size effects and the diminishing impact of the terminal hydrogens. Note that the order of the asymptotic values for the different model chemistries is consistent with our discussion of basis set and functional effects at the end of Section IIIA.

While this behavior is typical for non-conjugated polymers, systems with a conjugated π-electron backbone may only show extensivity for very long oligomers, *i.e.*, when the system becomes larger than its electronic correlation length. Fig. 5 shows the α/n values for the conjugated prototype polymer PA as a function of oligomer size. Unlike PE (which reaches a constant α/n for $n = 4$), the α/n of PA does not converge in the plotted range up to $n = 11$, but rather it increases with increasing number of monomer units throughout which the valence electrons are correlated. As the oligomer size increases, so do the conjugation and polarizability. The π-system is subject to response in its entirety (rather than its individual constituent monomer units), which thus cooperatively amplifies the polarizability values it yields.

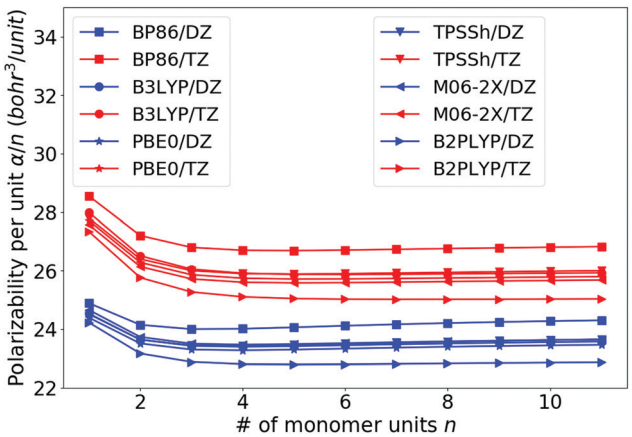

Fig. 4 Polarizability α per number of monomer units n of polyethylene (PE) with varying oligomer chain length computed using different model chemistries. DZ results are shown in blue, TZ results in red.

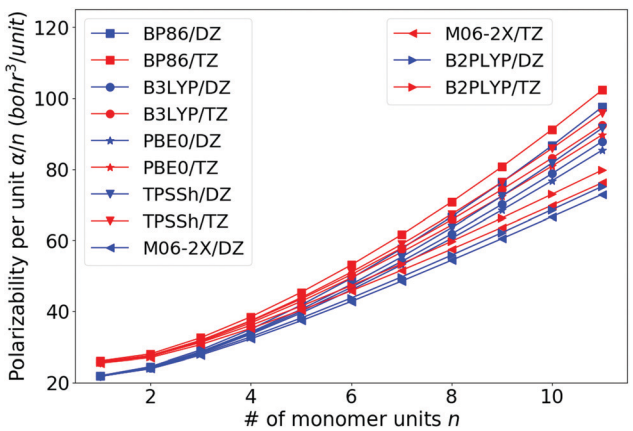

Fig. 5 Polarizability α per number of monomer units n of polyacetylene (PA) with varying chain length computed using different model chemistries. DZ results are shown in blue, TZ results in red.

Fig. 5 offers another interesting observation, *i.e.*, that for conjugated systems, the basis set effect (reflecting the more inflexible DZ orbital expansion) stops being the dominant factor and that the functional effect (reflecting exact-exchange-driven localization) starts dominating the order of the results for longer chains. Instead of a clean separation by basis set (all TZ values above DZ), we now approach an order by functional. This trend is consistent with the regression slopes discussed in Section IIIA. Finally, we note that while the individual monomer units of PE and PA have comparable polarizabilities, the α/n values of PA grow dramatically and become much larger than those of PE.

For a clearer comparison of the conjugated and non-conjugated regimes, we study two conjugated example polymers – PB and PT – for which we break conjugation by introducing non-planarity as shown in Fig. 1. As expected, the α/n values for conjugated PB and PT increase rapidly with oligomer length (see Fig. 6 and 7), while for non-conjugated PB and PT, they increase significantly less and start to taper off. (The slight residual increase may be due to a weak conjugation between the π-system of the aromatic rings and the σ-framework of the perpendicular adjacent rings.)

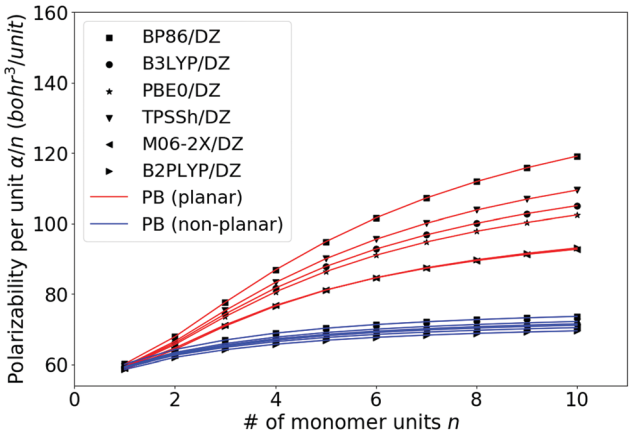

Fig. 6 Polarizability per monomer unit of planar and non-planar PB as a function of chain length, calculated from different model chemistries.

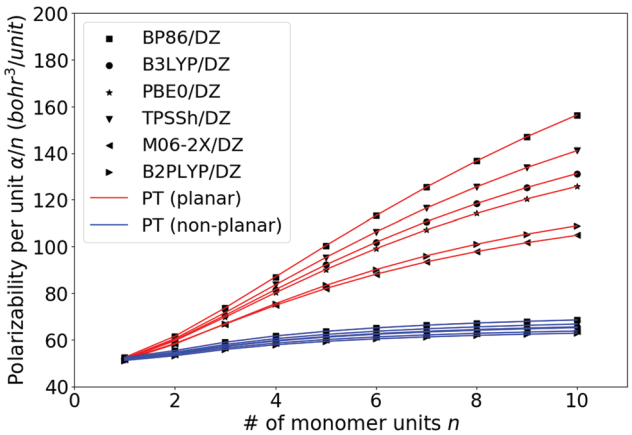

Fig. 7 Polarizability per monomer unit of planar and non-planar PT as a function of chain length, calculated from different model chemistries.

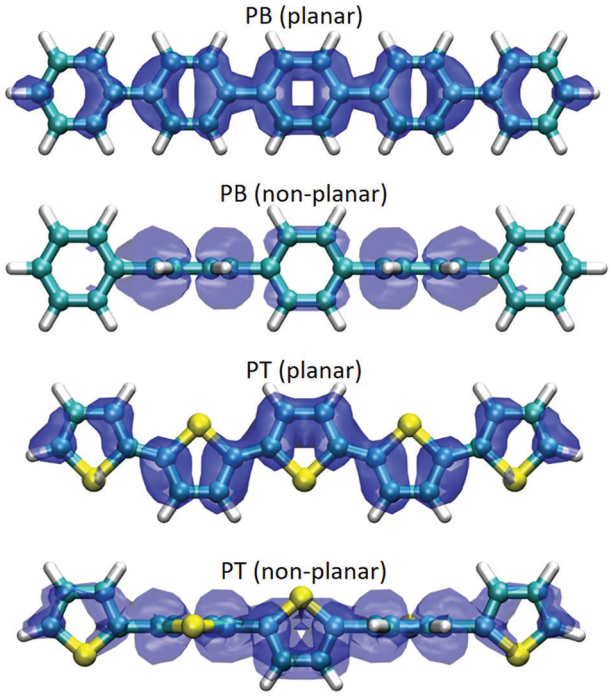

Fig. 8 Highest occupied molecular orbital electron density of planar and non-planar pentamers of PB and PT at PBE0/TZ level of theory.

To illustrate the change from conjugated to non-conjugated systems, we plot the electron density of the highest occupied molecular orbital (HOMO) for PB and PT pentamers (see Fig. 8; plots of the HOMOs themselves are provided in the ESI†). The conjugated systems feature a relatively evenly delocalized HOMO density that can readily facilitate substantial charge redistribution throughout the molecule. The non-conjugated systems in contrast have HOMO density that is more localized on the disconnected monomer units.

As the polarizability in conjugated systems increases non-linearly until extensivity is reached, we propose a non-linear fit to efficiently account for this behavior. (We stress that the linear extrapolation scheme introduced in ref. 30 still works in principle, but in practice, it may require calculations of extended oligomer sequences that would be prohibitive.) The α/n value shows a quasi-linear trend for smaller n before it asymptotically converges to a constant value for large n. A mathematical expression (eqn (1)) was proposed by Hurst *et al.*[54] to model this behavior.

$$\alpha = \exp\left(a + \frac{b}{n} + \frac{c}{n^2}\right) \quad (1)$$

However, it does not provide a good fit for long oligomers (*cf.* Fig. 9). We propose to add a 3rd-order term as outlined in eqn (2).

$$\alpha = \exp\left(a + \frac{b}{n} + \frac{c}{n^2} + \frac{d}{n^3}\right) \quad (2)$$

We can show that this new model improves the predictions significantly, in particular by capturing the correct asymptotic limit: Fig. 9 displays the PBE0/DZ results for PA, PB, and PT for

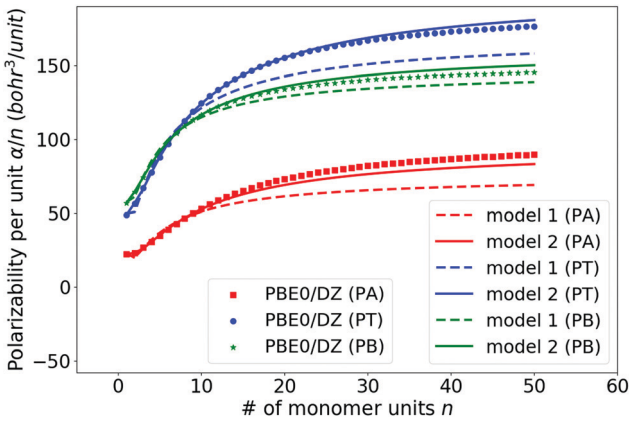

Fig. 9 Validation of the old (1) and new (2) polarizability models for long oligomers (up to n = 50) of PA, PB, and PT. The models are parametrized using a fit of PBE0/DZ data for short oligomers (up to n = 10).

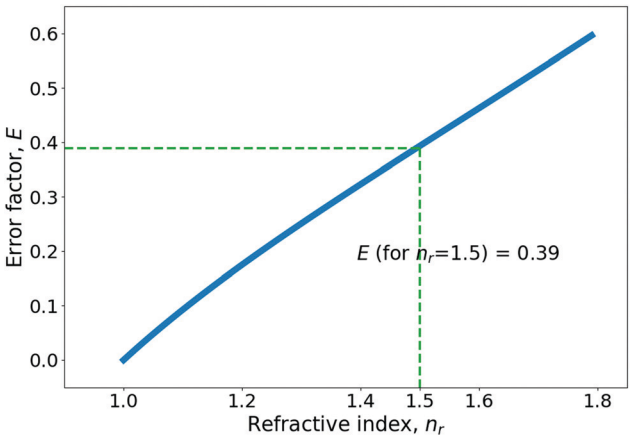

Fig. 10 Error factor (E) for RI values (n_r) ranging from 1 to 1.8.

n up to 50. (To save computing time, the structures were not fully optimized, but bond lengths and angles were selected based on the optimized geometry of the PBE0/DZ optimized n = 10 oligomer.) The model predictions are based on parameters fitted using the DFT results for relatively small oligomers up to n = 10. The new model is valid for all the three example polymers yielding very small asymptotic errors, while the model by Hurst *et al.* shows significant discrepancies in each case. We note that the model rapidly improves further as additional trainings points are provided.

C. Error propagation

As we remarked on in Section IIIA, the deviations between the polarizability results with respect to the reference is relatively large with MAPEs between 1.6–5.6% for TZ and 6.5–12.8% for DZ, while those of the RI predictions that build on these α values is much smaller with MAPEs (with respect to the experimental data) between 0.8–1.9% for TZ and 2.8–5.2% for DZ. To understand the error propagation from polarizability (and number density) to RI values, we consider the underlying Lorentz–Lorenz equation (eqn (3)).

$$\frac{(n_r^2-1)}{(n_r^2+2)}=\frac{4\pi}{3}N\alpha \tag{3}$$

$$\frac{dn_r}{n_r}=\frac{(n_r^2-1)(n_r^2+2)}{6n_r^2}\left(\frac{dN}{N}+\frac{d\alpha}{\alpha}\right) \tag{4}$$

$$E=\frac{(n_r^2-1)(n_r^2+2)}{6n_r^2} \tag{5}$$

We differentiate eqn (3) to obtain eqn (4) and subsequently the error factor (E) as shown in eqn (5). We observe that E is only dependent on the magnitude of the RI value. For RI values ranging from 1 to 1.8, the value of E ranges from 0 to 0.59, respectively, as plotted in Fig. 10. Hence, when the error in the number density is ignored, the error in the RI predictions, for RI around 1.5, should be in the order of 40% of the inherent error in the polarizabilities. For instance, if a lower-level method such as BP86/DZ is employed for the polarizability input instead of the PBE0/TZ reference (with MAPE of 6.5%), the additional error in the RI prediction would be about 2.5% points (the observed value is somewhat larger with 2.8% points). This analysis suggests that the use of more affordable model chemistries with larger polarizability errors can be justified as it only leads to a modest increase in the RI prediction errors. The same argument holds for using an extrapolation scheme with relatively short oligomer sequences below the extensibility threshold with reasonable asymptotic errors.

IV. Conclusions

In the work presented here, we benchmarked a range of different DFT model chemistries for the calculation of polarizability inputs for RI predictions *via* the Lorentz–Lorenz equation. We found that PBE0/def2-TZVP-D3 and B3LYP/def2-TZVP-D3 perform best with RI prediction MAPEs of less than 1%. They notably outperform several more modern and complex functionals, including B2PLYP. Amongst the less demanding approaches, BP86/def2-SVP-D3, B3PLY/def2-SVP-D3, PBE0/def2-SVP-D3, and TPSSh/def2-SVP-D3 emerge as viable alternatives with reasonable errors given the considerable cost savings, with BP86 being the most efficient option (alas also exhibiting the largest uncertainties). We could observe trends and systematic biases consistent with small basis set inflexibility and over-localization driven by exact exchange. Aside from this model chemistry assessment, we also revisited the oligomer extrapolation scheme to rapidly obtain polarizability values at the polymer limit. We augmented our linear approach that is highly efficient for non-conjugated polymers with a non-linear alternative for conjugated systems, building on a model by Hurst *et al.* We could show that this more flexible model can make accurate asymptotic predictions given data from relatively short oligomer sequences, which renders it dramatically more efficient than the linear scheme and more accurate than the Hurst model without additional cost. Finally, we conducted a formal analysis of the polarizability error propagation into the RI predictions and found a relatively modest impact, suggesting that reasonable errors in the polarizabilities due to the use of

lower-level model chemistries or approximate extrapolation schemes are acceptable. With these additional insights, we can make informed and rational (rather than *ad hoc*) decisions regarding the computation of polarizabilities, and tailor our RI modeling protocol to alleviate bottlenecks that limit its utility in the context of high-throughput studies. In subsequent work, we plan to deploy the adjusted protocol for the large-scale screening of organic polymer candidate libraries in order to identify concrete high-RI lead compounds.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This work was supported by start-up funds provided through the University at Buffalo (UB) and the National Science Foundation CAREER program under grant No. OAC-1751161. Computing time on the high-performance computing clusters '*Rush*', '*Alpha*', '*Beta*', and '*Gamma*' was provided by the UB Center for Computational Research (CCR). The work presented in this paper is part of MAFA's PhD thesis.[55] We are grateful to Profs. Chong Cheng and Michel Dupuis of UB for valuable discussions and insights.

References

1 T. Higashihara and M. Ueda, *Macromolecules*, 2015, **48**, 1915–1929.
2 G.-S. Liou, P.-H. Lin, H.-J. Yen, Y.-Y. Yu, T.-W. Tsai and W.-C. Chen, *J. Mater. Chem.*, 2010, **20**, 531–536.
3 T. T. Huang, C. L. Tsai, S. Tateyama, T. Kaneko and G. S. Liou, *Nanoscale*, 2016, **8**, 12793–12802.
4 T. Lei, J. Y. Wang and J. Pei, *Chem. Mater.*, 2014, **26**, 594–603.
5 S.-S. Sun, L. R. Dalton, S.-S. Sun and L. R. Dalton, *Introduction to Organic Electronic and Optoelectronic Materials and Devices (Optical Science and Engineering Series)*, CRC Press, Inc., Boca Raton, FL, USA, 2008.
6 J.-g. Liu and M. Ueda, *J. Mater. Chem.*, 2009, **19**, 8907–8919.
7 H. Liu, I. Blakey, W. E. Conley, G. George, D. J. Hill and A. K. Whittaker, *J. Micro/Nanolithogr., MEMS, MOEMS*, 2008, **7**, 023001.
8 H. Jintoku and H. Ihara, *Chem. Commun.*, 2014, **50**, 10611–10614.
9 J. J. Griebel, *et al.*, *Adv. Mater.*, 2014, **26**, 3014–3018.
10 A. Javadi, A. Shockravi, M. Kamali, A. Rafieimanesh and A. M. Malek, *J. Polym. Sci., Part A: Polym. Chem.*, 2013, **51**, 3505–3515.
11 S. Gazzo, G. Manfredi, R. Potzsch, Q. Wei, M. Alloisio, B. Voit and D. Comoretto, *J. Polym. Sci., Part B: Polym. Phys.*, 2016, **54**, 73–80.
12 J. Hachmann, T. L. Windus, J. A. McLean, V. Allwardt, A. C. Schrimpe-Rutledge, M. A. F. Afzal and M. Haghighatlari, *Framing the role of big data and modern data science in chemistry*, Technical Report, 2018, NSF CHE Workshop Report.
13 R. S. Sánchez-Carrera, S. Atahan, J. Schrier and A. Aspuru-Guzik, *J. Phys. Chem. C*, 2010, **114**, 2334–2340.
14 A. N. Sokolov, S. Atahan-Evrenk, R. Mondal, H. B. Akkerman, R. S. Sánchez-Carrera, S. Granados-Focil, J. Schrier, S. C. B. Mannsfeld, A. P. Zoombelt, Z. Bao and A. Aspuru-Guzik, *Nat. Commun.*, 2011, **2**, 437–438.
15 J. Hachmann, R. Olivares-Amaya, S. Atahan-Evrenk, C. Amador-Bedolla, R. S. Sánchez-Carrera, A. Gold-Parker, L. Vogt, A. M. Brockway and A. Aspuru-Guzik, *J. Phys. Chem. Lett.*, 2011, **2**, 2241–2251.
16 R. Olivares-Amaya, C. Amador-Bedolla, J. Hachmann, S. Atahan-Evrenk, R. S. Sánchez-Carrera, L. Vogt and A. Aspuru-Guzik, *Energy Environ. Sci.*, 2011, **4**, 4849–4861.
17 C. Amador-Bedolla, R. Olivares-Amaya, J. Hachmann and A. Aspuru-Guzik, Organic Photovoltaics, in *Informatics for Materials Science and Engineering: Data-driven Discovery for Accelerated Experimentation and Application*, ed. K. Rajan, Butterworth-Heinemann, Amsterdam, 2013, ch. 17, pp. 423–442.
18 J. Hachmann, R. Olivares-Amaya, A. Jinich, A. L. Appleton, M. A. Blood-Forsythe, L. R. Seress, C. Roman-Salgado, K. Trepte, S. Atahan-Evrenk, S. Er, S. Shrestha, R. Mondal, A. Sokolov, Z. Bao and A. Aspuru-Guzik, *Energy Environ. Sci.*, 2014, **7**, 698–704.
19 E. O. Pyzer-Knapp, C. Suh, R. Gómez-Bombarelli, J. Aguilera-Iparraguirre and A. Aspuru-Guzik, *Annu. Rev. Mater. Sci.*, 2015, **45**, 195–216.
20 S. A. Lopez, E. O. Pyzer-Knapp, G. N. Simm, T. Lutzow, K. Li, L. R. Seress, J. Hachmann and A. Aspuru-Guzik, *Sci. Data*, 2016, **3**, 160086.
21 T. D. Huan, A. Mannodi-Kanakkithodi, C. Kim, V. Sharma, G. Pilania and R. Ramprasad, *Sci. Data*, 2016, **3**, 160012.
22 V. Sharma, C. Wang, R. G. Lorenzini, R. Ma, Q. Zhu, D. W. Sinkovits, G. Pilania, A. R. Oganov, S. Kumar, G. A. Sotzing, S. A. Boggs and R. Ramprasad, *Nat. Commun.*, 2014, **5**, 4845.
23 A. Mannodi-Kanakkithodi, G. Pilania, T. D. Huan, T. Lookman and R. Ramprasad, *Sci. Rep.*, 2016, **6**, 20952.
24 A. Alexandridis, E. Chondrodima, K. Moutzouris and D. Triantis, *J. Mater. Sci.*, 2012, **47**, 883–891.
25 S. S. Park, S. Lee, J. Y. Bae and F. Hagelberg, *Chem. Phys. Lett.*, 2011, **511**, 466–470.
26 H. Redmond and J. E. Thompson, *Phys. Chem. Chem. Phys.*, 2011, **13**, 6872–6882.
27 G. Lisa, S. Curteanu and C. Lisa, *Environ. Eng. Manage. J.*, 2010, **9**, 483–487.
28 X. L. Yu, B. Yi and X. Y. Wang, *J. Comput. Chem.*, 2007, **28**, 2336–2341.
29 A. J. Holder, L. Ye, J. D. Eick and C. C. Chappelow, *QSAR Comb. Sci.*, 2006, **25**, 905–911.
30 M. A. F. Afzal, C. Cheng and J. Hachmann, *J. Chem. Phys.*, 2018, **148**, 241712.
31 G. Slonimskii, A. Askadskii and A. Kitaigorodskii, *Polym. Sci. USSR*, 1970, **12**, 556–577.
32 H. Drucker, C. J. Burges, L. Kaufman, A. J. Smola and V. Vapnik, Support vector regression machines, *Advances in Neural Information Processing Systems*, 1997, pp. 155–161.
33 A. J. Smola and B. Schölkopf, *Statistics and Computing*, 2004, vol. 14, pp. 199–222.

34 R. Parr and Y. Weitao, *Density-Functional Theory of Atoms and Molecules*, International Series of Monographs on Chemistry, Oxford University Press, 1994.
35 W. Koch and M. C. Holthausen, *A chemist's guide to density functional theory*, John Wiley & Sons, 2015.
36 J. P. Perdew, A. Ruzsinszky, J. Tao, V. N. Staroverov, G. E. Scuseria and G. I. Csonka, *J. Chem. Phys.*, 2005, **123**, 062201.
37 J. P. Perdew, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1986, **33**, 8822–8824.
38 A. D. Becke, *Phys. Rev. A: At., Mol., Opt. Phys.*, 1988, **38**, 3098–3100.
39 C. Lee, W. Yang and R. G. Parr, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1988, **37**, 785–789.
40 A. D. Becke, *J. Chem. Phys.*, 1993, **98**, 5648–5652.
41 C. Adamo and V. Barone, *J. Chem. Phys.*, 1999, **110**, 6158–6170.
42 J. Tao, J. P. Perdew, V. N. Staroverov and G. E. Scuseria, *Phys. Rev. Lett.*, 2003, **91**, 146401.
43 Y. Zhao and D. G. Truhlar, *Theor. Chem. Acc.*, 2008, **120**, 215–241.
44 Y. Zhao and D. G. Truhlar, *Acc. Chem. Res.*, 2008, **41**, 157–167, PMID: 18186612.
45 S. Grimme, *J. Chem. Phys.*, 2006, **124**, 034108.
46 F. Weigend, R. Ahlrichs, K. A. Peterson, T. H. Dunning, R. M. Pitzer and A. Bergner, *Phys. Chem. Chem. Phys.*, 2005, **7**, 3297.
47 S. Grimme, J. Antony, S. Ehrlich and H. Krieg, *J. Chem. Phys.*, 2010, **132**, 154104.
48 A. K. Rappe, C. J. Casewit, K. S. Colwell, W. A. Goddard and W. M. Skiff, *J. Am. Chem. Soc.*, 1992, **114**, 10024–10035.
49 N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch and G. R. Hutchison, *J. Cheminf.*, 2011, **3**, 33.
50 F. Neese, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2012, **2**, 73–78.
51 J. Hachmann, W. S. Evangelista and M. A. F. Afzal, *ChemHTPS 0.7* – An Automated Virtual High-Throughput Screening Program Suite for Chemical and Materials Data Generation, 2017, https://bitbucket.org/hachmanngroup/chemhtps.
52 J. Hachmann, M. A. F. Afzal, M. Haghighatlari and Y. Pal, *Mol. Simul.*, 2018, **44**, 921–929.
53 P. Mori-Sánchez, Q. Wu and W. Yang, *J. Chem. Phys.*, 2003, **119**, 11001–11004.
54 G. J. B. Hurst, M. Dupuis and E. Clementi, *J. Chem. Phys.*, 1988, **89**, 385–395.
55 M. A. F. Afzal, *From Virtual High-Throughput Screening and Machine Learning to the Discovery and Rational Design of Polymers for Optical Applications*, PhD thesis, University at Buffalo, 2018.

