Extrapolating DFT towards the complete basis set limit: Lessons from the PBE family of functionals

Peter Kraus*

School of Molecular and Life Sciences,
Curtin University,
GPO Box U1987, Perth 6845, WA, Australia

E-mail: peter.kraus@curtin.edu.au

Abstract

Extrapolation of density functional theory results from 2- and 3-ζ calculations is a promising method for extracting higher accuracy data from calculations of systems at the affordability limit. In this work, I present formulas for the determination of extrapolation parameters, that account for the make-up of the density functional approximation. The formulas are fitted to reproduce the complete basis set limit energies of PBE and related density functional approximations, using a set of 30 singlet diatomics. Their performance is extensively evaluated using standard benchmark datasets. The current formulas are shown to be transferable outside the PBE-family of functional approximations, with the resulting extrapolation parameters outperforming the previous, empirical values. A good performance of [2,3]-ζ extrapolations for interaction energies of systems with significant non-covalent character is confirmed, and holds even in systems of $\sim$100 atoms in size.

1 Introduction

Basis set extrapolation methods allow computational chemists to reliably approximate results of expensive calculations in larger basis sets by combining results obtained at the same level of theory from two or more cheaper calculations in smaller basis sets. This is particularly important in post-Hartree-Fock (HF) wavefunction theory, where the calculation of electronic correlation dominates the computational cost. It has long been known that the correlation energy (E^{corr}) approaches its infinite basis set limit ($E^{\mathrm{corr}}_{\infty}$) as a function of a cubic power law of the basis set size,

$$E^{\mathrm{corr}}_{\infty} = E^{\mathrm{corr}}_{X} - AX^{-\alpha} \tag{1}$$

where the extrapolation parameter $\alpha = 3$, X is the cardinal number of a finite correlation-consistent basis set, and A is a system-dependent parameter fitted to the E^{corr}_{X} data.[1] Notably, while this cubic power law is correct for all values of Xs, other empirically determined values of α may provide better results for $X \leq 3$.[2]

Extrapolation methods are not widely used in density functional theory (DFT) calculations. This may be due to several reasons: DFT calculations are computationally inexpensive compared to correlated WFT, reducing the applicability of such methods; there are many different empirical density functional approximations (DFAs) with each requiring its own α, meaning such approaches are seen as less general; reference datasets of complete basis set limit energies from DFT are not as common as their WFT counterparts, hindering the development of such methods; and even the true basis set limit convergence of HF was until recently a matter of debate, putting any DFT extrapolation attempts on an uncertain theoretical basis.

In a previous work,[3] I have shown that the extrapolation parameters α determined for a set of DFAs using numerical complete basis set calculations are surprisingly transferrable to other DFAs. However, the values of α are strongly dependent on the basis set family. For instance, extrapolated def2-[st]zvpd non-covalent interaction energies generally outperform their def2-qzvpd counterparts at 80% of the computational cost, which holds true even for double-hybrid density functional approximations (DHDFAs) as long as the correlated perturbation theory (PT) component is extrapolated appropriately. The best performing extrapolation function for DFT is the exponential–square root function (expsqrt), first proposed by Jensen:[4]

$$E_{\infty}^{\mathrm{fctl}} = E_{X}^{\mathrm{fctl}} - Ae^{-\alpha\sqrt{X}} \tag{2}$$

This function has since been proven to be the correct form for energy convergence in both HF and Kohn-Sham DFT with Gaussian basis sets.[5] However, several issues higlighted in the previous work remain. The determination of the values of the extrapolation parameters α, as described in Ref. 3, is highly method-specific due to the introduced averaging. Additionally, the values of α clearly scale with the inclusion of exact (HF) exchange in hybrid DFAs, but the introduction of formulas which correct the α by the percentage of exact exchange did not improve the results. Finally, the influence of the reduced correlation component in double hybrid exchange–correlation functionals upon introduction of PT correlation has not been investigated.

In the current work, I address the three above issues by systematically investigating of PBE-like DFAs. Compared to the set of 9 diatomics used in previous work,[3] a larger dataset of complete basis set energies from finite element calculations is used, comprising 30 singlet diatomic species, and including cations and anions. This dataset is evaluated for PBE and related functionals with varying fractions of HF exchange and PT correlation. I investigate trends over a large number of basis set families, including most of the modern basis set families used with DFT and WFT. This leads to a formula which is specific for each basis set family, but universal for all DFAs, taking into account all components of the DFA recipe. The extrapolation parameters α obtained from these formulas are then thoroughly examined using several benchmark databases.

2 Computational methods

Analogously to previous work, I use the following nomenclature for the total energy of a DFA (E^{DFA}) and its components:

$$E^{\mathrm{DFA}} = E^{\mathrm{fctl}} + \mathit{\Delta} E^{\mathrm{dh}} + \mathit{\Delta} E^{\mathrm{disp}} + \mathit{\Delta} E^{\mathrm{nl}} \tag{3}$$

The first component is the self-consistent-field (SCF) energy of the exchange-correlation functional (fctl), the second component is the double-hybrid (dh) correction, the third term corresponds to dispersion corrections, such as Grimme's -D3,[9] and the fourth term is a non-local correlation term, such as in the VV10 DFA,[10] or generalised as -NL correction.[11]

The first two terms can be further split into the following components:

$$E^{\mathrm{fctl}} = (1 - a_{\mathrm{x}})E^{\mathrm{x}} + a_{\mathrm{x}}E^{\mathrm{HFx}} + (1 - a_{\mathrm{c}})E^{\mathrm{c}} \tag{4}$$

$$\mathit{\Delta} E^{\mathrm{dh}} = a_{\mathrm{c}}E^{\mathrm{PTc}} \tag{5}$$

Note that the coefficients a_{x} and a_{c} correspond to the fraction of HF exchange (HFx) and PT correlation (PTc) in the DFA recipe. While a_{x} is constrained to ≤ 1,[6] no such rigorous constraint on a_{c} exists.[7] In practice, most common DHDFAs obey both constraints, with the notable exception of the spin-component-scaled double hybrids (DSD family).[8] For simplicity, in the following discussion it is assumed both constraints hold. The fit is always performed using the exact values of the coefficients in front of E^x and E^c in Eq. (4), even if the latter is not equal to $(1 - a_c)$. Finally, while in principle the $\mathit{\Delta} E^{\mathrm{nl}}$ term contributes to the SCF cycle, it has been conclusively shown to produce nearly identical results when applied as a post-SCF correction.[12] The post-SCF treatment is applied throughout this work.

The $\{a_{\mathrm{x}}, a_{\mathrm{c}}\}$ parameter space explored in this work consists of all 20 combinations of $a_{\mathrm{x}} \in$

$\{0.1, 0.3, 0.5, 0.7, 0.9\}$ with $a_c \in \{0.0, 0.2, 0.4, 0.6\}$ as well as the pure variant (i.e. $a_x = a_c = 0.0$). The parameter sets are chosen to cover the range used by most of the common single and double hybrids. The exchange and correlation functionals of Perdew, Burke, and Ernzerhof (PBE)[13] are used, as they represent a "non-empirical" DFA, even though it could be argued parameter-counting is a questionable metric of empiricity.[14] The generalization of the results obtained from this family of PBE-like functionals is discussed below.

Table 1: Set of 30 diatomic molecules, including their bond lengths (R) and the ℓ_{max} of the basis sets used in the finite element calculations.

Diatomic	**R** (Å)	ℓ_{max}	
CH^+	1.13085	15	11
OH^+	1.02890	17	13
OH^-	0.94246	15	13
FH	0.91696	17	13
C_2	1.24780	17	11
CN^+	1.17290	17	13
CN^-	1.17160	17	13
N_2	1.09434	17	11
NO^+	1.06206	17	13
CO	1.12821	17	13
CF^+	1.22875	19	15
OF^-	1.49228	21	17
F_2	1.41184	21	15
SiH^+	1.50410	25	19
SH^-	1.34993	27	21
HCl	1.29119	27	21
CP^-	1.58753	27	21
CS	1.53442	27	21
SiN^-	1.55578	27	21
NP	1.49085	27	21
SN^+	1.44000	27	21
SiO	1.50974	25	19
PO^+	1.41900	25	19
SF^-	1.70395	29	23
ClF	1.66162	29	23
SiS	1.93000	31	23
P_2	1.89340	31	23
PS^+	1.87200	31	23
SCl^-	2.14846	33	25
Cl_2	2.04262	33	25

The complete basis set energies of the 30 diatomic molecules listed in Table 1 are calculated using the `diatomic` component of the finite element code HelFEM,[15] using parameters determined with the `diatomic_cbasis` tool. Three anionic diatomics were difficult to converge using the following combinations of a_x and a_c: SF^- did not converge with $a_\mathrm{x} = 0.1$ and $a_\mathrm{c} = 0.6$, and OF^- as well as OH^- failed to converge with the pure functional, as well as for all four values of a_c with $a_\mathrm{x} = 0.1$. This is attributed to the diffuse nature of the anion exacerbated by the self-interaction error in PBE. Neither increasing the numerical integration radius from 40 Å to 120 Å, nor increasing the grid parameters led to convergence.

All other calculations were performed with Psi4 version 1.4a2.dev923,[16] or a development version of the program including basis set extrapolation routines for DFT implemented as part of previous[3] and current work. Calculations of the 30 diatomic molecules were carried out using the PK-supermatrix SCF algorithm to avoid issues with missing auxilliary basis sets, with a tightened energy convergence (10^{-10} Eh), and a large DFT quadrature with 150 radial and 974 angular points. The basis set families investigated in this work are listed in Table 2, significantly expanding on previous work where only four basis set families were compared. All basis sets were used as-included with Psi4, or downloaded from the Basis Set Exchange.[17,18] It is notable that the energy obtained with 4-ζ variants of cc-pwcv$X$z and aug-cc-pwcv$X$z is in many cases lower than that with the 5-ζ variants; this non-convergence does not happen with other basis sets.

Results of the calculations of the diet100 variant[37] of the GMTKN55 database[38] are reused from previous work, comprising 4 basis set families and 15 DFAs (see Table S1 for a detailed list); the WTMAD-2 criterion is used as figure of merit.[3] Additional calculations of the ASCDB database[39] were carried out using the standard density-fitted SCF algorithm, and the same (150, 974) grid, comprising 8 basis set families (marked with † in Table 2) and 3 PBE-like DFAs (PBE-D3(BJ),[13,40] PBE0-D3(BJ),[40,41] and PBE0DH-D3(BJ)[42,43]). Cases including elements for which basis sets are not defined within a given basis set family are excluded from the dataset of the corresponding basis set family. The ASCDB database allows for decomposition of the overall error metric into components of non-covalent interactions (NCI), thermochemistry

Table 2: Basis set families, the ζ range used in extrapolation, and whether the basis set includes tighter core d-functions (Core) and/or diffuse augmentation (Diffuse).

Family	ζ **range**	Core?	Diffuse?	Reference
cc-pv$X$z†	2 – 6	×	×	Dunning[19–22]
cc-pwcv$X$z†	2 – 5	✓	×	Dunning[19–21,23]
aug-cc-pv$X$z†	2 – 6	×	✓	Dunning[19–22]
aug-cc-pwcv$X$z†	2 – 5	✓	✓	Dunning[19–21,23]
def2-Xzvp†	2 – 4	×	×	Ahlrichs[24]
def2-Xzvpp	2 – 4	×	×	Ahlrichs[24]
def2-Xzvpd†	2 – 4	×	✓	Ahlrichs[24,25]
def2-Xzvppd	2 – 4	×	✓	Ahlrichs[24,25]
pc-N	1 – 5*	×	×	Jensen[26]
pcseg-$N^\dagger$	1 – 5*	×	×	Jensen[27]
aug-pc-N	1 – 5*	×	✓	Jensen[28]
aug-pcseg-$N^\dagger$	1 – 5*	×	✓	Jensen[27]
Xzapa-nr	2 – 6	×	✓	Ranasinghe[29]
Xzapa-nr-cv	2 – 6	✓	✓	Ranasinghe[29,30]
jorge-Xzp	2 – 6	×	×	Jorge[31–34]
jorge-aXzp	2 – 5	×	✓	Jorge[31,35,36]

* Note that for (aug-)pc-N and (aug-)pcseg-N families, $N \approx X + 1$.
† Basis set families used in the ASCDB benchmark.

(THCH), non-local effects (NLE), and unbiased calculations (UNBC),[39] allowing a more detailed insight into the performance of an extrapolation method than with the diet GMTKN55 subsets.[37]

The performance of basis set extrapolations for optimizing geometries of van der Waals complexes was evaluated using the NCDT database,[44] in its revised form,[45] using a development version of Psi4 with a tightened set of convergence criteria ($\Delta E < 1$ μEh, max $F < 15$ μEh/a_0, RMS $F < 10$ μEh/a_0, max $d < 0.6$ ma_0, RMS $d < 0.4$ ma_0, i.e. `interfrag_tight`) and an unpruned (150, 974) quadrature grid. The geometries were optimized using the def2-Xzvpd basis set family, used with revPBE-D3(BJ), M062X-D3, B97M-D3(BJ), B3LYP-D3(BJ), PBE0-D3, and ωB97X-D3(BJ), selected based on their performance in GMTKN55,[46] and previous NCDT results.[47]

To evaluate the performance of the basis set extrapolations in their most likely use case, that is in large systems with a significant non-covalent character, the interaction energies of the geometries from the L7 dataset[48] were evaluated using M062X-D3/def2-Xzvpd and B97M-V/pcseg-N.

This benchmark was designed to capture the dominant binding motifs in biological chemistry,[48] comprising 7 systems ranging in size between 48 and 112 atoms. The revised reference energies using correlation corrections obtained from DLPNO-CCSD(T_0)/cc-pv[dt]z calculations[49] are used instead of the original values. The calculations for the L7 dataset are performed with Psi4 version 1.4a2.dev923, using the default unpruned (75, 302) grid and the frozen core approximation for all double hybrid functionals.

The values of A and α in Eq. (2) for each of the 30 diatomics are obtained using a non-linear least squares fit of Eq. (6) using the `curve_fit` function of the Python library `scipy.optimize`.[50]

$$\ln(E_X^{\mathrm{fctl}} - E_\infty^{\mathrm{fctl}}) = \ln(A) - \alpha\sqrt{X} \tag{6}$$

The E_X^{fctl} values correspond to the energies from individual basis sets in each family listed in Table 2, with $X \approx \zeta$. The complete basis set limit value E_∞^{fctl} is the value obtained from HelFEM. The extrapolation parameters α are then averaged based on the system charge, obtaining the means ($\overline{\alpha}$) and medians ($\tilde{\alpha}$) for the cationic (α^+), neutral (α^0), anionic (α^-), and all (α) diatomic species. Finally, a linear fit to the mean values $\overline{\alpha}$ of all combinations of $\{a_\mathrm{x}, a_\mathrm{c}\}$ is performed, using Eq. (7), obtaining the intercept α_0, and two linear coefficients α_x and α_c.

$$\overline{\alpha} = \alpha_0 + \alpha_\mathrm{x} a_\mathrm{x} + \alpha_\mathrm{c} a_\mathrm{c} \tag{7}$$

3 Results

3.1 Extrapolation formulas

The results for each basis set family at all a_x with $a_\mathrm{c} = 0$ are presented in Fig. 1. The cationic outlier present in all basis sets is OH^+; the anionic outlier most visible in def2-Xzvp is OH^-. With

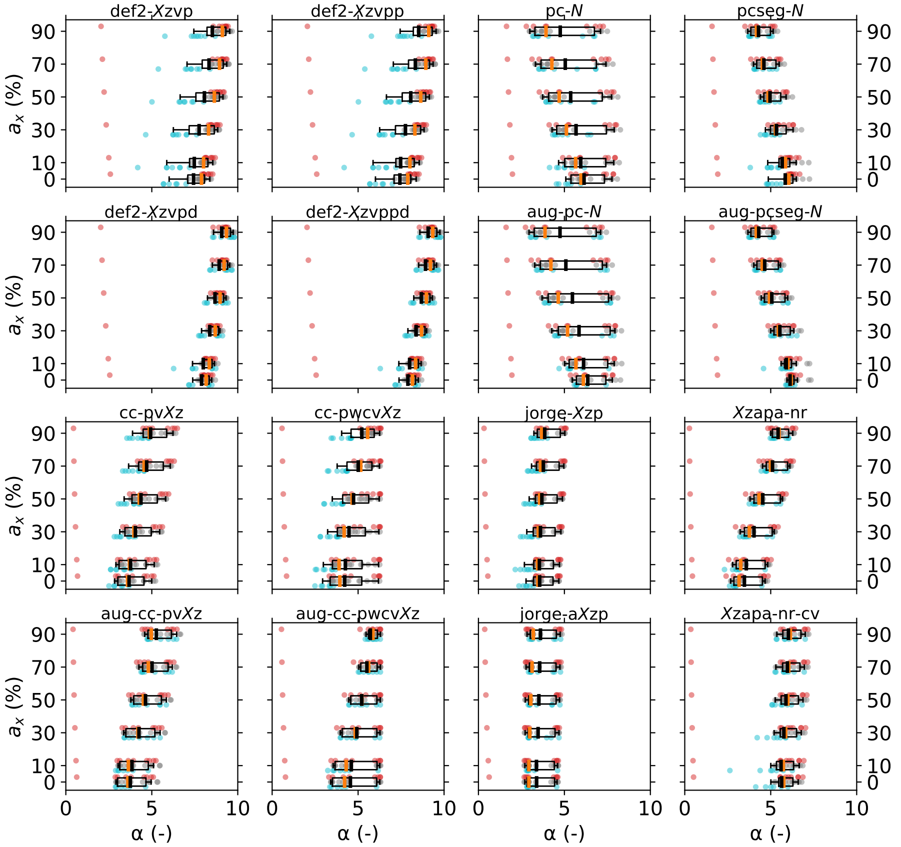

Figure 1: Extrapolation parameters α obtained with various basis set families at $a_c = 0$ and a_x as shown on the vertical axis. Colour coding corresponds to cationic (●), neutral (●), and anionic (●) species, with the box plots showing the overall median ($\tilde{\alpha}$, ▌), mean ($\overline{\alpha}$, ▌), quartiles, and the 5th and 95th percentile.

the exception of the Jensen basis set families, all $\overline{\alpha}$s show a positive correlation with a_x. The inclusion of HF exchange into the PBE recipe actually increases the overall self-interaction error (SIE) for hydrogenic atoms, as for the pure functional the exchange and correlation contributions to the SIE cancel each other out, and at non-zero a_x this error cancellation is lost.[51] However, for diatomics, the inclusion of HF exchange considerably improves the performance of the DFA in the SIE4x4 benchmark.[51] In any case, with a fraction of HF exchange, the electronic density is less delocalised. As the $\exp(-\alpha\sqrt{X})$ term accounts for the error in the convergence of energy with respect to basis set size X,[52] it is not surprising that for more compact densities the errors vanish quicker. This may also explain why no such correlation holds for a_c.

The augmentation of the basis sets by diffuse functions leads to a clear shift of the extrapolation parameters α for anionic species (●) to values in line with the neutral and cationic species. This trend has been discussed in the context of basis set extrapolations previously.[3,53] Here, I wish to reiterate the need for the use of diffuse functions whenever anionic systems are involved. The data show no statistically significant difference between $\overline{\alpha^+}$ and $\overline{\alpha^0}$.

The addition of tighter core *d*-electron functions pushes the extrapolation parameters α for many of the diatomics upwards towards ~ 6. This trend is most apparent for aug-cc-pv$X$z → aug-cc-pwcv$X$z augmentation, however it is also present in cc-pv$X$z → cc-pwcv$X$z, as well as for Xzapa-nr → Xzapa-nr-cv. It should be noted that the value of α that is optimal for predicting the error of the energy expectation value approaches $\pi\sqrt{3} \approx 5.44$ from below as basis set size approaches completeness.[52]

Table 3: Exchange and correlation dependence of $\overline{\alpha}$ for each basis set family studied. Coefficients are as per Eq. (7).

Family	α_0	α_x	α_c
cc-pv$X$z	3.622	1.511	0.005
cc-pwcv$X$z	4.157	1.192	-0.048
aug-cc-pv$X$z	3.676	1.887	0.139
aug-cc-pwcv$X$z	4.485	1.445	0.085
pc-N	6.172	-1.623	0.183
pcseg-N	5.883	-1.825	0.227
aug-pc-N	6.390	-1.874	0.260
aug-pcseg-N	6.166	-2.137	0.296
def2-Xzvp	7.406	1.266	-0.046
def2-Xzvpp	7.408	1.267	-0.046
def2-Xzvpd	7.925	1.370	0.101
def2-Xzvppd	7.927	1.371	0.101
jorge-Xzp	3.531	0.338	-0.011
jorge-aXzp	3.386	0.245	0.033
Xzapa-nr	3.306	2.525	-0.040
Xzapa-nr-cv	5.618	0.490	-0.016

3.2 ASCDB database

The mean unsigned errors (MUE) in the ASCDB database are shown for three widely used PBE-like functionals in Fig. 2. As expected, the pure functional PBE-D3(BJ) is outperformed by the single hybrid PBE0-D3(BJ). However, the double hybrid PBE0DH-D3(BJ) does not provide a systematic improvement over the single hybrid. Diffusely augmented basis set families outperform the corresponding non-augmented families in all cases, which is not surprising given the inclusion of non-covalent interactions in the ASCDB database. The Jensen basis set families (pcseg-*N* and aug-pcseg-*N*) systematically outperform both the original Dunning (cc-pv*X*z and aug-cc-pv*X*z) as well as the Ahlrichs (def2-*X*zvp and def2-*X*zvpd) sets. The weighted core-valence correlation consistent Dunning sets including tight *d*-functions (cc-pwcv*X*z and aug-cc-pwcv*X*z) offer comparable performance to the Jensen sets, albeit at a higher computational cost.

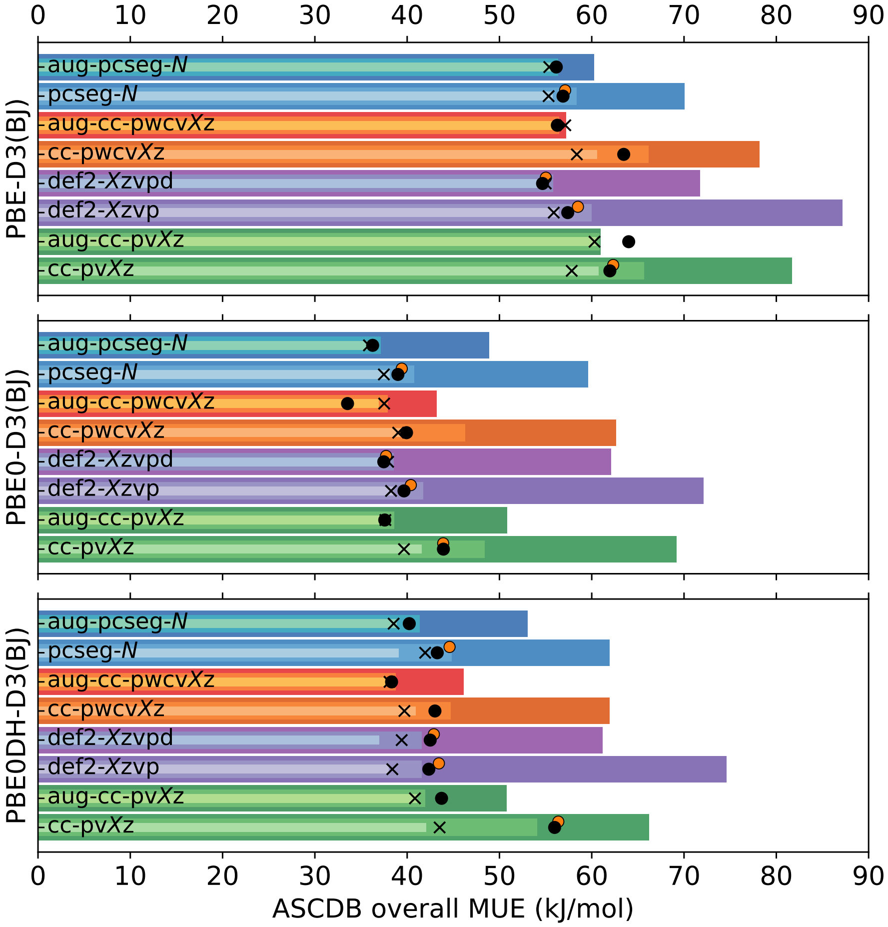

Figure 2: The mean unsigned errors of various functional and basis set combinations in the ASCDB database. Calculations with 2-, 3-, and 4-ζ basis sets shown as bars. Results from the [2,3]-ζ extrapolation (●) and [3,4]-ζ extrapolation ($\times$) from current work compared to previous [2,3]-ζ results (●), where available.

The performance of the [2,3]-ζ extrapolation (●) using the formulas from Table 3 is encouraging. The extrapolations generally outperform the 3-ζ calculations, with the following exceptions: The extrapolation using aug-cc-pv$X$z basis sets with PBE-D3(BJ) performs worse than the 2-ζ calculations; and the extrapolations using def2-Xzvp, def2-Xzvpd, cc-pv$X$z, or aug-cc-pv$X$z basis sets with PBE0DH-D3(BJ) perform worse than 3-ζ results. These exceptions will be analysed further below. Notably, the previously determined [2,3]-ζ extrapolation parameters (●)[3] perform worse than the current formulas in all cases. The results are especially encouraging as the revised formulas perform significantly better with the Jensen basis sets. Finally, the [3,4]-ζ extrapolation (×) rarely outperforms the 4-ζ results, and as such cannot be recommended.

Detailed results for the various subsets of the ASCDB database are included in the Supplementary Information (Figs. S1–S4). In the non-covalent interactions subset, [2,3]-ζ extrapolation performs significantly better than the 3-ζ calculations using all basis set families and functionals. Generally results comparable to 4-ζ quality can be expected. In the thermochemistry subset, the performance of the [2,3]-ζ extrapolation is more mixed. In the most extreme cases, the MUE increases with basis set size (e.g. PBE-D3(BJ) with aug-cc-pv$X$z or aug-cc-pwcv$X$z); the resulting poorer performance of extrapolated data can be attributed to the functional as opposed to the basis sets. That said, when 3-ζ and 4-ζ results are compared, the MUE in the thermochemistry data does not improve as significantly as for non-covalent interactions, and the [2,3]-ζ extrapolation is unlikely to increase the errors. In systems with significant non-local character, extrapolation methods offer no benefit; in fact it is this class of problems which causes the poor performance of PBE0DH-D3(BJ) when coupled with a [2,3]-ζ extrapolation using def2-Xzvp, def2-Xzvpd, aug-cc-pv$X$z, or aug-cc-pv$X$z basis sets. This is of course a more general issue with DFT as opposed to a basis set completeness error. Notably, the non-local effects subset is the only subset where the double-hybrid PBE0DH-D3(BJ) consistently outperforms the single hybrid PBE0-D3(BJ) when used with 4-ζ basis sets; the [2,3]-ζ extrapolation applied to cc-pwcv$X$z and aug-cc-pwcv$X$z yields results comparable to the 4-ζ data. Finally, in the unbiased benchmarks subset of ASCDB, the performance of the [2,3]-ζ extrapolation is also underwhelming. This might be correlated with the significant increase in performance of the 3-ζ basis sets compared to the 2-ζ results, and the comparably minor improvement upon a further increase in the basis set size to 4-ζ. The 2-ζ results are likely too inaccurate to improve the 3-ζ results by extrapolation.

3.3 GMTKN55 database

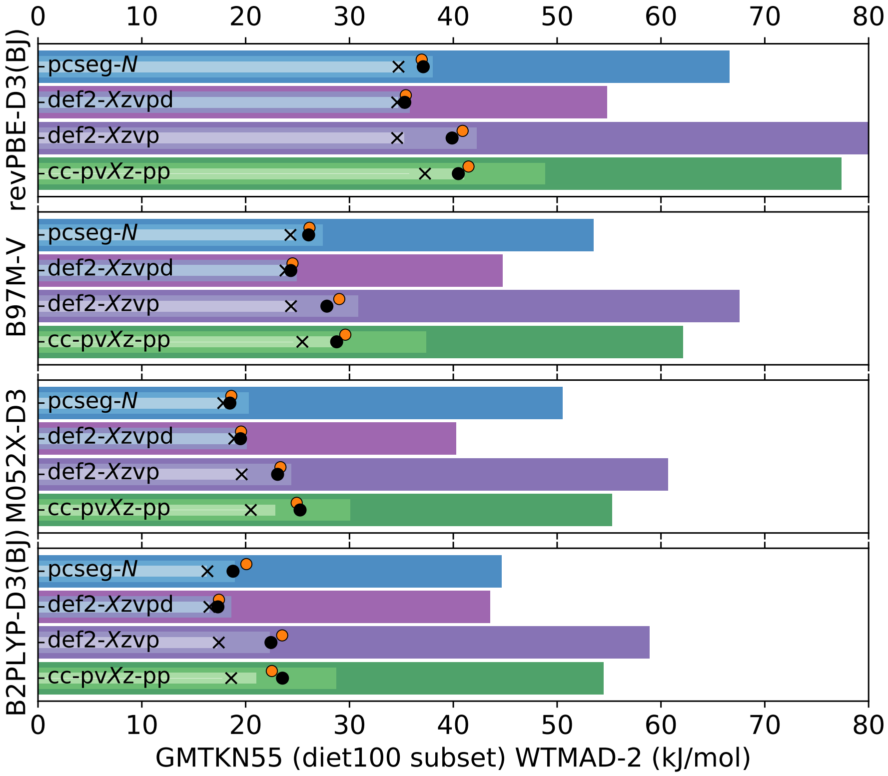

Figure 3: The weighted total mean absolute deviation (WTMAD-2) of various functional and basis set combinations in the diet100 subset of the GMTKN55 database. Calculations with 2-, 3-, and 4-ζ basis sets shown as bars. Results from the [2,3]-ζ extrapolation (•) and [3,4]-ζ extrapolation (×) from current work compared to previous [2,3]-ζ results (•), where available.

The results of basis set extrapolation with four representative DFAs for the diet100 subset of the GMTKN55 database is shown in Fig. 3. More detailed results, calculated with a wider range of functionals and the cc-pv$X$z-pp, def2-Xzvp, def2-Xzvpd, and pcseg-N basis set families, are shown in the Supplementary Information (Figs. S5). In the vast majority of cases, the current [2,3]-ζ extrapolation (•) outperforms or matches the previous extrapolation results (•), with the only exception being the cc-pv$X$z-pp results (see e.g. M052X-D3 and B2PLYP-D3(BJ) data). The four DFAs shown in Fig. 3 are all significantly different from the PBE family of functionals (the closest being revPBE, where the Lieb-Oxford bound is relaxed[54]), therefore this improvement in performance confirms the transferrability of the extrapolation formulas, which are based on PBE-like DFAs. Additionally, in all but two cases (DSD-BLYP-D3(BJ) with def2-Xzvp and pcseg-N basis sets), the current [2,3]-ζ extrapolation performs at least as well as 3-ζ calculations. As also shown in previous work,[3] the extrapolation using def2-[s,t]zvpd basis sets matches def2-qzvpd performance in this GMTKN55 subset. This is encouraging, as diffusely-augmented basis sets are necessary for proper description of anionic systems, and the def2-Xzvpd sets are defined for the whole periodic table. The performance of the [2,3]-ζ extrapolation with pcseg-N basis sets is generally also improved, especially with DHDFAs. However, as also noted previously,[3] the pcseg-2 results (3-ζ quality) generally outperform the [2,3]-ζ extrapolation with def2-Xzvp and cc-pv$X$z-pp and are often comparable to the cc-pvqz-pp results. The choice of an appropriate basis set family therefore remains crucial and cannot be remedied by basis set extrapolation.

3.4 NCDT database

Given the good performance of extrapolation schemes for interaction energies of systems with significant non-covalent character, a similarly good performance may be expected for obtaining geometries of non-covalent complexes with such schemes. The results for the NCDT database, presented as the mean absolute deviations of selected bonds in each complex, summed over all 17 complexes, are shown in Fig. 4 for six density functional approximations, three of which have been recommended based on the GMTKN55 benchmarks,[46] and two (B3LYP-D3(BJ) and PBE0-D3) have performed well in previous NCDT benchmarks.[47]

To bring perspective to the extrapolation results (indicated by an the gray bar in Fig. 4), let us start with a few comments on the general performance of the DFAs and trends in the dataset. Firstly, with the exception of M062X-D3, all DFAs struggle with optimization of the interfragment angle of the $HCN\cdots H_2CO$ structure; in the case of B3LYP-D3(BJ) this complex alone accounts for ∼25% of the total MAD. Secondly, the three functionals chosen based on their GMTKN55 performance perform significantly worse than both B3LYP-D3(BJ) and PBE0-D3, which remain

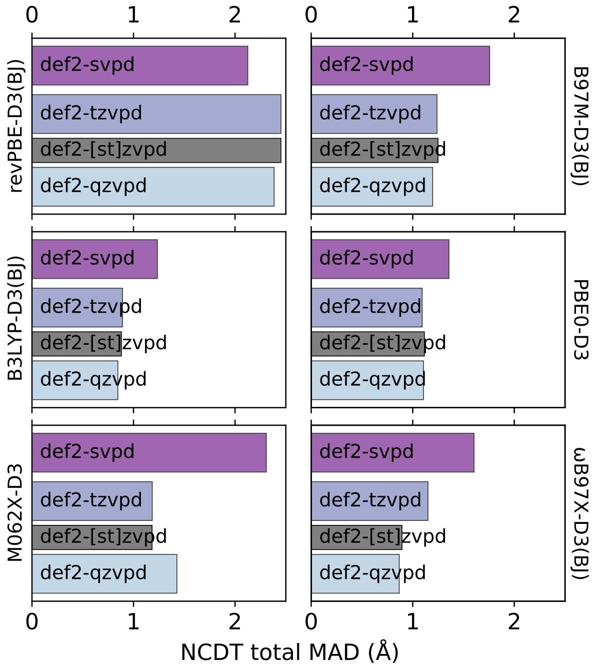

Figure 4: The total of the mean absolute deviations in the optimized structures of the 17 complexes in the NCDT database. The performance of the def2-[st]zvpd extrapolation is indicated by the gray bar.

good methods for structure optimization. However, the comparison is not quite fair, as the range-separated hybrids ωB97M-V and ωB97X-V, which were previously found to offer excellent performance with both GMTKN55 and NCDT, are here excluded from the study as analytical gradients for the non-local correction are not available in Psi4. Accordingly, the ωB97X-D3(BJ) variant, for which analytic gradients are available, is also evaluated and with a 4-ζ basis it shows results on par with B3LYP-D3(BJ). Finally, in most cases, the improvement from def2-svpd to def2-tzvpd is significant, while the difference between def2-tzvpd and def2-qzvpd is marginal, if any. With revPBE-D3(BJ), increasing the basis set size decreases the agreement with the reference data, which is likely an issue with the DFA. The M062X-D3 functional was trained using a 3-ζ basis set, therefore a decrease in performance with a 4-ζ basis is not unexpected. The comparably poor performance of the basis set extrapolation can then be attributed to the small potential for improvement in geometries past 3-ζ basis sets, and the fact that more accurate interaction energies do not necessarily translate to more accurate geometries. Basis set extrapolations therefore cannot be recommended to obtain more accurate structures with DFT.

3.5 L7 database

The most likely use-case of extrapolation methods in DFT would be for systems where a calculation of 3-ζ quality would approach the limit of affordability. The systems in the L7 database are comparably large, between 48 and 112 atoms in size,[48] yet reference interaction energies of near-CCSD(T) quality are now available.[49] This dataset is therefore an appropriate proxy for the performance of methods in such applications.

The results with a variety of combinations of DFAs and basis set families are shown in Fig. 5. The three double hybrids (DSD-BLYP-D3(BJ), DSD-PBEP86-NL, and B2PLYP-D3(BJ)) perform poorly compared to the less expensive methods, with a 3-ζ MAE above 20 kJ/mol and a systematic overbinding regardless of the basis set family. This may be due to the exceptionally slow convergence of the double hybrid correction ΔE^{dh}, especially in the two spin-component-scaled double hybrids. Indeed, the reported MAE of MP2 extrapolated towards the complete basis set limit exceeds 30 kJ/mol.[49] However, a more detailed analysis of the poor performance of this class of methods is beyond the scope of this work.

As also shown in Fig. 5, the [2,3]-ζ extrapolations outperform the 3-ζ results in all cases, with significant improvements over 3-ζ results using B97M-V/pcseg-[12], B3LYP-D3(BJ)/def2-[st]zvpd, as well as the three double hybrids. Notably, none of the six methods presented in Fig. 5 are members of the PBE-family of functionals, confirming the good transferability of the extrapolation formulas developed in the current work.

4 Conclusion

A systematic way of extrapolating density functional theory results towards the complete basis set limit can be derived from data computed with the PBE family of density functional approximations. The proposed formulas adjust the extrapolation α for each DFA based on the admixture

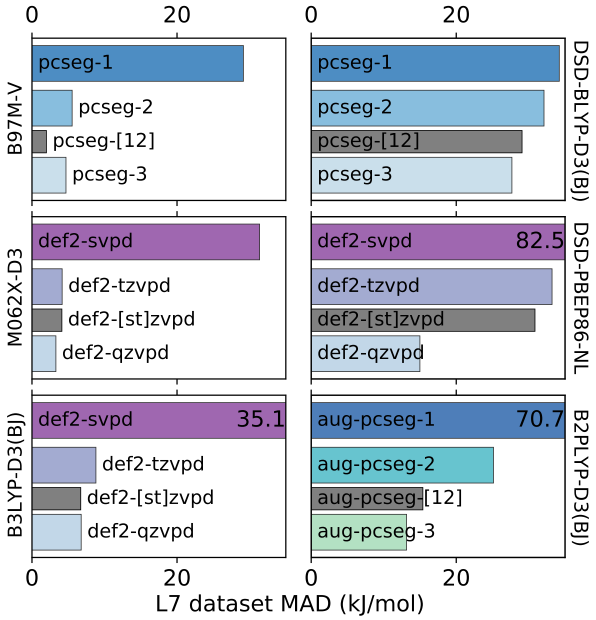

Figure 5: The mean absolute deviations in the interaction energies of the 7 structures in the L7 database. The performance of the [2,3]-ζ extrapolation methods is indicated by the gray bar.

of Hartree-Fock exchange and second-order perturbation theory correlation into the functional recipe. The extrapolation parameters α are also dependent on the basis set family used in the extrapolation. The use of diffusely-augmented basis sets is strongly recommended for anionic species.

A set of complete basis set energies for 30 singlet diatomic molecules has been calculated with 21 PBE-related DFAs using the finite element code HelFEM. The dataset is complemented by energies of the same 30 diatomic molecules, calculated with 16 basis set families and the same 21 PBE-related DFAs. This comprehensive set was used to fit linear formulas for scaling the extrapolation parameter α based on the fraction of HF exchange a_x and PT correlation a_c in each of the 21 DFAs. The formulas for calculating the parameter α are determined individually for each basis set familiy, while the extrapolation parameters α derived from such formulas are DFA as well as basis set family specific.

The current, systematically derived, extrapolation parameters outperform the previous, empirically averaged values. For single point energies in the ASCDB database, the performance of [2,3]-ζ extrapolations exceeds 3-ζ and generally approaches 4-ζ results in systems with significant non-covalent character. For thermochemistry calculations the [2,3]-ζ results at worst match and often outperform 3-ζ results. The improvement over 3-ζ results in interaction energies for the L7 database of large systems is also significant. On the contrary, for geometry optimizations of non-covalent complexes, the [2,3]-ζ extrapolation barely outperforms 3-ζ results and therefore cannot be recommended. The basis set extrapolation methods are implemented in a development version of Psi4, with extrapolated analytic gradient calculations as well as extrapolated numerical Hessians available.

Acknowledgement

I would like to thank Susi Lehtola for his help with HelFEM; Jan Gerit Brandenburg, Lars Goerigk, Amir Karton, Julian Gale, and Laura McKemmish for their comments and feedback; and the Forrest Research Foundation for funding. This work was supported by resources provided by the Pawsey Supercomputing Centre (project f97) and the National Computational Infrastructure (project f97), with funding from the Australian Government and the Government of Western Australia.

Supporting Information Available

Supporting information available: List of methods used with the diet100 variant of GMTKN55; Figures of results of the subsets of the ASCDB database; Figures of results of the diet100 subset of the GMTKN55 database. Additional supporting information, including the scripts generating the inputs, parsing the outputs, and creating figures are available on Zenodo under DOI: 10.5281/zenodo.4783007.

Table of contents graphic

References

(1) Helgaker, T.; Klopper, W.; Koch, H.; Noga, J. Basis-set convergence of correlated calculations on water. *The Journal of Chemical Physics* **1997**, *106*, 9639–9646.

(2) Truhlar, D. G. Basis-set extrapolation. *Chemical Physics Letters* **1998**, *294*, 45–48.

(3) Kraus, P. Basis set extrapolations for density functional theory. *J. Chem. Theory Comput.* **2020**, *16*, 5712–5722.

(4) Jensen, F. Estimating the Hartree—Fock limit from finite basis set calculations. *Theor Chem Acc* **2005**, *113*, 267–273.

(5) Shaw, R. A. The completeness properties of Gaussian-type orbitals in quantum chemistry. *Int J Quantum Chem* **2020**, *120*.

(6) Becke, A. D. A new mixing of Hartree–Fock and local density-functional theories. *The Journal of Chemical Physics* **1993**, *98*, 1372–1377.

(7) Grimme, S. Semiempirical hybrid density functional with perturbative second-order correlation. *Journal of Chemical Physics* **2006**, *124*, 034108–034108.

(8) Kozuch, S.; Martin, J. M. L. Spin-component-scaled double hybrids: An extensive search for the best fifth-rung functionals blending DFT and perturbation theory. *J. Comput. Chem.* **2013**, 2327–2344.

(9) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate *ab initio* parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *The Journal of Chemical Physics* **2010**, *132*, 154104.

(10) Vydrov, O. A.; Van Voorhis, T. Nonlocal van der Waals density functional: The simpler the better. *Journal of Chemical Physics* **2010**, *133*.

(11) Hujo, W.; Grimme, S. Performance of the van der waals density functional VV10 and (hybrid)GGA variants for thermochemistry and noncovalent interactions. *Journal of Chemical Theory and Computation* **2011**, *7*, 3866–3871.

(12) Najibi, A.; Goerigk, L. The nonlocal kernel in van der Waals density functionals as an additive correction: An extensive analysis with special emphasis on the B97M-V and ωB97M-V approaches. *J. Chem. Theory Comput.* **2018**, *14*, 5725–5738.

(13) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.

(14) Peverati, R. Fitting elephants in the density functionals zoo: Statistical criteria for the evaluation of density functional theory methods as a suitable replacement for counting parameters. *Int J Quantum Chem* **2020**,

(15) Lehtola, S. Fully numerical Hartree-Fock and density functional calculations. II. Diatomic molecules. *Int J Quantum Chem* **2019**, *119*.

(16) Smith, D. G. A.; Burns, L. A.; Simmonett, A. C.; Parrish, R. M.; Schieber, M. C.; Galvelis, R.; Kraus, P.; Kruse, H.; Di Remigio, R.; Alenaizan, A.; James, A. M.; Lehtola, S.; Misiewicz, J. P.; Scheurer, M.; Shaw, R. A.; Schriber, J. B.; Xie, Y.; Glick, Z. L.; Sirianni, D. A.; O'Brien, J. S.; Waldrop, J. M.; Kumar, A.; Hohenstein, E. G.; Pritchard, B. P.; Brooks, B. R.; Schaefer, H. F.; Sokolov, A. Y.; Patkowski, K.; DePrince, A. E.; Bozkaya, U.; King, R. A.; Evangelista, F. A.; Turney, J. M.; Crawford, T. D.; Sherrill, C. D. Psi4 1.4: Open-source software for high-throughput quantum chemistry. *J. Chem. Phys.* **2020**, *152*, 184108.

(17) Schuchardt, K. L.; Didier, B. T.; Elsethagen, T.; Sun, L.; Gurumoorthi, V.; Chase, J.; Li, J.; Windus, T. L. Basis Set Exchange: A Community Database for Computational Sciences. *J. Chem. Inf. Model.* **2007**, *47*, 1045–1052.

(18) Pritchard, B. P.; Altarawy, D.; Didier, B.; Gibson, T. D.; Windus, T. L. New Basis Set Exchange: An Open, Up-to-Date Resource for the Molecular Sciences Community. *J. Chem. Inf. Model.* **2019**, *59*, 4814–4820.

(19) Dunning, T. H. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *J. Chem. Phys.* **1989**, *90*, 1007–1022.

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

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

(22) Mourik, T. V.; Dunning, T. H. Gaussian basis sets for use in correlated molecular calculations. VIII. Standard and augmented sextuple zeta correlation consistent basis sets for aluminum through argon. *International Journal of Quantum Chemistry* **2000**, *76*, 205–221.

(23) Peterson, K. A.; Dunning, T. H. Accurate correlation consistent basis sets for molecular core–valence correlation effects: The second row atoms Al–Ar, and the first row atoms B–Ne revisited. *The Journal of Chemical Physics* **2002**, *117*, 10548–10560.

(24) Weigend, F.; Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297.

(25) Rappoport, D.; Furche, F. Property-optimized Gaussian basis sets for molecular response calculations. *Journal of Chemical Physics* **2010**, *133*, 0–11.

(26) Jensen, F. Polarization consistent basis sets. II. Estimating the Kohn–Sham basis set limit. *The Journal of Chemical Physics* **2002**, *116*, 7372–7379.

(27) Jensen, F. Unifying General and Segmented Contracted Basis Sets. Segmented Polarization Consistent Basis Sets. *J. Chem. Theory Comput.* **2014**, *10*, 1074–1085.

(28) Jensen, F. Polarization consistent basis sets. III. The importance of diffuse functions. *The Journal of Chemical Physics* **2002**, *117*, 9234–9240.

(29) Ranasinghe, D. S.; Petersson, G. A. CCSD(T)/CBS atomic and molecular benchmarks for H through Ar. *The Journal of Chemical Physics* **2013**, *138*, 144104.

(30) Ranasinghe, D. S.; Frisch, M. J.; Petersson, G. A. Core-core and core-valence correlation energy atomic and molecular benchmarks for Li through Ar. *The Journal of Chemical Physics* **2015**, *143*, 214110.

(31) Canal Neto, A.; Muniz, E.; Centoducatte, R.; Jorge, F. Gaussian basis sets for correlated wave functions. Hydrogen, helium, first- and second-row atoms. *Journal of Molecular Structure: THEOCHEM* **2005**, *718*, 219–224.

(32) Barbieri, P. L.; Fantin, P. A.; Jorge, F. E. Gaussian basis sets of triple and quadruple zeta valence quality for correlated wave functions. *Molecular Physics* **2006**, *104*, 2945–2954.

(33) Jorge, F.; Sagrillo, P.; de Oliveira, A. Gaussian basis sets of 5 zeta valence quality for correlated wave functions. *Chemical Physics Letters* **2006**, *432*, 558–563.

(34) Campos, C.; Ceolin, G.; Canal Neto, A.; Jorge, F.; Pansini, F. Gaussian basis set of sextuple zeta quality for hydrogen through argon. *Chemical Physics Letters* **2011**, *516*, 125–130.

(35) Fantin, P.; Barbieri, P.; Canal Neto, A.; Jorge, F. Augmented Gaussian basis sets of triple and quadruple zeta valence quality for the atoms H and from Li to Ar: Applications in HF, MP2, and DFT calculations of molecular dipole moment and dipole (hyper)polarizability. *Journal of Molecular Structure: THEOCHEM* **2007**, *810*, 103–111.

(36) de Oliveira, P.; Jorge, F. Augmented Gaussian basis set of quintuple zeta valence quality

for H and from Li to Ar: Applications in DFT calculations of molecular electric properties. *Chemical Physics Letters* **2008**, *463*, 235–239.

(37) Gould, T. 'Diet GMTKN55' offers accelerated benchmarking through a representative subset approach. *Phys. Chem. Chem. Phys.* **2018**, *20*, 27735–27739.

(38) Goerigk, L.; Hansen, A.; Bauer, C.; Ehrlich, S.; Najibi, A.; Grimme, S. A look at the density functional theory zoo with the advanced GMTKN55 database for general main group thermochemistry, kinetics and noncovalent interactions. *Phys. Chem. Chem. Phys.* **2017**, *19*, 32184–32215.

(39) Morgante, P.; Peverati, R. Statistically representative databases for density functional theory *via* data science. *Phys. Chem. Chem. Phys.* **2019**, *21*, 19092–19103.

(40) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* **2011**, *32*, 1456–1465.

(41) Adamo, C.; Barone, V. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *The Journal of Chemical Physics* **1999**, *110*, 6158–6158.

(42) Brémond, E.; Adamo, C. Seeking for parameter-free double-hybrid functionals: The PBE0-DH model. *The Journal of Chemical Physics* **2011**, *135*, 024106.

(43) Bousquet, D.; Brémond, E.; Sancho-García, J. C.; Ciofini, I.; Adamo, C. Non-parametrized functionals with empirical dispersion corrections: A happy match? *Theor Chem Acc* **2015**, *134*, 1602.

(44) Kraus, P.; Obenchain, D. A.; Frank, I. Benchmark-quality semiexperimental structural parameters of van der Waals complexes. *J. Phys. Chem. A* **2018**, *122*, 1077–1087.

(45) Kraus, P. Non-covalent dimers and trimers (NCDT) database version 2.1. 2020.

(46) Goerigk, L.; Mehta, N. A trip to the density functional theory zoo: Warnings and recommendations for the user. *Aust. J. Chem.* **2019**, *72*, 563.

(47) Kraus, P.; Frank, I. Density functional theory for microwave spectroscopy of noncovalent complexes: A benchmark study. *J. Phys. Chem. A* **2018**, *122*, 4894–4901.

(48) Sedlak, R.; Janowski, T.; Pitoňák, M.; Řezáč, J.; Pulay, P.; Hobza, P. Accuracy of Quantum Chemical Methods for Large Noncovalent Complexes. *J. Chem. Theory Comput.* **2013**, *9*, 3364–3374.

(49) Ballesteros, F.; Dunivan, S.; Lao, K. U. Coupled cluster benchmarks of large noncovalent complexes: The L7 dataset as well as DNA–ellipticine and buckycatcher–fullerene. *J. Chem. Phys.* **2021**, *154*, 154104.

(50) SciPy 1.0 Contributors,; Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; van der Walt, S. J.; Brett, M.; Wilson, J.; Millman, K. J.; Mayorov, N.; Nelson, A. R. J.; Jones, E.; Kern, R.; Larson, E.; Carey, C. J.; Polat, İ.; Feng, Y.; Moore, E. W.; VanderPlas, J.; Laxalde, D.; Perktold, J.; Cimrman, R.; Henriksen, I.; Quintero, E. A.; Harris, C. R.; Archibald, A. M.; Ribeiro, A. H.; Pedregosa, F.; van Mulbregt, P. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* **2020**, *17*, 261–272.

(51) Lonsdale, D. R.; Goerigk, L. The one-electron self-interaction error in 74 density functional approximations: A case study on hydrogenic mono- and dinuclear systems. *Phys. Chem. Chem. Phys.* **2020**,

(52) Kutzelnigg, W. Expansion of a wave function in a Gaussian basis. I. Local versus global approximation. *Int. J. Quantum Chem.* **2013**, *113*, 203–217.

(53) Varandas, A. J. Straightening the hierarchical staircase for basis set extrapolations: A low-cost approach to high-accuracy computational chemistry. *Annu. Rev. Phys. Chem.* **2018**, *69*, 177–203.

(54) Zhang, Y.; Yang, W. Comment on "Generalized gradient approximation made simple". *Phys. Rev. Lett.* **1998**, *80*, 890–890.