A guide to benchmarking enzymatically catalysed reactions: the importance of accurate reference energies and the chemical environment

Dominique A. Wappett · Lars Goerigk

Received: date / Accepted: date

Abstract We explore two significant factors on the outcomes of benchmark studies for enzymatically catalysed reactions, namely the level of theory of the benchmarks and the size of the model system used to represent the enzyme active site. For the benchmarks, we compare two potential alternatives to canonical coupled cluster results for situations where CCSD(T) is computationally too demanding: a strategy to estimate finite basis set coupled cluster values and the local-correlation DLPNO-CCSD(T) method at the complete basis set limit. We confirm the high accuracy of DLPNO-CCSD(T) used with tight thresholds. We also show that notable differences can be seen when using both sets of references for a benchmark study, with absolute deviations from the higher quality references generally smaller than those from lower-quality ones as well as changes in the ranking of the assessed methods. For geometries, we test three models for the active site of 4-oxalocrotonate tautomerase: one typical of the QM region that may be used in QM/MM studies, and two smaller variants that neglect the surrounding chemical environment. Benchmarking of 12 density functionals known to perform well on enzymatically catalysed reactions shows inconsistent performance of each method across the three models, contradicting the common idea that small representative systems can be used to accurately assess the applicability of low-level methods for larger biochemical applications. Our findings shall serve as a reminder on the standards that should be adhered to in benchmark studies, and as a guide for future studies, both on enzyme-related and other chemical problems.

Dominique A. Wappett
The University of Melbourne, Parkville, Australia

Lars Goerigk
The University of Melbourne, Parkville, Australia
E-mail: lgoerigk@unimelb.edu.au

1 Introduction

Enzymes have long been of keen interest to (bio)chemists due to their remarkable catalytic properties—not only are they especially efficient, they are also highly selective and allow for reaction conditions that are milder than those required with many inorganic catalysts. The treatment of enzymatically-catalysed reactions on the molecular level through computational techniques enables researchers to explore possible mechanisms and the particular structural factors that affect an enzyme's efficiency. This information can then be used to explain experimental data, as well as modify or design new enzymes for specific applications, such as treatment of pollutants and drug discovery [1–7]. While theoretically designed enzymes have yet to perform with the same efficiency as naturally-occurring ones [8], they can often be improved with the highly successful directed evolution method [9–12], the importance of which can be seen in the 2018 Nobel Prize in Chemistry which was awarded to 50 % to Prof. Frances Arnold. Complementarily, directed evolution works best from a good starting point, so synthetic enzyme design also benefits from the insights of computational studies.

Quantum-chemical methods grow ever more useful for the analysis of enzymatically-catalysed reactions due to improvements in computer hardware and software that allow for the treatment of larger systems. One approach involves "cluster" models [13] which treat the active site and surrounding enzyme environment inside a dielectric cavity to account for polarisation effects, but the models must often be large to include enough

of the chemically important environment. It can thus be more efficient to use a hybrid quantum mechanics/molecular mechanics (QM/MM) approach [14–18], which involves treating the active site with a higher (QM) level of theory, and using faster (MM) methods to calculate the indirect contributions of the rest of the enzyme. As this provides a more specific definition of the surrounding environment, the "QM region" that contains the active site can be smaller. Density functional approximations (DFAs) are a common choice of QM method, but semi-empirical or ab initio methods are also possible. A reliable QM method is necessary to ensure good results, but accuracy must often be balanced with computational cost, and thus benchmark studies focussing on enzyme active-site models can be beneficial as they enable computational (bio)chemists to make informed choices of appropriate QM methods.

Benchmarking is ubiquitous in computational chemistry, as it benefits both the method developer who seeks to make accurate DFAs and the general user who wishes to use them. The process involves selecting a test set, calculating the relevant energies or properties at a highly accurate level of theory, and then comparing lower level methods against the benchmarks to assess their performance. The success of a study is, thus, highly dependent on the quality of the benchmarks, and poor reference values will alter the perceived performance of the tested methods, which can significantly change the conclusions of the study [19, 20]. The choice of test set is also important, as studies are most useful when they are guided by the result one aims to achieve. When the goal is to find robust and widely applicable methods, one should use broad test sets like GMTKN55 [20–23], MGCDB84 [24] and Database 2015B [25], but when choosing a method for a specific application, a more targeted test set of relevant reactions often gives better guidance.

One such specific set for biochemically relevant reactions is our set of barrier heights (BHs) and reaction energies (REs) for enzymatically catalysed reactions [26, 27], which is an updated version of an earlier set published by Kromann et al. [28], and contains active-site models of varying sizes for five specific enzymatically-catalysed reactions. Our work on this set focussed on the importance of London-dispersion corrections in geometry optimisations to ensure that the intermolecular enzyme-substrate interactions are represented correctly, and then subsequently finding appropriate benchmark levels of theory for systems of intermediate size using DLPNO-CCSD(T)—domain based local pair natural orbital coupled cluster with singles, doubles and perturbative triples [29, 30]. This method is a common recommendation for a reliable second choice when canonical CCSD(T) [31], often referred to as the "gold standard" of chemical accuracy, is not achievable due to computational constraints. It reduces the cost significantly by prescreening the pair correlations and only including those above a certain "PNO threshold" threshold at each step [32], thereby decreasing the number of pairs treated at the coupled cluster level while retaining most of the conventional approach's accuracy [32, 33]. Therefore, we thoroughly tested the different default PNO thresholds, as well as the choice of basis sets used for the extrapolation to the complete basis set (CBS) limit, to find an appropriate balance between cost and accuracy and suggested different strategies for obtaining reliable benchmarks based on system size [26].

This method is a common recommendation for a reliable second choice when canonical CCSD(T) [31], often referred to as the "gold standard" of chemical accuracy, is not achievable due to computational constraints. It reduces the cost significantly by prescreening the pair correlations and only including those above a certain threshold at each step [32], thereby decreasing the number of pairs treated at the coupled cluster level while retaining most of the accuracy [32, 33].

Other recent studies [34, 35] have used smaller models that included considerably smaller portions of the specific enzyme environment, which allowed for the use of canonical CCSD(T). Paiva et al. [35] have used a set of four minimalistic models that represented biochemical reactions to benchmark DLPNO-CCSD(T) alongside a range of DFAs, and showed that the error of DLPNO-CCSD(T)/CBS—extrapolated from double- and triple-ζ atomic-orbital (AO) basis sets—was 0.56 kcal/mol with the NormalPNO thresholds and 0.40 kcal/mol with the TightPNO thresholds, significantly larger than the reported <0.25 kcal/mol general error of DLPNO-CCSD(T) for reaction energies when used with the TightPNO thresholds in ref [32]. The benchmarks used were estimated CCSD(T)/CBS values, i.e. generally second-order Møller-Plesset perturbation theory (MP2/CBS) extrapolated from triple- and quadruple-ζ basis sets and then corrected with the difference in correlation energy between CCSD(T) and MP2 at the triple-ζ level or lower. While various extrapolation strategies and basis sets were tested for these estimated benchmarks to find an appropriate balance of accuracy and efficiency [36–39], the error of this type of estimation scheme is around 0.6 kcal/mol when the correction is calculated with aug-cc-pVDZ [36, 40, 41], already larger than the observed deviations of DLPNO-CCSD(T)/CBS in their study. We have also shown in our own tests of basis sets that not all extrapolations are equal, and we have recommended against DLPNO-CCSD(T)/CBS values

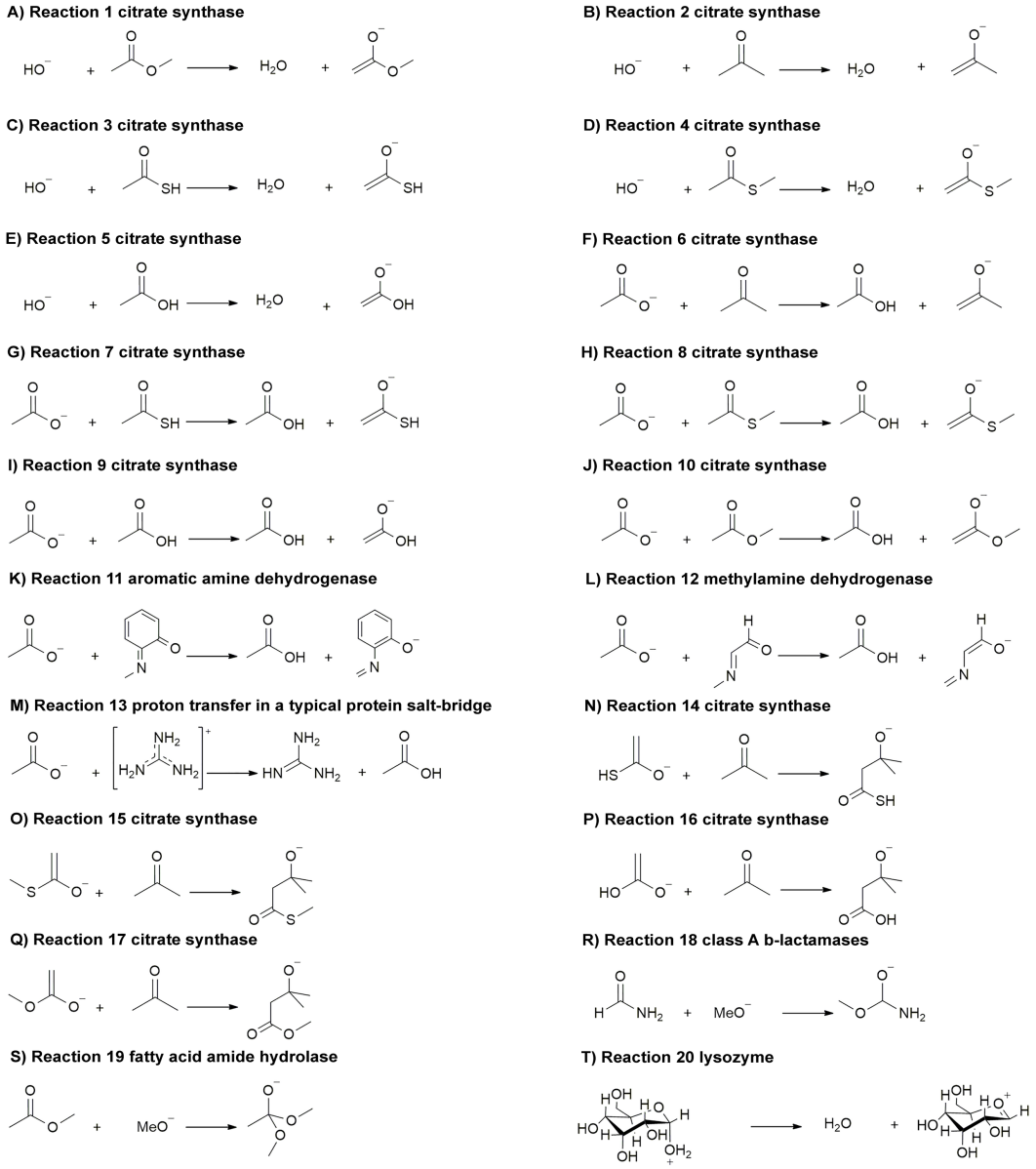

Fig. 1 Reactions in the enzyme reaction-energy test set. The reaction numbers and not the letters are used in the following discussion. The original figure was taken from ref [34] [Peer J 2020 (CC-BY licence)] and then modified to correct a typo in the product structure of reaction **20**.

extrapolated from double-/triple-ζ basis set pairs as they are often significantly different from values extrapolated from triple-/quadruple-ζ basis sets [26]. It is likely then that significantly lower errors will be seen when comparing DLPNO-CCSD(T)/CBS(TZ/QZ) to true CCSD(T)/CBS results, which is something that we will demonstrate herein.

Another recently published, small test set is that of Sirirak et al. [34], which takes the approach of modelling steps in enzyme reaction mechanisms through reactions of small molecules that represent common functional groups, with all reactant and product molecules being treated individually. Again, an estimated CCSD(T) benchmark was used by taking CCSD(T) [31]/aug-cc-pVDZ [42] numbers and correcting them with the difference between aug-cc-pVTZ [42] and aug-cc-pVDZ for spin-component-scaled MP2 (SCS-MP2) [43]. As well as the error introduced by estimating results, the use of a single, finite basis-set result is problematic for coupled cluster methods due to the slow convergence of the correlation energy [44], and therefore energies should be extrapolated to the CBS limit to get reliable results. Again we question the use of estimated results as benchmarks as the process can be unreliable, and any

error in the benchmarks will also affect the calculated errors of any methods tested against them—previous studies on pericyclic reactions and inorganic reactions have already shown how the quality of the reference values influences the outcome of Density Functional Theory (DFT) benchmarks [19, 20]. While quadruple-ζ level CCSD(T) results may not be achievable even with smaller models, estimation is unlikely to be the best second choice method for calculating reference values.

Based on these two studies and in light of the previous examples on how to conduct better benchmark studies, we pursue two main aims with this work. The first is to explore how various coupled cluster based levels of theory compare to each other, and then how the choice of benchmark affects the observed performance of a range of QM methods. For this purpose, we use the set of Sirirak et al. [34], which consists of 20 REs associated with reactions designed to be characteristic of steps in enzyme reaction mechanisms. The first 13 reactions in the set are proton transfer reactions, and the remaining seven are non-proton transfer reactions. All 20 reactions are shown in figure 1, with the enzyme used to catalyse each reaction listed alongside its scheme.

The second factor we aim to explore is how the choice of active-site model structures affects benchmarking. We have previously explored how the geometry-optimisation level of theory, particularly the inclusion of London-dispersion corrections, changes the calculated BHs and REs of a set of models for five enzymatically-catalysed reactions [26], and although it was not a specific focus of the work, the effect of increasing the model size could also be indirectly gauged from the reactions which had multiple models of different sizes. Herein, we consider how decreasing the size of the model changes the results, as many studies choose to use smaller models to test methods which will then be applied to larger ones.

There are multiple possibilities for how an enzyme's active site can be modelled—the enzyme RE test set we use herein uses small molecules that represent general functional groups of substrates and enzymes, while other studies have used models of the active site that contain the substrate and some parts of the surrounding enzyme environment like a cluster model or QM region [13, 26, 28, 45–47]. Smaller structures have the benefit of being less specific to a particular enzyme and computationally less demanding, but they can miss a large portion of potentially important non-covalent interactions and geometric factors when neighbouring amino-acid residues are not taken into account. The limitations of small models are most notable when the enzyme and substrate components are optimised and calculated separately. When benchmarking towards the goal of choosing a QM method for enzyme-related QM/MM studies, one seeks to find a method that will appropriately treat the enzyme-substrate interactions in the QM region, and so exclusion of these interactions brings into question whether conclusions drawn from testing smaller models apply to larger ones.

To explore the extent to which smaller, simplified models can impact the recommendations of a benchmark study, we use three models of the reaction catalysed by 4-oxalocrotonate tautomerase (4-OT), shown in figure 2. This enzyme converts unconjugated α-keto acids to their conjugated tautomers through a two-step mechanism, which involves a proton transfer from the substrate to the *N*-terminal catalytic proline residue (Pro1) in the first step, and another proton transfer from Pro1 back to the substrate in the second step. This reaction has been studied extensively with QM/MM methods [48–51], and is a clear example of the importance of the chemical environment, as the negatively charged substrate is strongly stabilised through hydrogen bonding and electrostatic effects, particularly in the charge-separated intermediate. The smallest recommended model that adequately captures this stabilisation consists of the substrate, Pro1 residue, three arginine residues and two water molecules (for further details see the description of "Model A" in ref [51]). While a small QM region is inappropriate for mechanistic studies of 4-OT and the results will not be consistent with experimental data of the whole enzyme, one can still question the extent to which a reduced model can impact the performance of methods in a benchmark study, where the calculated reference data will also be influenced by the deficiencies of the model.

When exploring both of these factors, we use a select set of DFAs to conduct multiple small benchmark studies—against two sets of references for the RE test set, and with three different models for the 4-OT reaction—in order to demonstrate how the perceived performance of the DFAs depends on the references and active-site model. We are confident this study serves as a guide for those seeking to benchmark model systems of enzymatically-catalysed reactions for QM/MM applications, in their

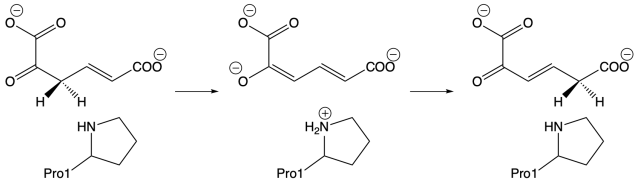

Fig. 2 Tautomerisation catalysed by 4-oxalocrotonate tautomerase (4-OT)

choices of both reference values and model systems. In fact, many findings of the present study can also be transferred to benchmarking of other scenarios and our guide can be seen as a standard that could be adopted in such studies.

In the following, we list the relevant computational details before we proceed with updating the reference values for the enzyme RE test set. We then use these and the old reference values to conduct an example benchmark study. We also briefly discuss the importance of using dispersion corrections with DFAs, comparing dispersion-uncorrected functionals to their DFT-D3(0) and -D3(BJ) corrected variants [52, 53]. We then turn to the model systems of the 4-OT active site, and again conduct an example benchmark study in order to assess the similarity of the results between the models.

2 Computational details

2.1 Calculation of reference values

Herein, we follow the protocol for calculating reference values determined in our previous work on enzymatically catalysed reactions. We therefore refrain from a detailed study of basis-set effects and other factors that affect the results, as they have all been discussed in ref [26]. Benchmark REs for all systems were obtained at the DLPNO-CCSD(T)/CBS level of theory, using the predefined TightPNO thresholds [32]. For the small molecules in the RE test set, the values were extrapolated from the augmented correlation consistent Dunning AO basis sets aug-cc-pVTZ and aug-cc-pVQZ [42]. The minimally augmented Ahlrichs-type basis sets ma-def2-TZVPP and ma-def2-QZVPP [54] were used for the new 4-OT models to ensure consistency with the original model, for which reference data at this level had already been published [26]. The resolution of the identity approximation for Coulomb integrals and chain of spheres approximation for exchange integrals (RIJ-COSX) [55] was also used with the Ahlrichs-type basis sets to speed up these calculations. Values were extrapolated to the CBS limit following the standard two-point extrapolation schemes for HF [56] and correlation [57] energies:

$$E_{SCF}^{(\infty)} = \frac{E_{SCF}^{(X)} \cdot exp(-\alpha\sqrt{Y}) - E_{SCF}^{(Y)} \cdot exp(-\alpha\sqrt{X})}{exp(-\alpha\sqrt{Y}) - exp(-\alpha\sqrt{X})}, \tag{1}$$

and

$$E_{Corr}^{(\infty)} = \frac{X^{\beta} \cdot E_{Corr}^{(X)} - Y^{\beta} \cdot E_{Corr}^{(Y)}}{X^{\beta} - Y^{\beta}}, \tag{2}$$

where X and Y are the cardinal numbers of the finite basis sets, $E_{SCF}^{(X/Y)}$ and $E_{Corr}^{(X/Y)}$ are the related HF (self consistent field, SCF) and correlation energies, respectively, and the ∞ indicates the CBS energies. α and β are basis set specific constants. For a triple-/quadruple-ζ extrapolation with the Dunning-type aug-cc-pVnZ basis sets these are 5.46 and 3.05, respectively; for the equivalent extrapolation with the Ahlrichs-type ma-def2-nZVPP basis sets they are 7.88 and 2.97 [58].

We calculated CCSD(T) results for reactions **3**, **5**, **7**, **9** and **13** from the RE test set shown in figure 1 using the aug-cc-pVTZ and aug-cc-pVQZ basis sets, and these results are also extrapolated using the above schemes. These particular reactions were chosen to compare the differences between various coupled-cluster approaches for calculating benchmarks as they are small enough to obtain quadruple-ζ level CCSD(T) results, and they cover the whole range of elements included in the set.

All coupled cluster calculations were conducted using ORCA version 4.2.1 [59, 60].

2.2 Construction of new model systems for 4-OT

Our previous work on the 4-OT reaction [26] used an active-site model, originally created by Sevastik and Himo [51], which we reoptimised at the PBEh-3c [61] level of theory. The model contained the substrate, Pro1 residue and some surrounding environment involved in stabilising the charge. Herein, we have used this model, which we refer to as the original model, to create two further models: a minimal model, which contains only the substrate and Pro1 residue, and a separated model, with individual substrate and Pro1 structures. Geometries involved in these two new models were also optimised at the PBEh-3c level of theory using ORCA version 4.2.1, with the "tight" setting for SCF convergence and the default setting for geometry convergence. The "grid3" and "finalgrid5" settings were used for ORCA's multigrid option.

2.3 Tests of density functional approximations

All methods tested in our exemplary benchmark studies are listed in table 1. We have chosen 12 DFAs from the best performers in the categories of Generalised Gradient Approximation (GGA), meta-GGA, hybrid and double-hybrid density functionals from our previous study on enzymatically catalysed reactions as our main set of DFAs, and for the enzyme RE test set we also take the results of the benchmark study conducted by Sirirak et al. and calculate the deviations of each

Table 1 QM methods tested in the example benchmark studies, along with their dispersion corrections.

Density functional approximations tested in this work		
Type	Method	Dispersion correction
GGA	OLYP[a] [62–64]	D3(BJ)[f] [65]
	PBE[a] [66]	D3(BJ) [53]
	revPBE-NL[a] [67, 68]	VV10[g] [68]
meta-GGA	B97M-V[a] [69]	VV10 [69]
	B97M-D3(BJ)[a] [22, 69]	D3(BJ) [22]
	SCAN[a] [70]	D3(BJ) [71]
hybrid	M062X[a] [72]	D3(0)[h] [65]
	ωB97M-V[a] [73]	VV10 [73]
	ωB97M-D3(BJ)[a] [22, 73]	D3(BJ) [22]
double-hybrid	SOS0-PBE0-2[a] [74]	D3(BJ) [21]
	DOD-SCAN[a] [75]	D3(BJ) [75]
	revDOD-PBE[a] [75]	D3(BJ) [75]
Additional methods and results taken from ref [34]		
Type	Method	Dispersion correction
semi-empirical	AM1 [76]	
	PM3 [77]	
	SCC-DFTB [78]	
GGA	BLYP[b] [63, 64, 79]	D3(0) [52]
	BP86[b] [79–81]	D3(0) [52]
meta-GGA	TPSS[b] [82]	D3(0) [52]
hybrid	B3LYP[b,c] [83, 84]	D3(0) [52]
	BHLYP[b] [85]	
ab-initio	MP2[b,d,e] [86]	
	SCS-MP2[d,e] [43]	
	CCSD(T)[d] [31]	

Basis sets used: [a]def2-QZVP [87], [b]6-311+G(d) [88], [c]6-31+G(d) [89], [d]aug-cc-pVDZ [42], [e]aug-cc-pVTZ [42].
[f]D3(BJ): DFT-D3 with Becke-Johnson damping. [52, 53]
[g]VV10: nonlocal van der Waals kernel.[90]
[h]D3(0): DFT-D3 with zero-damping. [52]

method from our new reference values. We additionally test BLYP and B3LYP when looking at DFT-D3-type London-dispersion corrections. All calculations, except those using the SCAN functional, were run with ORCA version 4.2.1 [59, 60] with the default settings for SCF and geometry convergence. TURBOMOLE version 7.4.1 [91–94] was used for the SCAN functional, with the recommended grid options "gridsize 4" and "radsize 40" [20, 70, 71] and a convergence criterion of $1 \times 10^{-7} E_h$. The RIJCOSX approximation was used with most functionals, and the frozen-core approximation was used with the double-hybrid functionals. Van-der-Waals DFAs used the nonlocal VV10-kernel in its post-SCF implementation; it was shown in ref [22] that this does not impact the results but halves the computational effort compared to the originally developed, full-SCF version. All functionals were evaluated with the def2-QZVP basis set [87]. All deviations in our following discussions were calculated as the difference between an assessed method and the reference value.

3 Results and discussion

3.1 On the quality of reference values

To provide a comparison between various CCSD(T) based approaches, table 2 presents REs for reactions **3**, **5**, **7**, **9** and **13** from the RE test set introduced in figure 1. The approaches tested are estimated (est.) CCSD(T)/aug-cc-pVTZ, conventional CCSD(T)/aug-cc-pVTZ, CCSD(T)/CBS, and DLPNO-CCSD(T)/CBS, where the CBS results are extrapolated from the aug-cc-pVTZ and aug-cc-pVQZ AO basis sets. The estimated numbers are taken from Sirirak et al.'s study in ref [34]—as already mentioned in Section 1, a given CCSD(T)/aug-cc-pVDZ total energy is corrected by adding the difference between SCS-MP2/aug-cc-pVTZ and SCS-MP2/ aug-cc-pVDZ. The est. CCSD(T)/aug-cc-pVTZ REs used herein have been calculated from the molecular energies provided in Sirirak et al.'s supporting information; the REs for the additional methods in the following benchmark study have also been calculated in the same way. Although these may differ from their listed REs, the published statistics for the study are consistent with the recalculated values.

Comparison of the estimated and actual CCSD(T)/aug-cc-pVTZ REs shows that estimation does not consistently replicate the results of the latter. The estimated method gives the exact same value for reaction **13**, but results in differences of almost 0.4 kcal/mol in reactions **3** and **7**. The differences between triple-ζ and extrapolated CCSD(T) results are larger, with differences of at least 0.6 kcal/mol seen for reactions **3**, **5** and **13**. As CCSD(T)/CBS results are our ideal benchmarks, the accuracy of the other approaches can be gauged by how closely they replicate the CCSD(T)/CBS values. The est. CCSD(T)/aug-cc-pVTZ approach shows the most noticeable absolute deviations, in the range of 0.53-0.97 kcal/mol. The difference between DLPNO-CCSD(T)/ CBS and CCSD(T)/CBS for these reactions is almost negligible, with all absolute differences being

Table 2 Reaction energies (kcal/mol) for reactions **3**, **5**, **7**, **9** and **13** of the RE test set (figure 1) calculated with different coupled-cluster-based levels of theory, and their deviations from CCSD(T)/CBS. TZ refers to the aug-cc-pVTZ basis set.

	RE3	RE5	RE7	RE9	RE13
est. CCSD(T)/TZ[a]	−28.68	−17.79	13.08	23.97	−111.81
deviation	*0.97*	*0.84*	*0.66*	*0.53*	*−0.60*
CCSD(T)/TZ	−29.05	−17.95	12.73	23.83	−111.81
deviation	*0.60*	*0.68*	*0.31*	*0.39*	*−0.60*
DLPNO-CCSD(T)/CBS	−29.62	−18.62	12.48	23.48	−111.23
deviation	*0.03*	*0.01*	*0.06*	*0.04*	*−0.02*
CCSD(T)/CBS	−29.65	−18.63	12.42	23.44	−111.21

[a]Values taken from ref [34]; see table S3 for details.

in the range of 0.01-0.06 kcal/mol. As predicted, this range is significantly lower than the 0.4 kcal/mol error of DLPNO-CCSD(T)/CBS (DZ/TZ) reported by Paiva et al. [35], and also lower than the error of the est. CCSD(T)/CBS references used in that study.

While none of the approaches tested here deviate by more than 1 kcal/mol—the generally accepted chemical accuracy limit for REs—for these five tested reactions, it is clear that the estimation strategy is not appropriate for calculating reference data for a benchmark study and that one should not rely on TZ data as a benchmark. We also see that the DLPNO-CCSD(T)/CBS (TZ/QZ) level of theory is the best alternative when CCSD(T)/CBS is computationally unfeasible, which is often the case when benchmarking models of enzymatically catalysed reactions and indeed is the case here with the saccharides in reaction **20** (figure 1), which contain 25 and 22 atoms for the reactant and product, respectively. We, thus, choose to use DLPNO-CCSD(T)/CBS(TZ/QZ) to update the benchmarks for this set, and all REs at this level of theory are listed alongside the est. CCSD(T)/aug-cc-pVTZ REs in table S3 in the electronic supplementary material for comparison. Although the results in table 2 show minimal differences between these two methods, absolute differences of >1 kcal/mol are seen for six of the 20 reactions in the set, and four of those, namely reactions **14** - **17**, have absolute differences of >2 kcal/mol. These large differences occur only in the non-proton transfer reactions, suggesting that the estimation strategy is even less appropriate when considering processes other than proton transfers.

When conducting a benchmark study, a DFA is judged by its ability to replicate a reference value, so the quality of the chosen reference value will directly affect the DFA's perceived accuracy. In figure 3 we present the mean absolute deviations (MADs) of each method in table 1 from two sets of references—est. CCSD(T)/aug-cc-pVTZ and DLPNO-CCSD(T)/CBS—to explore how the observed performance of the methods changes with the benchmark against which they are tested. We also show the est. CCSD(T)/aug-cc-pVTZ results in this plot and mention that the overall MAD of this method against our updated reference values is 1.1 kcal/mol. Again, we discourage its use as a reference in further studies.

Most of the MADs are lower against the DLPNO-CCSD(T)/CBS references, with the majority of the exceptions being MP2, SCS-MP2 and CCSD(T) results from the original study. Considering that the SCS-MP2 and CCSD(T)/aug-cc-pVDZ results were used in the calculation of the est. CCSD(T)/aug-cc-pVTZ values, it is understandable that their deviations from those older references are lower than against our updated ones. Interestingly, we also see that BP86-D3(0) and TPSS-D3(0) perform worse against the DLPNO-CCSD(T)/CBS references than the est. CCSD(T)/aug-cc-pVTZ ones; in the case of BP86, the DFT-D3(0) variant even ends up with a higher MAD than the uncorrected DFA. In passing, we note that the semi-empirical MO methods SCC-DFTB, AM1 and PM3 have some of the largest MAD decreases when using the new references, but proportional to the actual values these are not significant improvements and they are by far outperformed by all assessed DFAs. The largest difference between the two sets of references is seen for ωB97M-D3(BJ), which has an MAD of 3.3 kcal/mol against the est. CCSD(T)/aug-cc-pVTZ references and 2.1 kcal/mol against DLPNO-CCSD(T)/CBS—a decrease of 1.2 kcal/mol. We also see significant reductions in the other statistics for ωB97M-D3(BJ) when going to higher quality references, with its root mean square deviation (RMSD) dropping from 3.9 to 2.8 kcal/mol and the error range from 11.4 to 8.0 kcal/mol.

In general, we see that good methods—ones which have performed well in previous studies [20–22, 24]—perform even better against better references, and therefore have larger changes in the MADs when going from est. CCSD(T)/aug-cc-pVTZ to DLPNO-CCSD(T)/CBS references. B97M-V and B97M-D3(BJ), two meta-GGAs which have been shown to outperform many hybrid functionals [22, 26, 69], show reductions in the MADs of 0.5 and 0.7 kcal/mol respectively. For the hybrids, the decreases in the MADs range from 0.8 [M062X-D3(0)] to 1.2 [ωB97M-D3(BJ)] kcal/mol, and for the double hybrids the decreases range from 0.5 [SOS0-PBE0-2-D3(BJ)] to 0.9 kcal/mol [revDOD-PBE-D3(BJ)]. The expected trend of improved accuracy as one climbs the rungs of Jacob's Ladder is therefore more visible with the newly generated, more accurate benchmarks.

There are also differences in the ordering of the DFAs, particularly in the hybrids and double hybrids newly tested in this work. With the old references, ωB97M-V and SOS0-PBE0-2-D3(BJ) are the best of their respective categories, while ωB97M-D3(BJ) and revDOD-PBE-D3(BJ) outperform these with respect to the new references. For the GGA and meta-GGA DFAs, we see the same ranking with both sets of references, although the MADs of the meta-GGAs are slightly more spread out with the new ones. Overall, we conclude that est. CCSD(T)/aug-cc-pVTZ REs are not good enough references for benchmarking, and we stress the importance of using high-level reference values.

The ordering of the functionals also allows us to qualitatively compare this set to our previous bench-

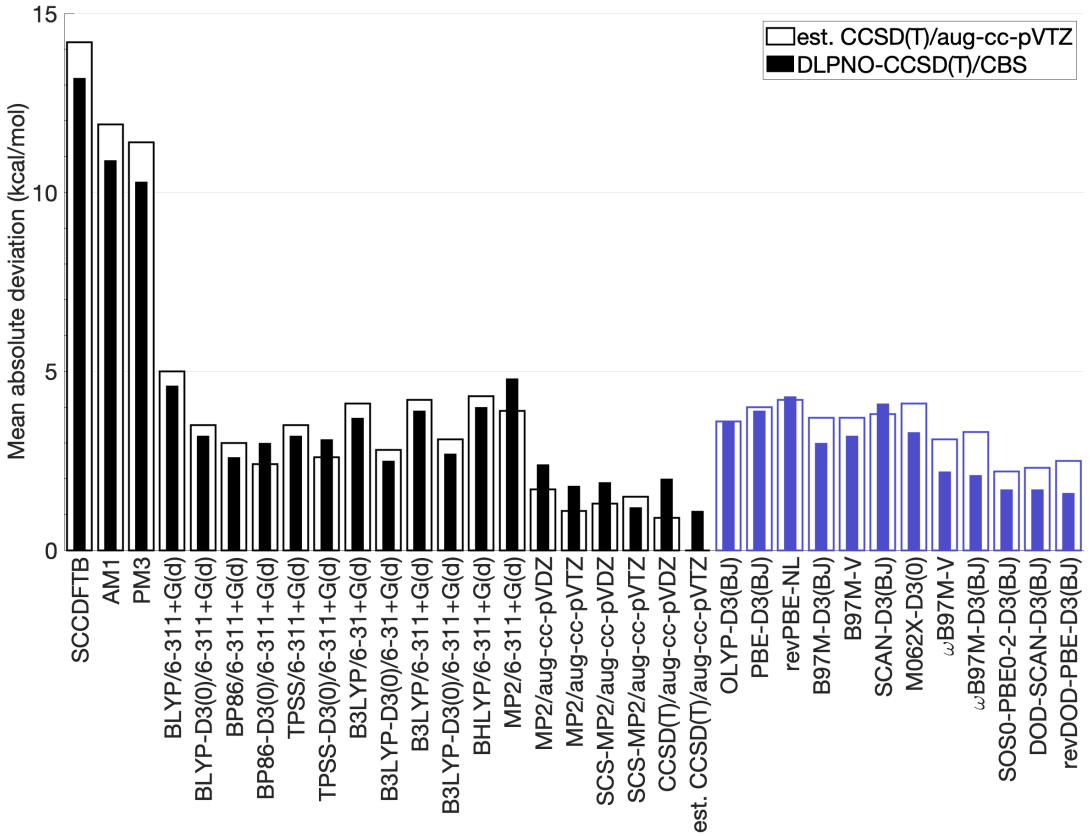

Fig. 3 Mean absolute deviations (kcal/mol) relative to two sets of references: est. CCSD(T)/aug-cc-pVTZ (outlines) and DLPNO-CCSD(T)/CBS (solid bars). Data for the methods shown in black has been taken from Sirirak et al. [34], methods shown in blue are new in this work. Unless otherwise noted, DFT methods were evaluated with the def2-QZVP basis set.

mark set of active-site models in ref [26]. In figure 3, the 12 DFAs newly chosen for this work are presented in order of their accuracy for our previous test set within their respective rungs of Jacob's ladder—for example, OLYP-D3(BJ) was the best performing GGA, followed by PBE-D3(BJ) and revPBE-NL; consequently, a small upwards trend within each rung is also expected for the present set. This behaviour is indeed seen for the GGAs and meta-GGAs, but the rankings of both the hybrids and double hybrids are reversed. While slight differences between test sets are to be expected, one would expect similar trends from two sets that are both designed to represent similar types of reactions. Therefore, this is an indication that the approach of modelling enzymatically catalysed reactions without the surrounding chemical environment may not be appropriate when trying to find low-level methods for subsequent QM/MM studies.

3.2 The effect of London-dispersion corrections

Having updated the references for the reactions shown in figure 1, we briefly detour from our main aims and update the test of dispersion corrections conducted by Sirirak et al. with our new reference data, additional DFAs, and the DFT-D3(BJ) correction, which was recommended as the default and is to be preferred over the original DFT-D3(0) correction by the developers [53]. In figure 4 we show the MADs and mean deviations (MDs) from the DLPNO-CCSD(T)/CBS benchmarks for each method in its uncorrected form and with the DFT-D3(0) and DFT-D3(BJ) dispersion corrections. We note that M062X-D3(0) is the only dispersion corrected form of M062X as the D3(BJ) correction has been shown to overcorrect for the Minnesota functionals [65, 95], and we do not use DFT-D3(0) with DOD-SCAN or revDOD-PBE because these functionals have been specifically parameterised for use with DFT-D3(BJ) [75].

The DFAs used with the Pople basis sets—as done by Sirirak et al.—overestimate the REs, resulting in

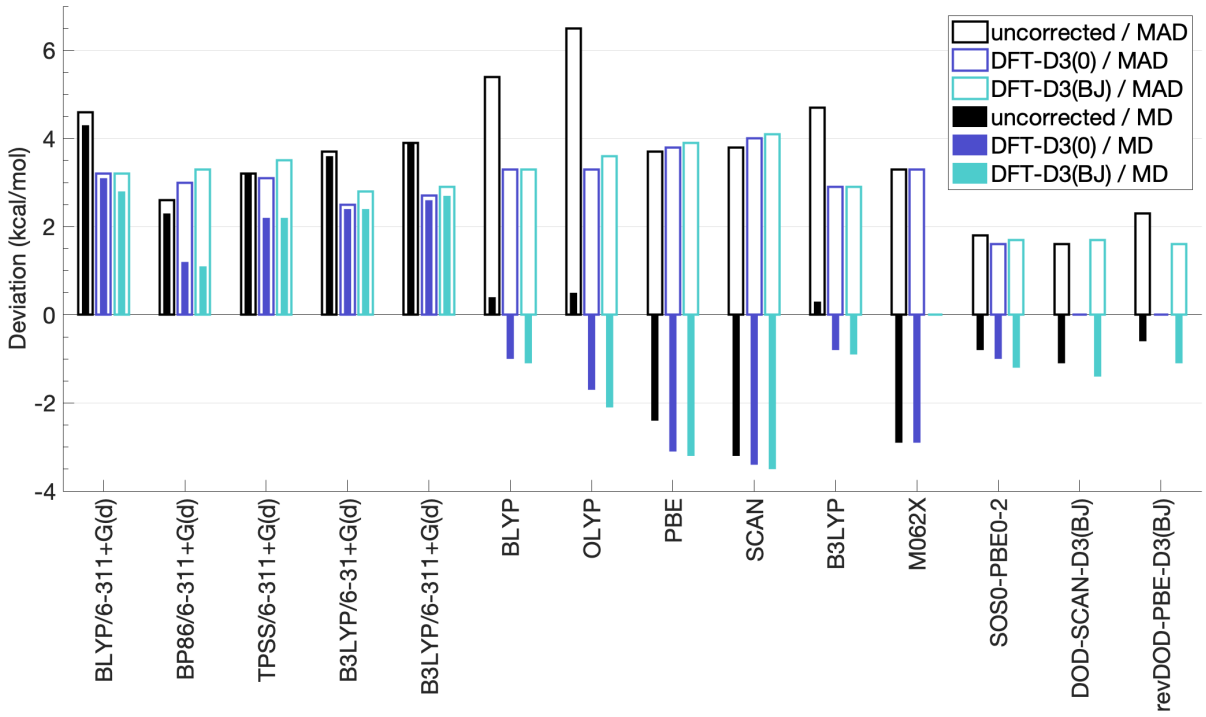

Fig. 4 Mean absolute deviations (outlines) and mean deviations (solid bars) of selected density functionals with the DFT-D3(0) and DFT-D3(BJ) dispersion corrections where applicable. Deviations are calculated from DLPNO-CCSD(T)/CBS results. All deviations are given in kcal/mol. Unless otherwise noted, the def2-QZVP basis set was used.

positive MDs, whereas our choice of the def2-QZVP basis set results in consistent underestimation of the REs. We additionally provide BLYP/def2-QZVP and B3LYP/def2-QZVP results here to confirm that this is caused by the basis set, not the specific functionals. Methods that consistently underestimate REs can result in the perception that dispersion corrections make the results worse, as the energies are further lowered and thus deviate more from the values one wishes to replicate. This perception is incorrect, however, and should not be considered as a reason not to use the corrections as they are merely revealing weaknesses in the underlying DFA that are otherwise cancelled out by the incorrect long-range behaviour [20]. This also applies to the higher MADs seen when DFT-D3(BJ) is applied compared to DFT-D3(0), as the Becke-Johnson damping function results recovers short-range dispersion effects and as such provides larger absolute dispersion energies than its DFT-D3(0) predecessor [53].

3.3 The importance of the enzyme environment

While the RE test set we have just explored is adequate for testing different benchmark levels of theory, the approach of using small, separate molecules cannot realistically represent a particular active site due to the missing surrounding enzyme environment. This is one of the potential reasons why we saw see trends in DFA performance in Section 3.1 that were inconsistent with our previous work on larger active-site models. In this section, we explore this aspect further by testing the inadequacies of smaller models using three models of the 4-OT reaction (see figure 2): a large active-site model—the original model from our previous benchmark study that includes neighbouring residues that stabilise the reaction, updated from a model created by Sevastik and Himo [51]—a minimal model that contains only the substrate and catalytic Pro1 residue, and a model that involves separate structures for the substrate and Pro1 to mimic the approach used for the RE test set discussed earlier. Geometries of the original and minimal models are shown in figure 5, and DLPNO-CCSD(T)/CBS REs for all three models are given in table 3. In Sevastik and Himo's study of the mechanism of 4-OT, it was stated that small models that did not account for the surrounding environment would give unrealistic energies. Indeed, the REs obtained by us show that the minimal and separated models give significantly different results to the original one.

The main difficulty in modelling the 4-OT reaction is the charge separation in the intermediate, and

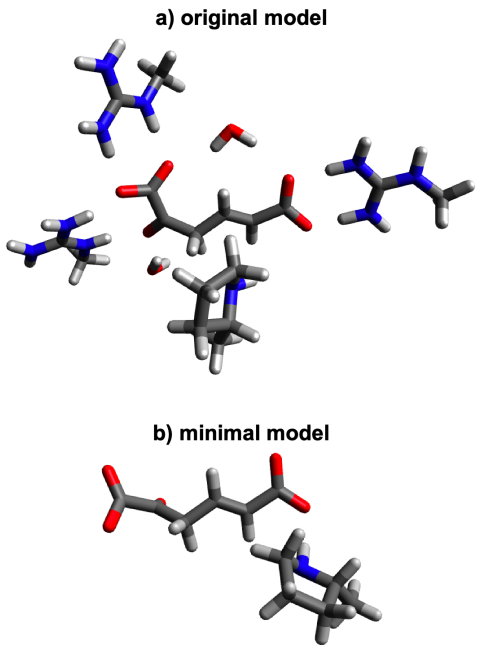

Fig. 5 Reactant structures of the 4-OT original and minimal models. C atoms are shown in grey, H in white, N in blue and O in red.

Table 3 Reaction energies (REs) (kcal/mol) at the DLPNO-CCSD(T)/CBS level of theory for the three models of 4-OT.

Model	RE-A[a]	RE-B[b]	Overall RE
Original	−4.90	1.25	−3.65
Minimal	5.45	−6.57	−1.12
Separated	276.40	−275.12	1.28

[a] reaction energy for the first step (see figure 2)
[a] reaction energy for the second step (see figure 2)

this is the main cause of the differences between the two smaller and the original models. The first problem caused by the charge separation is the extremely large REs for the separated model, where the newly formed charges are not stabilised at all. We note that although anionic species can potentially be problematic for the ma-def2-nZVPP basis sets as they have only been minimally augmented with diffuse functions, this is unlikely to be a factor in this case because CBS results extrapolated from the aug-cc-pVTZ/QZ basis sets are very similar. The original model accounts for the charge-separated intermediate best by including three cationic arginine residues to stabilise all three negative charges, but even without these the minimal model provides some stabilisation through electrostatic interactions between the substrate and protonated Pro1 residue. The strength of the stabilisation can be quantified by taking the difference between the separated and minimal models for each structure, and we find that combining the structures results in a lowering of the total energy of 282.39 kcal/mol for the intermediate when these electrostatic effects are involved, while the reactant and product energies only decrease by 11.44 and 13.84 kcal/mol, respectively (see table S14).

The second problem caused by the charge separation is represented by the signs of the REs and how they reverse between the original and minimal models. The intermediate form of the substrate is conjugated all the way along the molecule, and therefore is thermodynamically favoured over the reactant and product as long as the charges are appropriately balanced. This results in a negative first RE (RE-A) and positive second RE (RE-B) for the original model. When the model contains none of the surrounding environment, however, the instability of the intermediate due to the charge separation dominates, and so RE-A is positive and RE-B is negative for both the minimal and separated models. We note that the use of an implicit solvent model to help stabilise the charge in the minimal model does not change this trend; in fact, the results for the solvent-corrected minimal model are closer to that of the separated model than the original model. Further details of the solvent tests can be found in the electronic supplementary material (table S15).

Before continuing the discussion of the 4-OT models, a comparison with the enzyme RE test set discussed in the previous subsections is appropriate, as the issue of charge separation is also seen for reaction **13**, which is the only reaction in the set that involves a cationic species reacting with an anionic one. The RE for this reaction (−111.23 kcal/mol) is the highest-magnitude RE in the set, and it is almost four times that of the second highest-magnitude RE (reaction **3** with RE=−29.62 kcal/mol). The strongly negative RE in reaction **13** is due to the neutral products being significantly more stable than the ionic reactants, especially when they cannot interact with each other through electrostatic effects; this is similar to RE-B of the separated 4-OT model, which is also large and negative as the system leaves the charge separated intermediate state. This allows us to question the suitability of the current form of reaction **13** for the assessment of methods for subsequent use in enzymatically catalysed reactions.

Aside from the charge-separation problem involving the intermediate, the overall REs in the 4-OT models are also influenced by the lack of charge stabilisation provided by the chemical environment, showing a trend of increasing thermodynamic stability of the product relative to the reactant with larger model size. The overall RE of the separated model is purely the difference between the two tautomers of the substrate as the proline energies are identical and cancel each other. Conse-

quently, the product tautomer is seen to be slightly less stable than the reactant one. While the product is the tautomer with the larger π-system due to conjugation between the double bond and the pyruvate group, the position of the double bond in the reactant allows for conjugation with the carboxylate group, and the additional minor resonance form provided may stabilise that negative charge in the absence of any external stabilisation. In the minimal model, Pro1 provides some stabilisation through ion-dipole interactions, and the arginine residues in the original model provide even more through electrostatic effects. When the charge is even somewhat stabilised, the conjugated tautomer then becomes the thermodynamically favoured product as expected.

Overall, one can see that there are significant differences between the models, and that the smaller ones have an incomplete representation of the chemistry within the active site, particularly with respect to charged species. The approach of treating each component separately is especially problematic; while it is unlikely to be considered by those conducting mechanistic studies—since it is not suited to the calculation of transition states and barrier heights—we nonetheless strongly recommend that it be avoided even in simplified, preliminary studies that are designed to guide further computational enzyme studies.

Having identified the main differences between the three 4-OT models, we continue by testing our set of DFAs on them to determine if and how strongly the deficiencies of the smaller models would impact the conclusions of a QM benchmark study. In figure 6 we present MADs and MDs for each model, calculated with the 12 DFAs picked from our previous work [26]. Since each model only has two associated REs, these statistics are not a reliable indicator of the performance of each DFA for subsequent applications, however, the numbers exemplify the impact of the different models on the DFA rankings.

While almost all MDs are negative, not all REs are underestimated. The general trend is that RE-A is underestimated while RE-B is overestimated, but this is not consistent between models nor methods. For example, PBE-D3(BJ) exhibits this behaviour for the separated (deviations for RE-A and RE-B are −5.56 and 2.70 kcal/mol, respectively) and original (deviations of −9.85 and 7.45 kcal/mol) models, but both REs are underestimated with the minimal model (deviations of −0.68 and −0.38 kcal/mol). RE-B of the minimal model is also underestimated by revPBE-NL, SCAN-D3(BJ) and DOD-SCAN-D3(BJ) (deviations of −0.21, −0.09 and −0.06 kcal/mol, respectively), but all other assessed DFAs overestimate it.

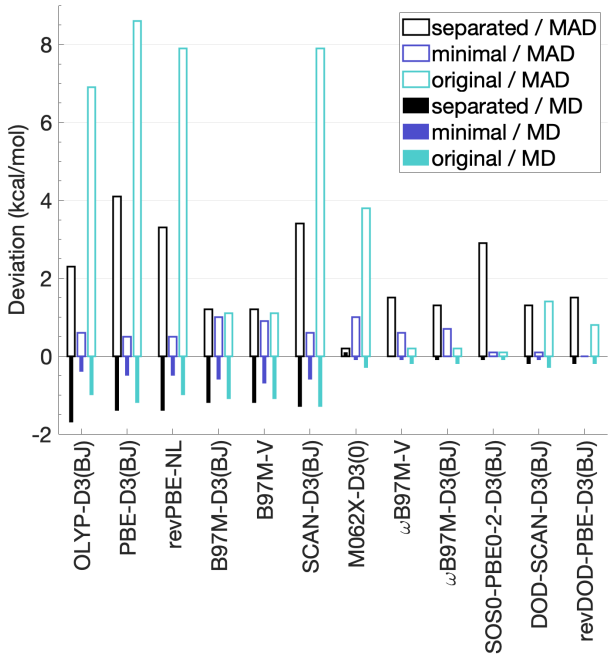

Fig. 6 Mean absolute deviations (outlines) and mean deviations (solid bars) of selected DFAs for the three models of the 4-OT reaction. The def2-QZVP basis set was used in all cases.

The minimal model gives the lowest MADs of the three 4-OT models for all but three DFAs; interestingly, these are all hybrid functionals. All MADs for this model are less than or equal to 1 kcal/mol, which would imply that any of these DFAs, including the GGAs, could be an appropriate choice for applications when the systems are similar to this model. The same cannot be said when considering the other models, however, as the MADs for the other two models are generally larger, with the worst combination being GGAs applied to the original model (for example, PBE-D3(BJ) gives the largest MAD at 8.6 kcal/mol).

Based on the MADs shown here, one can say that the best DFAs for the original model are SOS0-PBE0-2-D3(BJ), ωB97M-V, and ωB97M-D3(BJ); for the minimal model, the three double-hybrid DFAs perform best; and for the separated model, M062X-D3(0), B97M-V, and B97M-D3(BJ). The only DFA that appears in the top three for more than one model is SOS0-PBE0-2-D3(BJ), but it has the 4th highest MAD for the separated model. Significant differences can also be seen in the ordering of the DFAs within their rungs of Jacob's Ladder—based on the original model, OLYP-D3(BJ) is the best GGA, however it has the highest MAD of the three GGAs for the minimal model. Similarly, SCAN-D3(BJ) is the best meta-GGA when using the minimal

model, but the worst for the original model. The overall impression from comparing these three models in this way is that the conclusions drawn for one model do not necessarily hold for the others.

While benchmark studies do not necessarily require perfectly realistic models, especially when the references are independent of any experimental data, the models must reflect the results one aims to achieve. It is understandable that many studies involve relatively small models because the computational demands of treating larger ones can be high, but the comparison between our three example models shows that the results of the smaller systems are not consistent with those of the largest model. When the goal is to choose a DFA for QM/MM analysis of an enzyme, the QM region will necessarily include the most important parts of the enzyme's active site, and therefore a simplified study on a smaller model may wrongly identify and recommend DFAs that are not the most appropriate for the actual application. Therefore, we strongly recommend to anyone seeking to benchmark for their own computational treatments of enzyme-related problems that a study should use the whole QM region, or at least the largest model for which one can obtain high level reference values while retaining the necessary chemical environment.

4 Summary and conclusions

Herein, we explored two important factors that influence the outcomes of benchmark studies—namely the quality of the benchmarks themselves as well as the choice of underlying structures—in the context of studying enzymatically-catalysed reactions, and choosing QM methods for QM/MM analysis thereof. Through the use of a set of 20 reactions of small molecules designed to represent common steps in enzyme mechanisms, we showed that DLPNO-CCSD(T)/CBS with the TightPNO thresholds is a good alternative to conventional CCSD(T)/CBS when computational resources are limited, while estimation of high-level data using corrected low-level numbers and single basis set results such as CCSD(T)/aug-cc-pVTZ should be avoided. A selection of QM methods, mainly density functional approximations, was then applied to the enzyme RE test set, and the deviations from estimated CCSD(T)/ aug-cc-pVTZ and DLPNO-CCSD(T)/CBS benchmarks were assessed. We saw that most methods had lower mean absolute deviations when compared to DLPNO-CCSD(T)/CBS, and the biggest decreases occurred for DFT methods that had performed well in other studies. The fact that good methods perform better with improved reference data is further proof of their general reliability and robustness as also pointed out in refs [19] and [20]. The ordering of the tested methods is also dependent on the choice of references, particularly for hybrid and double-hybrid DFAs. Using our new DLPNO-CCSD(T)/CBS reference values, we also briefly compared some of the pure DFAs with their DFT-D3(0) and DFT-D3(BJ) corrected variants, and confirmed the importance of using a dispersion correction in DFT studies.

We also used a two-step tautomerisation reaction, catalysed by the enzyme 4-oxalocrotonate tautomerase, to explore if smaller active-site models could give similar benchmarking results to larger ones despite their chemical deficiencies. The three tested models of this enzyme includes an active-site model that contained some neighbouring amino acid residues, a simplified model that contained only the substrate and the catalytic proline residue which was directly involved in the proton transfer steps, and a separated model where the substrate and proline residue were treated separately. Considering the high-level REs calculated to be used as the benchmarks, it is clear that ignoring the chemical environment neglects significant intermolecular interactions, particularly when the important components are also separated from each other. This is especially problematic in charge separated systems that are strongly stabilised by electrostatic effects. Benchmarking twelve DFAs on all three models showed very little consistency between the results for each model, both in the ordering of functionals and actual magnitudes of the deviations. As such, one must ensure that the model used is large enough to adequately represent the chemistry that will be studied, in order to get the most relevant recommendations.

Overall, the main point we wish to stress to those seeking to conduct their own benchmark studies on enzymatically-catalysed reactions is to not be unnecessarily "cheap". Lower quality references and smaller model systems will undoubtedly make the benchmarking process faster, but they can easily make poor-performing methods look just as appealing as good ones. Instead, one should always aim to use the highest quality references and largest model systems that are computationally feasible, in order to get the most accurate results not only in their benchmark studies, but also from their applications by using an appropriate method that has been chosen for the right reasons. While our study focussed on enzymatically catalysed reactions, we hope that our insights can also serve as a guideline for other future chemical benchmark studies.

Acknowledgements D.A.W. acknowledges an Australian Government Research Training Program Scholarship. We are thankful for the allocation of computing resources by the Na-

tional Computational Infrastructure (NCI) National Facility within the National Computational Merit Allocation Scheme (Project No. fk5).

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. Welborn VV, Head-Gordon T (2018) Computational design of synthetic enzymes. Chem Rev 119(11):6613–6630, DOI 10.1021/acs.chemrev.8b00399
2. Piazzetta P, Marino T, Russo N, Salahub DR (2015) Direct hydrogenation of carbon dioxide by an artificial reductase obtained by substituting rhodium for zinc in the carbonic anhydrase catalytic center. A mechanistic study. ACS Catal 5(9):5397–5409, DOI 10.1021/acscatal.5b00185
3. Sousa SF, Fernandes PA, Ramos MJ (2012) Computational enzymatic catalysis – clarifying enzymatic mechanisms with the help of computers. Phys Chem Chem Phys 14(36):12431–12441, DOI 10.1039/c2cp41180f
4. Sousa JPM, Neves RPP, Sousa SF, Ramos MJ, Fernandes PA (2020) Reaction mechanism and determinants for efficient catalysis by DszB, a key enzyme for crude oil bio-desulfurization. ACS Catal 10(16):9545–9554, DOI 10.1021/acscatal.0c03122
5. Mulholland AJ (2005) Modelling enzyme reaction mechanisms, specificity and catalysis. Drug Discovery Today 10(20):1393–1402, DOI 10.1016/s1359-6446(05)03611-1
6. Świderek K, Tuñón I, Moliner V (2013) Predicting enzymatic reactivity: from theory to design. Wiley Interdiscip Rev: Comput Mol Sci 4(5):407–421, DOI 10.1002/wcms.1173
7. Paiva P, Sousa SF, Ramos MJ, Fernandes PA (2018) Understanding the catalytic machinery and the reaction pathway of the malonyl-acetyl transferase domain of human fatty acid synthase. ACS Catal 8(6):4860–4872, DOI 10.1021/acscatal.8b00577
8. Korendovych IV, DeGrado WF (2014) Catalytic efficiency of designed catalytic proteins. Curr Opin Struct Biol 27:113–121, DOI 10.1016/j.sbi.2014.06.006
9. Bloom J, Meyer M, Meinhold P, Otey C, MacMillan D, Arnold F (2005) Evolving strategies for enzyme engineering. Curr Opin Struct Biol 15(4):447–452, DOI 10.1016/j.sbi.2005.06.004
10. Brustad EM, Arnold FH (2011) Optimizing non-natural protein function with directed evolution. Curr Opin Chem Biol 15(2):201–210, DOI 10.1016/j.cbpa.2010.11.020
11. Renata H, Wang ZJ, Arnold FH (2015) Expanding the enzyme universe: accessing non-natural reactions by mechanism-guided directed evolution. Angew Chem, Int Ed 54(11):3351–3367, DOI 10.1002/anie.201409470
12. Arnold FH (2017) Directed evolution: bringing new chemistry to life. Angew Chem, Int Ed 57(16):4143–4148, DOI 10.1002/anie.201708408
13. Siegbahn PEM, Himo F (2009) Recent developments of the quantum chemical cluster approach for modeling enzyme reactions. J Biol Inorg Chem 14(5):643–651, DOI 10.1007/s00775-009-0511-y
14. Senn HM, Thiel W (2009) QM/MM methods for biomolecular systems. Angew Chem, Int Ed 48(7):1198–1229, DOI 10.1002/anie.200802019
15. van der Kamp MW, Mulholland AJ (2013) Combined quantum mechanics/molecular mechanics (QM/MM) methods in computational enzymology. Biochemistry 52(16):2708–2728, DOI 10.1021/bi400215w
16. Chung LW, Sameera WMC, Ramozzi R, Page AJ, Hatanaka M, Petrova GP, Harris TV, Li X, Ke Z, Liu F, Li HB, Ding L, Morokuma K (2015) The ONIOM method and its applications. Chem Rev 115(12):5678–5796, DOI 10.1021/cr5004419
17. Zheng M, Waller MP (2016) Adaptive quantum mechanics/molecular mechanics methods. Wiley Interdiscip Rev: Comput Mol Sci 6(4):369–385, DOI 10.1002/wcms.1255
18. Dziedzic J, Mao Y, Shao Y, Ponder J, Head-Gordon T, Head-Gordon M, Skylaris CK (2016) TINKTEP: A fully self-consistent, mutually polarizable QM/MM approach based on the AMOEBA force field. J Chem Phys 145(12):124106, DOI 10.1063/1.4962909
19. Karton A, Goerigk L (2015) Accurate reaction barrier heights of pericyclic reactions: Surprisingly large deviations for the CBS-QB3 composite method and their consequences in DFT benchmark studies. J Comput Chem 36(9):622–632, DOI 10.1002/jcc.23837
20. Goerigk L, Hansen A, Bauer C, Ehrlich S, Najibi A, Grimme S (2017) 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 19(48):32184–32215, DOI 10.1039/C7CP04913G

21. Mehta N, Casanova-Páez M, Goerigk L (2018) Semi-empirical or non-empirical double-hybrid density functionals: which are more robust? Phys Chem Chem Phys 20(36):23175–23194, DOI 10.1039/c8cp03852j
22. Najibi A, Goerigk L (2018) 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 14(11):5725–5738, DOI 10.1021/acs.jctc.8b00842
23. Najibi A, Goerigk L (2020) DFT-D4 counterparts of leading meta-generalized-gradient approximation and hybrid density functionals for energetics and geometries. J Comput Chem 41(30):2562–2572, DOI 10.1002/jcc.26411
24. Mardirossian N, Head-Gordon M (2017) Thirty years of density functional theory in computational chemistry: an overview and extensive assessment of 200 density functionals. Mol Phys 115(19):2315–2372, DOI 10.1080/00268976.2017.1333644
25. Yu HS, He X, Li SL, Truhlar DG (2016) MN15: A Kohn–Sham global-hybrid exchange–correlation density functional with broad accuracy for multi-reference and single-reference systems and noncovalent interactions. Chem Sci 7(8):5032–5051, DOI 10.1039/c6sc00705h
26. Wappett DA, Goerigk L (2019) Toward a quantum-chemical benchmark set for enzymatically catalyzed reactions: important steps and insights. J Phys Chem A 123(32):7057–7074, DOI 10.1021/acs.jpca.9b05088
27. Wappett DA, Goerigk L (2020) Erratum to "toward a quantum-chemical benchmark set for enzymatically catalyzed reactions: Important steps and insights". J Phys Chem A 124(5):1062–1062, DOI 10.1021/acs.jpca.0c00425
28. Kromann JC, Christensen AS, Cui Q, Jensen JH (2016) Towards a barrier height benchmark set for biologically relevant systems. PeerJ 4:e1994, DOI 10.7717/peerj.1994
29. Riplinger C, Sandhoefer B, Hansen A, Neese F (2013) Natural triple excitations in local coupled cluster calculations with pair natural orbitals. J Chem Phys 139(13):134101, DOI 10.1063/1.4821834
30. Riplinger C, Pinski P, Becker U, Valeev EF, Neese F (2016) Sparse maps?a systematic infrastructure for reduced-scaling electronic structure methods. ii. linear scaling domain based pair natural orbital coupled cluster theory. J Chem Phys 144(2):024109
31. Raghavachari K, Trucks GW, Pople JA, Head-Gordon M (1989) A fifth-order perturbation comparison of electron correlation theories. Chem Phys Lett 157(6):479–483
32. Liakos DG, Sparta M, Kesharwani MK, Martin JML, Neese F (2015) Exploring the accuracy limits of local pair natural orbital coupled-cluster theory. J Chem Theory Comput 11(4):1525–1539, DOI 10.1021/ct501129s
33. Liakos DG, Guo Y, Neese F (2019) Comprehensive benchmark results for the domain based local pair natural orbital coupled cluster method (DLPNO-CCSD(t)) for closed- and open-shell systems. J Phys Chem A 124(1):90–100, DOI 10.1021/acs.jpca.9b05734
34. Sirirak J, Lawan N, der Kamp MWV, Harvey JN, Mulholland AJ (2020) Benchmarking quantum mechanical methods for calculating reaction energies of reactions catalyzed by enzymes. PeerJ Phys Chem 2:e8, DOI 10.7717/peerj-pchem.8
35. Paiva P, Ramos MJ, Fernandes PA (2020) Assessing the validity of DLPNO-CCSD(t) in the calculation of activation and reaction energies of ubiquitous enzymatic reactions. J Comput Chem 41(29):2459–2468, DOI 10.1002/jcc.26401
36. Pereira AT, Ribeiro AJM, Fernandes PA, Ramos MJ (2017) Benchmarking of density functionals for the kinetics and thermodynamics of the hydrolysis of glycosidic bonds catalyzed by glycosidases. Int J Quantum Chem 117(18):e25409, DOI 10.1002/qua.25409
37. Neves RPP, Fernandes PA, Varandas AJC, Ramos MJ (2014) Benchmarking of density functionals for the accurate description of thiol–disulfide exchange. J Chem Theory Comput 10(11):4842–4856, DOI 10.1021/ct500840f
38. Ribeiro AJM, Ramos MJ, Fernandes PA (2010) Benchmarking of DFT functionals for the hydrolysis of phosphodiester bonds. J Chem Theory Comput 6(8):2281–2292, DOI 10.1021/ct900649e
39. Brás NF, Perez MAS, Fernandes PA, Silva PJ, Ramos MJ (2011) Accuracy of density functionals in the prediction of electronic proton affinities of amino acid side chains. J Chem Theory Comput 7(12):3898–3908, DOI 10.1021/ct200309v
40. Friedrich J (2015) Efficient calculation of accurate reaction energies—assessment of different models in electronic structure theory. J Chem Theory Comput 11(8):3596–3609, DOI 10.1021/acs.jctc.5b00087
41. Papajak E, Truhlar DG (2012) What are the most efficient basis set strategies for correlated wave function calculations of reaction energies and barrier heights? J Chem Phys 137(6):064110, DOI 10.1063/1.4738980

42. Kendall RA, Dunning TH, Harrison RJ (1992) Electron affinities of the first-row atoms revisited. systematic basis sets and wave functions. J Chem Phys 96(9):6796–6806, DOI 10.1063/1.462569
43. Grimme S (2003) Improved second-order Møller–Plesset perturbation theory by separate scaling of parallel- and antiparallel-spin pair correlation energies. J Chem Phys 118(20):9095–9102, DOI 10.1063/1.1569242
44. Jensen F (2017) Introduction to Computational Chemistry, 3rd edn. John Wiley & Sons, Incorporated, arXiv:1011.1669v3
45. Kozlowski PM, Kumar M, Piecuch P, Li W, Bauman NP, Hansen JA, Lodowski P, Jaworska M (2012) The cobalt–methyl bond dissociation in methylcobalamin: New benchmark analysis based on density functional theory and completely renormalized coupled-cluster calculations. J Chem Theory Comput 8(6):1870–1894, DOI 10.1021/ct300170y
46. Siegbahn PEM, Blomberg MRA (1999) Density functional theory of biologically relevant metal centers. Annu Rev Phys Chem 50(1):221–249, DOI 10.1146/annurev.physchem.50.1.221
47. Larsson ED, Dong G, Veryazov V, Ryde U, Hedegård ED (2020) Is density functional theory accurate for lytic polysaccharide monooxygenase enzymes? Dalton Trans 49(5):1501–1512, DOI 10.1039/c9dt04486h
48. Cisneros GA, Liu H, Zhang Y, Yang W (2003) Ab initio QM/MM study shows there is no general acid in the reaction catalyzed by 4-oxalocrotonate tautomerase. J Am Chem Soc 125(34):10384–10393, DOI 10.1021/ja029672a
49. Tuttle T, Keinan E, Thiel W (2006) Understanding the enzymatic activity of 4-oxalocrotonate tautomerase and its mutant analogues: a computational study. J Phys Chem B 110(39):19685–19695, DOI 10.1021/jp0634858
50. Tuttle T, Thiel W (2007) Substrate orientation in 4-oxalocrotonate tautomerase and its effect on QM/MM energy profiles. J Phys Chem B 111(26):7665–7674, DOI 10.1021/jp0685986
51. Sevastik R, Himo F (2007) Quantum chemical modeling of enzymatic reactions: the case of 4-oxalocrotonate tautomerase. Bioorg Chem 35(6):444–457, DOI 10.1016/j.bioorg.2007.08.003
52. Grimme S, Antony J, Ehrlich S, Krieg H (2010) A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. J Chem Phys 132(15):154104, DOI 10.1063/1.3382344
53. Grimme S, Ehrlich S, Goerigk L (2011) Effect of the damping function in dispersion corrected density functional theory. J Comput Chem 32(7):1456–1465, DOI 10.1002/jcc.21759
54. Zheng J, Xu X, Truhlar DG (2010) Minimally augmented Karlsruhe basis sets. Theor Chem Acc 128(3):295–305, DOI 10.1007/s00214-010-0846-z
55. Izsák R, Neese F (2011) An overlap fitted chain of spheres exchange method. J Chem Phys 135(14):144105, DOI 10.1063/1.3646921
56. Karton A, Martin JML (2005) Comment on: "estimating the Hartree–Fock limit from finite basis set calculations" [Jensen F (2005) Theor Chem Acc 113:267]. Theor Chem Acc 115(4):330–333, DOI 10.1007/s00214-005-0028-6
57. Halkier A, Helgaker T, Jørgensen P, Klopper W, Koch H, Olsen J, Wilson AK (1998) Basis-set convergence in correlated calculations on Ne, N_2, and H_2O. Chem Phys Lett 286(3-4):243–252, DOI 10.1016/s0009-2614(98)00111-0
58. Neese F, Valeev EF (2010) Revisiting the atomic natural orbital approach for basis sets: Robust systematic basis sets for explicitly correlated and conventional correlated ab initio methods? J Chem Theory Comput 7(1):33–43, DOI 10.1021/ct100396y
59. Neese F (2011) The ORCA program system. Wiley Interdiscip Rev: Comput Mol Sci 2(1):73–78, DOI 10.1002/wcms.81
60. Neese F (2017) Software update: the ORCA program system, version 4.0. Wiley Interdiscip Rev: Comput Mol Sci 8(1):e1327, DOI 10.1002/wcms.1327
61. Grimme S, Brandenburg JG, Bannwarth C, Hansen A (2015) Consistent structures and interactions by density functional theory with small atomic orbital basis sets. J Chem Phys 143(5):054107, DOI 10.1063/1.4927476
62. Handy NC, Cohen AJ (2001) Left-right correlation energy. Mol Phys 99:403–412
63. Lee C, Yang W, Parr RG (1988) Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. Phys Rev B 37(2):785–789, DOI 10.1103/physrevb.37.785
64. Miehlich B, Savin A, Stoll H, Preuss H (1989) Results obtained with the correlation energy density functionals of Becke and Lee, Yang and Parr. Chem Phys Lett 157(3):200–206, DOI 10.1016/0009-2614(89)87234-3
65. Goerigk L, Grimme S (2011) A thorough benchmark of density functional methods for general main group thermochemistry, kinetics, and noncovalent interactions. Phys Chem Chem Phys

13(14):6670–6688, DOI 10.1039/c0cp02984j
66. Perdew JP, Burke K, Ernzerhof M (1996) Generalized gradient approximation made simple. Phys Rev Lett 77(18):3865–3868, DOI 10.1103/physrevlett.77.3865
67. Zhang Y, Yang W (1998) Comment on "generalized gradient approximation made simple". Phys Rev Lett 80:890–890
68. Hujo W, Grimme S (2011) Performance of the van der Waals density functional VV10 and (hybrid)GGA variants for thermochemistry and noncovalent interactions. J Chem Theory Comput 7(12):3866–3871, DOI 10.1021/ct200644w
69. Mardirossian N, Head-Gordon M (2015) Mapping the genome of meta-generalized gradient approximation density functionals: the search for B97M-V. J Chem Phys 142(7):074111, DOI 10.1063/1.4907719
70. Sun J, Ruzsinszky A, Perdew J (2015) Strongly constrained and appropriately normed semilocal density functional. Phys Rev Letters 115(3):036402, DOI 10.1103/physrevlett.115.036402
71. Brandenburg JG, Bates JE, Sun J, Perdew JP (2016) Benchmark tests of a strongly constrained semilocal functional with a long-range dispersion correction. Phys Rev B 94(11):115144, DOI 10.1103/physrevb.94.115144
72. Zhao Y, Truhlar DG (2007) The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. Theor Chem Acc 120(1-3):215–241, DOI 10.1007/s00214-007-0310-x
73. Mardirossian N, Head-Gordon M (2016) ωB97M-V: A combinatorially optimized, range-separated hybrid, meta-GGA density functional with VV10 nonlocal correlation. J Chem Phys 144(21):214110, DOI 10.1063/1.4952647
74. Alipour M (2016) Seeking for spin-opposite-scaled double-hybrid models free of fitted parameters. J Phys Chem A 120(20):3726–3730, DOI 10.1021/acs.jpca.6b03406
75. Santra G, Sylvetsky N, Martin JML (2019) Minimally empirical double-hybrid functionals trained against the GMTKN55 database: revDSD-PBEP86-D4, revDOD-PBE-D4, and DOD-SCAN-D4. J Phys Chem A 123(24):5129–5143, DOI 10.1021/acs.jpca.9b03157
76. Dewar MJS, Zoebisch EG, Healy EF, Stewart JJP (1985) Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. Journal of the American Chemical Society 107(13):3902–3909, DOI 10.1021/ja00299a024
77. Stewart JJP (1989) Optimization of parameters for semiempirical methods II. applications. J Comput Chem 10(2):221–264, DOI 10.1002/jcc.540100209
78. Elstner M, Porezag D, Jungnickel G, Elsner J, Haugk M, Frauenheim T, Suhai S, Seifert G (1998) Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties. Phys Rev B 58:7260–7268, DOI 10.1103/PhysRevB.58.7260
79. Becke AD (1988) Density-functional exchange-energy approximation with correct asymptotic behavior. Phys Rev A 38(6):3098–3100, DOI 10.1103/physreva.38.3098
80. Perdew JP (1986) Density-functional approximation for the correlation energy of the inhomogeneous electron gas. Phys Rev B 33(12):8822–8824, DOI 10.1103/physrevb.33.8822
81. Perdew JP (1986) Erratum: Density-functional approximation for the correlation energy of the inhomogeneous electron gas. Phys Rev B 34(10):7406–7406, DOI 10.1103/physrevb.34.7406
82. Tao J, Perdew JP, Staroverov VN, Scuseria GE (2003) Climbing the density functional ladder: nonempirical meta–generalized gradient approximation designed for molecules and solids. Phys Rev Letters 91(14):146401, DOI 10.1103/physrevlett.91.146401
83. Becke AD (1993) Density-functional thermochemistry. iii. the role of exact exchange. J Chem Phys 98:5648–5652, DOI 10.1063/1.464913
84. Stephens PJ, Devlin FJ, Chabalowski CF, Frisch MJ (1994) Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields. J Phys Chem 98(45):11623–11627
85. Becke AD (1993) A new mixing of Hartree–Fock and local density-functional theories. J Chem Phys 98(2):1372–1377, DOI 10.1063/1.464304
86. Møller C, Plesset MS (1934) Note on an approximation treatment for many-electron systems. Phys Rev 46(7):618–622, DOI 10.1103/physrev.46.618
87. Weigend F, Ahlrichs R (2005) 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 7(18):3297–3305, DOI 10.1039/b508541a
88. Hehre WJ (1976) Ab initio molecular orbital theory. Acc Chem Res 9(11):399–406, DOI 10.1021/ar50107a003

89. Hehre WJ, Ditchfield R, Pople JA (1972) Self-consistent molecular orbital methods. XII. further extensions of gaussian-type basis sets for use in molecular orbital studies of organic molecules. J Chem Phys 56(5):2257–2261, DOI 10.1063/1.1677527
90. Vydrov OA, Voorhis TV (2010) Nonlocal van der Waals density functional: The simpler the better. J Chem Phys 133(24):244103, DOI 10.1063/1.3521275
91. Furche F, Ahlrichs R, Hättig C, Klopper W, Sierka M, Weigend F (2014) Turbomole. Wiley Interdiscip Rev: Comput Mol Sci 4(2):91–100, DOI 10.1002/wcms.1162
92. Treutler O, Ahlrichs R (1995) Efficient molecular numerical integration schemes. J Chem Phys 102(1):346–354, DOI 10.1063/1.469408
93. Eichkorn K, Treutler O, Öhm H, Häser M, Ahlrichs R (1995) Auxiliary basis sets to approximate Coulomb potentials. Chem Phys Lett 240(4):283–290, DOI 10.1016/0009-2614(95)00621-a
94. Eichkorn K, Weigend F, Treutler O, Ahlrichs R (1997) Auxiliary basis sets for main row atoms and transition metals and their use to approximate Coulomb potentials. Theor Chem Acta 97(1-4):119–124, DOI 10.1007/s002140050244
95. Goerigk L (2015) Treating London-dispersion effects with the latest minnesota density functionals: problems and possible solutions. J Phys Chem Lett 6:3891–3896, DOI 10.1021/acs.jpclett.5b01591