Volume 19 | Number 48 | 28 December 2017 | Pages 32147–32744

PCCP

Physical Chemistry Chemical Physics

rsc.li/pccp

ISSN 1463-9076

ROYAL SOCIETY OF CHEMISTRY

PAPER

Lars Goerigk, Stefan Grimme *et al.*

A look at the density functional theory zoo with the advanced GMTKN55 database for general main group thermochemistry, kinetics and noncovalent interactions

PCCP

PAPER

View Article Online
View Journal | View Issue

Cite this: *Phys. Chem. Chem. Phys.*, 2017, **19**, 32184

A look at the density functional theory zoo with the advanced GMTKN55 database for general main group thermochemistry, kinetics and noncovalent interactions†

Lars Goerigk,* [a] Andreas Hansen, [b] Christoph Bauer, [b] Stephan Ehrlich,‡ [b] Asim Najibi [a] and Stefan Grimme* [b]

We present the GMTKN55 benchmark database for general main group thermochemistry, kinetics and noncovalent interactions. Compared to its popular predecessor GMTKN30 [Goerigk and Grimme *J. Chem. Theory Comput.*, 2011, **7**, 291], it allows assessment across a larger variety of chemical problems—with 13 new benchmark sets being presented for the first time—and it also provides reference values of significantly higher quality for most sets. GMTKN55 comprises 1505 relative energies based on 2462 single-point calculations and it is accessible to the user community *via* a dedicated website. Herein, we demonstrate the importance of better reference values, and we re-emphasise the need for London-dispersion corrections in density functional theory (DFT) treatments of thermochemical problems, including Minnesota methods. We assessed 217 variations of dispersion-corrected and -uncorrected density functional approximations, and carried out a detailed analysis of 83 of them to identify robust and reliable approaches. Double-hybrid functionals are the most reliable approaches for thermochemistry and noncovalent interactions, and they should be used whenever technically feasible. These are, in particular, DSD-BLYP-D3(BJ), DSD-PBEP86-D3(BJ), and B2GPPLYP-D3(BJ). The best hybrids are ωB97X-V, M052X-D3(0), and ωB97X-D3, but we also recommend PW6B95-D3(BJ) as the best conventional global hybrid. At the meta-generalised-gradient (meta-GGA) level, the SCAN-D3(BJ) method can be recommended. Other meta-GGAs are outperformed by the GGA functionals revPBE-D3(BJ), B97-D3(BJ), and OLYP-D3(BJ). We note that many popular methods, such as B3LYP, are not part of our recommendations. In fact, with our results we hope to inspire a change in the user community's perception of common DFT methods. We also encourage method developers to use GMTKN55 for cross-validation studies of new methodologies.

Received 20th July 2017,
Accepted 17th October 2017

DOI: 10.1039/c7cp04913g

rsc.li/pccp

1 Introduction

Kohn–Sham (KS) density functional theory[1] (DFT) has without doubt shaped modern molecular quantum chemistry like no other methodology in recent years, and it is now the by far most frequently applied approach used by computational chemists.[2] Its implementation into standard quantum-chemical software packages has also made it easily accessible to the non-expert that may want to support their experimental results with computational insights. While exact in nature, the true exchange–correlation functional in DFT remains elusive, and instead the user has to rely on density functional approximation (DFAs). As a consequence, hundreds of DFAs have been developed, each with their own advantages and disadvantages. This means that despite DFT calculations being technically easy to perform, it is no trivial task to choose the right DFA for a specific problem.

Benchmarking—an assessment of DFA results against a set of reference data—has become a vital tool in method development and validation, which provides invaluable information to method users. Early benchmark sets focussed mostly on heat of formations (HoFs), electron affinities (EAs), ionisation potentials (IPs) and proton affinities (PAs) of relatively small molecules

[a] *School of Chemistry, The University of Melbourne, Parkville, Australia. E-mail: lars.goerigk@unimelb.edu.au; Tel: +61-3-83446784*
[b] *Universität Bonn, Mulliken Center for Theoretical Chemistry, Bonn, Germany. E-mail: grimme@thch.uni-bonn.de; Fax: +49-228-739064; Tel: +49-228-732544*

† Electronic supplementary information (ESI) available: Details on the new MB16-43 set. Damping parameters for the DFT-D3 dispersion correction. Comparison between the new and old reference values for ALX6. Dispersion-corrected *vs.* -uncorrected results. Analysis of best and worst MAD/RMSD. Weighted total mean absolute deviations. Statistical results for all test sets and DFAs. See DOI: 10.1039/c7cp04913g

‡ Present address: Schrödinger GmbH, Dynamostr. 13, 68161 Mannheim, Germany.

with experimental reference values; popular examples are the G2-1,[3] G2/97,[4] G3/99,[5] and G3/05[6] test sets. Those sets rely heavily on HoFs, which are derived from total atomisation energies (TAEs). While TAEs are without doubt an important test case for any new method, it was also pointed out in 2011 that the results of TAE benchmark studies are not necessarily representative for the treatment of thermochemical problems.[7] In 2017, Bartlett and co-workers automatically created 11 247 reaction-energy benchmark values from the 140 TAEs in the W4-11 set[8] and they confirmed the aforementioned conclusion from 2011.[9] Moreover, they showed that the usually expected error cancellations between the product and reactant species in reaction-energy calculations do not always happen, but that errors can in fact be amplified. This demonstrates the need to specifically design various benchmark sets covering reaction energies.

Others also share the latter view, which is why most sets developed after the turn of the millennium introduced chemically more relevant problems, such as reaction energies, barrier heights (BHs) or noncovalent interactions (NCIs). Some sets would still depend on experimental reference data, but gradually, zero-point vibrational-energy (ZPVE) exclusive, non-relativistic high-level *ab initio* values became the preferred benchmarks. Some popular examples are Truhlar's BH test sets HTBH38[10] and NHTBH38,[11] the isomerisation set ISO34,[12] and Hobza's sets for inter- and intramolecular NCI energies. In particular, we would like to mention the S22 set from 2006[13] and the Benchmark Energy and Geometry DataBase (BEGDB), which provides online access to Hobza's other seminal contributions to this field.[14]

Carrying out studies on a single benchmark set may give a biased picture of a DFA's accuracy and applicability, and using a different benchmark set may even give an opposing outcome. One way of introducing an unbiased approach to benchmarking is Korth and Grimme's MB08-165 set[15] of randomly generated artificial molecules (AMs), which turned out to be a good indicator for a DFA's robustness.[7,15] A different strategy to identify a robust method was first rigorously carried out by Truhlar, who compiled databases that consist of different benchmark sets covering various chemical problems.[16–21] The latest of those compilations is the "Database 2015B", which comprises 471 molecular and atomic data (relative energies).[22] A DFA is deemed robust if it performs equally well over different benchmark sets. In that case, there is a higher probability that it can also be safely employed to new problems. Also, two of us developed benchmark databases that focussed on general main-group thermochemistry, kinetics, and noncovalent interactions (GMTKN) that we dubbed GMTKN24[23] and GMTKN30.[24] Alternative large databases that took parts of GMTKN30 and combined them with other benchmark sets have also been recently promoted by Mardirossian and Head-Gordon.[25,26]

Our herein presented work focusses entirely on our GMTKN databases. The first version, GMTKN24, contained 24 benchmark sets and was initially used to establish that non-empirically derived DFAs belonging to the generalised-gradient-approximation (GGA) or meta-GGA classes were not necessarily better equipped to describe thermochemistry than (semi-)empirical ones.[23] In 2011, GMTKN24 was extended by six additional sets and renamed GMTKN30. While it was initially used to cross-validate the PWPB95 double-hybrid DFA, two of us later used GMTKN30 to test 45 DFAs from all five rungs of Jacob's Ladder, with the lowest rung being the local-density approximation (LDA), followed by GGAs or non-separable gradient approximations (NGAs),[27] meta-GGAs/NGAs, hybrid DFAs, and double-hybrid DFAs. Double hybrids are one representative for the fifth rung, as they incorporate information from virtual KS orbitals through a second-order Møller–Plesset-type (MP2) term.[28,29] The motivation behind Jacob's Ladder was to be able to classify the chaotic "zoo" of DFAs according to their underlying components, with the expectation that a higher rung would guarantee a more accurate outcome. The 2011 GMTKN30 study showed that this is indeed the case, with double hybrids being the most reliable and accurate functionals, clearly outperforming related MP2-type methods.[7] While hybrid functionals followed in this ranking, the difference between meta-GGAs and GGAs was not very pronounced. As no surprise came the finding that LDAs were not competitive at all for molecular chemistry.

The large GMTKN30 study also provided the user community with clear guidelines on which functionals to trust and which to avoid.[7] Based on GMTKN30, the best DFAs turned out to be all double hybrids, with the exception of non-empirical double hybrids.[29] In particular, Kozuch and Martin's DSD[30–32] functionals and the PWPB95[24]-D3[33] functional were the clear winners.[7,29] The best hybrid DFAs were Truhlar's M062X-D3,[34] M052X-D3,[35] and PW6B95-D3[36] DFAs, with the latter showing generally larger robustness and less technical issues, such as occasionally occurring strong quadrature-grid and convergence problems.[7] With the exception of ωB97X-D,[37] range-separated hybrids showed no better behaviour than global hybrids. Most important was probably the finding that the most popular DFA by far, namely B3LYP,[38,39] turned out to be the worst of 23 hybrids for the calculation of reaction energies.[7] Some GGAs turned out to be competitive with meta-GGAs, and the revPBE-D3[40] and B97-D3[41] approaches were recommended. Furthermore the study demonstrated that London-dispersion effects do also influence reaction energies and BHs, contrary to the common perception that they are weak and, thus, negligible in those cases.[7] The large GMTKN30 study also indicated that Minnesota DFAs needed to be long-range dispersion-corrected despite the wide-spread belief that they already incorporate dispersion effects;[7] subsequent studies[42,43] confirmed these early indications (also see ref. 25 for a related, more recent study).

Shortly after their introduction, GMTKN24 and GMTKN30 became very popular tools in method development and evaluation, with selected examples being ref. 44–51. However, despite this success, some questions became evident to us over time. Are 30 benchmark sets really sufficient to assess a method's robustness? Would the overall picture change if we added new sets? Are the system sizes covered in GMTKN30 still representative of current problems? Is the letter "K" in GMTKN—namely BHs—underrepresented with only two test sets? Are all reference values reliable? Particularly the last question is important. It was inspired by a 2015 study, which

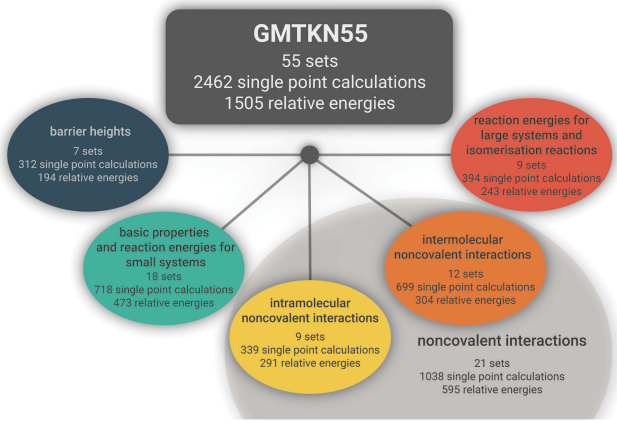

Fig. 1 The new GMTKN55 database and its various categories.

showed that the popular CBS-QB3[52,53] composite approach used to generate reference BHs of pericyclic reactions (BHPERI set) had a surprisingly large mean absolute deviation (MAD) of 2.1 kcal mol^{-1} with respect to the explicitly-correlated Weizmann protocols W1-F12 and W2-F12.[54] This value turned out to be larger than the MADs for double-hybrid DFAs, and thus evaluating DFT methods against CBS-QB3 numbers became questionable.[55] Indeed, DFA rankings based on the older reference values changed significantly when the newer ones were used.

Herein, we intend to answer those questions, and we present the advanced and improved GMTKN55 database, which now covers 55 test sets (Fig. 1). With seven sets for BHs, that category is much better represented than in the two predecessors. We present 13 completely new sets to become part of GMTKN55. In addition, existing sets were modified or the database was further extended by sets (partially) taken from the literature, for which we present new reference values in most cases. For most of the sets that remain identical to GMTKN30, we also present updated reference values. Contrary to related large databases, we tried to increase the accuracy of the reference values substantially by relying on Weizmann composite protocols whenever possible. In GMTKN55, we also provide a higher number of large systems, such as C_{60} isomers, and we also include more ionic systems. An overview of the new database is given in Table 1. A detailed description of each of the 55 test sets follows in Section 2.

Having established the new database, we then proceed with a detailed evaluation of DFAs to re-assess whether the old recommendations for GMTKN30 are still valid, and how approaches published since then compare to them as well as methods that were not considered in 2011. As outlined in Section 3, we carefully selected 83 out of 217 DFAs for this task. Contrary to other comprehensive benchmark studies that avoided testing double hybrids,[19,22,26] we again include those herein. Moreover, all 83 DFAs are dispersion corrected, which enables us to conduct a more consistent analysis compared to other studies.[19,22,26] New DFT-D3[33,44] dispersion-correction damping parameters are presented for 35 methods for the first time. In Section 4, we suggest how to best analyse GMTKN55, before we proceed with a detailed discussion in Section 5. We will conclude with an update of our previous recommendations for each of the four highest rungs on Jacob's Ladder to guide the DFT user and method developer for future applications.

2 Description of the GMTKN55 database

The new GMTKN55 database covers 55 different test sets and for each tested method, the user has to conduct 2462 single-point calculations to obtain 1505 relative energies for subsequent statistical analysis, as opposed to GMKTKN30 with 1218 systems and 841 relative energies. As shown in Fig. 1, these 55 sets can be distributed amongst four categories. Similarly to the GMTKN24 and GMTKN30 predecessors, the first category comprises what we call "basic properties", and it includes standard benchmarking problems, such as TAEs, EAs, or IPs. Contrary to its predecessors, though, we also include reactions between smaller systems, as is reflected in the name for this category, which in total comprises 18 test sets. Reactions between larger systems and isomerisations are covered by the second category (nine sets in total). A drawback of GMTKN24 and GMTKN30 is that BHs are underrepresented with only two benchmark sets in the old "basic properties" category. GMTKN55 now features seven different test sets that are collected in a separate category. 21 sets directly allow assessing a method's ability to describe NCIs. This category can be further divided into two subcategories focussing on intermolecular (12 sets) or intramolecular interactions (nine sets), respectively.

Table 1 lists all 55 benchmark sets including a short description, the number of data points for each set, the number of single-point calculations that need to be carried out, and the nature of the reference data. For users familiar with the GMTKN30 predecessor, Table 1 also summarises which sets were modified, left unchanged or newly added.

Out of the 55 sets, only three are identical to GMTKN30 (G21EA, G21IP, ACONF). 15 sets have the same geometries as before, but updated reference values, which were either published elsewhere or that were calculated for this work. These are the sets: NBPRC, G2RC, BH76RC, DARC, RSE43, BSR36, ISO34, BH76, BHPERI, ADIM6, S22, HEAVY28, WATER27, IDISP, and SCONF. Five sets are extensions or modifications of existing sets and, in addition, we determined new reference values for these. To reflect these changes, we gave those sets new names. The set formerly known as PA becomes PA26, AL2X becomes AL2X6, ISOL22 was extended to ISOL24, and PCONF is now known as PCONF21. The DC13 set is an extension of DC9 and it also contains parts of the O3ADD6 set, which has become obsolete and will no longer be a part of GMTKN55. Six sets were replaced by new ones: the W4-11 set replaces W4-08, the newly developed SIE4x4 is a replacement for SIE11, the new ALK8 set replaces ALK6, we present the new MB16-43 set as a substitute for MB08-165, the new RG18 set replaces RG6, and AMINO20x4 makes CYCONF obsolete. 26 sets are an addition to GMTKN55. Out of those, 13 were taken from the literature without any or with only minor modifications: ALKBDE10, FH51, CDIE20,

Table 1 Description of the subsets within the GMTKN55 database and changes compared to its predecessor GMTKN30

Set	Description	Changes with respect to GMTKN30	#[a]	$\overline{\lvert\Delta E\rvert}$[b]	Ref. method
Basic properties and reaction energies for small systems					
W4-11[8]	Total atomisation energies	Replaces W4-08[56]	140 (152)	306.91	[c]
G21EA[3,23]	Adiabatic electron affinities	None	25 (50)	33.62	[d]
G21IP[3,23]	Adiabatic ionisation potentials	None	36 (71)	257.61	[d]
DIPCS10	Double-ionisation potentials of closed-shell systems	New[e]	10 (20)	654.26	[f]
PA26	Adiabatic proton affinities (incl. of amino acids)	Extension of PA;[23,57,58] new ref.[e]	26 (52)	189.05	[f,g,h]
SIE4x4[c]	Self-interaction-error related problems	New;[e] replaces SIE11[23]	16 (23)	33.72	[g]
ALKBDE10[59]	Dissociation energies in group-1 and -2 diatomics	New from lit.	10 (20)	100.69	[d]
YBDE18[60]	Bond-dissociation energies in ylides	New from lit.; new ref.[e]	18 (29)	49.28	[g]
AL2x6	Dimerisation energies of AlX_3 compounds	Modification of AL2X; new ref.[e]	6 (11)	35.88	[f,g]
HEAVYSB11	Dissociation energies in heavy-element compounds	New[e]	11 (22)	58.02	[h]
NBPRC[23,24,61]	Oligomerisations and H_2 fragmentations of NH_3/BH_3 systems H2 activation reactions with PH_3/BH_3 systems	New ref.[62]	12 (21)	27.71	[f,g]
ALK8	Dissociation and other reactions of alkaline compounds	New;[e] replaces ALK6[33]	8 (17)	62.60	[h]
RC21	Fragmentations and rearrangements in radical cations	New[e]	21 (41)	35.70	[f]
G2RC[4,23]	Reaction energies of selected G2/97 systems	New ref.[e]	25 (47)	51.26	[g]
BH76RC[23]	Reaction energies of the BH76[10,11,23] set	New ref.[e]	30	21.39	[g]
FH51[63,64]	Reaction energies in various (in-)organic systems	New from lit.	51 (87)	31.01	[i]
TAUT15	Relative energies in tautomers	New[e]	15 (25)	3.05	[f]
DC13[18,23,28,65–73]	13 difficult cases for DFT methods	Extension of DC9;[23] new ref.[e]	13 (30)	54.98	[f,g,h,i,k,l]
Reaction energies for large systems and isomerisation reactions					
MB16-43	Decomposition energies of artificial molecules	New;[e] replaces MB08-165[15]	43 (58)	414.73	[f]
DARC[23,74]	Reaction energies of Diels–Alder reactions	New ref.[62]	14 (22)	32.47	[f]
RSE43[75]	Radical-stabilisation energies	New ref.[e]	43 (88)	7.60	[f]
BSR36[76,77]	Bond-separation reactions of saturated hydrocarbons	New ref.[e]	36 (38)	16.20	[k]
CDIE20[78]	Double-bond isomerisation energies in cyclic systems	New from lit.	20 (36)	4.06	[f]
ISO34[12]	Isomerisation energies of small and medium-sized organic molecules	New ref.[e]	34 (63)	14.57	[f]
ISOL24[79]	Isomerisation energies of large organic molecules	Extension of ISOL22;[24] new ref.[e]	24 (48)	21.92	[k]
C60ISO[80]	Relative energies between C_{60} isomers	New from lit.	9 (10)	98.25	[m]
PArel	Relative energies in protonated isomers	New[e]	20 (31)	4.63	[h]
Reaction barrier heights					
BH76[10,11,23]	Barrier heights of hydrogen transfer, heavy atom transfer, nucleophilic substitution, unimolecular and association reactions	New ref.[e]	76 (86)	18.61	[g]
BHPERI[23,81–83]	Barrier heights of pericyclic reactions	New ref.[55]	26 (61)	20.87	[f,g]
BHDIV10	Diverse reaction barrier heights	New[e]	10 (20)	45.33	[f,g]
INV24[84]	Inversion/racemisation barrier heights	New from lit.	24 (48)	31.85	[f,g,k]
BHROT27	Barrier heights for rotation around single bonds	New[e]	27 (40)	6.27	[f,g]
PX13[85]	Proton-exchange barriers in H_2O, NH_3, and HF clusters	Modified set from lit.	13 (29)	33.36	[f]
WCPT18[86]	Proton-transfer barriers in uncatalysed and water-catalysed reactions	Modified set from lit.	18 (28)	34.99	[n]
Intermolecular noncovalent interactions					
RG18	Interaction energies in rare-gas complexes	New;[e] replaces RG6[33]	18 (25)	0.58	[o]
ADIM6[33]	Interaction energies of *n*-alkane dimers	New ref.[e]	6 (12)	3.36	[f]
S22[13]	Binding energies of noncovalently bound dimers	New ref.[87]	22 (57)	7.30	[l]
S66[88]	Binding energies of noncovalently bound dimers	New from lit.	66 (198)	5.47	[l]
HEAVY28[33]	Noncovalent interaction energies between heavy element hydrides	New ref.[e]	28 (38)	1.24	[h]
WATER27[89]	Binding energies in $(H_2O)_n$, $H^+(H_2O)_n$ and $OH^-(H_2O)_n$	New ref.[90]	27 (30)	81.14	[p]
CARBHB12	Hydrogen-bonded complexes between carbene analogues and H_2O, NH_3, or HCl	New[e]	12 (36)	6.04	[g]
PNICO23[91]	Interaction energies in pnicogen-containing dimers	Modifies set from lit.; new ref.[e]	23 (69)	4.27	[f,g]
HAL59[92,93]	Binding energies in halogenated dimers (incl. halogen bonds)	Combination of two sets from lit.	59 (105)	4.59	[l]
AHB21[94]	Interaction energies in anion–neutral dimers	New from lit.	21 (63)	22.49	[h,i]
CHB6[94]	Interaction energies in cation–neutral dimers	New from lit.	6 (18)	26.79	[h,l]
IL16[94]	Interaction energies in anion–cation dimers	New from lit.	16 (48)	109.04	[l]
Intramolecular noncovalent interactions					
IDISP[12,23,24,95,96]	Intramolecular dispersion interactions	New ref.[e]	6 (13)	14.22	[k]
ICONF	Relative energies in conformers of inorganic systems	New[e]	17 (27)	3.27	[f]

Table 1 (*continued*)

Set	Description	Changes with respect to GMTKN30	#[a]	$\overline{\lvert\Delta E\rvert}$[b]	Ref. method
ACONF[97]	Relative energies of alkane conformers	None	15 (18)	1.83	*q*
AMINO20x4[98]	Relative energies in amino acid conformers	Replaces CYCONF[99]	80 (100)	2.44	*p*
PCONF21[100,101]	Relative energies in tri- and tetrapeptide conformers	Extension of PCONF;[23,100] new ref.[e]	18 (21)	1.62	*k*
MCONF[102]	Relative energies in melatonin conformers	New from lit.; new ref.[e]	51 (52)	4.97	*k*
SCONF[23,103]	Relative energies of sugar conformers	New ref.[e]	17 (19)	4.60	*k*
UPU23[104]	Relative energies between RNA-backbone conformers	New from lit.	23 (24)	5.72	*m*
BUT14DIOL[105]	Relative energies in butane-1,4-diol conformers	New from lit.; new ref.[e]	64 (65)	2.80	*f*

[a] Number of relative energies and single-point calculations (in parentheses), except for BH76RC, for which the single-point calculations carried out for BH76 are sufficient. [b] Averaged absolute relative energy (kcal mol^{-1}). [c] W4.[106] [d] (Theoretically back-corrected) exp. [e] This work. [f] W1-F12.[54] [g] W2-F12.[54] [h] CCSD(T)[107]/CBS. [i] CCSD(T)-F12[108]/CBS. [j] W3.2.[106] [k] DLPNO-CCSD(T)/CBS.[109] [l] Estimated CCSD(T)/CBS. [m] DLPNO-CCSD(T)/CBS*.[104] [n] W2.2.[106] [o] CP-corrected[110] CCSD(T)/CBS. [p] Estimated CCSD(T)-F12/CBS. [q] W1h-val.[111]

C60ISO, INV24, PX13, WCPT18, S66, HAL59, AHB21, CHB6, IL16, and UPU23. Four sets were taken from the literature, but we present new reference values for them: YBDE18, PNICO23, MCONF, and BUT14DIOL. Finally, we add nine entirely new test sets under the names DIPCS10, HEAVYSB11, RC21, TAUT15, PArel, BHDIV10, BHROT27, CARBHB12, and ICONF.

Detailed descriptions of each test set follow below. Similarly to its predecessors, all reference values in GMTKN55 are non-relativistic and ZPVE exclusive, and in the majority of the cases obtained from all-electron treatments. All structures and other information can be obtained from a dedicated website.[112]

2.1 Basic properties and reaction energies for small systems

2.1.1 The new W4-11 set. The W4-11 set for TAEs by Karton *et al.*[8] is an extension of their older W4-08 database.[56] It contains the same 99 small molecules as W4-08, with the addition of 41 medium-sized organic molecules, for which London-dispersion can play an additional stabilising role.[50] 99 molecules consist of first-row elements, 19 are second-row species, 21 are mixed first- and second-row species; additionally, the set also contains dihydrogen. 16 systems exhibit some multi-reference character.[8] All reference values are based on the highly accurate W4 composite level of theory,[106] for which a 95% confidence interval of about 0.17 kcal mol^{-1} has been reported.[8] These reference TAEs range from 2.7 kcal mol^{-1} (Be_2) to 1007.9 kcal mol^{-1} (propane), with the average value being 306.91 kcal mol^{-1}. 152 single-point calculations have to be carried out to treat W4-11.

2.1.2 The unchanged G21EA set. The G21EA set for 25 adiabatic EAs taken from the G2-1 set[3] was left unchanged and is the same as for the GMTKN24 and GMTKN30 predecessors. The EAs range from −0.2 kcal mol^{-1} (NO) to 89.5 kcal mol^{-1} (CN). The average absolute electron affinity for this set is 33.62 kcal mol^{-1}, for which 50 single-point calculations need to be carried out.

2.1.3 The unchanged G21IP set. Also, the G21IP set with 36 adiabatic IPs taken from G2-1 has been left unchanged.[3,23] These IPs range from 118.5 kcal mol^{-1} (Na) to 401.7 kcal mol^{-1} (F), with an average of 257.61 kcal mol^{-1}. In total, 71 single-point energies have to be calculated to evaluate this set.

2.1.4 The new DIPCS10 set. Herein, we present the new DIPCS10 set which contains 10 double-ionisation potentials of closed-shell systems, namely ethylene, ethane, cyclobutadiene, diazene, ammonia, formaldehyde, dihydrogen sulphide, phosphine, magnesium, and beryllium. All molecular structures were obtained at the TPSS[113]-D3(BJ)[33,44]/def2-TZVP[114] level of theory and all reference values at the explicitly-correlated W1-F12 level. The only exception to the latter are the reference values for Mg and Be, for which all-electron CCSD(T)/cc-pVC5Z[115] calculations were carried out due to technical reasons. The 10 double-ionisation potentials range from 522.1 kcal mol^{-1} (Be) to 776.5 kcal mol^{-1} (NH_3), with an average of 654.26 kcal mol^{-1}. In total, 20 single-point energies have to be calculated to evaluate this set.

2.1.5 The new PA26 set. The PA26 set for PAs is an extension of the PA set from GMTKN24 and GMTKN30. The original 12 systems were kept, however, their reference values were changed from W1[57] to W2-F12 (ammonia, water, ethyne, silane, phosphine, dihydrogen sulphide, hydrogen chloride, and dihydrogen) and from estimated CCSD(T)/complete-basis-set (CBS) numbers to W1-F12 (ethene, butadiene, hexatriene, and octatetraene).[62] In addition, we extended this set by 14 molecules, for which reference values were obtained at the CCSD(T)/CBS(def2-TZVPP/def2-QZVPP[114])//PBEh-3c[116] level of theory: C_2F_6, ethanol, CH_3COOH, glycine, phenol, acetylsalicylic acid, CH_2S, Si_2H_6, cysteine, phosphapyrrole, and the four nucleobases adenine, thymine, guanine, and cytosine. The new set contains 26 PAs that range from 106.2 (H_2) to 236.0 kcal mol^{-1} (guanine), with an average value of 189.05 kcal mol^{-1}. A total of 52 single-point calculations have to be carried out for this set.

2.1.6 The new SIE4x4 set. In GMTKN24, the SIE11 test set for self-interaction-error (SIE) related problems was introduced.[23] However, closer analysis shows that "pure" one-electron SIE-effects are hard to capture with this set, and therefore we herein introduce the new SIE4x4 set as a replacement. It contains four positively charged dimers (H_2^+, He_2^+, $(NH_3)_2^+$, and $(H_2O)_2^+$) for which their dissociation energies at four different points along their dissociation potential energy curves are calculated: at their respective inter-monomeric equilibrium distance, 1.25 times, 1.5 times and 1.75 times their equilibrium distance. The relevant structures were optimised at the SCS-MP2[117]/def2-TZVPP level

of theory, and all reference values are W2-F12 energies, which are identical to the Hartree–Fock (HF)/CBS values for H_2^+. The dissociation energies range from 4.9 ($(NH_3)_2^+$ at 1.75 times its equilibrium distance) to 64.4 kcal mol^{-1} (H_2^+ in its equilibrium distance). The average absolute reference energy is 33.72 kcal mol^{-1}. 23 single-point calculations need to be carried out for this set.

2.1.7 The new ALKBDE10 set. In 2015, Yu and Truhlar presented a study on bond-dissociation energies (BDEs) of polar diatomics containing either a group-1 or group-2 atom.[59] We adopted ten of those dimers from their study to form the new ALKBDE10 test set: BeF, BeO, CaO, HF, KF, LiF, KF, LiF, LiO, MgO, MgS, and NaO. All structures are based on the PW6B95-D3(BJ)/def2-QZVPP level of theory. The ZPVE-corrected experimental reference values were taken from the original study. The BDEs range from 56.7 kcal mol^{-1} (MgS) to 139.2 kcal mol^{-1} (LiF) with an average value of 100.69 kcal mol^{-1}. 20 single-point calculations have to be carried out for this set.

2.1.8 The new YBDE18 set. In 2012, Truhlar and co-workers presented the YBDE18 benchmark database for BDEs in 18 ylidic systems to assess DFT-based methods for the treatment of this important class of organic reagents.[60] We adopt this set with the same MP2/aug-cc-pVTZ[118] geometries. The original study used the CCSD(T)-F12a[108]/VTZ-F12 level of theory to obtain reference values, herein we present new values based on the W2-F12 composite approach. We observed absolute differences between both levels of theories of up to 0.94 kcal mol^{-1} (NF_3-CH_2 system). Compared to the new W2-F12 values, the original CCSD(T)-F12a/VTZ-F12 values had a mean deviation of -0.36 kcal mol^{-1} and a mean absolute deviation (MAD) of the same magnitude, indicating a systematic underestimation. The new W2-F12 BDEs range from 12.31 kcal mol^{-1} (NF_3-$C(BH_2)_2$) to 91.14 kcal mol^{-1} (F_2S-CH_2) with an average value of 49.28 kcal mol^{-1}. 29 single-point calculations have to be carried out for this set.

2.1.9 The updated AL2X6 set. AL2X6 is an updated version of the AL2X set in GMTKN24 and GMTKN30. The original set contained binding energies of seven dimers of alane derivatives with theoretically back-corrected experimental data compiled by Johnson *et al.*[74] Herein, we present new reference values for the W2-F12 level of theory with the exception of $Al_2(CH_3)_6$, for which W1-F12 was technically more feasible. The largest difference between W2-F12 and W1-F12 for the other systems was 0.3 kcal mol^{-1} for Al_2Cl_6. As W*n*-F12 treatments are currently not defined for bromine, we deleted the Al_2Br_6 system from this benchmark set and renamed the resulting set AL2X6. The largest difference between the old and new reference values was 2.7 kcal mol^{-1} for Al_2H_6 and the MAD for the old values compared to the new ones is significant at 1.4 kcal mol^{-1}. The new binding energies in AL2X6 range from 23.1 kcal mol^{-1} ($Al_2(CH_3)_6$) to 51.6 kcal mol^{-1} (Al_2F_6) with an average value of 35.88 kcal mol^{-1}. 11 single-point calculations need to be carried out for this set.

2.1.10 The new HEAVYSB11 set. The new HEAVYSB11 benchmark set contains 11 homolytic bond-cleavage energies of covalently bound dimers of the heavy element hydrides GeH_3, SH, SeH, of the methylated heavy elements $Sn(CH_3)_3$, $Pb(CH_3)_3$, $TeCH_3$, $P(CH_3)_2$, $As(CH_3)_2$, $Sb(CH_3)_2$, and of Cl and Br. All molecular structures were obtained at the PBE0[119,120]-D3(BJ)/def2-TZVP level of theory. The reference data were computed at the CCSD(T)/CBS[121,122] (aug-cc-pwCVTZ-PP/aug-cc-pwCVQZ-PP)[123,124] level. The exception are the reactions involving $SnCH_3$ and $PbCH_3$, for which the diffuse functions were neglected. The bond-dissociation energies range from 43.79 ($[Sb(CH_3)_2]_2$) to 73.82 kcal mol^{-1} ($(GeH_3)_2$) with an average of 58.02 kcal mol^{-1}. A total of 22 single-point calculations need to be carried out for this set.

2.1.11 The updated NBPRC set. The NBPRC[23,61] set for 12 oligomerisation and H_2 fragmentation reactions of NH_3/BH_3 systems, as well as binding and dihydrogen-splitting energies for frustrated Lewis-pair (FLP) models keeps its general form introduced for GMTKN30. However, we replaced the old estimated CCSD(T)/CBS references with newer W2-F12 ones, except for the hydrogen-splitting reactions involving BCl_3-$P(CH_3)_6$ and BF_3-$P(CH_3)_6$, for which W1-F12 numbers were obtained.[62] The largest difference between the old and new reference values is 1.04 kcal mol^{-1} ($3NH_2BH_2 \rightarrow (NH_2BH_2)_3$). The MAD for the old values with respect the newer ones is 0.42 kcal mol^{-1}. The new reaction energies in this set range from -48.9 kcal mol^{-1} ($3NH_2BH_2 \rightarrow (NHBH)_3 + 3H_2$) to 40.4 kcal mol^{-1} ($PH_3BH_3 + H_2 \rightarrow [PH_4]^+[BH_4]^-$), with an averaged absolute energy of 27.71 kcal mol^{-1}. 21 single-point calculations need to be carried out for this set.

2.1.12 The new ALK8 set. The new ALK8 set replaces the ALK6 set in GMTNK30. It contains two decomposition reactions of alkaline metal complexes into their dimers (taken from the previous ALK6[33] set) complemented by six reactions of additional alkaline compounds (mainly lithium organyls). The new CCSD(T)/CBS(aug-cc-pwCVTZ/aug-cc-pwCVQZ)[125]//PBEh-3c reference values range from 25.30 kcal mol^{-1} for the reaction $Na^+ + LiNaH_2 \rightarrow Li^+ + Na_2H_2$ up to 131.13 kcal mol^{-1} for the reaction $Li_4(CH_3)_4 \rightarrow 4LiCH_3$, with an averaged absolute energy of 62.60 kcal mol^{-1}. 17 single-point calculations need to be carried out for this set.

2.1.13 The new RC21 set. RC21 is a new benchmark set created in the context of the QCEIMS[126] mass-spectrum simulation project. It comprises 21 reaction energies of organic radical cation fragmentation and rearrangement reactions of various types (α- and heterocyclic cleavage, benzylium formation and tropylium rearrangement of the toluene radical cation, allylic and non-activated bond cleavage, McLafferty rearrangement, H_2O elimination from the isocrotonic acid radical cation, retro-Diels–Alder reaction, and CO elimination from the phenol radical cation). The reference values are obtained with the W1-F12 protocol based on PBE0-D3(BJ)/def2-TZVP optimised geometries. They range from -6.72 kcal mol^{-1} for the McLafferty rearrangement of 2-pentanone up to 126.56 kcal mol^{-1} for the HCN loss from pyridine, with an averaged absolute energy of 35.70 kcal mol^{-1}. 41 single-point calculations need to be carried out for this set.

2.1.14 The updated G2RC set. The composition of the G2RC subset, with 25 reactions whose reactants and products had been taken from the G2/97 set of HoFs,[4] remains the same

as in GMTKN24 and GMTKN30. However, we replaced the original back-corrected experimental reference data with new W2-F12 values. The biggest difference between the old and new reference values was found for the last reaction in the set (3.41 kcal mol^{-1} for $Li_2 + F_2 \rightarrow 2LiF$). The MAD for the old values relative to the new ones is 0.9 kcal mol^{-1}. The reaction energies cover an energy range from −2.18 kcal mol^{-1} ($CH_3CHO \rightarrow CO + CH_4$) to −216.11 kcal mol^{-1} ($Li_2 + F_2 \rightarrow 2LiF$), with an averaged absolute energy of 51.26 kcal mol^{-1}. A total of 47 single-point calculations need to be carried out for this set.

2.1.15 The updated BH76RC set. As in GMTKN24 and GMTKN30, the BH76RC set contains 30 reaction energies for the reactions investigated in the BH76 barrier-height test set. The original reference values were based on either W1 values or other 'best theoretical estimates' (see ref. 10 and 11 for more details). Herein, we provide more consistent values based on the W2-F12 composite approach. Differences between these values can be as large as 1.64 kcal mol^{-1} ($H + PH_3 \rightarrow H_2 + PH_2$), while the MAD for the old values compared to the new ones is 0.51 kcal mol^{-1}. The reaction energies range from −103.28 kcal mol^{-1} ($H + F_2 \rightarrow HF + F$) to 3.69 kcal mol^{-1} ($H + N_2 \rightarrow HN_2$), with an averaged absolute energy of 21.39 kcal mol^{-1}. As the systems are taken from BH76, no additional single-point calculations have to be carried out if BH76 is also analysed.

2.1.16 The new FH51 set. The FH51 set contains 51 reaction energies in small inorganic and organic systems and was developed by Friedrich and Hänchen in 2013.[63] Herein, we use Friedrich's updated CCSD(T)-F12/CBS reference values from 2015,[64] which have an averaged absolute value of 31.01 kcal mol^{-1}. The energies in FH51 range from −150.81 kcal mol^{-1} ($C_6H_{12}O + 2H_2O_2 \rightarrow$ ethyl-γ-butyrolactone + $3H_2O$) to −0.18 kcal mol^{-1} ($C_3H_7CO_2H + NH_3 \rightarrow C_3H_7CONH_2 + H_2O$). 87 single-point calculations need to be carried out for this set.

2.1.17 The new TAUT15 set. TAUT15 is a set of 15 chemically different tautomerisations: keto–enol tautomerism of acetyl acetone and malone aldehyde, beta-imine ketone tautomerism, heterocycle and nucleobase tautomerism, the tautomerism of 2-hydroxypyridine *vs.* pyridone, relative energies of low-lying cytosine and guanine tautomers, N-heterocycle tautomerism, relative energies of 1*H*-tetrazole *vs.* 2*H*-tetrazole and proxy- *vs.* tele-methylimidazole, respectively, and thiol–thione tautomerism of thioformic acid and 2-pyridinethiol. The reference values obtained at the W1-F12//SCS(1.1,0.6)-MP2[127]/def2-QZVP level range from −5.44 kcal mol^{-1} up to 13.03 kcal mol^{-1}, with an average absolute energy of 3.05 kcal mol^{-1}. 25 single-point calculations need to be carried out for this set.

2.1.18 The extended DC13 set. In GMTKN24 and GMTKN30, we used the DC9 test set, which was a compilation of nine reactions taken from the literature that were known to be difficult for DFT methods.[23,28,65–69] The reference values were based on different levels of theory or experimental numbers. Herein, we present the extended DC13 set that contains the same nine reactions with updated, entirely theoretical reference values and four additional reactions. For the original nine reactions, different levels of theories were employed depending on system size. W3.2 was used for the decomposition of Be_4 into beryllium atoms, W2-F12 for the formation of S_8 from S_2 as well as for the 1,3-dipolar cycloaddition between ethene and diazomethane. The W1-F12 composite approach was used for the the tautomeric 2-pyridone/2-hydroxypyridine system, the isomerisation from hepta-1,2,3,5,6-hexaene to hepta-1,3,5-triyne, and the carbo-[3]-oxacarbon isomerisation. The DLPNO-CCSD(T)/TightPNO[128]/CBS(aug-cc-pVTZ/aug-cc-pVQZ) level provides new reference values for the isomerisation between two $(CH)_{12}$ isomers. The reference value for the dimerisation of tetramethyl-ethene to octamethyl-cyclobutane was updated from SCS-MP2/CBS to CCSD(T)/CBS(cc-pVTZ/cc-pVQZ).

In 2016, Manna and Martin carried out a detailed study on C_{20} isomers and confirmed that those are very challenging systems, as their electronic structures can differ quite strongly, which often prevents error-cancellation effects when calculating the related isomerisation energies.[73] Two of those isomers—C_{20} in its cage and bowl configurations—were already a part of the original DC9 set. Manna and Martin obtained a reliable and accurate isomerisation energy of −8.2 kcal mol^{-1} based on estimated CCSD(T)/CBS values (MP2/CBS(aug-cc-pVQZ/aug-cc-pV5Z) corrected with the correlation-energy difference between CCSD(T)/cc-pVTZ and MP2/cc-pVTZ). Unfortunately, the authors used geometries that had been obtained at a level of theory different from the structures used in DC9. However, we prefer to keep the original DC9 structures to allow users that have their own local version of GMTKN30 to easily upgrade it to GMTKN55. To solve this dilemma, we estimated the influence of the geometry on the isomerisation energy with the help of DLPNO-CCSD(T)/TightPNO/cc-pVTZ calculations on both sets of geometries. The resulting correction value of 0.5 kcal mol^{-1} was then added to the reference value proposed by Manna and Martin, thus, yielding a value of −7.7 kcal mol^{-1}, which we will use for the extended DC13 set in GMTKN55.

The changes between the old and new references range from being marginal (0 kcal mol^{-1} for the tautomeric 2-pyridone/2-hydroxypyridine system) to being sizeable (5 kcal mol^{-1} for the reaction involving S_8 and 5.6 kcal mol^{-1} for the C_{20} isomers).

As a tenth system, we include an example from Karton and Martin's study on 45 isomerisation energies in C_8H_8 systems.[71] As GMTKN55 itself already contains a large number of isomerisation energies for hydrocarbons, we decided to include the most difficult reaction from the Karton and Martin study in the DC13 set. This is the 41st reaction in Karton and Martin's set with a W1-F12 reaction energy of 109.92 kcal mol^{-1}. The GMTKN24 and GMTKN30 databases contained the O3ADD6 set that considered the addition of ozone to either ethane or ethyne.[72] We decided to leave this set out of GMTKN55, as its analysis is difficult because it involves a mixture of reaction energies, BHs, and association energies. Instead, we took the two reaction energies, re-evaluated them at the W2-F12 level of theory and included them as eleventh and twelfth entries in DC13. The thirteenth and final reaction was taken from a similar test set by Zhao and Truhlar,[18] namely the reaction of hexachlorobenzene with hydrogen chloride to dichlorine and benzene. For this reaction, we present new data at the W1-F12 level of theory.

The reaction energies in the extended DC13 set range from −106.0 kcal mol^{-1} (S_8 reaction) to 152.6 kcal mol^{-1} (reaction 13) with an averaged absolute reaction energy of 54.98 kcal mol^{-1}. A total of 30 single-point calculations need to be carried out for this set.

2.2 Reaction energies for large systems and isomerisation reactions

2.2.1 The new MB16-43 set. In 2010, Korth and Grimme introduced the idea of ''mindless benchmarking'', *i.e.*, reactions of randomly created structures are used to compile a test set without any subconscious bias from the developers.[15] The resulting MB08-165 test set contained 165 randomly created systems that consisted of eight atoms each. MB08-165 turned out to be insightful in the assessment of a method's robustness and, thus, formed a crucial component in the GMTKN24 and GMTKN30 test sets. Systems containing only eight atoms, however, may not be representative for usual applications, and therefore we replace MB08-165 with our newly developed MB16-43 set, which contains the decomposition of 43 AMs—each containing 16 atoms—into element hydrides (for elements of the groups 1, 2, and 13–15) or diatomics (for the other elements). An example reaction is: $2H_6B_2N_2O_2FNaAlCl + 4H_2 \rightarrow 4BH_3 + 2N_2 + 2O_2 + F_2 + 2NaH + 2AlH_3 + Cl_2$. More details on the formation of the AMs and determining the final composition of the MB16-43 set can be found in the ESI.† The 43 reference energies were obtained at the W1-F12 level of theory. The resulting reaction energies range from −362.98 to 1290.73 kcal mol^{-1}, with an averaged absolute reaction energy of 414.73 kcal mol^{-1}. A total of 58 single-point calculations need to be carried out for this set.

2.2.2 The updated DARC set. The DARC set with 14 Diels–Alder reactions has been updated for this work with new reference values. Instead of the original estimated CCSD(T)/CBS,[23] we propose to now use W1-F12 values.[62] The differences can be as large as 1.6 kcal mol^{-1} (reaction of ethene with butadiene) and the MAD for the old values relatively to the new ones is 0.54 kcal mol^{-1}. The reaction energies range from −60.8 kcal mol^{-1} (reaction of ethyne with butadiene) to −14.0 kcal mol^{-1} (reaction of furan with maleine to form the endo-product). The averaged absolute reaction energies for this set is 32.47 kcal mol^{-1}, and a total of 22 single-point calculations need to be carried out.

2.2.3 The updated RSE43 set. The RSE43[75] set contains 43 radical-stabilisation energies (RSEs), with an RSE being the reaction energy for the hydrogen abstraction in hydrocarbons by a methyl radical. The reference values for this set are herein updated from estimated CCSD(T)/CBS to W1-F12 values, with a slight change in the averaged absolute RSE from 7.5 to 7.60 kcal mol^{-1}. While this seems like an overall small change, we note that for individual systems the difference can be as large as 7.7 kcal mol^{-1} (CCl_3CH_2). The new RSEs range from −26.4 kcal mol^{-1} ($H_2NCHCOOH$) to 6.9 kcal mol^{-1} (CCl_3CH_2). A total of 88 single-point calculations need to be carried out for this set.

2.2.4 The updated BSR36 set. We keep the BSR36 set of 36 bond-separation reactions of saturated hydrocarbons, as introduced by Steinmann *et al.*[76] and then later updated with estimated CCSD(T)/CBS reference values by Krieg and Grimme.[77] Contrary to GMTKN30, however, we replace the estimated CCSD(T)/CBS values with rigorously extrapolated ones (DLPNO-CCSD(T)/TightPNO/CBS(aug-cc-pVTZ/aug-cc-pVQZ)). The MAD of the old reference values with respect to the newer ones is 0.5 kcal mol^{-1}, and individual differences get as big as 1.6 kcal mol^{-1} in some cases. The new reaction energies range from 2.24 kcal mol^{-1} (ring system 1) to 48.82 kcal mol^{-1} (cage system 4), with an average of 16.20 kcal mol^{-1}. 38 single-point calculations need to be carried out for this set.

2.2.5 The new CDIE20 set. In 2014, Yu and Karton presented the DIE60 set with 60 double-bond isomerisation reactions in linear, branched and cyclic conjugated dienes. As our new GMTKN55 database already contains various isomerisation reactions, particularly for linear and branched systems, we have compiled a subset of 20 reactions in cyclic dienes from Yu and Karton's data (reactions 20–22, 25, 26, 28, 29, 40, 43–49, 51, 52, 56, 57 and 60 in ref. 78) We dub this subset CDIE20. The geometries and W1-F12 reference values are the same as in the original DIE60 set. These energies range from −5.9 to 8.6 kcal mol^{-1} with an averaged absolute value of 4.06 kcal mol^{-1} for the entire set. 36 single-point calculations need to be carried out for this set.

2.2.6 The updated ISO34 set. The popular ISO34[12] set with 34 organic isomerisation reactions remains a part of GMTKN55, however, we herein present new W1-F12 reference values that replace the original back-corrected experimental ones. The largest absolute difference was observed for reaction 33 (3.08 kcal mol^{-1}), and the MAD for the old values with respect to the new ones is 0.71 kcal mol^{-1}. The new isomerisation energies range from 1.06 kcal mol^{-1} (reaction 4) to 66.06 kcal mol^{-1} (reaction 27), with an average value of 14.57 kcal mol^{-1}. 63 single-point calculations need to be carried out for this set.

2.2.7 The extended ISOL24 set. In 2010, Huenerbein *et al.* presented a set of 24 isomerisation reactions of large molecules of general, ''real-life'' interest, such as a steroid, a dye, or a sugar.[79] For GMTKN30, two of those systems—dubbed systems 1 and 4 in the original publication—were initially discarded because it was feared their size would prevent black-box benchmark studies; the resulting set was named ISOL22.[24] Over the past 6 years, computational architectures have improved, and we no longer see a reason to exclude those two systems. Thus, we reintroduce them herein and rename the set ISOL24. Note, that the reactant and product of reaction 22 are interchanged compared to the original ISOL24 set. Furthermore, we replace the original SCS-MP3[129]/CBS reference values with DLPNO-CCSD(T)/TightPNO/CBS(def2-TZVPP/def2-QZVPP) ones. The MAD for the old reference values of ISOL22 with respect to our new ones is with 1.29 kcal mol^{-1} significant. The isomerisation energies of this updated set range from 0.14 kcal mol^{-1} (system 22) to 71.01 kcal mol^{-1} (system 1), with an average value of 21.92 kcal mol^{-1}. A total of 48 single-point calculations need to be carried out for this set.

2.2.8 The new C60ISO set. In 2017, Sure *et al.* published the C60ISO benchmark[80] consisting of relative energies for the five energetically most stable C_{60} isomers, and for five

additional C_{60} isomers with different geometrical features. The latter are up to 166 kcal mol^{-1} higher in energy compared to the lowest C_{60}-I_h isomer and the average of the respective nine relative energies is 98.25 kcal mol^{-1}. This test set aims to identify methods that are accurate for modelling carbon nano-structures. The reference relative energies were obtained at the DLPNO-CCSD(T)/CBS*//PBE-D3(BJ)/def2-TZVP level of theory.

2.2.9 The new PArel set. In GMTKN30, only absolute PAs were included. Since relative PAs play an important role in chemistry, particularly in biochemical applications, we compiled a new benchmark set of 20 relative energies between 31 tautomers of the protonated forms of the nucleobases adenosine and thymine, a methyl hexofuranoside, *para*-aminobenzoic acid, $H_4P_2O_7$, S_4O_4, $C_2H_2F_4$, and C_2Cl_4. The structures of the most stable protomers were searched with GFN-*x*TB[130] and subsequently optimised with PBEh-3c. The CCSD(T)/CBS(def2-TZVPP,def2-QZVPP) reference values for the relative energies of the respective protomers range from 0.60 kcal mol^{-1} (*para*-aminobenzoic acid) to 11.20 kcal mol^{-1} (protonated S_4O_4), with an average value of 4.63 kcal mol^{-1}.

2.3 Reaction barrier heights

2.3.1 The updated BH76 set. The BH76 subset is a combination of the HTBH38[10] and NHTBH38[11] sets by Truhlar and co-workers, and was also used in GMTKN24 and GMTKN30. It contains 76 forward and reverse BHs for 38 different hydrogen-transfer, heavy-atom transfer, nucleophilic-substitution, unimolecular and association reactions. As outlined above for BH76RC, we herein adopt new W2-F12 reference values. While most reference values only change marginally (the MAD for the old values is only 0.4 kcal mol^{-1}), we also observed outliers. For instance, the forward barrier for the hydrogen-transfer reaction between NH_2 and the ethyl radical is underestimated by 2.3 kcal mol^{-1} by the old reference value. The new averaged absolute BH is 18.61 kcal mol^{-1}.

2.3.2 The updated BHPERI set. The original BHPERI set used in GMTKN24 and GMTKN30 contained 26 BHs of pericyclic reactions that had been obtained with the popular CBS-QB3[52,53] composite approach.[23,81–83] In 2015, Karton and Goerigk derived new W2-F12 and W1-F12 reference data, and they demonstrated that the MAD for CBS-QB3 was with 2.1 kcal mol^{-1} much larger than the typical error for double-hybrid DFT methods.[55] We therefore adopt their new reference values for GMTKN55. They range from 0.5 kcal mol^{-1} to 35.3 kcal mol^{-1}, resulting in an average BH of 20.87 kcal mol^{-1}. 61 single-point calculations need to be carried out for this set.

2.3.3 The new BHDIV10 set. Our newly composed benchmark set BHDIV10 has the scope to also assess BHs of larger and chemically more diverse reactions. It includes 10 chemically interesting transition states for, *e.g.*, the B–N dewar benzene formation, H_2 activation with an FLP model complex, a Si/$BHCl_2$ exchange reaction, C_2H_2 trimerisation to benzene, and the 1,3-silyl shift in allylsilane. The reference values were obtained with W2-F12 or W1-F12 (for BH 3 and 7) based on PBEh-3c geometries, including the continuum solvation model COSMO[131] for the optimisation of the species involved in reactions 5 and 6. They range from 13.64 kcal mol^{-1} for the barrier of the CO_2 activation reaction with an FLP model complex up to 96.17 kcal mol^{-1} for the barrier of the reaction of methane with ethene, with an average of 45.33 kcal mol^{-1}. A total of 20 single-point calculations need to be carried out for this test set.

2.3.4 The new INV24 set. In 2016, Goerigk and Sharma presented the first comprehensive benchmark set for BHs in inversion and racemisation reactions (INV24).[84] It comprises 24 BHs for inversion in triatomic, pyramidal, cyclic, helical and bowl-shaped systems calculated at the W2-F12, W1-F12 or DLPNO-CCSD(T)/TightPNO/CBS(def2-TZVPP/def2-QZVPP) levels of theory ranging from 4.5 to 79.7 kcal mol^{-1}. The average barrier is 31.85 kcal mol^{-1}. 48 single-point calculations need to be carried out for this set.

2.3.5 The new BHROT27 set. Our new BHROT27 set is the first comprehensive benchmark set allowing investigation of barriers for rotation around single bonds. Together with INV24, it is one of two BH test sets for processes that do not involve any bond breaking or formation. The set contains 27 barriers for 14 molecules. All geometries were obtained at the TPSS-D3(BJ)/def2-TZVP level of theory. The BHs are based on the W2-F12 and W1-F12 levels of theory and they range from 1.01 to 17.24 kcal mol^{-1} with an average value of 6.27 kcal mol^{-1}. 40 single-point calculations need to be carried out for this set.

2.3.6 The new PX13 set. In 2012, Radom and co-workers studied the complexation energies of 13 water, ammonia and hydrogen-fluoride clusters of varying sizes, as well as the BHs for proton transfer in them.[85] Reference values for all energies were based on the W1-F12 level of theory. The reference points for all 13 barriers were the separated monomers. Herein we adopt a modified version of this set, where the BHs have the clusters in their minimum-energy configuration as reference points. Those new barriers were calculated from the difference between the complexation energies and barriers presented in the original paper. The new reference energies range from 14.6 kcal mol^{-1} ($[HF]_5$ cluster) to 59.3 kcal mol^{-1} ($[NH_3]_2$ cluster) with an average value of 33.36 kcal mol^{-1}. 29 single-point calculations need to be carried out for this set.

2.3.7 The new WCPT18 set. In 2012, Karton *et al.*[86] presented 27 accurate BHs (WCPT27) for nine proton-transfer tautomerisation reactions including carbonyls, imines, propene, and thiocarbonyls, either uncatalysed, or catalysed by one or two water molecules. The latter case is omitted in GMTKN55 since the formation of a hydrogen bond between the water molecules is already assessed separately in the WATER27 set and hence only 18 BHs of the original set were considered in GMTKN55 (denoted WCPT18). The BHs are calculated with respect to the seperated species. The W2.2//B3LYP/A'VTZ reference values range from 2.68 kcal mol^{-1} (catalysed reaction 2) to 81.24 kcal mol^{-1} (uncatalysed reaction 6), with an average value of 34.99 kcal mol^{-1}. Note that the WCPT27 article reported those reference values for the reverse reactions.[86] 28 single-point calculations need to be carried out for this set.

2.4 Intermolecular noncovalent interactions

2.4.1 The new RG18 set. The new RG18 set of interaction energies in rare-gas complexes replaces the RG6 set in GMTKN30. Besides the interaction energies of the Ne, Ar and Kr dimers and trimers, the tetramers of Ne and Ar, the hexamer of Ne, and complexes with HF, ethyne, ethane and benzene are included. The high-level reference interaction energies were calculated with counterpoise (CP) corrected CCSD(T)/CBS(aug-cc-pVTZ/aug-cc-pVQZ) for all dimer interaction energies and CCSD(T)/CBS(aug-cc-pwCVTZ/aug-cc-pwCVQZ) (all electrons correlated) for trimer, tetramer, and hexamer interaction energies (CP correction not applicable). Ne_2, Ar_2 and Kr_2 geometries were taken from the RG6 set, monomer geometries of the complexes were taken from the W4-11 set, all other geometries were optimised at the TPSS-D3(BJ)/def2-TZVP level of theory. They range from 0.08 kcal mol^{-1} (Ne_2) to 1.51 kcal mol^{-1} (Ar_4), with an average value of 0.58 kcal mol^{-1}, which is the smallest among all sets in GMTKN55. A total of 25 single-point calculations need to be carried out for this set.

2.4.2 The updated ADIM6 set. The ADIM6 test set in GMTKN30 contains binding energies of six alkane dimers ranging from the ethane to the *n*-heptane dimers. For GMTKN55 we replaced the original estimated CCSD(T)/CBS[33] with new W1-F12 values. The MAD for the original values is only 0.05 kcal mol^{-1}, indicating that in this case the original reference values were already of sufficient accuracy. The binding energies in ADIM6 range from 1.34 to 5.55 kcal mol^{-1} with an average value of 3.36 kcal mol^{-1}. 12 single-point calculations need to be carried out for this set.

2.4.3 The updated S22 set. Since its introduction in 2006,[13] the reference values for the famous S22 set for interaction energies in noncovalently bound dimers have been updated several times. In GMTKN30, we used the estimated CCSD(T)/CBS values calculated by Sherrill and co-workers in 2010, also often referred to as the S22A set.[132] Herein, we update those numbers and adopt their revised reference values (S22B).[87] The binding energies range from 0.53 kcal mol^{-1} (methane dimer) to 18.75 kcal mol^{-1} (formic-acid dimer) and have an average value of 7.30 kcal mol^{-1}. 57 single-point calculations have to be carried out for this set.

2.4.4 The new S66 set. In 2011, Hobza and co-workers presented the S66 set as a larger "cousin" of S22 with 66 NCI energies obtained at the estimated CCSD(T)/CBS level of theory.[88] In 2011, some of us thoroughly tested DFT methods for S66 and concluded that this set ideally complemented the data presented in the NCI section of GMTKN30.[42] Therefore, we decided to make this set a part of GMTKN55. The interaction energies range from 2.82 kcal mol^{-1} (benzene–ethene dimer) to 19.49 kcal mol^{-1} (acetamide–uracil dimer) with an average value of 5.47 kcal mol^{-1}. 198 single-point calculations need to be carried out for this set.

2.4.5 The updated HEAVY28 set. Herein, we present new reference values for the HEAVY28[33] test set with 28 binding energies in non-covalently bound heavy-element-hydride dimers. While the old values were based on CP-corrected, estimated CCSD(T)/CBS data, we herein used the same strategy as for HEAVYSB11 to obtain CCSD(T)/CBS numbers, with the only difference that also a CP correction was employed. The changes, however, turned out to be marginal and the MAD for the old references with respect to the new ones is only 0.08 kcal mol^{-1}. The interaction energies in HEAVY28 range from 0.52 kcal mol^{-1} [$(TeH_2)_2$] to 3.35 kcal mol^{-1} (TeH_2-NH_3), with an averaged interaction energy of 1.24 kcal mol^{-1}. A total of 38 single-point calculations need to be conducted for this set.

2.4.6 The updated WATER27 set. The first reference values for the WATER27 set for binding energies of neutral, positively, and negatively charged water clusters were presented by Bryantsev *et al.*[89] These were estimated CCSD(T)/CBS values for all clusters except for the four largest $(H_2O)_{20}$ clusters, for which only MP2/CBS data were presented. Herein, we replace all 27 values by more accurate estimated CCSD(T)-F12/CBS numbers presented by Martin and co-workers in 2017.[90] The MAD of the old with respect to the new values is 1.00 kcal mol^{-1}, which is a sizeable difference. The largest deviations are seen for the $(H_2O)_{20}$ clusters (changes between 2.23 to 8.82 kcal mol^{-1}), as well as for the reaction from $(H_2O)_8$ (in S_4 symmetry) to $(H_3O)^+(H_2O)_6(OH)^-$ (change of 1.1 kcal mol^{-1}). Note that a reference value for the latter was not directly given in ref. 90, but it was obtained from the published total energies.[133] The smallest (revised) interaction energy in WATER27 is that for the water dimer (4.97 kcal mol^{-1}) and the largest for the $(H_2O)_{20}$ cluster in its edge-sharing form (209.08 kcal mol^{-1}). The average interaction energy for WATER27 is 81.14 kcal mol^{-1}. This set involves 30 single-point calculations.

2.4.7 The new CARBHB12 set. The CARBHB12 set is a newly compiled benchmark set for interaction energies of 12 hydrogen-bonded complexes between singlet carbene and its $CClCH_3$, SiH_2, and 1,3-dimethylimidazol-2-ylidene (NHC) analogues with H_2O, NH_3, and HCl, respectively. This special but important class of intermolecular NCIs was missing in GMTKN30. The high-level reference interaction energies were obtained with the W2-F12 protocol (without CP and geometry relaxation) based on PBEh-3c optimised geometries. The structure of the HCl–NHC complex is a result of an intermediate step in the optimisation because the minimum corresponded to a full HCl–NHC proton transfer. The reference values range from 1.21 kcal mol^{-1} (NH_3–SiH_2) to 16.30 kcal mol^{-1} (HCl–NHC), with an average value of 6.04 kcal mol^{-1}. 36 single-point calculations need to be carried out for this set.

2.4.8 The new PNICO23 set. In 2015, Setiawan *et al.* studied pnicogen-pnicogen interactions in 36 noncovalently bound homo and hetero dimers of the form R_3E-ER_3 and R_3E-$E'R'_3$ with E, E′ = N, P, or As, and R, R′ = H, BH_2, CH_3,CN, NH_2, F, Cl, Br, and I.[91] Geometries and binding energies were presented and discussed at the ωB97X-D/aug-cc-pVTZ level of theory. While the nature of such interactions is worthwhile to be explored, accurate reference numbers need to be obtained for those systems. For that purpose we discarded all dimers containing As, Br, or I and took the geometries of the remaining 23 dimers from the 2015 study and obtained binding energies at the W2-F12 level of theory. For three systems that level was not feasible and we employed W1-F12 instead (H_3N–$P(CN)_3$, H_3N–$PH(CN)_2$, $(PHFCH_3)_2$). The interaction energies range from

0.64 kcal mol^{-1} ($(PF_3)_2$) to 10.97 kcal mol^{-1}, with an average value of 4.27 kcal mol^{-1}. The resulting set is dubbed PNICO23, for which a total of 69 single-point calculations need to be carried out.

2.4.9 The new HAL59 set. The new HAL59 set for binding energies in halogenated noncovalently bound dimers is a combination of parts of two existing sets in the literature: XB51[92] and X40.[93] We took the geometries and estimated CCSD(T)/CBS reference numbers for XB51 published by Kozuch and Martin in 2013, but excluded the Pd and Li species, which resulted in 45 systems. We then complemented these dimers with halogenated aromatic systems from Hobza's X40 set; these were systems 11, 12, and 19–30 with references obtained at the estimated CCSD(T)/CBS level. We have, thus, created one balanced set that explores 59 interaction energies, which range from 0.29 kcal mol^{-1} (FI-FCCH) to 20.34 kcal mol^{-1} (FI-pyridine), with an average value of 4.59 kcal mol^{-1}. A total of 105 single-point calculations need to be carried out for this set.

2.4.10 The new AHB21 set. With the exception of some systems in the WATER27 set, the GMTKN24 and GMTKN30 databases only explored NCIs in neutral species. Herein, we introduce a more comprehensive description of these interactions by introducing three test sets that also contain charged species. The first of these sets is AHB21 introduced by Jansen, Herbert and co-workers in 2015.[94] It contains the interaction energies between anionic and neutral monomers in 21 noncovalently bound dimers. We took the geometries and CCSD(T)-(F12)/CBS reference values from the original study without any modification. The interaction energies range from −7.97 kcal mol^{-1} (N_3^--NH_3) to −65.68 kcal mol^{-1} (F^--HF), with an average absolute interaction energy of 22.49 kcal mol^{-1}. A total of 63 single-point calculations need to be carried out for this set.

2.4.11 The new CHB6 set. CHB6 has also been taken from ref. 94. It contains six cationic–neutral dimers with interaction energies obtained at the CCSD(T)/CBS level of theory, and estimated CCSD(T)/CBS level for three alkali-benzene complexes. The interaction energies range from −17.83 kcal mol^{-1} (K^+-H_2O) to −39.09 kcal mol^{-1} (Li^+-C_6H_6), with an average absolute interaction energy of 26.79 kcal mol^{-1}. A total of 18 single-point calculations need to be carried out for this set.

2.4.12 The new IL16 set. Also IL16 is taken from the same study as AHB21 and CHB6. It contains 16 model dimers representative of typical cation–anion pairs in ionic liquids, hence the acronym "IL". In fact, those geometries were originally published by Izgorodina and co-workers as part of their larger IL-2013 set,[134] however for IL16 a different estimated CCSD(T)/CBS scheme was applied compared to the earlier study in 2013. The resulting interaction energies range from −87.42 kcal mol^{-1} to −120.80 kcal mol^{-1}, with an averaged absolute interaction energy of 109.04 kcal mol^{-1}. A total of 48 single-point calculations need to be carried out for this set.

2.5 Intramolecular noncovalent interactions

2.5.1 The updated IDISP set. The IDISP[12,23,24,95,96] set for intramolecular interactions in large hydrocarbon molecules was already used in GMTKN24 and GMTKN30. The set includes the dimerisation reaction of anthracene, the folding of two longer alkane chains (tetradecane and docosane), the isomerisation of octane and undecane, and the formation of [2.2]paracyclophane. The new DLPNO-CCSD(T)/($T_{CutPairs}$ = 10^{-5} E_h)/CBS(aug-cc-pVTZ/aug-cc-pVQZ) reference values range from −1.21 kcal mol^{-1} to −60.28 kcal mol^{-1}, with an averaged absolute relative energy of 14.22 kcal mol^{-1}. They differ from the GMTKN30 reference values by 1.07 kcal mol^{-1} on average with a significant maximum deviation of 2.36 kcal for the folding of docosane. 13 single-point calculations need to be carried out for this set.

2.5.2 The new ICONF set. To also assess intramolecular interactions in inorganic molecules, we compiled a new set of inorganic conformers denoted ICONF. It includes conformers of N_3H_5, N_4H_6, $N_3P_3H_{12}$, Si_5H_{12}, Si_6H_{12}, P_7H_7, S_4O_4, S_8, $H_2S_2O_7$, and $H_4P_2O_7$. The ICONF set consists of 17 relative conformer energies obtained with the W1-F12 protocol and TPSS-D3(BJ)/def2-TZVP optimised geometries in the range of 0.10 kcal mol^{-1} to 12.16 kcal mol^{-1}, with an average value of 3.27 kcal mol^{-1}. 27 single-point calculations need to be carried out for this set.

2.5.3 The unchanged ACONF set. The ACONF set for alkane conformers was also used in GMTKN24 and GMTKN30. The set includes 15 relative energies of *n*-butane, *n*-pentane and *n*-hexane conformers. It was originally published by Gruzman *et al.*,[97] who presented accurate W1h-val reference values that range from 0.60 kcal mol^{-1} (between two butane conformers) to 4.93 kcal mol^{-1} (between two hexane conformers), with an average value of 1.83 kcal mol^{-1}. A total of 18 single-point calculations need to be carried out for this set.

2.5.4 The new AMINO20x4 set. In 2014, Popelier, Jensen, and co-workers published the YMPJ database containing conformers of the 20 proteogenic amino acids, each capped with peptide bonds at the N and C termini and with neutral side chains.[135] Martin and co-workers then took those structures, reoptimised them and presented highly reliable estimated CCSD(T)-F12/CBS data.[98] With more than 500 conformers, using the complete set would have been overbearing compared to the others. To maintain a balance between the various sets, we therefore only adopt the five energetically lowest-lying conformers for each amino acids, thus, resulting in 100 single-point calculations and 80 relative energies. We dub this subset of Martin and co-workers' database AMINO20x4. The relative conformational energies range from 0.06 kcal mol^{-1} (between two lysine conformers) to 7.37 kcal mol^{-1} (between two histidine conformers), with an average relative energy of 2.44 kcal mol^{-1}.

2.5.5 The extended PCONF21 set. The PCONF21 set is an extension of the PCONF set[100] of ten relative energies between eleven phenylalanyl-glycyl-glycine tripeptide conformers, which was already included in GMTKN24 and GMTNK30. The additional data points are eight relative energies of ten tetrapeptide conformers taken from the TPCONF benchmark set of Goerigk *et al.*[101] They have the form ACE–ALA–X–ALA–NME, where ALA is alanine, X is either glycine (GLY) or serine (SER), ACE is an acetyl group, and NME a methylamide group. The tetrapeptides reflect biologically relevant backbone conformations, namely those of parallel and anti-parallel β-sheets, right-handed and left-handed α-helices, and the polyproline-II helix. For both the

tri- and tetrapeptides, estimated CCSD(T)-(F12)/CBS were originally used. We replaced those with DLPNO-CCSD(T)/TightPNO/CBS(aug-cc-pVTZ/aug-cc-pVQZ) numbers while keeping the original geometries. The resulting relative energies range from 0.02 kcal mol^{-1} (conformer 444 of the original PCONF set) up to 2.74 kcal mol^{-1} (the SER-containing tetrapeptide in its β-sheet form), with an average value of 1.62 kcal mol^{-1}. On average, they differ from the old PCONF and TPCONF reference values by 0.26 kcal mol^{-1} (maximum deviation 0.62 kcal mol^{-1}) and 0.30 kcal mol^{-1} (maximum deviation 0.52 kcal mol^{-1}), respectively, resulting also in a new energy ordering of the some tripeptide conformers.

2.5.6 The new MCONF set. In 2013, Martin and co-workers[102] published a benchmark set of relative energies for 52 conformers of melatonin. These are dominated by quadrupole–dipole, aromatic–amide interactions, and weak intramolecular hydrogen bonds, which are important in many biomolecules. The SCS-MP2/cc-pVTZ structures of the original publication were adopted, but the reference values ((CCSD(T)/cc-pVTZ(p on H) − MP2/cc-pVTZ(p on H) + MP2-F12/cc-pVTZ-F12)[136]) were updated. The new reference values obtained at the DLPNO-CCSD(T)/TightPNO/CBS(aug-cc-pVTZ/aug-cc-pVQZ) level range from 0.39 kcal mol^{-1} up to 8.75 kcal mol^{-1}, with an average value of 4.97 kcal mol^{-1}. They differ from the old reference values by 0.25 kcal mol^{-1} on average with a maximum deviation of 0.40 kcal mol^{-1}.

2.5.7 The updated SCONF set. The SCONF set of 14 relative energies of 3,6-anhydro-4-*O*-methyl-D-galactitol (AnGol15) conformers and three relative energies of β-D-glucopyranose conformers[103] also forms part of GMTKN55, however, we updated the estimated CCSD(T)/CBS values introduced together with GMTKN24[23] with new DLPNO-CCSD(T)/TightPNO/CBS(aug-cc-pVTZ/aug-cc-pVQZ) ones. The MAD of the old values with respect to the new ones is with 0.32 kcal mol^{-1} significant for a conformational-energy test set. The updated SCONF has an average relative energy of 4.60 kcal mol^{-1}; 19 single-point calculations need to be carried out.

2.5.8 The new UPU23 set. In 2015, Kruse *et al.* created a benchmark set of 46 uracil dinucleotides denoted as UpU46, which represents all backbone conformational families of RNA.[104] It is an important test for methods used to study the conformational ranking of nucleic acids and biomolecules in general. Since the molecules of the UpU46 set are relatively large, we decided to include only 23 relative energies between 24 randomly chosen uracil dinucleotide structures from the original benchmark in the GMTNK55 database in order to keep the computational effort reasonable. Hence, we name this set UPU23 in GMTKN55. We adopted the original reference values, which were obtained at the DLPNO-CCSD(T)/CBS*//TPSS-D3(BJ)/def2-TZVP(COSMO) level of theory. They range from 0.57 kcal mol^{-1} to 14.41 kcal mol^{-1}, with an average value of 5.72 kcal mol^{-1}.

2.5.9 The new BUT14DIOL set. In 2014, Kozuch, Bachrach, and Martin presented a benchmark set of 65 conformers of butane-1,4-diol, with the majority of them having strong intramolecular hydrogen bonds.[105] Reference energies were obtained at the CCSD(T)-F12b/cc-pVTZ-F12 level of theory. We adopted these structures, but updated the reference values with W1-F12 data that range from 0.15 to 4.70 kcal mol^{-1}, with an average value of 2.80 kcal mol^{-1}. In GMTKN55, we name this set BUT14DIOL.

3 Computational details

3.1 Selected density functional approximations

We selected the DFAs for the subsequent benchmark study on GMTKN55 based on the following reasoning. DFAs that had performed well for GMTKN30—see, for instance, those mentioned in the introduction—were included again in this study. The same was also true for methods that experience high popularity, such as PBE,[137] BP86,[138–140] BLYP,[138,141,142] B3LYP, PBE0,[119,120] BHLYP,[143] or TPSSh.[144] On the other hand, methods with low accuracy for GMTKN30[7,29] were excluded from this study, as it was unlikely that their performance relative to other DFAs would significantly improve for GMTKN55. Our previous advice on those functionals will, thus, remain valid; examples are LDA functionals,[145–147] PBEsol,[148] SSB,[149] M06HF,[150] "non-empirical" and "one-parameter" double hybrids,[29] or the range-separated hybrids CAM-B3LYP[151] and LC-ωPBE.[152] The well-performing range-separated hybrid ωB97X-D[37] with a DFT-D2[41] dispersion-correction term was replaced by two newer versions that had been reparametrised with the DFT-D3[33] correction with zero-damping (DFT-D3(0)) or the nonlocal van der Waals (vdW) VV10[153] kernel; these versions are called ωB97X-D3[154] and ωB97X-V,[155] respectively.

Additional DFAs are either those that had not been published by the time of the GMTKN30 study—for instance, SCAN[156] or Minnesota-type DFAs published since 2011—or older approaches available in Gaussian or ORCA that had never been tested for GMTKN30 before, such as the HCTH family,[157] DFAs based on PBE-hole exchange,[158] or M08HX.[159] In 2011, we reported an incompatibility between dispersion corrections and the VSXC[160] meta-GGA functional as well as severe self-consistent-field (SCF) convergence problems for most systems in GMTKN30.[7] Unusual problems for NCIs have already been reported for SOGGA11,[161] and we additionally observed convergence problems for GMTKN55, which made its routine application unfeasible. The same can be reported for the GLYP[141,142,162] GGA functional, which was consequently also excluded from this study.

In this article, we will, thus, investigate the 83 dispersion-corrected DFAs listed in Table 2: 19 GGAs/NGAs, 9 meta-GGAs/NGAs, 48 hybrids, and 7 double hybrids. Note that in some cases different versions of a dispersion correction were applied to the same underlying exchange–correlation functional approximations: ωB97X-D3 and ωB97X-V, VV10 and rPW86PBE-D3(BJ), and B3LYP-D3(BJ) and B3LYP-NL.[46] We therefore assess 80 unique exchange–correlation DFAs. In addition, we also analysed the results for dispersion-uncorrected DFAs and for DFAs corrected with the D3(0) variant, even though D3(BJ) should be the preferred one in most cases. This results in a total of 217 DFA variations, whose results are all presented on the GMTKN55 website.[112]

A recent related study promoted the study of 200 DFAs, however, that number also includes a mixture of dispersion-uncorrected and various dispersion-corrected versions of the same underlying DFA.[26] When correcting for this, the 200 DFAs break down to 91 unique exchange–correlation DFAs—10 of which were ruled out from our study, as mentioned above—as well as the HF wave function method.

3.2 Dispersion corrections

Contrary to other related studies, all 83 methods were consistently assessed with a dispersion-correction term. For the majority of functionals, this was achieved with the DFT-D3(BJ) dispersion correction with Becke–Johnson damping,[203,204] which has become the default version due to its physically more correct behaviour for short- and medium-range distances between two interacting non-covalently bound fragments. In some cases—for instance, the majority of older Minnesota functionals—this damping function turned out to be incompatible due to short-range double-counting effects. In those instances, the older zero-damping version was applied. Three functionals were tested with a non-local vdW kernel (VV10, B3LYP-NL, and ωB97X-V), while the APFD functional was developed together with its own correction. The recommended dispersion-correction for each DFA is listed in Table 2 alongside the reference that presented the relevant dispersion-correction parameters for the first time.

For 35 DFAs, we had to determine DFT-D3 parameters for the first time. While we did this for both versions, we still recommend DFT-D3(BJ) in most cases, except for PW91P86,

Table 2 Overview of the 83 dispersion-corrected DFAs tested in this study

Name	Dispersion correction[a]	Program[b]	Name	Dispersion correction[a]	Program[b]
GGA			PW1PW[119,120,163]	D3(0)[c]	ORCA
PBE[137]	D3(BJ)[44]	TM	MPW1KCIS[11]	D3(BJ)[c]	G09
PBEhPBE[158]	D3(BJ)[c]	G09	MPWKCIS1K[11]	D3(BJ)[c]	G09
revPBE[40]	D3(BJ)[44]	ORCA	PBE0[119,120]	D3(BJ)[44]	TM
RPBE[164]	D3(BJ)[c]	ORCA	PBEh1PBE[119,120,158]	D3(BJ)[c]	G09
PW91[163]	D3(BJ)[165]	G09	PBE1KCIS[17]	D3(BJ)[c]	G09
BLYP[138,141,142]	D3(BJ)[44]	TM	X3LYP[166]	D3(BJ)[c]	ORCA
BP86[138–140]	D3(BJ)[44]	TM	O3LYP[167]	D3(BJ)[c]	ORCA
BPBE[137,138]	D3(BJ)[7]	ORCA	B97-1[168]	D3(BJ)[c]	G09
OPBE[137,169]	D3(BJ)[7]	ORCA	B97-2[170]	D3(BJ)[c]	G09
OLYP[141,142,169]	D3(BJ)[7]	ORCA	B98[171]	D3(BJ)[c]	G09
XLYP[141,142,166]	D3(BJ)[c]	ORCA	HISS[172]	D3(BJ)[c]	G09
mPWLYP[141,142,173]	D3(BJ)[7]	ORCA	HSE03[174]	D3(BJ)[c]	G09
PW91P86[139,140,163]	D3(0)[c]	ORCA	HSE06[174,175]	D3(BJ)[176]	G09
mPWPW91[163,173]	D3(BJ)[c]	ORCA	TPSSh[144]	D3(BJ)[7]	ORCA
rPW86PBE[137,177]	D3(BJ)[44]	ORCA	revTPSSh[144,178]	D3(BJ)[c]	G09
B97-D3(BJ)[41]	D3(BJ)[44]	TM	TPSS0[179]	D3(BJ)[44]	ORCA
HCTH/407[157]	D3(BJ)[c]	G09	revTPSS0[178,179]	D3(BJ)[c]	G09
N12[27]	D3(0)[43]	G09	TPSS1KCIS[180]	D3(BJ)[c]	G09
VV10[153]	VV10[153]	ORCA	BMK[181]	D3(BJ)[7]	G09
			τHCTHhyb[182]	D3(BJ)[c]	G09
Meta-GGA			M05[183]	D3(0)[7]	G09
PKZB[184]	D3(0)[c]	G09	M052X[35]	D3(0)[7]	G09
TPSS[113]	D3(BJ)[44]	ORCA	M06[34]	D3(0)[7]	TM
revTPSS[178]	D3(BJ)[c]	G09	M062X[34]	D3(0)[7]	TM
SCAN[156]	D3(BJ)[185]	TM[d]	M08HX[159]	D3(0)[c]	G16
τHCTH[182]	D3(BJ)[c]	G09	M11[186]	D3(BJ)[43]	G09
M06L[187]	D3(0)[7]	TM	SOGGA11X[188]	D3(BJ)[43]	G09
M11L[189]	D3(0)[43]	G09	N12SX[190]	D3(BJ)[43]	G09
MN12L[191]	D3(BJ)[43]	G09	MN12SX[190]	D3(BJ)[43]	G09
MN15L[21]	D3(0)[c]	G16	MN15[22]	D3(BJ)[c]	G16
			LC-ωhPBE[192]	D3(BJ)[c]	G16
Hybrid			ωB97X-D3[154]	D3(0)[154]	ORCA
B3LYP[38,39]	D3(BJ)[44]/VV10[e 46]	TM/ORCA	ωB97X-V[155]	VV10[155]	ORCA[d]
B3PW91[38]	D3(BJ)[7]	ORCA	APFD[193]	APFD[193]	G09
B3P86[38,138–140]	D3(BJ)[c]	ORCA			
BHLYP[143]	D3(BJ)[7]	TM	Double hybrid		
B1P86[119,120,138–140]	D3(BJ)[c]	ORCA	B2PLYP[28]	D3(BJ)[7]	ORCA
B1LYP[119,120,138,141,142]	D3(BJ)[c]	ORCA	B2GPPLYP[56]	D3(BJ)[7]	ORCA
B1B95[194]	D3(BJ)[7]	ORCA	MPW2PLYP[195]	D3(BJ)[c]	ORCA
MPW1B95[196]	D3(BJ)[7]	ORCA	PWPB95[24]	D3(BJ)[7]	ORCA
PW6B95[36]	D3(BJ)[44]	TM	DSD-BLYP[30]	D3(BJ)[7]	ORCA
MPWB1K[196]	D3(BJ)[7]	ORCA	DSD-PBEP86[31]	D3(BJ)[31]	ORCA
mPW1LYP[119,120,141,142,173]	D3(0)[c]	ORCA	DSD-PBEB95[32]	D3(BJ)[32]	ORCA
mPW1PW91[119,120,173]	D3(BJ)[c]	ORCA			

[a] Type of dispersion correction. The cited reference presented the respective parameters for the first time. D3(BJ): DFT-D3 with Becke–Johnson damping.[33,44] D3(0): DFT-D3 with zero-damping.[33] VV10: nonlocal van der Waals kernel, as presented for the VV10 functional.[153] APFD: spherical-atom dispersion term.[193] [b] ORCA: ORCA 3.0.3 or ORCA 4.0.0.[197,198] TM: TURBOMOLE 7.1.1.[199,200] G09: GAUSSIAN09 Revision D.01 or E.02.[201] G16: GAUSSIAN16 Revision A.03.[202] [c] Determined for this work. [d] Local version. [e] This DFA is called "B3LYP-NL".[46]

PKZB, MN15L, mPW1LYP, PW1PW, and M08HX, for which DFT-D3(0) is recommended. In those cases, no DFT-D3(BJ) parametrisation was achievable because of over-binding tendencies of the functionals and partial coverage of the dispersion interactions due to artificially built-in mid-range correlation effects. In the case of MN15L and MN15, the determined DFT-D3 damping parameters had the smallest influence on their overall performance; whether this is a positive or negative aspect will discussed below in the results section.

For DFT-D3(BJ), three parameters were fitted in a least-squares sense, while for DFT-D3(0) only two parameters had to be determined. The fit was carried out on a training set that contained the S66x8,[88] S22x5,[205] and NCIBLIND[206] sets, which consider noncovalently bound dimers in their equilibrium and non-equilibrium geometries. A total of 718 data points were included in this set which are mostly not overlapping with GMTKN55 sets. This is contrary to the first DFT-D3 parametrisations that relied on sets that were also part of GMTKN30;[33] for a third suggested training set for DFT-D3, see ref. 207. The D3(BJ) and D3(0) parameters of all DFAs used in this work are listed in Tables S3 and S4 (ESI†).

3.3 Technical details

Table 2 lists the programs that were used to assess the DFAs. The entire assessment of GMTKN55 and the parametrisations of the DFT-D3 corrections were carried out with the Ahlrichs-type split-valence quadruple-ζ Gaussian atomic-orbital (AO) basis set def2-QZVP. This basis set was chosen not only because it makes results comparable to previous GMTKN30 studies, but also because this basis set provides results close to the CBS limit for most properties; it, thus, provides a picture of a DFA's "true" performance without relying on error-compensation effects, as discussed elsewhere.[48] As in the case of the GMTKN24 and GMTKN30 studies,[7,23,24] the def2-QZVP set for oxygen was augmented with Dunning's diffuse s and p functions for WATER27.[118]

Diffuse s and p functions were applied to all non-hydrogen atoms in the G21EA, AHB21 and IL16 sets; diffuse s functions were applied to H. The resulting basis set is henceforth called aug'-def2-QZVP. Core-electrons of heavy elements in some systems of HEAVY28, HEAVYSB11, and HAL59 were replaced with the def2-ECP effective-core-potentials.[114] Note that herein we do not carry out a basis-set dependence study for smaller AO basis sets. The expected basis-set dependence for conventional and double-hybrid DFAs has already been established for GMTKN30,[7,26] and repeating this analysis would not provide any new information.

All TURBOMOLE DFT calculations were carried out with the module RIDFT[208] by employing the resolution-of-the-identity method to the Coulomb integrals (RI-J); auxiliary basis functions were taken from the TURBOMOLE library.[209,210] TURBOMOLE's multi-grid option "m4" was applied for the numerical integration of the exchange–correlation potentials.[209] Note that other studies, including GMTKN30 studies,[7] extensively elaborated on the grid-dependence of some DFAs.[25,211–214] Herein, we use quadrature grids that are feasible for routine applications to provide more viable guidelines. The SCF convergence criterion was set to 10^{-7} E_h. The SCAN functional may show slow convergence for the radial quadrature grid, and therefore TURBOMOLE's option "radsize" = 40 was used together with a grid size of 4.

All (meta-)GGA functionals in ORCA were also treated with the RI-J approximation, whereas hybrids and the hybrid parts of double hybrids were treated with the chain-of-sphere approximation[215] to evaluate exchange integrals (RIJCOSX). ORCA's default settings were used in the latter case. The second-order perturbative treatment for the double hybrids was also carried out with the RI approximation and the appropriate auxiliary basis functions.[216] Contrary to previous studies, we also employed the frozen-core approximation for double hybrids to prevent basis-set superposition errors in the treatment of core–core electron correlation.[217] All SCF calculations in ORCA were carried out with ORCA's quadrature grid "3", followed by a non-iterative step with the larger grid "4". These options are similar to TURBOMOLE's multi-grid strategy. The SCF convergence criterion was set to ORCA's "tightscf" option, which is similar to the option chosen for TURBOMOLE and GAUSSIAN. The nonlocal correction in VV10, B3LYP-NL and ωB97X-V was employed post-SCF and with ORCA's van der Waals grid "vdwgrid2".

All GAUSSIAN calculations were carried out with the standard quadrature "fine grid". The SCF convergence criterion was set to 10^{-7} E_h.

MOLPRO2015.1[218,219] was used to obtain reference values for the various Weizmann composite schemes mentioned in Section 2. Some W1-F12 calculations were also carried out with TURBOMOLE's CCSDF12[220] module for computational efficiency reasons: the hexane and heptane dimers in ADIM6, the hexachlorobenzene reaction in DC13, and all calculations for MB16-43. The same TURBOMOLE module was also used for CCSD(T) calculations to obtain reference values for the HEAVYSB11 and HEAVY28 sets as well as for the tetramethyl-ethene reaction in DC13 (see Section 2 for details).

All DLPNO-CCSD(T)[109] calculations were carried out with the sparse-maps version[221] implemented in ORCA 4.0.0. Except for the CBS* calculations (see below) and the large systems in IDISP, the "TightPNO" setup[128] was used, which corresponds to the following threshold values: $T_{CutPairs} = 10^{-5}$ E_h, $T_{CutPNO} = 10^{-7}$, $T_{CutDO} = 5 \times 10^{-3}$, and the use of the full MP2 guess. For the extrapolation to the CBS limit, the original scheme with an exponent of 1.63 proposed by Halkier and Helgaker[121,122] was employed for HF, while an exponent of 3 was used for the correlation energy. Extrapolations were based on either the def2-TZVPP and def2-QZVPP, or aug-cc-pVTZ and aug-cc-pVQZ basis sets (as indicated above). In the DLPNO-CCSD(T)/CBS* calculations, only the electron pair cut-off was tightened to $T_{CutPairs} = 10^{-5}$ E_h while all other threshold values were kept at their respective defaults. The CBS* basis set extrapolation is based on a multiplicative extrapolation scheme and the def2-TZVP basis set together with the MP2/CBS(cc-pVDZ/cc-pVTZ) energy as well as scaling factor to account for missing diffuse functions (for details, see ref. 104). This protocol was developed specifically for accurate reference calculation of larger systems

where a DLPNO-CCSD(T)/TightPNO/CBS calculation is not computationally feasible. The resulting uncertainty of the reference values is, however, slightly larger than for the previous setup.

4 How to analyse GMTKN55

As is common practice, the statistical data that we calculate for each benchmark set and DFA comprise mean deviations (MDs), MADs, root-mean-square deviations (RMSDs), and deviation spans. To be consistent with our previous GMTKN30 studies, we will base the analysis in Section 5 primarily on MADs to determine the best and worst DFAs for each test set. While this analysis could also be carried out with RMSDs, we report those values only in the ESI† alongside the other aforementioned statistical values (Section S7). In fact, the RMSDs reported therein would give a similar overall picture of DFA performance.

The average relative absolute energies $\left(\overline{|\Delta E|}\right)$ shown in Table 1 can be as small as 0.58 kcal mol^{-1} (RG18 set) and as large as 414.43 kcal mol^{-1} (MB16-43) or 654.26 kcal mol^{-1} (DIPCS10). Consequently, MADs for a benchmark set with a large $\overline{|\Delta E|}$ are expected to be larger than for a set with a relatively small $\overline{|\Delta E|}$. For instance, MADs for MB16-43 usually exceed 15 kcal mol^{-1} for most dispersion-corrected hybrid DFAs. When considering the magnitude of the reaction energies—the largest reaction energy is 1290.74 kcal mol^{-1}—such seemingly large MADs are appropriate. To better compare different benchmark sets with each other, we initially tested two strategies. Firstly, we calculated the percentage deviation for each reaction to obtain mean and mean absolute percentage deviations (MPDs and MAPDs) over a benchmark set. While such an analysis had turned out to be very useful in the past for detecting the underbinding tendency of Minnesota functionals for the NCIs in the S66x8 set,[42,43] MAPDs for other sets in GMTKN55 turned out to be less robust and very sensitive to outliers. One example is the energy difference between the two lowest-lying tripeptide conformers in PCONF, which is only 0.02 kcal mol^{-1} according to the reference level of theory. Even the best DFA for this set—DSD-BLYP-D3(BJ) with an MAD of only 0.23 kcal mol^{-1}—predicts an energy difference of 0.16 kcal mol^{-1}. The percentage deviation is, thus, 702.5%. Even though this is the only outlier, this value distorts the MAPD to 56.8%, which does not seem to represent the overall excellent performance of this DFA.

Having ruled out MAPDs, we instead calculated normalised MADs (NMADs) as the ratio of a DFA's MAD and the test set's $\overline{|\Delta E|}$. The NMAD for DSD-BLYP-D3(BJ) for PCONF21, for instance, is only 0.14. The interested reader can find the NMADs for all assessed DFAs and benchmark sets in the ESI† alongside the other statistical values that we introduced above.

To identify robust DFAs and to enable a ranking of the assessed methods, the so-called weighted total mean absolute deviation (WTMAD) was introduced for GMTKN24 and GMTKN30.[23,24] Each benchmark set was assigned a ''difficulty factor'', which was calculated as the ratio of the MADs of BLYP and B2PLYP-D2, *i.e.*, between a GGA without dispersion and a very good method with dispersion. The test set's MAD was then multiplied by this difficulty factor and further scaled by the number of data points in the respective set. Finally, the average over all those weighted MADs was taken, resulting in a WTMAD. While this scheme does seem arbitrary, it was also verified that schemes without such weight factors gave similar DFA trends, thus, confirming the validity of the conclusions drawn from WTMADs.[23] For instance, the best WTMADs reported for GMTKN30 were for the three DSD methods that we also assess herein (1.3 kcal mol^{-1}).[29]

Also for GMTKN55, we propose using WTMADs to separate reliable from less accurate methods. In preliminary studies, we tested a total of 11 different schemes, based on MAPDs, NMADs, MADs or RMSDs, and each provided a similar picture. For instance, all schemes correctly reproduced the Jacob's Ladder idea. We narrowed those 11 schemes down to two, which we will use simultaneously to underline the reliability of our DFA recommendations.

In the first scheme, dubbed WTMAD-1, each benchmark set is weighted by a factor w that aims at aligning benchmark sets with largely differing $\overline{|\Delta E|}$ values. The MAD of a test set is scaled up by a factor of w = 10 if the set's average absolute reaction energy is below 7.5 kcal mol^{-1}. The MAD of a test set is scaled down by a factor of w = 0.1 if the set's average absolute reaction energy is larger than 75 kcal mol^{-1}. For the remaining sets, w was set to unity. The WTMAD-1 is then simply calculated as an average value over the 55 sets:

$$\text{WTMAD-1} = \frac{\sum_{i}^{55} w_i \cdot \text{MAD}_i}{55}. \quad (1)$$

While the weight factors for WTMAD-1 are arbitrarily defined, the alternative WTMAD-2 scheme uses the ratio between average of all 55 $\overline{|\Delta E|}$ values (56.84 kcal mol^{-1}) and the $\overline{|\Delta E|}$ for the respective test set as a weight factor. This value is then scaled by the number of relative energies N in that particular set before it is combined with the MAD. The WTMAD-2 is then calculated as the sum over all 55 weighted MADs and divided by the total number of relative energies in GMTKN55 (1505 data points):

$$\text{WTMAD-2} = \frac{1}{\sum_{i}^{55} N_i} \cdot \sum_{i}^{55} N_i \cdot \frac{56.84 \text{ kcal mol}^{-1}}{\overline{|\Delta E|}_i} \cdot \text{MAD}_i. \quad (2)$$

While eqn (1) and (2) allow calculating WTMADs for the entire GMTKN55 database, it is straightforward to apply those schemes to each of GMTKN55's categories (see Fig. 1). To obtain a WTMAD-1, the sum of relevant WTMADs is then simply divided by the number of test sets in that category. Likewise, a WTMAD-2 is obtained by division by the number of data points in the respective category.

The WTMADs for each DFA over the entire GMTKN55 and its categories are listed in the ESI.† While those values formally carry the unit kcal mol^{-1}, they should not be mistaken as an indicator for a method's average error. However, the values can be used as a way to score DFAs and to rank them. Also, we suggest that new DFAs developed in the future can be measured

against the DFAs tested herein by comparing their WTMADs. In particular, we challenge developers to beat the best WTMADs presented herein.

5 Benchmark study on GMTKN55

5.1 Example of the benefit of using new reference values

In Section 2, we outlined how we had updated the reference values of most benchmark sets. In some cases, those values only changed marginally, in others we noted significant differences. One example for a significant change had already been given for the BHPERI set in 2015, as also mentioned in the introduction.[55] When compared against W2-F12 and W1-F12 numbers, the original CBS-QB3 reference values turned out to be less accurate than double-hybrid DFAs. Moreover, some DFAs that were recommended for BHPERI in previous studies[7] turned out to be the least accurate in their respective DFA class when compared with the updated reference values.

Herein, we further underline the importance of using new reference values with one additional example. Table S5 in the ESI† shows the MADs for all 83 dispersion-corrected DFAs with respect to the old and the new reference values for the AL2X6 test set. In Section 2, we mentioned that the average absolute change in the reference reaction energies was 1.4 kcal mol^{-1} when adopting the new benchmark. From the values presented in Table S5 (ESI†), we calculated the average MAD for each rung on Jacob's Ladder. For AL2X6 with the original reference values, it turns out that the average MAD for GGA/NGA functionals is 4.39 kcal mol^{-1} and that meta-GGAs/NGAs perform on average slightly worse with a value of 4.53 kcal mol^{-1}. When using our new benchmark, the average MADs for both rungs do not only decrease, but meta-GGAs/NGAs become on average slightly better than GGAs/NGAs (3.65 *vs.* 3.86 kcal mol^{-1}). The average MAD for hybrids drops significantly from 3.66 to 2.83 kcal mol^{-1} with the new values, and also double hybrids become even more accurate (improvement from 1.71 to 1.18 kcal mol^{-1}).

Most striking, however, is the ranking of the DFAs. For instance, we observe a significant change when analysing the best three DFAs for this set. According to the old reference values, the range-separated Minnesota functional MN12SX-D3(BJ) is the best method with an MAD of 0.72 kcal mol^{-1}. It is followed by BHLYP-D3(BJ) (0.76 kcal mol^{-1}) and B2GPPLYP-D3(BJ) (1.01 kcal mol^{-1}). It comes somewhat as a surprise that the best two DFAs are hybrids and that double hybrids are outperformed. However, this picture changes entirely when the new reference values are applied. The best two performers are the DSD-type double hybrids DSD-PBEP86-D3(BJ) (MAD = 0.31 kcal mol^{-1}) and DSD-BLYP (MAD = 0.54 kcal mol^{-1}). They are followed by the hybrid PW6B95-D3(BJ) with an MAD of 0.61 kcal mol^{-1}. Even more striking is that MN12SX-D3(BJ) is now only the 18th-best DFA with an MAD of 1.35 kcal mol^{-1}; BHLYP is in the 23rd position (MAD = 1.56 kcal mol^{-1}).

The BHPERI example from the literature[55] as well as our discussion of AL2X6 demonstrate how recommendations based on DFA rankings can change substantially when more accurate reference values are used. We therefore advocate to use all our new reference values published herein in future studies.

5.2 The need for dispersion corrections

The fact that conventional DFAs need a London-dispersion correction to describe interaction energies in noncovalently bound complexes as well as relative conformational energies is well known; see ref. 222 for a detailed review. An exception to this rule are DFAs that contain a nonlocal vdW kernel, such as VV10 or ωB97X-V.[153,155] Over the past years, it has been demonstrated that dispersion corrections also positively influence geometries of organic and biomolecular systems[44,223–228] as well as the description of reaction energies and BHs.[50,84,222,229] This topic was also discussed for GMTKN30 in 2011.[7] These are only some of many examples in the literature, and yet, we still observe the common trend to rely on uncorrected DFAs in many computational organic chemistry applications; for a discussion on the shortcomings of such an approach, see *e.g.* ref. 48. In light of this discrepancy between theoretical insight and what methods are actually used by some in the user community, it seems necessary to re-emphasise the importance of dispersion corrections for the computational treatment of reactions.

Fig. 2 shows the effect of the DFT-D3 correction on the WTMAD-1 values of four DFAs—each is a representative of its corresponding rung on Jacob's Ladder—for the first three categories in GMTKN55 and the entire database. The actual values and results for the NCI categories are shown in the ESI† alongside WTMAD-2 values, which show the same trend. DFT-D3 decreases the WTMAD-1 values even in non-NCI categories with improvements of up to 36% for isomerisations and reactions of large systems (B3LYP). Note that even for smaller systems, where smaller dispersion contributions are expected, we still see sizeable reductions of 25–28% (BLYP and B3LYP).

The results for BHs may indicate that dispersion is less important for those and that dispersion corrections may, in fact, deteriorate the outcome. The latter is seemingly the case for BLYP, for which the WTMAD-1 increases from 5.19 to 5.53 kcal mol^{-1}. This behaviour, however, has a simple explanation. GGAs/NGAs suffer from the SIE to a much larger extent than hybrids. As a consequence, many reaction barriers are underestimated. For instance the MD of BLYP for BH76 is −8.32 kcal mol^{-1}. As the transition state for most reactions in BH76 can be seen as a "complex" of two interacting fragments (the two reactants or products), a dispersion correction stabilises the transition state more than it does the separate reactants (or products). Therefore, a functional that already underestimates a barrier, will do so even more when dispersion energy is added. In fact, the MD for BLYP-D3(BJ) is −9.23 kcal mol^{-1}. The blame for this does not lie with the dispersion correction. The better uncorrected result can be explained with error-compensation effects between the lack of incorporating dispersion and the SIE. A consequence is, therefore, that BLYP is simply not reliable enough to treat such problems. Indeed, dispersion corrections do improve WTMAD-1 values for the other DFAs in Fig. 2 Note that smaller improvements are seen than for the other two categories of

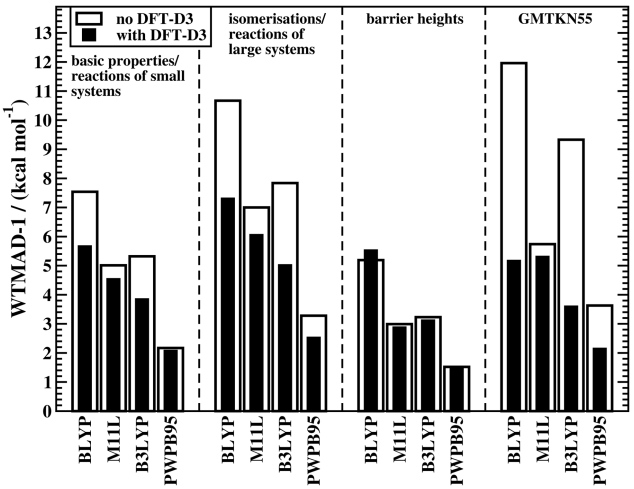

Fig. 2 The effect of dispersion corrections on WTMAD-1 values (kcal mol^{-1}) for the thermochemistry and kinetics categories of GMTKN55 and for the entire database.

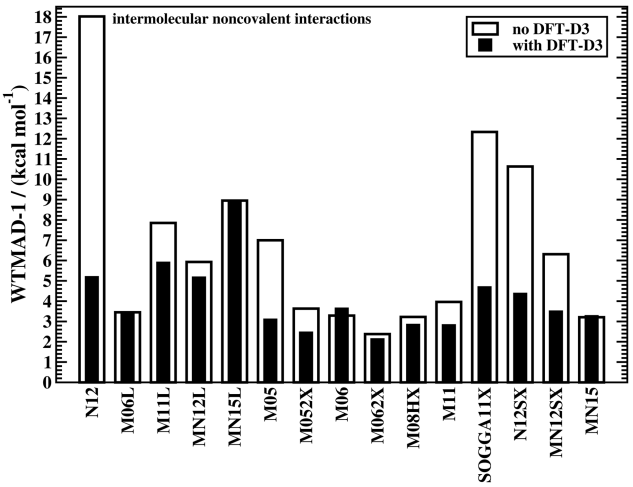

Fig. 3 The effect of dispersion corrections on WTMAD-1 values (kcal mol^{-1}) of Minnesota-type DFAs for intermolecular NCIs.

GMTKN55. That being said, it was out pointed elsewhere that, for instance, in the INV24 set, dispersion corrections influence inversion barriers of helical and bowl-shaped systems by up to 2 kcal mol^{-1}.[84] Indeed, we observe a reduction of the MAD in INV24 for B3LYP from 1.87 to 1.05 kcal mol^{-1}.

It comes as no surprise that the WTMADs for the entire GMTKN55 database also improve when dispersion corrections are employed (Fig. 2), the same was already reported for GMTKN30.[7] As a consequence, we will continue our analysis solely with dispersion-corrected methods given in Table 2; the statistics for uncorrected DFAs are shown in the ESI.†

A common strategy in DFT development is to empirically fit a DFA without any nonlocal vdW kernels to a training set of NCI energies. Minnesota DFAs, beginning with M05 and culminating in MN15, are popular examples of this idea. In our GMTKN30 study, we outlined how dispersion corrections can improve the M05 and M06 classes of DFAs; a nonlocal correction for M06L and DFT-D3 parameters for most of the other Minnesota methods were introduced in ref. 43. Herein, we also present DFT-D3 parameters for M08HX and the MN15 class of DFAs (Section 3.2).

The effect of DFT-D3 on all 15 tested Minnesota methods is depicted in Fig. 3, which displays WTMAD-1 values for the intermolecular interaction category (the actual numbers and WTMAD-2 values are shown in the ESI†). In most cases, DFT-D3 does indeed improve the description of the systems in that category, with some improvements being in the 44–71% range (N12, M05, SOGGA11X, N12SX, MN12SX). Note that M06, MN15L, and MN15 are the only Minnesota DFAs that do not seem to benefit from the DFT-D3 correction. This has been explained with the fact that those methods would describe complexes in their equilibrium geometries well, as the regime of overlapping electron clouds of interacting fragments may be partially described by the DFA itself.[19] However, for non-equilibrium distances it was conclusively shown that dispersion interactions are severely underestimated by most of the herein assessed Minnesota methods;[42,43] also see ref. 25 for another study analysing Minnesota DFAs for NCIs.

Because of the way Minnesota DFAs had been parametrised, some double-counting effects were reported with the Becke–Johnson version of DFT-D3 for some. These effects may also show up for some conformational energies. WTMADs for the category of intramolecular interactions shown in the ESI† confirm this trend. As such problems are very rare with other DFAs, it indicates that choosing Minnesota DFAs for NCIs is not a generally recommended strategy. However, Fig. 2 also shows how reaction energies and BHs are improved for the M11L method with DFT-D3. To be consistent with our observation and the fact that those methods do not properly describe London dispersion, we will only report dispersion-corrected results for Minnesota DFAs in the following sections, while uncorrected results can be found in the ESI.†

5.3 Discussion of GMTKN55 and its categories

Having established the benefit of our new reference values and the importance of using a London-dispersion correction, we proceed with our analysis of the 83 dispersion-corrected DFAs applied to GMTKN55. We will do so first for the separate categories of GMTKN55 before we conclude with a comprehensive look at the entire database to provide final recommendations to the DFT user. Our main aim in each category will be to identify the most robust DFAs that perform best over a series of benchmark sets. With a few exceptions, we will refrain from giving recommendations for each of the 55 individual test sets. However, the best and worst DFAs for each set are shown in the ESI† (Tables S8–S11).

We usually first carry out a ''best-worst'' analysis by counting how many times a DFA gives the best MAD and how many times the worst. This turned out to be a good estimate of DFA robustness in the past.[7] For instance, a method that yields the best and worst MAD an equal number of times should be regarded as less robust and reliable than one that never gives the worst MAD. In a second step, we then analyse the WTMADs

as defined in Section 4 to look at our results from a different perspective. A DFA that appears in both analyses will then be recommended as a reliable method for the assessed category. We will give such recommendations for each of the four highest rungs of Jacob's Ladder.

5.3.1 Basic properties and reactions of small systems. This category comprises 18 test sets, of which some are particularly noteworthy. The W4-11 set of challenging TAEs, for instance, shows a large spread in the MADs for 2nd-rung DFAs from 4.73 kcal mol^{-1} (B97-D3(BJ)) to 20.87 kcal mol^{-1} (PW91P86-D3(0)). Note that not all hybrids necessarily perform better; the MAD for BHLYP-D3(BJ) has a value that is very similar to that of PW91P86-D3(0). Our new SIE4x4 set unsurprisingly reveals that GGAs/NGAs suffer the most from SIE. N12-D3(0) has the lowest MAD for this DFA rung, but with 21.63 kcal mol^{-1} it is still very high. Double hybrids with large fractions of Fock exchange show the smallest MADs for SIE4x4, with DSD-PBEP86-D3(BJ) being the best of all 83 methods (MAD = 5.04 kcal mol^{-1}). The same double hybrid is also the best of all tested methods for the extended DC13 set of difficult reactions. Its MAD of 2.55 kcal mol^{-1} is significantly lower than those of the best methods for the other DFA rungs: N12-D3(0) (7.77 kcal mol^{-1}), SCAN-D3(BJ) (6.87 kcal mol^{-1}), and MN15-D3(BJ) (5.09 kcal mol^{-1}) (ESI†). Note that the latter has a lower MAD than the B2PLYP-D3(BJ) double hybrid, which has the largest MAD of rung-5 DFAs (6.77 kcal mol^{-1}). The worst-performing GGA for this set is OPBE-D3(BJ) (19.74 kcal mol^{-1}), while the meta-GGA τHCTH-D3(BJ) and the global hybrid mPW1LYP-D3(0) are the worst in their respective DFA classes with MADs of 10.97 and 11.91 kcal mol^{-1}, respectively.

To get a better overview of the large amount of obtained data, we first analyse GGAs and NGAs and identify the best and worst performing representatives (Fig. 4a). While no single method has the best MAD in the absolute majority of cases, we are still able to identify two GGAs that clearly distinguish themselves from others (also see Table S8, ESI†): B97-D3(BJ) yields the best MAD in four cases (W4-11, G21EA, G21IP, and RC21), while revPBE-D3(BJ) does so in three cases (HEAVYSB11, BH76RC, and TAUT15). Both DFAs have in common that they never give the worst MAD. This is in contrast to N12-D3(0), which performs best in four cases (SIE4x4, AL2X6, FH51, and DC13), but also performs worst in two cases (G21IP and TAUT15). Methods that clearly under-perform are HCTH/407-D3(BJ) (4 times) and OPBE-D3(BJ) (5 times). Interestingly, combining OPTX exchange with LYP rather than PBE correlation improves this picture and OLYP-D3(BJ) is the worst GGA in only one case (SIE4x4), while it is the best-performing one for the ALKBDE10 set. The popular PBE-D3(BJ) and BLYP-D3(BJ) are never the best GGAs, and in fact they give the worst MAD in one case each (BH76RC and DIPCS10, respectively). The popular BP86-D3(BJ) method never provides the best nor worst MAD, and for now it is not possible to assess its overall performance without consulting its WTMADs.

Fig. 5a shows WTMAD averages for each assessed DFA class. The figure confirms the expected trend for Jacob's Ladder, namely that rung-2 DFAs are on average less accurate than higher rungs. It also demonstrates that both our suggested WTMAD schemes predict the same trend, but that WTMAD-2 values are higher for every class than WTMAD-1 ones. The averaged WTMAD-1 for rung 2 for basic properties and reactions of small systems is 5.70 kcal mol^{-1} and the WTMAD-2 is 6.60 kcal mol^{-1} (also see Table S12, ESI†). While those numbers are averages, individual WTMAD-1 values for GGAs/NGAs range from 4.70 to 7.77 kcal mol^{-1}, and WTMAD-2 values from 5.54 to 8.23 kcal mol^{-1} (Tables S20 and S21, ESI†).

The best three GGAs/NGAs according to the WTMAD-1 and WTMAD-2 schemes are shown in Table 3. Both schemes determine revPBE-D3(BJ) to be the best GGA. The BPBE-D3(BJ) method, which neither gives the best nor the worst MAD, performs on average very well and it comes in second place for WTMAD-1, closely followed by BP86-D3(BJ). Based on WTMAD-2, BPBE-D3(BJ) has a slightly lower value than N12-D3(0) and comes in third place. However, the information gathered in Fig. 4 hinted at N12-D3(0) being less robust. Note that B97-D3(BJ) has the seventh-best WTMAD-1 (5.26 kcal mol^{-1}) and the fourth-best WTMAD-2 (5.98 kcal mol^{-1}), while BP86-D3(BJ) has the seventh-best WTMAD-2 with 6.30 kcal mol^{-1}. The two largest WTMADs shown in Table S13 (ESI†) confirm our previous analysis, namely that OPBE-D3(BJ) and HCTH/407-D3(BJ) should be avoided when studying the chemical problems considered in this category. At this stage, we can, thus, conclude that revPBE-D3(BJ) or BPBE-D3(BJ) are most likely the best choices for this category.

For rung 3 of Jacob's ladder, there are only two DFAs that never give the worst MAD (Fig. 4a). In fact, MN15L-D3(0) gives the best MAD in five cases (W4-11, G21IP, PA26, SIE4x4, and RC21) and TPSS-D3(BJ) in three (DIPCS10, ALKBDE10, and NBPRC). In general, meta-GGAs/NGAs are more accurate than rung-2 methods with averaged WTMADs of 4.91 kcal mol^{-1} (WTMAD-1) and 5.43 kcal mol^{-1} (WTMAD-2). That being said, the least accurate meta-GGAs have larger WTMADs than some GGAs/NGAs, with τHCTH-D3(BJ) having the largest WTMAD-1 with 6.82 kcal mol^{-1} and PKZB-D3(0) the largest WTMAD-2 with 6.71 kcal mol^{-1} (Table S13, ESI†). Also the results in Fig. 4a indicate that these two methods are among the worst-performing, as they have the largest MADs in four cases each. The best meta-GGAs/NGAs in Table 3 are all Minnesota methods, with MN15L-D3(0) having the lowest WTMADs according to both schemes (WTMAD-1 = 3.42 kcal mol^{-1}, WTMAD-2 = 4.01 kcal mol^{-1}). It is the only DFA for this functional rung that we recommend at this stage, as the other methods occasionally deliver the worst MADs (Fig. 4a).

Given the much larger number of assessed hybrid DFAs, the analysis of best and worst MADs for this category does not allow drawing as clear a picture as before, and there are indeed many approaches that neither give the best nor the worst MAD, but they may still be regarded as reliable. Nevertheless, a figure similarly to Fig. 4 is provided in the ESI.† DFAs that seem to be noteworthy are PW6B95-D3(BJ), M062X-D3(0), M08HX-D3(0), MN15-D3(BJ), SOGGA11X-D3(BJ), and ωB97X-V; they never give the worst MAD, but each of them provides the best in one to four cases. Out of those methods, the WTMAD-1 scheme places

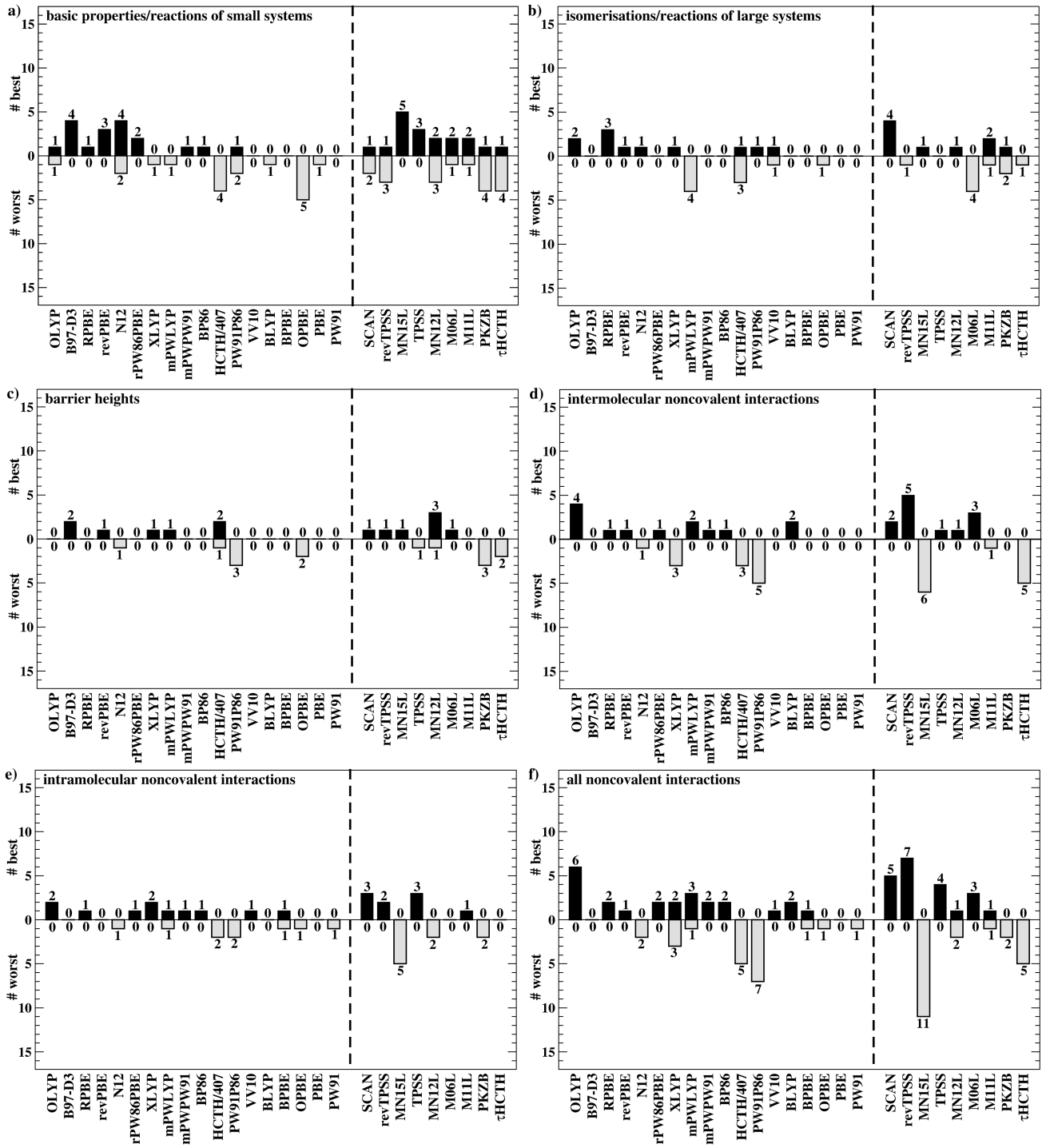

Fig. 4 Analysis of how many times a DFA yields the worst and best MAD in each of the categories of GMTKN55 (images a–f). The analysis was carried out separately for GGAs/NGAs (left part in each image) and meta-GGAs/NGAs (right part in each image). All DFAs are dispersion corrected, but the suffix "D3" was omitted from the labels for better clarity, unless it is needed to avoid ambiguity.

M08HX-D3(0), ωB97X-V, and M062X-D3(0) in the top three (Table 3). The WTMAD-2 scheme, confirms that M08HX-D3(0) and M062X-D3(0) do indeed seem to be competitive in this category, as they are predicted to be the best two hybrids. The WTMAD-2 values of the following DFAs are very close together, and while ωB97X-V ranks on seventh place, its WTMAD-2 of 3.34 kcal mol^{-1} is not too far away from that of MN15-D3(BJ) (2.95 kcal mol^{-1}). Based on their WTMADs shown in Table S13 (ESI†) and the number of times they provide the worst MAD (Fig. S3, ESI†), the worst-performing hybrids are B97-2-D3(BJ), O3LYP-D3(BJ), and the two revTPSS-based hybrids revTPSSh-D3(BJ) and revTPSS0-D3(BJ). In fact, those methods are outperformed by

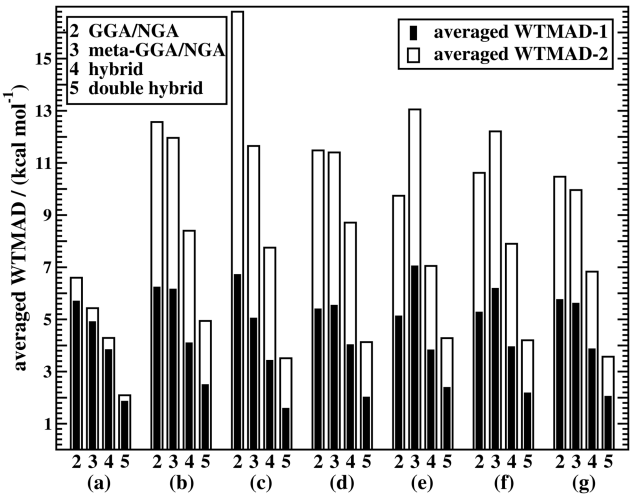

Fig. 5 Averaged WTMAD-1 and WTMAD-2 values (kcal mol^{-1}) for rungs 2–5 on Jacob's Ladder. The values are displayed for the various categories of GMTKN55, namely basic properties and reactions of small systems (a), isomerisations and reactions of large systems (b), barrier heights (c), intermolecular noncovalent interactions (d), intramolecular noncovalent interactions (e), and all noncovalent interactions (f). Values are also shown for the entire GMTKN55 database (g). Only dispersion-corrected DFAs were considered in this analysis.

Table 3 The best three DFAs for each of the four highest rungs on Jacob's Ladder for basic properties and reactions of small systems according to WTMAD-1 and WTMAD-2 values (kcal mol^{-1})

Rung	WTMAD-1	WTMAD-2
GGA/NGA	revPBE-D3(BJ) (4.70)	revPBE-D3(BJ) (5.54)
	BPBE-D3(BJ) (5.03)	N12-D3(0) (5.76)
	BP86-D3(BJ) (5.11)	BPBE-D3(BJ) (5.82)
Meta-GGA/NGA	MN15L-D3(0) (3.42)	MN15L-D3(0) (4.01)
	M06L-D3(0) (4.42)	MN12L-D3(BJ) (4.44)
	M11L-D3(0) (4.54)	M11L-D3(0) (4.89)
Hybrid	M08HX-D3(0) (2.48)	M062X-D3(0) (2.73)
	ωB97X-V (2.63)	M08HX-D3(0) (2.75)
	M062X-D3(0) (2.66)	MN15-D3(BJ) (2.95)
Double hybrid	DSD-PBEP86-D3(BJ) (1.46)	DSD-PBEP86-D3(BJ) (1.69)
	DSD-BLYP-D3(BJ) (1.63)	DSD-BLYP-D3(BJ) (1.88)
	DSD-PBEB95-D3(BJ) (1.70)	DSD-PBEB95-D3(BJ) (1.89)

the best (meta-)GGA/NGA approaches. While this indicates that not every hybrid is necessarily better than lower-rung DFAs, we also note that the averaged WTMADs for hybrids are about 1 kcal mol^{-1} lower than for rung 2 (Fig. 5).

The popular hybrids PBE0-D3(BJ) and B3LYP-D3(BJ) never give the best nor the worst MAD, and they rank in 26th and 29th position with WTMAD-1 values around 3.8 kcal mol^{-1}, which is around the average WTMAD-1 for hybrids. B3LYP-NL is in 27th place with almost the same WTMAD-1 as the D3-corrected version. WTMAD-2 values provide the same picture (ESI†). With an averaged WTMAD-1 of 1.87 kcal mol^{-1} and an averaged WTMAD-2 of 2.09 kcal mol^{-1}, double hybrids are by far superior than hybrids (Fig. 5). Even the largest WTMADs for double hybrids—for instance MPW2PLYP-D3(BJ) and B2PLYP-D3(BJ) with WTMAD-1 values 2.24 kcal mol^{-1}—are lower than the WTMADs of the best hybrids. In fact, the WTMADs of all seven assessed double hybrids are very close to one another, which makes a best-worst analysis of their MADs less insightful. In fact, any of the assessed double hybrids can safely be applied to the test sets considered so far. That being said, the by far best double hybrids in this category—and also the best DFAs out of all 83—are the three DSD methods (Table 3). We will see next if the DFAs recommended at this stage also perform well in the remaining categories.

5.3.2 Isomerisations and reactions of large systems. Given the difficulty of its molecules, MB16-43 can be considered as a tough challenge to any electronic-structure method, in the same way as its predecessor MB08-165 allowed gauging a method's robustness. When closely inspecting the MADs and RMSDs for this set, the generally large deviations merit special attention. For instance, even the best DFA DSD-PBEP86-D3(BJ) has an MAD of 6.46 kcal mol^{-1}. When normalised against the averaged absolute reaction energy for this set, this reduces to a nearly perfect NMAD of 0.02. Only six DFAs have an MAD smaller than 10 kcal mol^{-1}, with five of them being double hybrids and PW6B95-D3(BJ) being the only hybrid (MAD = 8.97 kcal mol^{-1}). Note that also for MB08-165 this method was the best hybrid.[7,159]

As both sets were created independently and without any user bias, this reconfirms the PW6B95's high robustness. While some Minnesota approaches have relatively low MADs close to 12 kcal mol^{-1} (MN12SX, and SOGGA11X), the two MN15-type DFAs both have nearly identical high MADs above 20 kcal mol^{-1}, thus indicating that the hybrid version is not more robust than the meta-NGA one. Interestingly, the two promising ωB97X-V and ωB97X-D3 methods both have MADs above 32 kcal mol^{-1}. The best second-rung DFA is SCAN-D3(BJ) with an MAD of 17.77 kcal mol^{-1}, while M06L-D3(0) is the worst with an MAD of 63.27 kcal mol^{-1}. The worst DFA of all is HCTH/407-D3(BJ) with an MAD of 76.52 kcal mol^{-1} (NMAD = 0.18).

While double hybrids are methods that tend to perform better than hybrids,[7,29] we like to point out that the new C60ISO set seems to be more challenging for them. In fact, the best double hybrid PWPB95-D3(BJ) has an MAD of 3.48 kcal mol^{-1}, which is double of that of the best hybrid (and the best DFA for this set) PW6B95-D3(BJ) (MAD = 1.65 kcal mol^{-1}). This shows the importance of developing benchmark sets with larger, more realistic (and unsaturated in this case) systems to identify needs for further development. That highly conjugated systems can be challenging for double hybrids was also shown for C_{20} and C_{24} isomers in ref. 73.

Due to the nine benchmark sets in this category having larger average absolute reaction energies, the individual weights used in the two WTMAD schemes differ more from one another and WTMAD-2 numbers turn out to be significantly larger than WTMAD-1 ones. That being said, both schemes still provide the same trends. The average WTMADs all reproduce the Jacob's Ladder scheme and double hybrids are the best approaches in this category (Fig. 5). Also, the best three DFAs on each rung are the same in both schemes (Table 4).

Fig. 4b points at RPBE-D3(BJ) and OLYP-D3(BJ) as potentially good approaches, as they are the only GGAs that give the best MADs in three and two cases, respectively: RPBE-D3(BJ) is the best for RSE43, BSR36 and ISO34, while OLYP-D3(BJ) performs best for DARC and ISOL24. revPBE-D3(BJ) performs equally well for RSE43, while N12-D3(0) yields the same MAD as RPBE-D3(BJ) for ISO34. In fact, both WTMAD schemes rank OLYP-D3(BJ) as the best GGA followed by N12-D3(0) and RPBE-D3(BJ) for WTMAD-1, whereas the order of the last two is reversed for WTMAD-2 (Table 4). The worst-performing rung-2 DFAs are mPWLYP-D3(BJ), OPBE-D3(BJ), BLYP-D3(BJ), and rPW86PBE-D3(BJ). Again, we see that OPTX exchange combined with PBE correlation does not seem to be a good match, even though this combination is popular in some areas.[149]

The only meta-GGA that is worth being mentioned according to Fig. 4b is SCAN-D3(BJ), which has the best MAD for four benchmark sets (MB16-43, DARC, ISOL24, and PArel). It outperforms Minnesota DFAs; in fact, M06L-D3(0) has the worst MAD in four cases (MB16-43, DARC, CDIE20, and ISOL24). Also both WTMAD schemes place SCAN-D3(BJ) at the top with a large gap before the second-ranking method. SCAN-D3(BJ) has a WTMAD-1 value of 4.55 kcal mol^{-1} followed by revTPSS-D3(BJ) with 5.52 kcal mol^{-1}; the WTMAD-2 for SCAN-D3(BJ) is 7.86 kcal mol^{-1}, whereas the next DFA M11L-D3(0) has a value of 10.46 kcal mol^{-1} (Table 4).

While PW6B95-D3(BJ) and M052X-D3(0) both give the best MADs in two cases (MB16-43 and C60ISO for the first, DARC and CDIE20 for the latter), it is interesting to note that the first does not appear in the list of top three WTMADs, while the latter is ranked as the best hybrid (Table 4). In 2nd and 3rd position follow BMK-D3(BJ) and M08HX-D3(0). The three worst hybrids are M05-D3(0), O3LYP-D3(BJ), and mPW1LYP-D3(0) (see ESI†). B3LYP-D3(BJ)'s WTMAD-1 is with 5.02 kcal mol^{-1} in 44th place and by nearly 1 kcal mol^{-1} worse than the average WTMAD-1 of 4.10 kcal mol^{-1} for this DFA class (Fig. 5). We do note however, that using the nonlocal VV10 kernel improves the WTMAD-1 to 4.00 kcal mol^{-1} for B3LYP-NL, even though this value still hovers around the average.

Table 4 The best three DFAs for each of the four highest rungs on Jacob's Ladder for isomerisations and reactions of large systems according to WTMAD-1 and WTMAD-2 values (kcal mol^{-1})

Rung	WTMAD-1	WTMAD-2
GGA/NGA	OLYP-D3(BJ) (4.81) N12-D3(0) (5.10) RPBE-D3(BJ) (5.17)	OLYP-D3(BJ) (9.67) RPBE-D3(BJ) (10.02) N12-D3(0) (10.30)
Meta-GGA/NGA	SCAN-D3(BJ) (4.55) revTPSS-D3(BJ) (5.52) M11L-D3(0) (6.06)	SCAN-D3(BJ) (7.86) M11L-D3(0) (10.46) MN15L-D3(0) (10.58)
Hybrid	M052X-D3(0) (2.62) BMK-D3(BJ) (2.74) M08HX-D3(0) (2.82)	M052X-D3(0) (5.20) M08HX-D3(0) (5.40) BMK-D3(BJ) (5.45)
Double hybrid	DSD-PBEB95-D3(BJ) (1.78) DSD-PBEP86-D3(BJ) (1.80) DSD-BLYP-D3(BJ) (2.22)	DSD-PBEB95-D3(BJ) (3.28) DSD-PBEP86-D3(BJ) (3.91) DSD-BLYP-D3(BJ) (4.32)

Even though their performance may not be the best for C60ISO, double hybrids again feature the lowest WTMADs. This time, however, MPW2PLYP-D3(BJ) and B2PLYP-D3(BJ) are outperformed by the best three hybrids (WTMAD-1 = 3.36 kcal mol^{-1}) (ESI†). The best three double hybrids—and also the best three DFAs in this category—are Martin's DSD methods with both WTMAD schemes ranking them in the same order: DSD-PBEB95-D3(BJ), DSD-PBEP86-D3(BJ), and DSD-BLYP-D3(BJ) (Table 4).

5.3.3 Barrier heights. Of the seven sets in the BH category, five cover reactions with bond-breaking and -formation processes, while two do not formally involve bond breaking, but instead the rotation around single bonds (BHROT27) or shape inversion (INV24). It comes at no surprise that the difference between the Jacob's Ladder rungs are larger for the first five than for the latter two. It is well known that SIE plays a particular importance in elongated bonds, as they occur in transition states of bond-breaking/formation reactions. Indeed, the best (meta-)GGAs/NGAs are by far not competitive enough when compared with hybrids and double hybrids. For instance, the best GGA for the new set of diverse BHs (BHDIV10) is B97-D3(BJ) with an MAD of 5.83 kcal mol^{-1} and an MD of −5.08 kcal mol^{-1}, which indicates systematic underestimation of the barriers.

This picture seems to improve for the next rung with the best meta-NGA being MN12L-D3(BJ) (MAD = 2.03), however, this is still not satisfying enough when comparing it with the best hybrid ωB97X-V (MAD = 0.85 kcal mol^{-1}) and the best double hybrid DSD-PBEB95-D3(BJ) (MAD = 0.83 kcal mol^{-1}) (ESI†). Contrary to that, the results for BHROT27 indicate that the best DFAs in each rung show fairly similar accuracy with revPBE-D3(BJ) having an MAD of only 0.37 kcal mol^{-1}, while DSD-PBEP86-D3(BJ) has a value of 0.21 kcal mol^{-1} (ESI†). Error-compensation effects between the minimum-energy structures and the transition states are the likely reason for the good performance of second-rung methods for this test set. As already reported elsewhere, inversion barriers also show smaller differences between DFAs, however, the Jacob's Ladder picture is still reproduced.[84] The best GGA B97-D3(BJ), for instance, has an MAD of 1.80 kcal mol^{-1}, which is reduced to 0.69 kcal mol^{-1} for B2PLYP-D3(BJ) (see ESI†). As also pointed out before, all double hybrids behave very similarly for this set and all their MADs are below the chemical-accuracy threshold of 1 kcal mol^{-1}.[84]

The averaged WTMADs in Fig. 5 are significantly higher for (meta-)GGAs/NGAs than for (double-)hybrids. For instance, average WTMAD-2 values are 16.80 kcal mol^{-1} for the 2nd, 11.64 kcal mol^{-1} for the 3rd, 7.75 kcal mol^{-1} for the 4th, and 3.51 kcal mol^{-1} for the 5th rung. While more detailed results for rungs 2 and 3 are shown in Fig. 4c and Table 5, we do not make specific recommendations for them, as hybrids and double hybrids should be used for BHs to obtain reliable results.

The lowest WTMAD-1 values are observed for SOGGA11X-D3(BJ), ωB97X-V, and M08HX-D3(0), while M08HX-D3(0), BMK-D3(BJ), and MN12SX-D3(BJ) have the lowest WTMAD-2 values (Table 5). The BMK method was originally designed to describe kinetics, as the letter "K" indicates. However, we did not

observe it to be the best DFA for any of the seven individual test cases (see ESI†). Instead, the MPWB1K method—also developed for kinetics—gave the best MAD in two cases when dispersion corrected (PX13 and WCPT18). It ranks as the fifth hybrid in the WTMAD-2 scheme, and as seventh in the WTMAD-1 list (see ESI†). Another hybrid specifically designed for the calculation of BHs—MPWKCIS1K-D3(BJ)—does not appear to be of any value and it ranks as 13th (WTMAD-2) and 14th (WTMAD-1). Both schemes evaluate the following three hybrids as the least accurate: O3LYP-D3(BJ), MPW1KCIS-D3(BJ), and TPSSh-D3(BJ).

BHLYP-D3(BJ), which is popular for BHs due to its large fraction of Fock-exchange, cannot be recommend at all for this purpose; it is in 26th position in the WTMAD-1 and in 21st in the WTMAD-2 list. It is almost needless to mention that also B3LYP should not be used (position 34 for B3LYP-D3(BJ) and 40 for B3LYP-NL in the WTMAD-2 picture, see ESI†).

The best double hybrids all deliver MADs that are 1 kcal mol^{-1} (BH76 set) or lower (all other sets), which means they can provide results with chemical accuracy. The best double hybrid for BHs is DSD-PBEB95-D3(BJ), which yields the best MAD in the majority of cases (BH76, DSD-PBEB95, PX13, and WCPT18). Double hybrids relying on B95 correlation seem to be particularly good in this category, as PWPB95 gives the best MAD for a fifth set, namely BHPERI (also see ref. 55). Both WTMAD schemes place DSD-PBEB95-D3(BJ) in first position with values of 1.02 kcal mol^{-1} (WTMAD-1) and 2.26 kcal mol^{-1} (WTMAD-2). These values are significantly lower than the second-best method DSD-BLYP-D3(BJ) with 1.45 and 3.04 kcal mol^{-1}, respectively. DSD-BLYP-D3(BJ) is closely followed by PWPB95-D3(BJ) (WTMAD-1 = 1.50 kcal mol^{-1}) or B2GPPLYP-D3(BJ) (WTMAD-2 = 3.24 kcal mol^{-1}).

5.3.4 Intermolecular noncovalent interactions. 12 sets deal with interaction energies in noncovalently bound complexes. In Section 5.2, we already demonstrated the necessity of using dispersion corrections to obtain better results, this includes Minnesota approaches. That being said, among the 83 tested dispersion-corrected DFAs, there remains a wide error spread among the results. This can already be seen for the averaged WTMADs in Fig. 5 where again double hybrids are on average the most accurate methods, followed by hybrids. Very important is the finding that there is on average no benefit from using rung-3 methods. Their average WTMADs are slightly higher than for rung-2 DFAs. In fact, the best three meta-GGAs M06L-D3(0), revTPSS-D3(BJ), and TPSS-D3(BJ) (see WTMADs in Table 6) are all outperformed by the best GGAs, which according to WTMAD-1 values are BLYP-D3(BJ), OLYP-D3(BJ), and B97-D3(BJ), while WTMAD-2 values place B97-D3(BJ), revPBE-D3(BJ) and OLYP-D3(BJ) in the top 3 (Table 6).

According to Fig. 4d, OLYP-D3(BJ) gives the best MAD for four test sets (ADIM6, HEAVY28, CARBHB12, and HAL59). BLYP-D3(BJ), on the other hand, provides spectacularly low MADs for the popular S22 and S66 sets (0.25 and 0.17 kcal mol^{-1}, respectively), which are commonly used to assess a method's performance to describe NCIs. Among the worst-performing GGAs we find HCTCH/407-D3(BJ), PW91P86-D3(0), PBEhPBE-D3(BJ)—where normal PBE exchange has been replaced by PBE-hole exchange—and XLYP-D3(BJ). The latter will be of importance in the next section.

Table 5 The best three DFAs for each of the four highest rungs on Jacob's Ladder for barrier heights according to WTMAD-1 and WTMAD-2 values (kcal mol^{-1})

Rung	WTMAD-1	WTMAD-2
GGA/NGA	B97-D3(BJ) (5.15) BLYP-D3(BJ) (5.53) rPW86PBE-D3(BJ) (5.56)	B97-D3(BJ) (13.15) HCTH/407-D3(BJ) (13.76) rPW86PBE-D3(BJ) (15.17)
Meta-GGA/NGA	M11L-D3(0) (2.88) M06L-D3(0) (3.31) MN15L-D3(0) (3.51)	M11L-D3(0) (5.47) MN15L-D3(0) (5.49) MN12L-D3(BJ) (5.74)
Hybrid	SOGGA11X-D3(BJ) (1.77) ωB97X-V (1.91) M08HX-D3(0) (1.99)	M08HX-D3(0) (3.33) BMK-D3(BJ) (3.73) MN12SX-D3(BJ) (3.74)
Double hybrid	DSD-PBEB95-D3(BJ) (1.02) DSD-BLYP-D3(BJ) (1.45) PWPB95-D3(BJ) (1.50)	DSD-PBEB95-D3(BJ) (2.26) DSD-BLYP-D3(BJ) (3.04) B2GPPLYP-D3(BJ) (3.24)

Table 6 The best three DFAs for each of the four highest rungs on Jacob's Ladder for intermolecular noncovalent interactions according to WTMAD-1 and WTMAD-2 values (kcal mol^{-1})

Rung	WTMAD-1	WTMAD-2
GGA/NGA	BLYP-D3(BJ) (3.20) OLYP-D3(BJ) (3.33) B97-D3(BJ) (3.42)	B97-D3(BJ) (5.95) revPBE-D3(BJ) (6.19) OLYP-D3(BJ) (7.03)
Meta-GGA/NGA	M06L-D3(0) (3.41) revTPSS-D3(BJ) (3.58) TPSS-D3(BJ) (4.15)	revTPSS-D3(BJ) (6.70) M06L-D3(0) (7.37) TPSS-D3(BJ) (7.59)
Hybrid	ωB97X-V (1.45) PW6B95-D3(BJ) (2.01) M062X-D3(0) (2.13)	ωB97X-V (3.03) PW6B95-D3(BJ) (4.22) BHLYP-D3(BJ) (4.46)
Double hybrid	PWPB95-D3(BJ) (1.75) B2PLYP-D3(BJ) (1.86) DSD-BLYP-D3(BJ) (1.94)	B2PLYP-D3(BJ) (3.78) DSD-PBEB95-D3(BJ) (3.90) DSD-BLYP-D3(BJ) (3.92)

ωB97X-V (WTMAD-1 = 1.45 kcal mol^{-1}, WTMAD-2 = 3.03 kcal mol^{-1}) is by far the best hybrid for intermolecular NCIs. We report a large gap between this and the second-best hybrid PW6B95-D3(BJ) (WTMAD-1 = 2.01 kcal mol^{-1}, WTMAD-2 = 4.22 kcal mol^{-1} in Table 6). While M062X-D3(0) follows in third place in the WTMAD-1 ranking, it is very surprising to see BHLYP-D3(BJ) appear as third-best hybrid in the WTMAD-2 list. Unexpectedly, BHLYP-D3(BJ) gives the best MAD in three cases (RG18, ADIM6, and HAL59), while ωB97X-V is the best hybrid in only two instances (S66 and PNICO23). ωB97X-V's MAD for S66 is with 0.12 kcal mol^{-1} the best of all methods, even outperforming double hybrids—the best double hybrid is DSD-BLYP-D3(BJ) with an MAD of 0.17 kcal mol^{-1}. If a user is unable to apply the nonlocal vdW correction, we recommend ωB97X-D3 as an alternative to ωB97X-V. It also gives the best MAD for two sets (S22 and AHB21). It also ranks fourth among WTMAD-1 results and fifth for WTMAD-2. Also noteworthy is the result for the IL16 set for ion pairs mimicking ionic liquids. B3LYP-NL is—together with

HSE06-D3(BJ)—the best hybrid with an MAD of 0.31 kcal mol^{-1}, which is significantly lower than the value of 0.76 kcal mol^{-1} for B3LYP-D3(BJ). Based on our WTMADs in the ESI,† we discourage from using the following hybrids: O3LYP-D3(BJ), TPSS1KCIS-D3(BJ), HSE03-D3(BJ), and BMK-D3(BJ).

The best three hybrids are competitive with the three double hybrids that have the largest WTMADs: MPW2PLYP-D3(BJ), B2GPPLYP-D3(BJ), and DSD-PBEP86-D3(BJ). That being said, those DFAs are still by far more accurate than the majority of the other DFAs. The best three double hybrids are listed in Table 6, and they are PWPB95-D3(BJ), B2PLYP-D3(BJ), and DSD-BLYP-D3(BJ) according to WTMAD-1 results, and B2PLYP-D3(BJ), DSD-PBEB95-D3(BJ) and DSD-BLYP-D3(BJ) according to the WTMAD-2 scheme.

Next, we investigate if these findings can also be transferred to intramolecular interactions.

5.3.5 Intramolecular noncovalent interactions. The last category of GMTKN55 is a very important one, as intramolecular NCIs are always present. They do not only have to be treated properly when investigating the conformational space of a molecule, but they also play a role in thermochemistry, particularly when reactants and products differ in molecular size or shape.

The top-3 list of GGAs/NGAs is somewhat surprising, as both WTMAD schemes list XLYP-D3(BJ) as the best GGA, even though it was among the four worst methods for intermolecular interactions (Table 7). revPBE-D3(BJ) and B97-D3(BJ) follow closely, however, and should probably be preferred, as we will discuss in the next section.

Averaged WTMAD-1 values for rung-3 DFAs are by 2 kcal mol^{-1} worse than for rung-2 ones. Closer inspection reveals that this is mostly due to the documented problem that double-counting effects of medium-range interactions can occur between a dispersion correction and the Minnesota meta-GGAs/NGAs due to their highly-parametrised nature and the way those parameters had been obtained (also see Fig. S1) (ESI†).[7,43] Interestingly, MN15L-D3(0) gives the worst MAD in 5 cases. In Section 3.2, we observed how it was nearly unaffected by the dispersion correction. However, it seems that this does not mean that it is capable of describing dispersion effects accurately, as the errors are relatively high. This means that there seems to be an underlying problem with MN15L for these types of interaction. In fact, the worst two methods among meta-GGAs/NGAs are MN15L-D3(0) and MN12L-D3(0). The only meta-GGA that seems to be competitive with GGAs is SCAN-D3(BJ) with a WTMAD-1 value of 3.61 kcal mol^{-1}, compared to the best GGA XLYP-D3(BJ) with a value of 4.06 kcal mol^{-1}.

Averaged WTMADs demonstrate again that hybrids outperform the lower rungs, but that they are themselves outperformed by double hybrids. The best hybrid for these interactions is ωB97X-V, which gives the best MADs among all 83 tested DFAs for the ACONF and BUT14DIOL test sets with nearly perfect values of 0.03 and 0.04 kcal mol^{-1}, respectively (ESI†). In our ranking of hybrids, this method is followed by two unexpected candidates, namely revTPSS0-D3(BJ) and B97-1-D3(BJ). We note that those methods are by far not common in quantum-chemical software and they do not perform particularly well in any of the previously discussed categories. The first conventional hybrids in the lists of WTMADs are B3LYP-D3(BJ) and BHLYP-D3(BJ). In fact, the WTMAD-2 list places the latter in fourth position together with ωB97X-D3. This comes again as a surprise and we are not aware that BHLYP-D3(BJ) has ever been recommended for, *e.g.*, the calculation of conformational energies.

Among double hybrids, the DSD-type methods DSD-BLYP-D3(BJ) and DSD-PBEP86-D3(BJ) as well as the general-purpose double hybrid B2GPPLYP-D3(BJ) can be recommended, whereas DSD-PBEB95-D3(BJ) is ranked as the worst double hybrid (WTMAD-1 = 3.47 kcal mol^{-1}, WTMAD-2 = 6.70 kcal mol^{-1}) (ESI†).

Repeating the WTMAD analysis without the IDISP set, gives the same top-5 DFAs for each rung, which means that our recommendations above are also valid for the treatment of conformers only. In summary, the discussion of this section has revealed some surprises and differences between intra- and intermolecular noncovalent interactions and we will address these in the following section.

5.3.6 All noncovalent interactions. Separating the NCI category of GMTKN55 into two separate parts had the advantage to identify problems with a DFA's robustness. For instance, XLYP-D3(BJ) is the best-performing GGA for intramolecular, but one of the worst for intermolecular interactions. ωB97X-V, on the other hand, is the best hybrid for both interaction types, which hints at its overall robustness. Here, we combine the two types of NCI categories to identify methods that provide a reliable description of both. In this a part of our discussion, we therefore focus on running an analysis across all 21 NCI sets.

By doing so, we realise that OLYP-D3(BJ) seems to outperform other GGAs/NGAs based on our best-worst analysis; it gives the best MAD in 6 out of the 21 benchmark sets considered herein (Fig. 4f). In fact, it overall ranks as the third-best DFA in this class on the WTMAD-1 list (Table 8), while it is in fourth position in the WTMAD-2 ranking (see ESI†). Ultimately, however, the WTMADs also indicate that BLYP-D3(BJ), B97-D3(BJ), and revPBE-D3(BJ) can be used, which are the GGAs that we recommend for this category. Interestingly, those three methods were also among the best four GGAs that were

Table 7 The best three DFAs for each of the four highest rungs on Jacob's Ladder for intramolecular noncovalent interactions according to WTMAD-1 and WTMAD-2 values (kcal mol^{-1})

Rung	WTMAD-1	WTMAD-2
GGA/NGA	XLYP-D3(BJ) (4.06) revPBE-D3(BJ) (4.10) B97-D3(BJ) (4.20)	XLYP-D3(BJ) (7.42) B97-D3(BJ) (7.84) revPBE-D3(BJ) (7.99)
Meta-GGA/NGA	SCAN-D3(BJ) (3.61) revTPSS-D3(BJ) (4.43) TPSS-D3(BJ) (4.74)	SCAN-D3(BJ) (6.61) revTPSS-D3(BJ) (7.06) TPSS-D3(BJ) (8.36)
Hybrid	ωB97X-V (2.29) revTPSS0-D3(BJ) (2.69) B97-1-D3(BJ) (2.77)	ωB97X-V (3.62) revTPSS0-D3(BJ) (4.77) B97-1-D3(BJ) (4.82)
Double hybrid	DSD-BLYP-D3(BJ) (1.87) B2GPPLYP-D3(BJ) (1.90) DSD-PBEP86-D3(BJ) (2.08)	DSD-BLYP-D3(BJ) (3.15) B2GPPLYP-D3(BJ) (3.21) DSD-PBEP86-D3(BJ) (3.46)

Table 8 The best three DFAs for each of the four highest rungs on Jacob's Ladder for all noncovalent interactions test sets in GMTKN55 according to WTMAD-1 and WTMAD-2 values (kcal mol^{-1})

Rung	WTMAD-1	WTMAD-2
GGA/NGA	BLYP-D3(BJ) (3.70) B97-D3(BJ) (3.75) OLYP-D3(BJ) (3.77)	B97-D3(BJ) (6.87) revPBE-D3(BJ) (7.07) BLYP-D3(BJ) (7.56)
Meta-GGA/NGA	revTPSS-D3(BJ) (3.94) SCAN-D3(BJ) (4.06) TPSS-D3(BJ) (4.40)	revTPSS-D3(BJ) (6.88) SCAN-D3(BJ) (7.58) TPSS-D3(BJ) (7.96)
Hybrid	ωB97X-V (1.81) ωB97X-D3 (2.56) PW6B95-D3(BJ) (2.56)	ωB97X-V (3.32) BHLYP-D3(BJ) (4.66) ωB97X-D3 (4.70)
Double hybrid	DSD-BLYP-D3(BJ) (1.91) B2PLYP-D3(BJ) (2.00) B2GPPLYP-D3(BJ) (2.05)	DSD-BLYP-D3(BJ) (3.55) B2GPPLYP-D3(BJ) (3.75) B2PLYP-D3(BJ) (3.78)

recommended in the GMTKN30 study meaning that extending GMTKN30 and considering more methods did not alter the recommendations.[7]

Again, the average WTMADs for GGAs/NGAs are smaller than for meta-GGAs/NGAs (Fig. 5). Particularly MN15L-D3(0) underperforms in this class and it gives the worst MAD in 11 out of the 21 cases. revTPSS-D3(BJ), on the other hand, is the best meta-GGA in 7 cases (Fig. 4f), followed by SCAN-D3(BJ) (5 cases). Nevertheless, their WTMADs are higher than those of the best three GGAs (Table 8). Thus, there is no overall advantage in using those for the treatment of NCI energies alone.

Again, we recommend to use DFAs of at least hybrid quality for this category. ωB97X-V clearly outperforms the other hybrids. It is the best hybrid in 5 cases and it also is the best in both WTMAD rankings of hybrids. It is more accurate than ωB97X-D3, which however is a worthwhile alternative. Among the conventional global hybrids we identify PW6B95-D3(BJ) and, surprisingly, BHLYP-D3(BJ) as good methods. The best Minnesota DFA for noncovalent interactions is M052X-D3(0), which ranks in fourth position. None of the newer developments can be recommended; for instance, the best post-2008 Minnesota DFAs are N12SX-D3(BJ) and MN15-D3(BJ), which hover around the 30th place in both WTMAD schemes (ESI†). MN12SX -D3(BJ) is one of the worst-ranking DFAs (see ESI†). We also note that the APFD approach, which had been recommended as an alternative to DFT-D3 and other dispersion corrections, cannot compete with them at all. It ranks 41st on the WTMAD-1 and 27th on the WTMAD-2 list.

Ultimately, double hybrids are to be preferred for the treatment of NCI problems. The average WTMAD-1 value, for instance, improves from 3.95 to 2.18 kcal mol^{-1} when compared to hybrids (Fig. 5). For this DFA class, we particularly recommend the DSD-BLYP-D3(BJ), B2PLYP-D3(BJ), and B2GPPLYP-D3(BJ) approaches (Table 8).

5.3.7 Looking at the entire GMTKN55 database. So far, we have discussed the different categories of GMTKN55 separately and given individual recommendations and warnings for them. This is of particular value to users that are dealing with very specialised problems; for instance, they may solely be concerned with calculating NCI energies. However, our discussion has also revealed a known problem of modern DFT, *i.e.*, that a recommendation given for one category may not necessarily be reproduced for another. We mentioned this already for XLYP-D3(BJ) in the inter- and intramolecular interactions sections, but also for MN15L-D3(0), which performed very well for basic properties and reactions of small systems, but was the worst rung-3 DFA for NCIs. Moreover, many real-life applications cannot be clearly categorised. For instance, when calculating a reaction energy, an accurate treatment of intramolecular NCIs may also become crucial. Therefore, we now proceed to the most important part of the discussion, which is a comprehensive analysis across the entire GMTKN55 database, which will lead to the clear recommendations on which DFAs to use and which to avoid when dealing with new problems.

Fig. 6 shows our best-worst perspective for each of the considered rungs on Jacob's Ladder across all test sets.

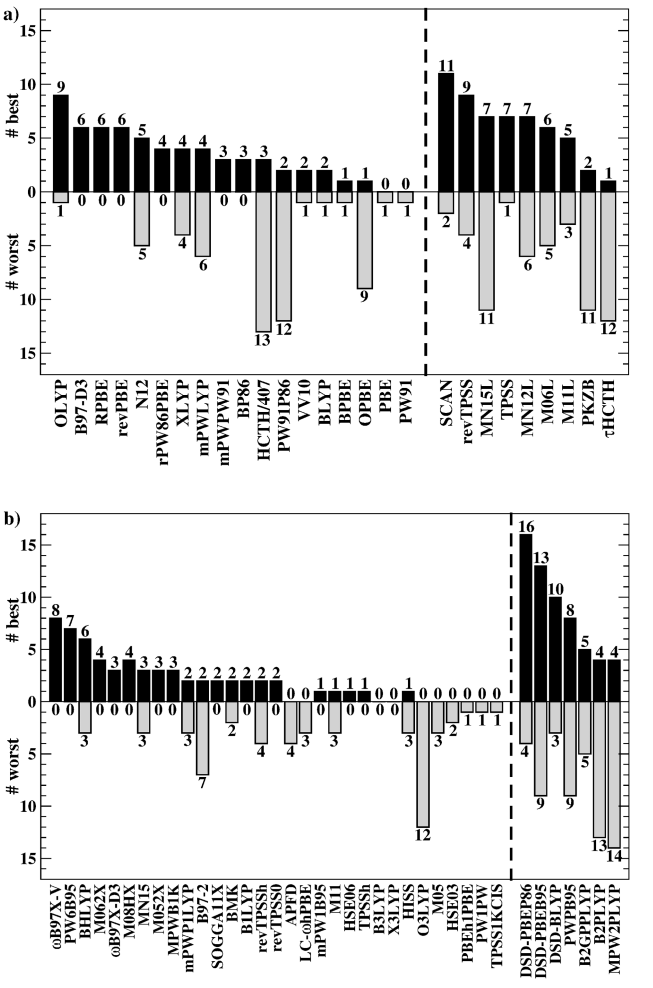

Fig. 6 Analysis of how many times a DFA yields the worst and best MAD for GMTKN55. All DFAs were dispersion corrected, but the suffix "D3" was omitted from the labels for better clarity, unless it is needed to avoid ambiguity. The analysis was carried out separately for each of the four highest rungs on Jacob's Ladder: GGA/NGA (a, left), meta-GGA/NGA (a, right), hybrid (b, left), and double hybrid (b, right).

Among the second rung, OLYP-D3(BJ) gives the best MADs in most of the cases, followed by B97-D3(BJ), revPBE-D3(BJ) and RPBE-D3(BJ). N12-D3(0) gives the best MAD in 5 cases, but also fails for the same number of test sets. Usage of OPBE-D3(BJ), PW91P86-D3(0), and HCTH/407-D3(BJ) should be avoided, as they each provide the worst MADs in 9 to 13 cases. Among the third rung, SCAN-D3(BJ) distinguishes itself from all other DFAs, as it gives the best MAD in 11 cases. MN15L-D3(0) does not seem to be overly robust. While it gives the best MAD 7 times, it also gives the worst in 11 cases. We also discourage from using PKZB-D3(0) and τHCTH-D3(BJ). Among the hybrids, ωB97X-V and PW6B95-D3(BJ) are noteworthy. They never give the worst MAD and instead offer the best outcome in 8 or 7 cases, respectively. While BHLYP-D3(BJ) seems to be surprisingly accurate for NCIs, we also note that it is less robust than the previously mentioned two methods, as it gives the worst MAD in 3 cases. O3LYP-D3(BJ) is the worst of the assessed hybrids, and it gives the highest MAD in 13 cases. Among the double hybrids, DSD-PBEP86-D3(BJ) and DSD-BLYP-D3(BJ) seem to be the most robust, while the statistics for DSD-PBEB95-D3(BJ) suffer from its disappointing performance for NCIs (best MAD in 13 cases, and worst in 9). The first-generation double hybrids B2PLYP-D3(BJ), and MPW2PLYP-D3(BJ) seem to be less accurate than the others, reflecting the progress that has been achieved in this field since their introduction. However, when comparing all 83 DFAs simultaneously, double hybrids never give the worst MAD (Fig. S4) (ESI†).

Next, we will verify if similar conclusions can be drawn from WTMADs. First of all, we confirm again that our results presented herein reproduce the Jacobs's Ladder scheme, indicating that GMTKN55 is indeed a good representative of contemporary chemical problems. The averaged WTMAD-1 for GGAs/NGAs is 5.76 kcal mol^{-1}, which is closely followed by a value of 5.62 kcal mol^{-1} for meta-GGAs/NGAs (Fig. 5). This is followed by a large reduction to 3.87 kcal mol^{-1} for hybrids and further improvement to 2.05 kcal mol^{-1} for double hybrids. The WTMAD-2 data reflect the same trend.

In the previous categories, we saw that, overall, the WTMAD-1 and WTMAD-2 schemes agreed on the best DFAs, but that they differed in functional order. Sometimes a method that was ranked as third best, may have been described as only the fourth or fifth best. However, when carrying out a comprehensive analysis of the entire set, both schemes provide the same top 3 for each DFA rung (Table 9). Fig. 7a shows a histogram for all 83 WTMAD-2 values assigned to bins of 1 kcal mol^{-1} width that shows an approximate normal-distribution behaviour. Such behaviour is even more pronounced when histograms are separately drawn for hybrid and GGA/NGA DFAs (Fig. 7b). Fig. 8 provides a graphical overview of all WTMAD-2 values, with the red bars indicating the best-performing DFAs for each rung; their actual numbers are shown in Table 9, where they are also compared with WTMAD-1 results. A figure similar to Fig. 8, but based on WTMAD-1, is shown in the ESI.† The overall order of the 83 dispersion-corrected DFAs barely changes when comparing the two different schemes. This indicates the reliability of the WTMADs introduced in this work.

Table 9 The best three DFAs for each of the four highest rungs on Jacob's Ladder for GMTKN55 according to WTMAD-1 and WTMAD-2 values (kcal mol^{-1})

Rung	WTMAD-1	WTMAD-2
GGA/NGA	revPBE-D3(BJ) (4.66) OLYP-D3(BJ) (4.75) B97-D3(BJ) (4.92)	revPBE-D3(BJ) (8.27) B97-D3(BJ) (8.55) OLYP-D3(BJ) (8.71)
Meta-GGA/NGA	SCAN-D3(BJ) (4.67) revTPSS-D3(BJ) (4.69) M06L-D3(0) (4.86)	SCAN-D3(BJ) (7.86) revTPSS-D3(BJ) (8.50) M06L-D3(0) (8.61)
Hybrid	ωB97X-V (2.32) ωB97X-D3 (2.71) M052X-D3(0) (2.73)	ωB97X-V (3.98) M052X-D3(0) (4.61) ωB97X-D3 (4.77)
Double hybrid	DSD-PBEP86-D3(BJ) (1.80) DSD-BLYP-D3(BJ) (1.81) B2GPPLYP-D3(BJ) (1.95)	DSD-BLYP-D3(BJ) (3.08) DSD-PBEP86-D3(BJ) (3.14) B2GPPLYP-D3(BJ) (3.26)

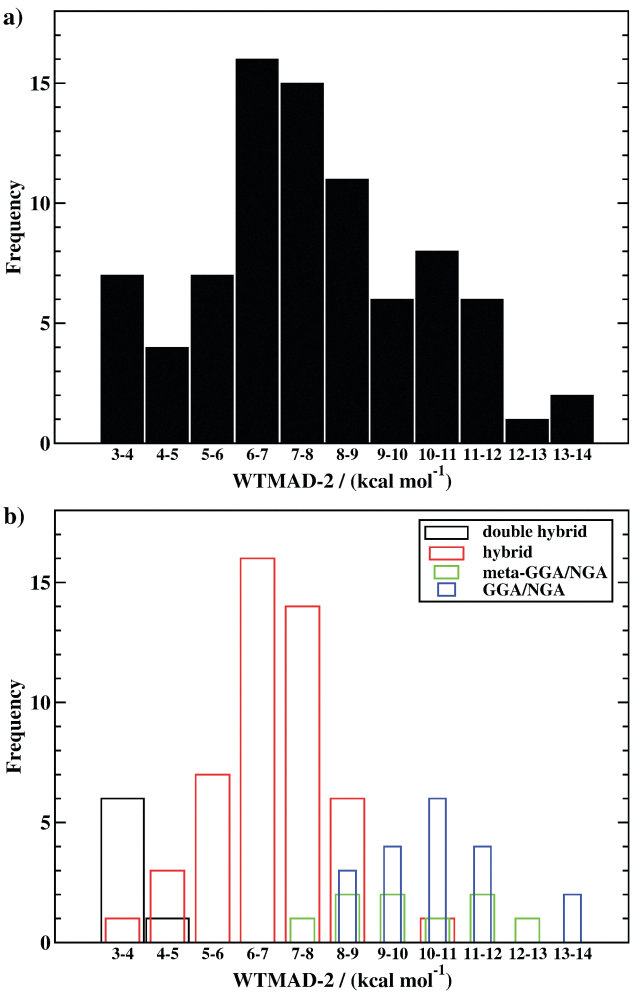

Fig. 7 Histograms (1 kcal mol^{-1} bins) showing the WTMAD-2 distributions for all 83 dispersion-corrected DFAs (a) and for each rung of Jacob's Ladder (b).

The results presented in Fig. 8 and Table 9 present our final recommendations. The best DFA of the entire study is DSD-BLYP-D3(BJ) (WTMAD-2 = 3.08 kcal mol^{-1}), closely followed by

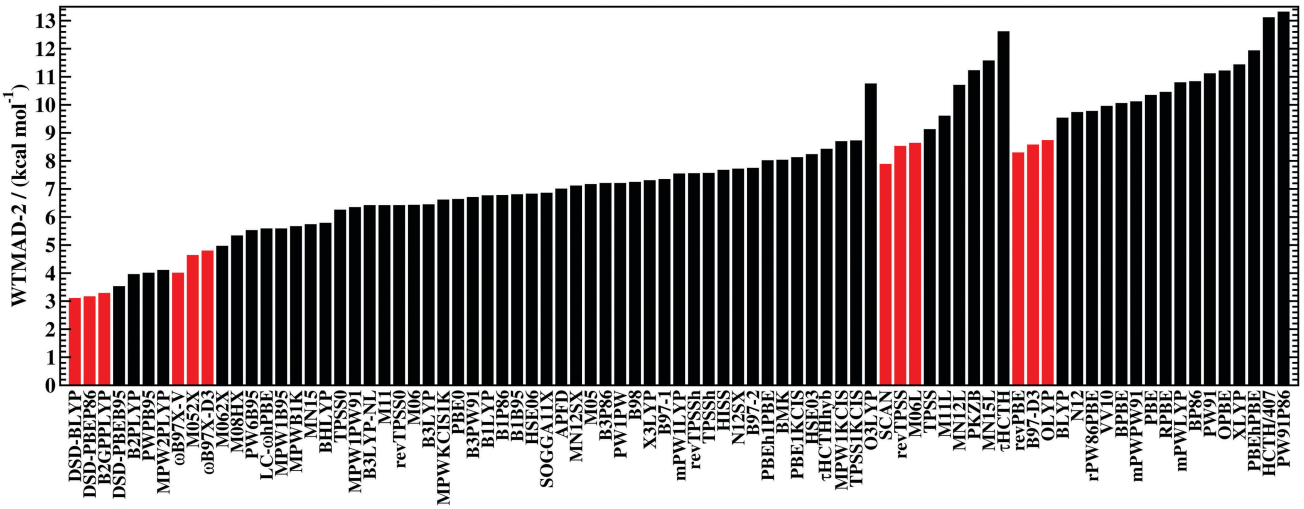

Fig. 8 Final WTMAD-2 values over the entire GMTKN55 for all assessed 83 dispersion-corrected DFAs (kcal mol^{-1}). The red bars indicate the three best approaches on their respective rung of Jacob's Ladder. The suffix "D3" was omitted in all cases, unless it is needed to avoid ambiguity.

DSD-PBEP86-D3(BJ) (3.14 kcal mol^{-1}), and B2GPPLYP-D3(BJ) (3.26 kcal mol^{-1}). Note that in the case of WTMAD-1, the two DSD methods share the nearly same value (1.80–1.81 kcal mol^{-1}). This recommendation closely resembles the previous result for GMTKN30, with the only difference that DSD-PBEB95-D3(BJ) shared the same WTMAD as the other two DSD methods.[29] Overall, any of the tested double hybrids should be preferred over most hybrids, with the only exception being the best hybrid ωB97X-V, which has a slightly better performance than the first-generation MPW2PLYP-D3(BJ) double hybrid (WTMAD-2 = 3.98 *vs.* 4.08 kcal mol^{-1}) (ESI†).

The second- and third-best hybrids are M052X-D3(0) and ωB97X-D3, but with WTMAD-2 values that are by 0.6–08 kcal mol^{-1} higher than for ωB97X-V (Table 9). In Fig. 8, it is shown how ωB97X-D3 is followed by M062X-D3(0) and M08HX-D3(0). Next follows the first conventional hybrid PW6B95-D3(BJ), which is neither range-separated nor does it depend on a large number of parameters; even ωB97X-V has 10 empirical parameters, which is 4 more than the PW6B95 exchange-correlation DFA. While the Minnesota DFAs perform very well for thermochemistry, we also have to take into account that the treatment of NCIs with them can be difficult. PW6B95-D3(BJ) may therefore be a more robust alternative with WTMAD-1 = 2.93 kcal mol^{-1} and WTMAD-2 = 5.50 kcal mol^{-1}.

The results for hybrids so far share resemblance with the recommendations for GMTKN30 made in 2011. Then, M062X-D3(0) was the best hybrid, followed by M052X-D3(0) and PW6B95-D3(0) as the best conventional, global hybrid. The older ωB97X-D, which was based on the DFT-D2 correction, followed. Considering that DFT-D2 has been replaced with DFT-D3 or the VV10 kernel, it is no surprise that ωB97X-V and ωB97X-D3 outperform the other methods, and overall we can reconfirm our initial recommendations from 2011. This is particularly noteworthy, as both the composition of the database and the way we calculate WTMADs have changed. Our results for hybrids are also noteworthy as it seems that none of the developments made on the Minnesota DFAs after 2008 reflect an overall improvement for main-group thermochemistry, when considering a large-enough database. Among those newer methods, MN15-D3(BJ) ranks the highest, however, it is only in tenth position among the hybrids. N12-SX-D3(0) ranks among the ten worst-performing hybrids.

B3LYP-D3(BJ) ranks only as the 18th-best hybrid in the WMTAD-2 scheme, followed closely by PBE0-D3(BJ) (20th). X3LYP-D3(BJ), which has been promoted as one of the best hybrids,[166,230] only ranks in 33rd position (WTMAD-2 = 7.28 kcal mol^{-1}). Note that this value increases to 14.07 kcal mol^{-1} when the dispersion correction is discarded, and we do not see any evidence that supports claims made in favour of X3LYP.[166,230] The worst assessed hybrid is O3LYP-D3(BJ); with a WTMAD-2 of 10.72 kcal mol^{-1} it is worse than the best (meta-)GGAs/NGAs.

For rung 3, we recommend SCAN-D3(BJ) (WTMAD-2 = 7.86) followed by revTPSS-D3(BJ) (WTMAD-2 = 8.50 kcal mol^{-1}) and M06L-D3(0) (WTMAD = 8.61 kcal mol^{-1}). For this rung, we observe the biggest change compared to previous recommendations, but that is mostly because in 2011 the GMTKN30 study only considered three meta-GGAs. The worst method on this rung is τHCTH-D3(BJ) with a WTMAD-2 value of 12.59 kcal mol^{-1}. When comparing the best meta-GGAs with the best GGAs, however, we note that only SCAN-D3(BJ) is noteworthy, as the best GGA revPBE-D3(BJ) has a lower WTMAD-2 result than revTPSS-D3(BJ) (8.27 kcal mol^{-1}). revPBE-D3(BJ) is followed by B97-D3(BJ) and OLYP-D3(BJ) in the list of best GGAs (Fig. 6).

Our final WTMAD-based recommendations are further backed up by the fact that they closely resemble our conclusions drawn earlier based on counting the number of best and worst MADs. We are, therefore, confident about the validity of our conclusions.

6 Summary and conclusions

We have presented the GMTKN55 benchmark database for general main group thermochemistry, kinetics and noncovalent

interactions, which is an updated and extended version of its predecessor GMTKN30. Compared to GMTKN30, it allows assessment of a larger variety of chemical problems with 55 test sets in total, 13 of which were presented here for the first time. It involves 2462 single-point calculations that are combined to 1505 relative energies, for most of which we presented new, higher-quality reference values. Indeed, we demonstrated how those newer reference values may change the outcome of a density functional approximation (DFA) ranking when compared with values of lower quality. We therefore recommend adopting the reference values herein. They can be conveniently accessed from a dedicated website that also made all geometries available for download and provides all reported data.[112]

Once again, we were able to demonstrate the importance of London dispersion in thermochemical problems and we re-emphasised the necessity of using (mostly long-range) dispersion corrections in conjunction with DFAs, even in applications that go beyond the determination of noncovalent interaction (NCI) energies. Our work also joins others that disproved the common perception that one can include London-dispersion effects into a DFA lacking nonlocal correlation terms by mere empirical parametrisation. In particular, we showed how the popular Minnesota class of DFAs benefits from dispersion corrections, also for systems in their equilibrium geometries where the electron clouds of two interacting fragments can overlap, contrary to claims in the literature.[19] However, those DFAs turned out to be less robust for the treatment of NCIs and ultimately the user fares better with using conventional DFAs corrected either with an additive dispersion correction, such as DFT-D3, or a nonlocal van der Waals kernel, such as in VV10 or ωB97X-V.

To demonstrate the benefits of the new GMTKN55 database, we conducted a comprehensive benchmark study of DFAs belonging to the four highest rungs on Jacob's Ladder. Contrary to other studies on large benchmark databases,[19,22] it was particularly important for us to base all assessed DFAs on an equal footing and we discussed only dispersion-corrected results. For this purpose, we determined and presented new parameters for the DFT-D3 dispersion correction for 35 DFAs.

We carried out a detailed study of DFAs, with numerical quadrature grids that are used in common applications and a large basis set to eliminate artificial AO basis-set related error-compensation effects to show a DFA's ''true performance''. In total, we assessed 217 variations of dispersion-corrected and -uncorrected DFAs, and then carried out a detailed analysis of 83 of them to identify robust and reliable approaches: 19 GGAs or NGAs (rung 2 of Jacob's Ladder), nine meta-GGAs/NGAs (rung 3), 48 hybrids (rung 4), and seven double hybrids (rung 5).

We divided the test sets of GMTKN55 into four main categories that we first discussed separately to identify DFAs for specific applications: basic properties and reactions of small systems, isomerisation reactions and reactions of large systems, barrier heights, and NCIs. The latter was further divided into a part dealing with intermolecular interactions, and a second part comprising intramolecular interactions.

Subsequently, we carried out a comprehensive analysis of all combined 55 sets, for which we confirmed the Jacob's Ladder scheme with higher-rung methods being more accurate. This comprehensive analysis is valuable for future real-life applications, which may touch more than one of our four categories of GMTKN55. Our final recommendations for robust, reliable and accurate DFAs are:

• Rung 2: revPBE-D3(BJ), followed by B97-D3(BJ), and OLYP-D3(BJ). If NCIs are the main area of interest, BLYP-D3(BJ) can also be used.

• Rung 3: SCAN-D3(BJ), revTPSS-D3(BJ), and M06L-D3(0). However, SCAN-D3(BJ) is the only meta-GGA studied herein that is not outperformed by the best GGAs. Note that our study focussed on energy differences for main-group molecules, and that meta-GGAs may be better than GGAs for bond lengths or main-group solids.

• Rung 4: ωB97X-V, M052X-D3(0), and ωB97X-D3. Note that while thermochemical properties are well described by the Minnesota DFA M052X-D3(0), the treatment of NCIs can pose a problem. For users that cannot apply these three methods for technical reasons, we recommend the global hybrid PW6B95-D3(BJ).

• Rung 5: DSD-BLYP-D3(BJ), DSD-PBEP86-D3(BJ), and B2GPPLYP-D3(BJ). In general, double hybrids should be the method of choice and one should refer to lower DFA rungs only if the application of double hybrids is not feasible. On modern computer architectures and with efficient implementations of the MP2 approach, double-hybrid calculations can be carried out routinely on large systems. Note that if none of the recommended double hybrids are feasible, the PWPB95-D3(BJ) method could be applied in conjunction with a Laplace transform[231] approach that brings down the formal scaling behaviour.[24] PWPB95-D3(BJ) was previously also shown to provide good results with triple-ζ AO basis sets that are statistically very similar to quadruple-ζ ones.[7]

Popular approaches, such as B3LYP, PBE0, X3LYP, BP86, PBE, and TPSS, exhibit average performance and we do not see any reasons to recommend them, despite the fact that they are available in every major molecular quantum-chemistry code. In fact, it seems that many DFAs in common programs have become obsolete for the treatment of main-group thermochemistry. Finally, it is particularly noteworthy that the majority of our recommendations are not reflected by the annual ''DFT poll''.[232] Based on the results of the previous year, the current 2017 poll only features five of the herein recommended DFAs in its list of the 20 most popular methods: revPBE, OLYP, M062X, B97-D2, and B2PLYP, *i.e.*, without or with older dispersion corrections. On the other hand, DFAs made the top-20 that did not perform well for GMTKN55: B3LYP, B3LYP-D2, BP86, HSE06, PBE, PBE0, PW91, and TPSSh.[232] When comparing that list with our analysis, one comes to the conclusion that popularity does not imply accuracy. With our study, we would like to inspire a change in the user's perception of DFT-based methods. This can be better achieved if more of the recommended methods are accessible to a broader community, which is why we would like to encourage program developers to implement the recommended approaches.

However, it should also be noted that the DFA assessment problem becomes even more complex when transition-metal

containing systems are considered for which in particular rung-2 or rung-3 methods are viable alternatives to the higher-rung methods. The setup of a reliable benchmark of similar quality and extension as GMTKN55 for transition-metal complexes is currently not possible in our opinion, but encouraging steps in this direction were published recently.[233,234]

We are aware that the "zoo" of DFT-based methods is large and continues to grow. While we could not consider every possible DFA in this study,[235–237] we are confident that we have provided a valuable starting point for future studies that would also assess those. Our two newly introduced schemes to represent a DFA's robustness and accuracy with one number (weighted total mean absolute deviation) allowed us to provide a comprehensive ranking of methods and we challenge method developers to identify new approaches that yield lower WTMAD-1/2 values. Work in our groups is also currently pointing into that direction including accuracy-competitive low-cost approaches.

Conflicts of interest

There are no conflicts of interest to declare.

Acknowledgements

We are in debt to Asst. Prof. Amir Karton for generously providing us with some of his W*n*-F12 reference values and for insightful discussions on the Weizmann composite protocols. We thank Prof. Jan Martin for providing us with a reference value for the 27th reaction in the WATER27 set. We are also grateful to Prof. Donald G. Truhlar for providing us with the molecular geometries of the YBDE18 set. We kindly thank Jakob Seibert for creating the TOC graphic of the article, Christoph Bannwarth for fruitful discussions, and Prof. John Perdew for useful feedback on this study. LG acknowledges funding through an Australian Research Council Discovery Early Career Researcher Award (project ID: DE140100550) and allocation of generous computing resources by the Victorian Life Science Computation Initiative/ Melbourne Bioinformatics (project ID: RA0005) and by the National Computational Infrastructure (NCI) National Facility within the National Computational Merit Allocation Scheme (project ID: fk5). This work was also funded by the Deutsche Forschungsgemeinschaft in the framework of the "Gottfried Wilhelm Leibniz-Preis" to SG. AN acknowledges an Australian Government Research Training Program Scholarship and a Melbourne Research Scholarship.

References

1 W. Kohn and L. J. Sham, *Phys. Rev.*, 1965, **140**, A1133–A1138.
2 A. D. Becke, *J. Chem. Phys.*, 2014, **140**, 18A301.
3 L. A. Curtiss, K. Raghavachari, G. W. Trucks and J. A. Pople, *J. Chem. Phys.*, 1991, **94**, 7221–7230.
4 L. A. Curtiss, K. Raghavachari, P. C. Redfern and J. A. Pople, *J. Chem. Phys.*, 1997, **106**, 1063–1079.
5 L. A. Curtiss, K. Raghavachari, P. C. Redfern, V. Rassolov and J. A. Pople, *J. Chem. Phys.*, 1998, **109**, 7764–7776.
6 L. A. Curtiss, P. C. Redfern and K. Raghavachari, *J. Chem. Phys.*, 2005, **123**, 124107.
7 L. Goerigk and S. Grimme, *Phys. Chem. Chem. Phys.*, 2011, **13**, 6670–6688.
8 A. Karton, S. Daon, J. M. L. Martin and B. Ruscic, *Chem. Phys. Lett.*, 2011, **510**, 165–178.
9 J. T. Margraf, D. S. Ranasinghe and R. J. Bartlett, *Phys. Chem. Chem. Phys.*, 2017, **19**, 9798–9805.
10 Y. Zhao, B. J. Lynch and D. G. Truhlar, *Phys. Chem. Chem. Phys.*, 2005, **17**, 43–52.
11 Y. Zhao, N. González-García and D. G. Truhlar, *J. Phys. Chem. A*, 2005, **109**, 2012–2018.
12 S. Grimme, M. Steinmetz and M. Korth, *J. Org. Chem.*, 2007, **72**, 2118–2126.
13 P. Jurečka, J. Sponer, J. Cerny and P. Hobza, *Phys. Chem. Chem. Phys.*, 2006, **8**, 1985–1993.
14 J. Rezac, P. Jurecka, K. E. Riley, J. Cerny, H. Valdes, K. Pluhackova, K. Berka, T. Rezac, M. Pitonak, J. Vondrasek and P. Hobza, *Collect. Czech. Chem. Commun.*, 2008, **73**, 1261–1270.
15 M. Korth and S. Grimme, *J. Chem. Theory Comput.*, 2009, **5**, 993–1003.
16 B. J. Lynch and D. G. Truhlar, *J. Phys. Chem. A*, 2003, **107**, 8996–8999.
17 Y. Zhao and D. G. Truhlar, *J. Chem. Theory Comput.*, 2005, **1**, 415–432.
18 Y. Zhao and D. G. Truhlar, *Theor. Chem. Acc.*, 2008, **120**, 215–241.
19 R. Peverati and D. G. Truhlar, *Philos. Trans. R. Soc., A*, 2014, **372**, 20120476.
20 H. S. Yu, W. Zhang, P. Verma, X. He and D. G. Truhlar, *Phys. Chem. Chem. Phys.*, 2015, **17**, 12146–12160.
21 H. S. Yu, X. He and D. G. Truhlar, *J. Chem. Theory Comput.*, 2016, **12**, 1280–1293.
22 H. S. Yu, X. He, S. L. Li and D. G. Truhlar, *Chem. Sci.*, 2016, **7**, 5032–5051.
23 L. Goerigk and S. Grimme, *J. Chem. Theory Comput.*, 2010, **6**, 107–126.
24 L. Goerigk and S. Grimme, *J. Chem. Theory Comput.*, 2011, **7**, 291–309.
25 N. Mardirossian and M. Head-Gordon, *J. Chem. Theory Comput.*, 2016, **12**, 4303–4325.
26 N. Mardirossian and M. Head-Gordon, *Mol. Phys.*, 2017, **115**, 2315–2372.
27 R. Peverati and D. G. Truhlar, *J. Chem. Theory Comput.*, 2012, **8**, 2310–2319.
28 S. Grimme, *J. Chem. Phys.*, 2006, **124**, 034108.
29 L. Goerigk and S. Grimme, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2014, **4**, 576–600.
30 S. Kozuch, D. Gruzman and J. M. L. Martin, *J. Phys. Chem. C*, 2010, **114**, 20801–20808.
31 S. Kozuch and J. M. L. Martin, *Phys. Chem. Chem. Phys.*, 2011, **13**, 20104–20107.

32 S. Kozuch and J. M. L. Martin, *J. Comput. Chem.*, 2013, **34**, 2327–2344.
33 S. Grimme, J. Antony, S. Ehrlich and H. Krieg, *J. Chem. Phys.*, 2010, **132**, 154104.
34 Y. Zhao and D. G. Truhlar, *Theor. Chem. Acc.*, 2008, **120**, 215–241.
35 Y. Zhao, N. E. Schultz and D. G. Truhlar, *J. Chem. Theory Comput.*, 2006, **2**, 364–382.
36 Y. Zhao and D. G. Truhlar, *J. Phys. Chem. A*, 2005, **109**, 5656–5667.
37 J.-D. Chai and M. Head-Gordon, *Phys. Chem. Chem. Phys.*, 2008, **10**, 6615–6620.
38 A. D. Becke, *J. Chem. Phys.*, 1993, **98**, 5648–5652.
39 P. J. Stephens, F. J. Devlin, C. F. Chabalowski and M. J. Frisch, *J. Phys. Chem.*, 1994, **98**, 11623–11627.
40 Y. Zhang and W. Yang, *Phys. Rev. Lett.*, 1998, **80**, 890.
41 S. Grimme, *J. Comput. Chem.*, 2006, **27**, 1787–1799.
42 L. Goerigk, H. Kruse and S. Grimme, *ChemPhysChem*, 2011, **12**, 3421–3433.
43 L. Goerigk, *J. Phys. Chem. Lett.*, 2015, **6**, 3891–3896.
44 S. Grimme, S. Ehrlich and L. Goerigk, *J. Comput. Chem.*, 2011, **32**, 1456–1465.
45 H. Eshuis and F. Furche, *J. Phys. Chem. Lett.*, 2011, **2**, 983–989.
46 W. Hujo and S. Grimme, *J. Chem. Theory Comput.*, 2011, **7**, 3866–3871.
47 M. Korth and W. Thiel, *J. Chem. Theory Comput.*, 2011, **7**, 2929–2936.
48 H. Kruse, L. Goerigk and S. Grimme, *J. Org. Chem.*, 2012, **77**, 10824–10834.
49 D. Bousquet, E. Bremond, J. C. Sancho-Garcia, I. Ciofini and C. Adamo, *J. Chem. Theory Comput.*, 2013, **9**, 3444–3452.
50 L. Goerigk, *J. Chem. Theory Comput.*, 2014, **10**, 968–980.
51 B. Chan, L. Goerigk and L. Radom, *J. Comput. Chem.*, 2016, **37**, 183–193.
52 J. A. Montgomery, M. J. Frisch, J. W. Ochterski and G. A. Petersson, *J. Chem. Phys.*, 1999, **110**, 2822–2827.
53 J. A. Montgomery Jr., M. J. Frisch, J. W. Ochterski and G. A. Petersson, *J. Chem. Phys.*, 2000, **112**, 6532–6542.
54 A. Karton and J. M. L. Martin, *J. Chem. Phys.*, 2012, **136**, 124114.
55 A. Karton and L. Goerigk, *J. Comput. Chem.*, 2015, **36**, 622–632.
56 A. Karton, A. Tarnopolsky, J. F. Lamere, G. C. Schatz and J. M. L. Martin, *J. Phys. Chem. A*, 2008, **112**, 12868–12886.
57 S. Parthiban and J. M. L. Martin, *J. Chem. Phys.*, 2001, **114**, 6014–6029.
58 Y. Zhao and D. G. Truhlar, *J. Phys. Chem. A*, 2006, **110**, 10478–10486.
59 H. Yu and D. G. Truhlar, *J. Chem. Theory Comput.*, 2015, **11**, 2968–2983.
60 Y. Zhao, H. T. Ng, R. Peverati and D. G. Truhlar, *J. Chem. Theory Comput.*, 2012, **8**, 2824–2834.
61 S. Grimme, H. Kruse, L. Goerigk and G. Erker, *Angew. Chem., Int. Ed.*, 2010, **49**, 1402–1405.
62 A. Karton, private communication.
63 J. Friedrich and J. Hänchen, *J. Chem. Theory Comput.*, 2013, **9**, 5381–5394.
64 J. Friedrich, *J. Chem. Theory Comput.*, 2015, **11**, 3596–3609.
65 S. Grimme, C. Mück-Lichtenfeld, E.-U. Würthwein, A. W. Ehlers, T. P. M. Goumans and K. Lammertsma, *J. Phys. Chem. A*, 2006, **110**, 2583–2586.
66 M. Piacenza and S. Grimme, *J. Comput. Chem.*, 2004, **25**, 83–99.
67 H. L. Woodcock, H. F. Schaefer III and P. R. Schreiner, *J. Phys. Chem. A*, 2002, **106**, 11923–11931.
68 P. R. Schreiner, A. A. Fokin, R. A. Pascal and A. de Meijere, *Org. Lett.*, 2006, **8**, 3635–3638.
69 C. Lepetit, H. Chermette, M. Gicquel, J.-L. Heully and R. Chauvin, *J. Phys. Chem. A*, 2007, **111**, 136–149.
70 J. S. Lee, *J. Phys. Chem. A*, 2005, **109**, 11927–11932.
71 A. Karton and J. M. Martin, *Mol. Phys.*, 2012, **110**, 2477–2491.
72 Y. Zhao, O. Tishchenko, J. R. Gour, W. Li, J. J. Lutz, P. Piecuch and D. Truhlar, *J. Phys. Chem. A*, 2009, **113**, 5786–5799.
73 D. Manna and J. M. L. Martin, *J. Phys. Chem. A*, 2016, **120**, 153–160.
74 E. R. Johnson, P. Mori-Sánchez, A. J. Cohen and W. Yang, *J. Chem. Phys.*, 2008, **129**, 204112.
75 F. Neese, T. Schwabe, S. Kossmann, B. Schirmer and S. Grimme, *J. Chem. Theory Comput.*, 2009, **5**, 3060–3073.
76 S. N. Steinmann, G. Csonka and C. Carminboeuf, *J. Chem. Theory Comput.*, 2009, **5**, 2950–2958.
77 H. Krieg and S. Grimme, *Mol. Phys.*, 2010, **108**, 2655–2666.
78 L.-J. Yu and A. Karton, *Chem. Phys.*, 2014, **441**, 166–177.
79 R. Huenerbein, B. Schirmer, J. Moellmann and S. Grimme, *Phys. Chem. Chem. Phys.*, 2010, **12**, 6940–6948.
80 R. Sure, A. Hansen, P. Schwerdtfeger and S. Grimme, *Phys. Chem. Chem. Phys.*, 2017, **19**, 14296–14305.
81 V. Guner, K. S. Khuong, A. G. Leach, P. S. Lee, M. D. Bartberger and K. N. Houk, *J. Phys. Chem. A*, 2003, **107**, 11445–11459.
82 D. H. Ess and K. N. Houk, *J. Phys. Chem. A*, 2005, **109**, 9542–9553.
83 T. C. Dinadayalane, R. Vijaya, A. Smitha and G. N. Sastry, *J. Phys. Chem. A*, 2002, **106**, 1627–1633.
84 L. Goerigk and R. Sharma, *Can. J. Chem.*, 2016, **94**, 1133–1143.
85 A. Karton, R. J. O'Reilly, B. Chan and L. Radom, *J. Chem. Theory Comput.*, 2012, **8**, 3128–3136.
86 A. Karton, R. J. O'Reilly and L. Radom, *J. Phys. Chem. A*, 2012, **116**, 4211–4221.
87 M. S. Marshall, L. A. Burns and C. D. Sherrill, *J. Chem. Phys.*, 2011, **135**, 194102.
88 J. Řezáč, K. E. Riley and P. Hobza, *J. Chem. Theory Comput.*, 2011, **7**, 2427–2438.
89 V. S. Bryantsev, M. S. Diallo, A. C. T. van Duin and W. A. Goddard III, *J. Chem. Theory Comput.*, 2009, **5**, 1016–1026.
90 D. Manna, M. K. Kesharwani, N. Sylvetsky and J. M. L. Martin, *J. Chem. Theory Comput.*, 2017, **13**, 3136–3152.
91 D. Setiawan, E. Kraka and D. Cremer, *J. Phys. Chem. A*, 2015, **119**, 1642–1656.

92 S. Kozuch and J. M. L. Martin, *J. Chem. Theory Comput.*, 2013, **9**, 1918–1931.
93 J. Rezac, K. E. Riley and P. Hobza, *J. Chem. Theory Comput.*, 2012, **8**, 4285–4292.
94 K. U. Lao, R. Schäffer, G. Jansen and J. M. Herbert, *J. Chem. Theory Comput.*, 2015, **11**, 2473–2486.
95 T. Schwabe and S. Grimme, *Phys. Chem. Chem. Phys.*, 2007, **9**, 3397–3406.
96 S. Grimme, *Angew. Chem., Int. Ed.*, 2006, **45**, 4460–4464.
97 D. Gruzman, A. Karton and J. M. L. Martin, *J. Phys. Chem. A*, 2009, **113**, 11974–11983.
98 M. K. Kesharwani, A. Karton and J. M. L. Martin, *J. Chem. Theory Comput.*, 2016, **12**, 444–454.
99 J. J. Wilke, M. C. Lind, H. F. Schaefer III, A. G. Császár and W. D. Allen, *J. Chem. Theory Comput.*, 2009, **5**, 1511–1523.
100 D. Řeha, H. Valdes, J. Vondrasek, P. Hobza, A. Abu-Riziq, B. Crews and M. S. de Vries, *Chem. – Eur. J.*, 2005, **11**, 6803–6817.
101 L. Goerigk, A. Karton, J. M. L. Martin and L. Radom, *Phys. Chem. Chem. Phys.*, 2013, **15**, 7028–7031.
102 U. R. Fogueri, S. Kozuch, A. Karton and J. M. L. Martin, *J. Phys. Chem. A*, 2013, **117**, 2269–2277.
103 G. I. Csonka, A. D. French, G. P. Johnson and C. A. Stortz, *J. Chem. Theory Comput.*, 2009, **5**, 679–692.
104 H. Kruse, A. Mladek, K. Gkionis, A. Hansen, S. Grimme and J. Sponer, *J. Chem. Theory Comput.*, 2015, **11**, 4972–4991.
105 S. Kozuch, S. M. Bachrach and J. M. Martin, *J. Phys. Chem. A*, 2014, **118**, 293–303.
106 A. Karton, E. Rabinovich, J. M. L. Martin and B. Ruscic, *J. Chem. Phys.*, 2006, **125**, 144108.
107 K. Raghavachari, G. W. Trucks, J. A. Pople and M. Head-Gordon, *Chem. Phys. Lett.*, 1989, **157**, 479–483.
108 G. Knizia, T. B. Adler and H.-J. Werner, *J. Chem. Phys.*, 2009, **130**, 054104.
109 C. Riplinger, B. Sandhoefer, A. Hansen and F. Neese, *J. Chem. Phys.*, 2013, **139**, 134101.
110 S. F. Boys and F. Bernardi, *Mol. Phys.*, 1970, **19**, 553–566.
111 J. M. L. Martin and S. Parthiban, in *Quantum-Mechanical Prediction of Thermochemical Data*, ed. J. Cioslowski, Kluwer, Dordrecht, 2001, pp. 31–65.
112 http://www.thch.uni-bonn.de/GMTKN55.
113 J. Tao, J. P. Perdew, V. N. Staroverov and G. E. Scuseria, *Phys. Rev. Lett.*, 2003, **91**, 146401.
114 F. Weigend and R. Ahlrichs, *Phys. Chem. Chem. Phys.*, 2005, **7**, 3297–3305.
115 T. H. Dunning, Jr., *J. Chem. Phys.*, 1989, **90**, 1007–1023.
116 S. Grimme, J. G. Brandenburg, C. Bannwarth and A. Hansen, *J. Chem. Phys.*, 2015, **143**, 054107.
117 S. Grimme, *J. Chem. Phys.*, 2003, **118**, 9095–9102.
118 R. A. Kendall, T. H. Dunning and R. J. Harrison, *J. Chem. Phys.*, 1992, **96**, 6796–6806.
119 C. Adamo and V. Barone, *J. Chem. Phys.*, 1999, **110**, 6158–6170.
120 M. Ernzerhof and G. E. Scuseria, *J. Chem. Phys.*, 1999, **110**, 5029–5036.
121 A. Halkier, T. Helgaker, P. Jørgensen, W. Klopper and J. Olsen, *Chem. Phys. Lett.*, 1999, **302**, 437.
122 A. Halkier, T. Helgaker, P. Jørgensen, W. Klopper, H. Koch, J. Olsen and A. K. Wilson, *Chem. Phys. Lett.*, 1998, **286**, 243–252.
123 K. A. Peterson and T. H. Dunning, Jr., *J. Chem. Phys.*, 2002, **117**, 10548–10560.
124 K. A. Peterson and K. E. Yousaf, *J. Chem. Phys.*, 2010, **133**, 174116.
125 B. P. Prascher, D. E. Woon, K. A. Peterson, T. H. Dunning, Jr and A. K. Wilson, *Theor. Chem. Acc.*, 2011, **128**, 69–82.
126 S. Grimme, *Angew. Chem., Int. Ed.*, 2013, **52**, 6306–6312.
127 T. Risthaus, M. Steinmetz and S. Grimme, *J. Comput. Chem.*, 2014, **35**, 1509–1516.
128 D. G. Liakos, M. Sparta, M. K. Kesharwani, J. M. L. Martin and F. Neese, *J. Chem. Theory Comput.*, 2015, **11**, 1525–1539.
129 S. Grimme, *J. Comput. Chem.*, 2003, **24**, 1529.
130 S. Grimme, C. Bannwarth and P. Shushkov, *J. Chem. Theory Comput.*, 2017, **13**, 1989–2009.
131 A. Klamt and G. Schürmann, *J. Chem. Soc., Perkin Trans. 2*, 1993, 799–805.
132 T. Takatani, E. G. Hohenstein, M. Malagoli, M. S. Marshall and C. D. Sherrill, *J. Chem. Phys.*, 2010, **132**, 144104.
133 J. M. L. Martin, private communication.
134 S. Zahn, D. R. MacFarlane and E. I. Izgorodina, *Phys. Chem. Chem. Phys.*, 2013, **15**, 13664–13675.
135 Y. Yuan, M. J. L. Mills, P. L. A. Popelier and F. Jensen, *J. Phys. Chem. A*, 2014, **118**, 7876–7891.
136 J. G. Hill, S. Mazumder and K. A. Peterson, *J. Chem. Phys.*, 2010, **132**, 054108.
137 J. P. Perdew, K. Burke and M. Ernzerhof, *Phys. Rev. Lett.*, 1996, **77**, 3865–3868.
138 A. D. Becke, *Phys. Rev. A: At., Mol., Opt. Phys.*, 1988, **38**, 3098–3100.
139 J. P. Perdew, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1986, **33**, 8822–8824.
140 J. P. Perdew, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1986, **34**, 7406.
141 C. Lee, W. Yang and R. G. Parr, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1988, **37**, 785–789.
142 B. Miehlich, A. Savin, H. Stoll and H. Preuss, *Chem. Phys. Lett.*, 1989, **157**, 200–206.
143 A. D. Becke, *J. Chem. Phys.*, 1993, **98**, 1372–1377.
144 V. N. Staroverov, G. E. Scuseria, J. Tao and J. P. Perdew, *J. Chem. Phys.*, 2003, **119**, 12129–12137.
145 J. C. Slater, *Phys. Rev.*, 1951, **81**, 385–390.
146 S. J. Vosko, L. Wilk and M. Nusair, *Can. J. Phys.*, 1980, **58**, 1200–1211.
147 J. P. Perdew and Y. Wang, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1992, **45**, 13244–13249.
148 J. P. Perdew, A. Ruzsinszky, G. I. Csonka, O. A. Vydrov, G. E. Scuseria, L. A. Constantin, X. Zhou and K. Burke, *Phys. Rev. Lett.*, 2008, **100**, 136406.
149 M. Swart, M. Solà and F. M. Bickelhaupt, *J. Chem. Phys.*, 2009, **131**, 094103.
150 Y. Zhao and D. G. Truhlar, *J. Phys. Chem. A*, 2007, **110**, 13126–13130.

151 T. Yanai, D. P. Tew and N. C. Handy, *Chem. Phys. Lett.*, 2004, **393**, 51–57.
152 O. A. Vydrov and G. E. Scuseria, *J. Chem. Phys.*, 2006, **125**, 234109.
153 O. A. Vydrov and T. Van Voorhis, *J. Chem. Phys.*, 2010, **133**, 244103.
154 Y.-S. Lin, G.-D. Li, S.-P. Mao and J.-D. Chai, *J. Chem. Theory Comput.*, 2013, **9**, 263–272.
155 N. Mardirossian and M. Head-Gordon, *Phys. Chem. Chem. Phys.*, 2014, **16**, 9904–9924.
156 J. Sun, A. Ruzsinszky and J. P. Perdew, *Phys. Rev. Lett.*, 2015, **115**, 036402.
157 A. D. Boese and N. C. Handy, *J. Chem. Phys.*, 2001, **114**, 5497–5503.
158 M. Ernzerhof and J. P. Perdew, *J. Chem. Phys.*, 1998, **109**, 3313–3320.
159 Y. Zhao and D. G. Truhlar, *J. Chem. Theory Comput.*, 2008, **4**, 1849–1868.
160 T. V. Voorhis and G. E. Scuseria, *J. Chem. Phys.*, 1998, **109**, 400–410.
161 R. Peverati, Y. Zhao and D. G. Truhlar, *J. Phys. Chem. Lett.*, 2011, **2**, 1991–1997.
162 P. M. W. Gill, *Mol. Phys.*, 1996, **89**, 433–445.
163 J. P. Perdew, in *Proceedings of the 21st Annual International Symposium on the Electronic Structure ofSolids*, ed. P. Ziesche and H. Eschrig, Akademie Verlag, Berlin, 1991, p. 11.
164 B. Hammer, L. B. Hansen and J. K. Norskov, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1999, **59**, 7413–7421.
165 J. R. Reimers, D. Panduwinata, J. Visser, Y. Chin, C. Tang, L. Goerigk, M. J. Ford, M. Sintic, T.-J. Sum, M. J. J. Coenen, B. L. M. Hendriksen, J. A. A. W. Elemans, N. S. Hush and M. J. Crossley, *Proc. Natl. Acad. Sci. U. S. A.*, 2015, **112**, E6101–E6110.
166 X. Xu and W. A. Goddard, *Proc. Natl. Acad. Sci. U. S. A.*, 2004, **101**, 2673–2677.
167 W.-M. How, A. J. Cohen and N. C. Handy, *Chem. Phys. Lett.*, 2001, **341**, 319.
168 F. A. Hamprecht, A. J. Cohen, D. J. Tozer and N. C. Handy, *J. Chem. Phys.*, 1998, **109**, 6264–6271.
169 N. C. Handy and A. J. Cohen, *Mol. Phys.*, 2001, **99**, 403–412.
170 P. J. Wilson, T. J. Bradley and D. J. Tozer, *J. Chem. Phys.*, 2001, **115**, 9233–9242.
171 H. L. Schmider and A. D. Becke, *J. Chem. Phys.*, 1998, **108**, 9624–9631.
172 T. M. Henderson, A. F. Izmaylov, G. E. Scuseria and A. Savin, *J. Chem. Theory Comput.*, 2008, **4**, 1254–1262.
173 C. Adamo and V. Barone, *J. Chem. Phys.*, 1998, **108**, 664–675.
174 J. Heyd, G. E. Scuseria and M. Ernzerhof, *J. Chem. Phys.*, 2003, **118**, 8207–8215.
175 J. Heyd, G. E. Scuseria and M. Ernzerhof, *J. Chem. Phys.*, 2006, **124**, 219906.
176 J. Moellmann and S. Grimme, *J. Phys. Chem. C*, 2014, **118**, 7615–7621.
177 E. D. Murray, K. Lee and D. C. Langreth, *J. Chem. Theory Comput.*, 2009, **5**, 2754–2762.
178 J. P. Perdew, A. Ruzsinszky, G. I. Csonka, L. A. Constantin and J. Sun, *Phys. Rev. Lett.*, 2009, **103**, 026403.
179 S. Grimme, *J. Phys. Chem. A*, 2005, **109**, 3067–3077.
180 Y. Zhao, B. J. Lynch and D. G. Truhlar, *Phys. Chem. Chem. Phys.*, 2005, **7**, 43–52.
181 A. D. Boese and J. M. L. Martin, *J. Chem. Phys.*, 2004, **121**, 3405–3416.
182 A. D. Boese and N. C. Handy, *J. Chem. Phys.*, 2002, **116**, 9559–9569.
183 Y. Zhao, N. E. Schultz and D. G. Truhlar, *J. Chem. Phys.*, 2005, **123**, 161103.
184 J. P. Perdew, S. Kurth, A. Zupan and P. Blaha, *Phys. Rev. Lett.*, 1999, **82**, 2544–2547.
185 J. G. Brandenburg, J. E. Bates, J. Sun and J. P. Perdew, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2016, **94**, 115144.
186 R. Peverati and D. G. Truhlar, *J. Phys. Chem. Lett.*, 2011, **2**, 2810–2817.
187 Y. Zhao and D. G. Truhlar, *J. Chem. Phys.*, 2006, **125**, 194101.
188 R. Peverati and D. G. Truhlar, *J. Chem. Phys.*, 2011, **135**, 191102.
189 R. Peverati and D. G. Truhlar, *J. Phys. Chem. Lett.*, 2012, **3**, 117–124.
190 R. Peverati and D. G. Truhlar, *Phys. Chem. Chem. Phys.*, 2012, **14**, 16187–16191.
191 R. Peverati and D. G. Truhlar, *Phys. Chem. Chem. Phys.*, 2012, **14**, 13171–13174.
192 T. M. Henderson, A. F. Izmaylov, G. Scalmani and G. E. Scuseria, *J. Chem. Phys.*, 2009, **131**, 044108.
193 A. Austin, G. A. Petersson, M. J. Frisch, F. J. Dobek, G. Scalmani and K. Throssell, *J. Chem. Theory Comput.*, 2012, **8**, 4989–5007.
194 A. D. Becke, *J. Chem. Phys.*, 1996, **104**, 1040–1046.
195 S. Grimme and T. Schwabe, *Phys. Chem. Chem. Phys.*, 2006, **8**, 4398–4401.
196 Y. Zhao and D. G. Truhlar, *J. Phys. Chem. A*, 2004, **108**, 6908.
197 F. Neese, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2012, **2**, 73–78.
198 ORCA – an ab initio, density functional and semiempirical program package, V. 4.0.0, F. Neese, MPI für Chemische Energiekonversion, Mülheim A. D. Ruhr, Germany, 2017.
199 TURBOMOLE, V7. 1.1, 2016, developed by University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, 1989–2007, available from http://www.turbomole.com.
200 R. Ahlrichs, M. Bär, M. Häser, H. Horn and C. Kölmel, *Chem. Phys. Lett.*, 1989, **162**, 165–169.
201 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci and G. A. Peterssonet *al.*, *Gaussian 09 Revisions D.01 and E.01*, Gaussian, Inc., Wallingford CT, 2009.
202 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci and G. A. Peterssonet *al.*, *Gaussian 16 Revision A.03*, Gaussian, Inc., Wallingford CT, 2016.
203 A. D. Becke and E. R. Johnson, *J. Chem. Phys.*, 2005, **123**, 154101.

204 E. R. Johnson and A. D. Becke, *J. Chem. Phys.*, 2005, **123**, 024101.
205 L. Gráfová, M. Pitoák, J. Řezáč and P. Hobza, *J. Chem. Theory Comput.*, 2010, **6**, 2365–2376.
206 D. E. Taylor, J. G. Ángyán, G. Galli, C. Zhang, F. Gygi, K. Hirao, J. W. Song, K. Rahul, O. A. von Lilienfeld, R. Podeszwa, I. W. Bulik, T. M. Henderson, G. E. Scuseria, J. Toulouse, R. Peverati, D. G. Truhlar and K. Szalewicz, *J. Chem. Phys.*, 2016, **145**, 124105.
207 D. G. A. Smith, L. A. Burns, K. Patkowski and C. D. Sherrill, *J. Phys. Chem. Lett.*, 2016, **7**, 2197–2203.
208 K. Eichkorn, O. Treutler, H. Öhm, M. Häser and R. Ahlrichs, *Chem. Phys. Lett.*, 1995, **240**, 283–290.
209 K. Eichkorn, F. Weigend, O. Treutler and R. Ahlrichs, *Theor. Chem. Acc.*, 1997, **97**, 119–124.
210 F. Weigend, M. Häser, H. Patzelt and R. Ahlrichs, *Chem. Phys. Lett.*, 1998, **294**, 143–152.
211 E. R. Johnson, A. Becke, C. D. Sherrill and G. A. DiLabio, *J. Chem. Phys.*, 2009, **131**, 034111.
212 L. Fusti-Molnar, X. He, B. Wang and K. M. Merz Jr., *J. Chem. Phys.*, 2009, **131**, 065102.
213 C. A. Jiménez-Hoyos, B. G. Janesko and G. E. Scuseria, *Phys. Chem. Chem. Phys.*, 2008, **10**, 6621–6629.
214 S. E. Wheeler and K. N. Houk, *J. Chem. Theory Comput.*, 2010, **6**, 395–404.
215 R. Izsak and F. Neese, *J. Chem. Phys.*, 2011, **135**, 144105.
216 F. Weigend, A. Köhn and C. Hättig, *J. Chem. Phys.*, 2002, **116**, 3175–3183.
217 T. Schwabe, *J. Phys. Chem. A*, 2013, **117**, 2879–2883.
218 H.-J. Werner, P. J. Knowles, G. Knizia, F. R. Manby and M. Schütz, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2012, **2**, 242–253.
219 H.-J. Werner, P. J. Knowles, G. Knizia, F. R. Manby and M. Schütz *et al.*, *MOLPRO, version 2015.1, a package of ab initio programs*, 2015, see http://www.molpro.net.
220 C. Hättig, D. P. Tew and A. Köhn, *J. Chem. Phys.*, 2010, **132**, 231102.
221 C. Riplinger, P. Pinski, U. Becker, E. F. Valeev and F. Neese, *J. Chem. Phys.*, 2016, **144**, 024109.
222 S. Grimme, A. Hansen, J. G. Brandenburg and C. Bannwarth, *Chem. Rev.*, 2016, **116**, 5105–5154.
223 S. Grimme and P. R. S. Schreiner, *Angew. Chem., Int. Ed.*, 2011, **50**, 12639–12642.
224 S. Grimme and M. Steinmetz, *Phys. Chem. Chem. Phys.*, 2013, **15**, 16031–16042.
225 W. Hujo and S. Grimme, *J. Chem. Theory Comput.*, 2013, **9**, 308–315.
226 L. Goerigk and J. R. Reimers, *J. Chem. Theory Comput.*, 2013, **9**, 3240–3251.
227 L. Goerigk, C. A. Collyer and J. R. Reimers, *J. Phys. Chem. B*, 2014, **118**, 14612–14626.
228 S. Rösel, H. Quanz, C. Logemann, J. Becker, E. Mossou, L. Cañadillas-Delgado, E. Caldeweyher, S. Grimme and P. R. Schreiner, *J. Am. Chem. Soc.*, 2017, **139**, 7428–7431.
229 S. Grimme, R. Huenerbein and S. Ehrlich, *ChemPhysChem*, 2011, **12**, 1258–1261.
230 S. Zheng, S. Xu, G. Wang, Q. Tang, X. Jiang, Z. Li, Y. Xu, R. Wang and F. Lin, *J. Chem. Inf. Model.*, 2017, **57**, 1535–1547.
231 J. Almlöf, *Chem. Phys. Lett.*, 1991, **181**, 319–320.
232 DFT Poll 2017 organised by M. Swart, M. Bickelhaupt and M. Duran, see http://www.marcelswart.eu/dft-poll/2017.html#start.
233 J. Wang, L. Liu and A. K. Wilson, *J. Phys. Chem. A*, 2016, **120**, 737–746.
234 P. Verma, Z. Varga, J. E. M. N. Klein, C. J. Cramer, L. Que and D. G. Truhlar, *Phys. Chem. Chem. Phys.*, 2017, **19**, 13049–13069.
235 N. Mardirossian and M. Head-Gordon, *J. Chem. Phys.*, 2015, **142**, 074111.
236 N. Mardirossian and M. Head-Gordon, *J. Chem. Phys.*, 2016, **144**, 214110.
237 The B97M-V and ωB97M-V DFAs were technically not available to us at the time of this study, but results for them will be presented in the near future.