Journal Name

ARTICLE TYPE

Cite this: DOI: 00.0000/xxxxxxxxxx

Received Date
Accepted Date

DOI: 00.0000/xxxxxxxxxx

WTMAD-4: A Fair Weighting Scheme for GMTKN55[†]

Kyle R. Bryenton[a,b] and Erin R. Johnson[a,b,c*]

The GMTKN55 data set is a collection of standard benchmarks used in molecular quantum chemistry that spans small- and large-molecule thermochemistry, reaction barriers, and non-covalent interactions. Herein, we identify a flaw in the weighted mean absolute deviation (WTMAD) definitions commonly used to quantify performance of various electronic-structure methods for the GMTKN55 set, which under-weight some of its component benchmarks by orders of magnitude. A new WTMAD-4 metric is proposed, based on typical errors observed for a set of ten minimally empirical dispersion-corrected density-functional approximations (DFAs), ensuring fair treatment across all benchmarks. The performance of 115 DFAs is then reassessed using WTMAD-4 and we highlight a literature example where a DFA parametrised by minimising WTMAD-2 underperforms for benchmarks marginalised by that metric.

1 Introduction

Density-functional theory (DFT) methods are a pillar of modern computational chemistry. A principle goal in DFT development is to create functionals that are broadly accurate, transferable, and maintain a low computational cost. However, to assess their performance, thorough benchmarking of DFT methods against reliable experimental or correlated-wavefunction reference data sets is required.[1,2] Since considering a single benchmark may lead to biased results, compiled databases consisting of numerous, chemically diverse benchmarks are now becoming commonplace.[3–5] The major challenges in creating such benchmarks are obtaining suitable high-level reference data from theory or experiment, curating them to represent a meaningful subset of chemical space, and devising a fair weighting of the component data sets. With respect to this last point, factors such as the wide range of energy scales, quality of the reference data, relative chemical importance, and numbers of systems comprising each benchmark, all become confounding variables.

The GMTKN55 database[3] is a collection of 55 individual benchmarks spanning the thermochemistry of small and large molecules, reaction barriers, and both intramolecular and intermolecular non-covalent interactions (NCI). It has become a standard benchmark in DFT method development in recent years[6–10] due to the robust quality of the reference data and the diverse span of chemical systems. The interested reader is directed to Ref. 3 for detailed information regarding the individual benchmarks. Due to the wide range of energy scales of its components, the overall error for the GMTKN55 set is reported as a weighted mean absolute deviation (WTMAD), and several definitions for such a weighted error have been proposed.

The first proposed metric,[3] WTMAD-1, is not commonly used. In this scheme, each subset is assigned an arbitrary weight, denoted w_i, where $w_i = 10$ when $\overline{|\Delta E|}_i < 7.5$ kcal/mol, $w_i = 0.1$ when $\overline{|\Delta E|}_i > 75$ kcal/mol, and $w_i = 1$ otherwise. The WTMAD-1 is then calculated as

$$\text{WTMAD-1} = \frac{1}{N_{\text{bench}}} \sum_{i=1}^{N_{\text{bench}}} w_i \cdot \text{MAD}_i \,, \qquad (1)$$

where the sum runs over all $N_{\text{bench}} = 55$ component benchmarks, and MAD_i is the mean absolute deviation between the computed and reference data for the ith benchmark. Alternatively, the most widely used error metric for the GMTKN55 is the WTMAD-2, also introduced in Ref. 3, and defined as

$$\text{WTMAD-2} = \sum_{i=1}^{N_{\text{bench}}} \frac{N_i}{N_{\text{total}}} \cdot \frac{\overline{|\Delta E|}_{\text{mean}}}{\overline{|\Delta E|}_i} \cdot \text{MAD}_i \,. \qquad (2)$$

Here, N_i is the number of data points in the ith benchmark and $N_{\text{total}} = \sum_{i=1}^{N_{\text{bench}}} N_i = 1505$ is the total number of data points. $\overline{|\Delta E|}_i$ is the average reference energy for the ith benchmark, and $\overline{|\Delta E|}_{\text{mean}}$ is the average of all $\overline{|\Delta E|}_i$ values.

Error metrics typically are reported for seven categories: “Basic + Small”, “Iso + Large”, “Barriers”, “Intermolecular NCI”, “Intramolecular NCI”, “All NCI”, and GMTKN55 as a whole. When calculating the per-category values, a fixed canonical value of $\overline{|\Delta E|}_{\text{mean}} = 56.84$ kcal/mol is used across all categories, while

[a] Department of Physics and Atmospheric Science, Dalhousie University, 6310 Coburg Road, Halifax, Nova Scotia, Canada, B3H 4R2

[b] Department of Chemistry, Dalhousie University, 6243 Alumni Crescent, Halifax, Nova Scotia, B3H 4R2, Canada. E-mail: erin.johnson@dal.ca

[c] Yusuf Hamied Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge, CB2 1EW, United Kingdom.

[†] Electronic Supplementary Information (ESI) available, see DOI: 10.1039/cXCP00000x/

N_{total} is rescaled per category. However, due to improvements in the reference data, the value of $\overline{|\Delta E|}_{\text{mean}}$ has changed (perhaps unknowingly) from paper to paper, which has been a source of confusion in the literature.[7]

Previous studies have shown a disproportionate weighting of certain benchmarks within GMTKN55 when computing the WTMAD values. In particular, data in Ref. 7 (ESI, Tables S5 to S16) shows that the same benchmarks (such as BH76) were disproportionally weighted very highly across all functionals considered, while contributions from others (such as IL16) were two or more orders of magnitude smaller. This issue is partially due to the fact that some GMTKN55 subsets contain nearly one hundred reaction energies, while others contain fewer than ten. Because the WTMAD-2 weights the individual MAD_i values by the number of component reactions, it introduces a bias towards methods that perform well for the larger subsets. Consequently, Wittmann *et al.* recently proposed the WTMAD-3 weighting scheme, which is identical to WTMAD-2 except that it attenuates the weights for a subset to be no more than 1% of the total reactions considered.[8] The WTMAD-3 is calculated via

$$\text{WTMAD-3} = \sum_{i=1}^{N_{\text{bench}}} \frac{N_i^{\text{damp}}}{N_{\text{total}}} \cdot \frac{\overline{|\Delta E|}_{\text{mean}}}{\overline{|\Delta E|}_i} \cdot \text{MAD}_i\,, \tag{3}$$

where $N_i^{\text{damp}} = \min(0.01 \times N_{\text{total}}\,, N_i)$. Thus, for the full GMTKN55, this limits each subset's weight to 15 systems. However, as we will see, this reweighting does not fully correct for the disproportionate weighting of some GMTKN55 subsets.

A very recent benchmarking study by Liang and Head-Gordon defined a new error metric, the normalised error ratio (NER).[4] The NER of a particular functional for a given benchmark is the ratio of its error to the standard error for that benchmark. The NER values are then averaged over all component benchmarks to obtain a value for the full data set, ensuring each has equal weighting. However, the drawback of this approach is that the standard error is defined as the "average of the second, third, and fourth lowest errors among all tested hybrid functionals."[4] This means that the NER results must be adjusted for each individual benchmarking study to identify the second through fourth lowest errors. Additionally, it is possible that some of the functionals considered are overfit and yield abnormally low errors for certain benchmarks; thus, the standard errors for each individual benchmark may not be representative of what a typical or expected error would be.

In this work, the WTMAD-4 metric is proposed. While identical to WTMAD-1 in its construction, the weights are based on the magnitudes of expected errors for a small set of representative, minimally empirical functionals, rather than on the absolute energy scales. As a result, each benchmark contributes meaningfully and appropriately to the overall WTMAD-4. Data from previous DFA benchmarking studies[3,7–9] is reassessed using WTMAD-4.

2 Data Set

Data for 115 previously studied DC–DFAs[3,7–9] was obtained by extracting the MAD values reported for each of the 55 benchmarks within GMTKN55. These MAD values were then combined with the appropriate weights to generate WTMAD-*N* values, tabulated for all 115 functionals (and 78 non-dispersion-corrected functionals from Ref 3) in the ESI. Because the complete benchmark results for these 115 previously studied DC–DFAs are not available, and in many cases the specific reference values used are uncertain, their WTMAD-*N* values should be regarded as approximate, with an estimated uncertainty of ca. 1–2%. Such a small margin of error is not expected to meaningfully impact the statistics presented in this article.

To ensure all previously studied DC–DFAs were treated as consistently as possible, their WTMAD-2 and WTMAD-3 values were recomputed by combining the MADs reported in earlier work with the updated $\overline{|\Delta E|}_i$ values from the revised GMTKN55 reference data,[8,11] corresponding to $\overline{|\Delta E|}_{\text{mean}} = 57.82$ kcal/mol.[11] However, for consistency with previous works,[3,12] a fixed value of $\overline{|\Delta E|}_{\text{mean}} = 56.84$ kcal/mol is used when calculating WTMAD-2 and WTMAD-3. To obtain $\text{WTMAD}_{57.82}$ values, an appropriate multiplicative scale factor may be used. We note that neither WTMAD-1, nor the proposed WTMAD-4, depend on $\overline{|\Delta E|}_{\text{mean}}$, and thus do not suffer from this issue.

3 Results and Discussion

3.1 Analysis of WTMAD Contributions

To determine the contributions of each subset to the overall WTMAD-*N* values, we first calculated the mean MAD ($\overline{\text{MAD}_i}$) for each GMTKN55 subset by averaging the MAD values across the set of 115 dispersion-corrected functionals. These mean MADs were then used to compute a single representative WTMAD-*N* value via Eqns. 1-3. Using these WTMAD-*N*'s and the $\overline{\text{MAD}_i}$'s, the percent contributions of each component benchmark to the overall WTMAD-1, WTMAD-2, and WTMAD-3 were evaluated for the entire GMTKN55. The results are collected in Table 1 and show that none of these treatments fairly weight all subsets.

All three previously-proposed WTMAD schemes result in a number of benchmarks having a disproportionately large contribution to the overall WTMAD, while others have near-zero contribution. The largest disparity was observed for WTMAD-2, where BH76 contributes 9.89% of the total, while DIPCS10 and IL16 contribute only 0.04% and 0.06%, respectively; this is a ca. 200-fold difference in scale. Upon review, we determined that calculating weights based on the MAD relative to the reference energy, $\overline{|\Delta E|}_i$, was not representative. For example, IL16's average reference energy is 109.04 kcal/mol, which is 170 times larger than its mean MAD across our methods of 0.64 kcal/mol. There also exist cases where the inverse is true and the ratio is small, causing over-contribution from subsets such as BH76, which has an average reference energy of 18.61 kcal/mol and a mean MAD of 4.07 kcal/mol. Although dynamic weights allow for easy expansion and transferability to other benchmarks, we could not find a dynamic weighting scheme that fairly represented all systems within GMTKN55.

Table 1 Shown for all the GMTKN55 benchmarks are the numbers of component reactions (N_i), mean reference energies ($\overline{\lvert\Delta E\rvert}_i$), mean MADs across all 115 DC-DFAs ($\overline{\text{MAD}}_i$), and recommended WTMAD-4 weights (w_i). Using the data in columns 2–5, and with respective weights from Eqs. 1–5, the percent contributions of each benchmark to the total WTMAD-N were evaluated and are shown in the final four columns. Note that the contributions for each subset are not exactly equal for WTMAD-4 as the weights were determined using a much smaller data set of ten minimally empirical functionals.

					WTMAD-N			
Benchmark	N_i	$\overline{\lvert\Delta E\rvert}_i$	$\overline{\text{MAD}}_i$	w_i	1	2	3	4
			Basic + Small					
AL2X6	6	35.88	2.38	3.27	1.20	0.24	0.43	2.25
ALK8	8	62.60	4.91	1.30	2.48	0.37	0.68	1.85
ALKBDE10	10	100.69	5.09	1.06	0.26	0.30	0.55	1.55
BH76RC	30	21.39	2.23	2.55	1.13	1.86	1.69	1.64
DC13	13	54.98	7.22	0.78	3.64	1.02	1.84	1.63
DIPCS10	10	654.26	4.64	1.65	0.23	0.04	0.08	2.21
FH51	51	31.01	2.26	2.39	1.14	2.21	1.18	1.56
G21EA	25	33.62	2.69	2.15	1.36	1.19	1.30	1.67
G21IP	36	257.61	3.30	1.63	0.17	0.27	0.21	1.55
G2RC	25	51.26	4.49	1.12	2.27	1.30	1.42	1.46
HEAVYSB11	11	58.02	3.22	2.73	1.63	0.36	0.66	2.54
NBPRC	12	27.71	2.06	2.32	1.04	0.53	0.96	1.38
PA26	26	189.05	2.74	1.98	0.14	0.22	0.24	1.57
RC21	21	35.70	3.50	1.45	1.77	1.22	1.59	1.46
SIE4x4	16	33.72	13.08	0.47	6.60	3.69	6.30	1.77
TAUT15	15	3.05	1.06	5.80	5.35	3.10	5.63	1.78
W4-11	140	306.91	5.69	1.07	0.29	1.54	0.30	1.76
YBDE18	18	49.28	2.81	2.55	1.42	0.61	0.93	2.07
				Total:	32.12	20.07	25.99	31.70
			Iso + Large					
BSR36	36	16.20	2.22	2.03	1.12	2.93	2.23	1.30
C60ISO	9	98.25	5.89	1.19	0.30	0.32	0.58	2.03
CDIE20	20	4.06	0.96	5.75	4.85	2.81	3.84	1.59
DARC	14	32.47	3.59	1.44	1.81	0.92	1.67	1.50
ISO34	34	14.57	1.26	4.49	0.64	1.75	1.40	1.63
ISOL24	24	21.92	3.15	1.99	1.59	2.05	2.33	1.81
MB16-43	43	468.39	21.30	0.33	1.08	1.16	0.74	2.02
PArel	20	4.63	1.08	5.51	5.45	2.78	3.79	1.72
RSE43	43	7.60	1.43	5.27	0.72	4.81	3.06	2.18
				Total:	17.56	19.53	19.64	15.78
			Barriers					
BH76	76	18.61	4.07	1.53	2.05	9.89	3.55	1.79
BHDIV10	10	45.33	3.53	1.66	1.78	0.46	0.84	1.69
BHPERI	26	20.87	2.71	2.29	1.37	2.01	2.11	1.79
BHROT27	27	6.27	0.45	10.95	2.27	1.15	1.17	1.42
INV24	24	31.85	1.42	4.41	0.72	0.64	0.72	1.81
PX13	13	33.36	4.60	1.18	2.32	1.07	1.93	1.56
WCPT18	18	34.99	3.29	1.77	1.66	1.01	1.53	1.68
				Total:	12.17	16.23	11.85	11.74
			Intermolecular NCI					
ADIM6	6	3.36	0.37	36.78	1.87	0.39	0.71	3.93
AHB21	21	22.49	0.78	6.24	0.39	0.43	0.56	1.41
CARBHB12	12	6.04	0.87	5.50	4.39	1.03	1.87	1.38
CHB6	6	26.79	1.52	4.14	0.77	0.20	0.37	1.82
HAL59	59	4.59	0.70	9.63	3.53	5.35	2.48	1.95
HEAVY28	28	1.24	0.44	13.96	2.22	5.91	5.76	1.77
IL16	16	109.04	0.64	11.49	0.03	0.06	0.10	2.12
PNICO23	23	4.27	0.69	7.86	3.48	2.21	2.63	1.56
RG18	18	0.58	0.25	29.88	1.26	4.62	7.00	2.16
S22	22	7.30	0.39	13.54	1.97	0.70	0.87	1.52
S66	66	5.47	0.32	18.39	1.62	2.30	0.95	1.70
WATER27	27	81.14	3.78	1.43	0.19	0.75	0.76	1.56
				Total:	21.72	23.95	24.06	22.88
			Intramolecular NCI					
ACONF	15	1.83	0.14	61.19	0.71	0.68	1.24	2.47
Amino20x4	80	2.44	0.27	22.65	1.36	5.27	1.80	1.77
BUT14DIOL	64	2.80	0.25	30.02	1.26	3.40	1.45	2.17
ICONF	17	3.27	0.29	20.01	1.46	0.90	1.44	1.68
IDISP	6	14.22	2.49	2.50	1.26	0.62	1.13	1.80
MCONF	51	4.97	0.40	21.21	2.02	2.44	1.31	2.45
PCONF21	18	1.62	0.71	8.66	3.58	4.69	7.12	1.77
SCONF	17	4.60	0.40	19.05	2.02	0.88	1.41	2.20
UPU23	23	5.72	0.55	10.12	2.78	1.32	1.56	1.61
				Total:	16.45	20.20	18.46	17.92

3.2 Formulation of WTMAD-4

Our analysis of the disparities in WTMAD-2 led to the formulation of WTMAD-4, defined as

$$\text{WTMAD-4} = \frac{1}{N_{\text{bench}}} \sum_{i=1}^{N_{\text{bench}}} w_i^{\text{WTMAD-4}} \cdot \text{MAD}_i\,. \tag{4}$$

This weighting scheme is identical to WTMAD-1 in its construction, but the weights are chosen to be

$$w_i^{\text{WTMAD-4}} = \frac{100}{N_{\text{bench}}}\left(\frac{3.5}{\overline{\text{MAD}}_i}\right), \tag{5}$$

such that each benchmark contributes evenly to the overall WTMAD-4 on average (i.e. $1/55 = 1.82\%$ for each benchmark when the mean MAD values are used). Here, $\overline{\text{MAD}}_i$ is the mean MAD for benchmark i across all functionals in the fit set. A factor of 3.5 was also included to scale the WTMAD-4 for consistency with the typical magnitudes of WTMAD-2.

The definition of WTMAD-4 is similar in spirit to the NER.[4] However, instead of comparing to the hybrid functionals that give the second through fourth lowest errors, we compare to the mean MADs for a set of ten representative, minimally empirical hybrid functionals. While the WTMAD-4 weights could be evaluated using the mean MADs from the full set of 115 functionals (and results obtained by so doing are shown in the ESI), we recommend instead obtaining the weights from a small set of well-behaved and numerically stable DFAs that are available in most electronic-structure programs. In this way, the error metric will not require readjusting depending on the set of functionals considered, and new weights can easily be generated if the benchmark is expanded.

Minimally empirical, dispersion-corrected functionals are selected to avoid any excessively low (or high) errors for a particular benchmark that can result from overfitting. The ten functionals selected to determine the mean MAD values and WTMAD-4 weights via Eqn. 5 are: PBE0-D3(BJ), B3LYP-D3(BJ), B3P86-D3(BJ), B3PW91-D3(BJ), PW1PW91-D3(0), MPW1PW91-D3(BJ), BHLYP-D3(BJ), HSE06-D3(BJ), HISS-D3(BJ), and LC-ωhPBE-D3(BJ). This recommended set consists of decades-old functionals with predictable behaviours that are well-documented in the literature. Furthermore, inspection of their largest MADs across the GMTKN55 composite benchmarks shows no extreme outliers, in contrast to several of the other DC-DFAs in the full set of 115 examined in this article.

Overall, if the set of functionals used to evaluate the weights performs well for a particular benchmark, the mean MAD will be low and that set will be heavily weighted. Conversely, if the set performs poorly for a particular benchmark, the mean MAD will be high and that set will be lightly weighted. The results in Table 1 show no drastic over- or under-weighting of the GMTKN55 benchmarks except for ADIM6 (3.93%), for which there are some very large outliers in the full 115-DFA set, as will be discussed later. All of the other benchmarks contribute between 1.30-2.54% of the total WTMAD-4 using the mean MADs for all 115 DC-DFAs considered.

Table 1 also shows the resulting percent contributions to the total WTMAD-4 for each GMTKN55 category. The contributions from WTMAD-4 are fairly similar to WTMAD-1, and notably different from WTMAD-2/3. The increased percent contribution of the “Basic + Small” classification between WTMAD-2 and WTMAD-4 is due to the high proportion of benchmarks within this category. However, we consider this shift in weight towards

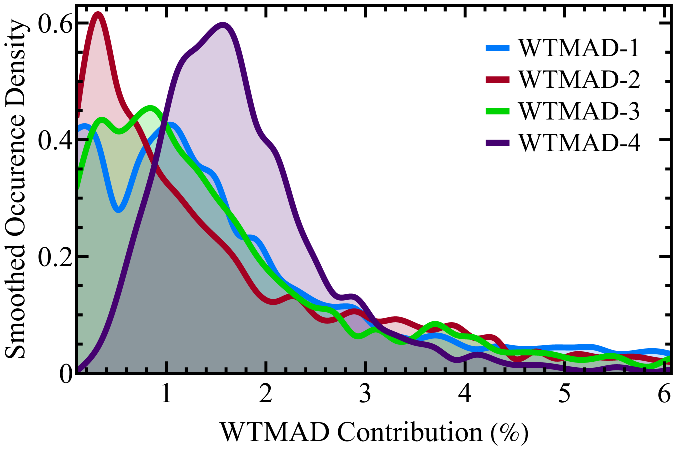

Fig. 1 Smoothed histograms showing the percent contributions of each GMTKN55 subset to the various WTMAD-N values for 115 DC-DFAs. The WTMAD-4 weights were obtained from Eq. 5 using the mean MAD values from ten minimally empirical functionals.

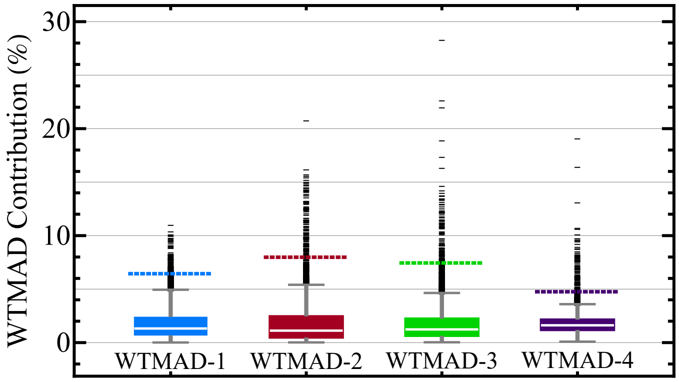

Fig. 2 Box-and-whisker plots showing the percent contributions of each GMTKN55 subset to the various WTMAD-N values for 115 DC-DFAs. For each WTMAD-N, the median is shown with a white line, the coloured boxes span the interquartile range (IQR) from the 25% to 75% quantiles. Whiskers extend to 1.5× the IQR, beyond which outliers are shown as individual points. The 98% quantile is indicated with a dashed line. The WTMAD-4 weights were obtained from Eq. 5 using the mean MAD values from ten minimally empirical functionals.

thermochemistry desirable and representative for general chemistry applications.

3.3 Validation using Literature Data

To validate our definition of WTMAD-4, we consider all data available in Refs. 3,7–9, which spans 115 DC-DFAs. Figure 1 shows smoothed histograms of the percent contributions to each WTMAD-N obtained for all 55 GMTKN55 data sets with 115 DC-DFAs, where the WTMAD-4 weights were obtained from our recommended subset of 10 well-behaved dispersion-corrected hybrid functionals. The curves show that all of the WTMAD schemes, except WTMAD-4, have many subsets with near-zero contributions and form distributions with slowly decaying tails. WTMAD-1 alone forms a bimodal distribution, likely due to it using only three bins for the weights based on average benchmark energies. Alternatively, WTMAD-4 forms a fast-decaying, skewed distribution that provides a more balanced weighting of all GMTKN55 subsets. A breakdown of the various WTMAD-N contributions by class of DFA is given in the ESI and shows good consistency across GGA, meta-GGA, hybrid, and double-hybrid functionals.

The same data shown in Figure 1 is recast in a box-and-whisker plot in Figure 2. Clearly, the WTMAD-4 has the most centralized median and a much tighter inter-quartile range (IQR), indicating a more consolidated spread of contributions towards the total WTMAD. The computed IQRs are 1.70, 1.97, 1.60, 0.97 for WTMAD-1 through WTMAD-4, respectively. Furthermore, the 98th quantile (indicated in Figure 2 using a dashed line) falls at 6.44, 7.98, 7.45, and 4.75% for WTMAD-1 through WTMAD-4, respectively, again highlighting the latter's more compact distribution.

There are only three large outliers with $>11\%$ contributions to the WTMAD-4, which all occur for ADIM6: MN15L-D3(0) (19.1%), M06-D3(0) (16.5%), and MN15-D3(BJ) (13.2%). Moreover, 7 of the 10 outliers with 9–11% contributions are again for ADIM6. The ADIM6 set corresponds to interaction energies of n-alkane dimers (bound only by London dispersion) and has quite a large WTMAD-4 weight due to the typically small errors ($\overline{\text{MAD}}$ of 0.37 kcal/mol) seen for this benchmark. Analysis of the signed errors shows these three Minnesota functionals systematically overbind the ADIM6 dimers, with MADs of 3.88, 1.82, and 1.32 kcal/mol, for MN15L-D3(0), M06-D3(0), and MN15-D3(BJ), respectively. It was noted in Ref. 3 that no DFT-D3(BJ) parameterisation was possible for MN15L due to the overbinding tendency of this highly empirical base functional, and that MN15L-D3(0) was the worst-performing meta-GGA for non-covalent interactions. Additionally, the overall GMTKN55 results did not benefit from the addition of a D3 correction only for M06, MN15L, and MN15, due to their mimicry of dispersion-like binding for van der Waals complexes near equilibrium, although they still neglect long-range London dispersion physics.[3]

Finally, one can consider the scale difference between the minimum-weighted and maximum-weighted GMTKN55 benchmarks for a given DC-DFA. The median scale difference across all 115 functionals is 315.6× for WTMAD-2, compared to 11.0× for WTMAD-4. Only five DC-DFAs give a WTMAD-4 scale difference greater than 43.1×, which is the lowest observed scale difference for WTMAD-2 across all methods considered.

3.4 Ranking of DFAs for GMTKN55

Table 2 shows the top performing of the 115 DC-DFAs considered for the GMTKN55, ranked according to both WTMAD-2 and WTMAD-4 values for comparison between the two weighting schemes. The results are divided into GGA, meta-GGA, hybrid, and double-hybrid functionals. Included in the ESI are the complete WTMAD-N rankings, as well as full lists of MADs for each benchmark, overall and category-specific partial WTMAD-N values, and the percent contributions to WTMAD-N for each functional.

From Table 2, the change in WTMAD scheme results in some order shifting for GGA functionals. While the top four DFAs are consistent, rPW86PBE-D3(BJ) is 5th ranked by WTMAD-2 vs. 7th by WTMAD-4, while MPWPW91-D3(BJ) is 5th ranked by WTMAD-4

Table 2 The top five (GGA/mGGA) and top ten (Hyb/D.Hyb) functionals for GMTKN55. On the left, sorted by WTMAD-2 values, on the right, sorted by WTMAD-4 values.

Functional	WTMAD-2 Sorted	WTMAD-4	Functional	WTMAD-2	WTMAD-4 Sorted
GGA					
revPBE-D3(BJ)	8.29	7.73	revPBE-D3(BJ)	8.29	7.73
B97-D3(BJ)	8.49	8.32	OLYP-D3(BJ)	8.76	8.16
OLYP-D3(BJ)	8.76	8.16	B97-D3(BJ)	8.49	8.32
BLYP-D3(BJ)	9.40	8.76	BLYP-D3(BJ)	9.40	8.76
rPW86PBE-D3(BJ)	9.67	8.94	MPWPW91-D3(BJ)	10.16	8.86
meta-GGA					
SCAN-D3(BJ)	7.96	7.95	revTPSS-D3(BJ)	8.44	7.73
revTPSS-D3(BJ)	8.44	7.73	SCAN-D3(BJ)	7.96	7.95
M06L-D3(0)	8.59	8.92	TPSS-D3(BJ)	9.10	7.95
TPSS-D3(BJ)	9.10	7.95	M06L-D3(0)	8.59	8.92
M11L-D3(0)	9.59	9.96	M11L-D3(0)	9.59	9.96
Hybrid					
ωB97X-V	3.93	4.40	ωB97X-V	3.93	4.40
M052X-D3(0)	4.62	5.19	PW6B95-D3(BJ)	5.50	5.14
ωB97X-D3(0)	4.73	5.38	M052X-D3(0)	4.62	5.19
M062X-D3(0)	4.92	5.22	M062X-D3(0)	4.92	5.22
M08HX-D3(0)	5.27	5.68	ωB97X-D3(0)	4.73	5.38
LC-ωhPBE-D3(BJ)	5.44	6.43	MPW1B95-D3(BJ)	5.55	5.48
PW6B95-D3(BJ)	5.50	5.14	M08HX-D3(0)	5.27	5.68
MPW1B95-D3(BJ)	5.55	5.48	MPW1PW91-D3(BJ)	6.33	5.82
MPWB1K-D3(BJ)	5.59	6.11	B1P86-D3(BJ)	6.78	5.92
BHLYP-D3(BJ)	5.65	7.02	PBE0-D3(BJ)	6.59	5.94
Double Hybrid					
DH24	1.72	2.23	DH24	1.72	2.23
revDH23	1.72	2.26	revDH23	1.72	2.26
XYG8	1.85	2.66	SOS-DH24	1.91	2.27
SOS-DH24	1.91	2.27	SOS-DH23	1.95	2.34
SOS-DH23	1.95	2.34	ωDOD-PBEP86-D3(BJ) ($\omega = 0.10$, $a_X = 0.69$)	2.20	2.50
ωDSD-PBEP86-D4 ($\omega = 0.16$, $a_X = 0.69$)	2.08	2.59	ωDOD-PBEP86-D3(BJ) ($\omega = 0.08$, $a_X = 0.72$)	2.21	2.52
ωDSD-PBEP86-D4 ($\omega = 0.13$, $a_X = 0.72$)	2.08	2.58	ωDOD-PBEP86-D4 ($\omega = 0.10$, $a_X = 0.69$)	2.17	2.53
ωDSD-PBEP86-D3(BJ) ($\omega = 0.13$, $a_X = 0.72$)	2.10	2.57	ωDOD-PBEP86-D4 ($\omega = 0.08$, $a_X = 0.72$)	2.18	2.53
ωDSD-PBEP86-D3(BJ) ($\omega = 0.16$, $a_X = 0.69$)	2.11	2.58	xDOD-PBEP86-D3(BJ) ($a_X = 0.72$)	2.23	2.53
xDSD-PBEP86-D4 ($a_X = 0.75$)	2.11	2.59	xDOD-PBEP86-D4 ($a_X = 0.72$)	2.19	2.54

vs. 9th by WTMAD-2. For the meta-GGAs, there is only slight reordering within the same top five functionals. However, the meta-GGA functionals tend to perform comparatively worse with respect to the GGAs with WTMAD-4 than with WTMAD-2. The results show no particular advantage in using the much more complicated meta-GGA functionals, as opposed to the simpler GGAs, for the GMTKN55.

In contrast, changing the WTMAD definition causes significant re-ordering of the hybrid functionals in Table 2. PW6B95-D3(BJ) is 7th ranked by WTMAD-2 (5.50), but improves to 2nd ranked with WTMAD-4 (5.14). Although ωB97X-V remains the top-ranked functional with both metrics, it gives a higher WTMAD-4 (4.40) than WTMAD-2 (3.93). Examining the MADs of the individual benchmarks for these two functionals, the largest differences arise from the "Iso. + Large" category. Here, ωB97X-V performs best for benchmarks with large WTMAD-2 weights and poorly for those with small weights, while the opposite is true for PW6B95-D3(BJ). In particular, ωB97X-V gives very large MADs for C60ISO and MB16-43 (13.74 and 32.51 kcal/mol), which have quite low WTMAD-2 weights. Conversely, PW6B95-D3(BJ) performs excellently for both C60ISO and MB16-43 (MADs of 1.65 and 8.97 kcal/mol), which are more fairly represented by WTMAD-4.

For double hybrid functionals, there is again significant reordering with the change from WTMAD-2 to WTMAD-4, which is attributed to the reduced weighting of BH76 and the increased weighting of other benchmarks such as IL16, DIPCS10, ALKBDE10, MB16-43, and WATER27. Notably, WTMAD-2 tends to prefer the spin-component scaled ωDSD double hybrids, while WTMAD-4 favours the ωDOD functionals, in which the same-spin correlation term is omitted, resulting in a lower computational cost.[7] Additionally, Table 2 shows that the XYG8 functional was ranked 3rd by WTMAD-2, but is ranked 17th by WTMAD-4. Compared to other double hybrids, XYG8 underperformed for DIPCS10, ALKBDE10, C60ISO, MB16-43, and WATER27; the increased WTMAD-4 weights for these benchmarks likely explains the large drop in XYG8's ranking. It is notable that the 8 parameters in XYG8 were fit to minimise the WTMAD-2.[13] In contrast, the empirical parameters in the revDH23 and DH24 functionals were also fit to the GMTKN55 set,[9,14] but these functionals are the leading performers with both WTMAD-2 and WTMAD-4, indicating that they may be more robust.

4 Summary

In this work, unintended behaviour of previous WTMAD metrics, used to quantify errors of various electronic-structure methods for the GMTKN55 thermochemical data set, was identified. Specifically, use of energy-based weights in WTMAD-2 and WTMAD-3 resulted in contributions of some benchmarks being more than 100× greater than others. The alternative WTMAD-4 metric, which uses error-based weights, was proposed herein to ensure that each benchmark within the GMTKN55 set is treated fairly. Ranking the performance of 115 DC-DFAs using WTMAD-4 reveals limitations of some functionals through reordering compared to WTMAD-2. Notably, its good performance on the BH76 reaction barriers, at the expense of the DIPCS10, ALKBDE10, C60ISO, MB16-43, and WATER27 benchmarks, resulted in a particularly low WTMAD-2 for the XYG8 functional, but a relatively higher WTMAD-4.

It is tempting to reason that, if particular a method gives a low WTMAD for the GMTKN55, it will perform well for all of its component benchmarks. However, this may not be the case if poor performance for one of more benchmarks is offset by good performance for the other subsets, and it is essential to consider data points that are the largest outliers. Due to unbalanced weighting, fitting to minimise WTMAD-2 may result in larger errors for benchmarks with very small contributions, while overfitting to BH76, as seen for XYG8. The more balanced weights in WTMAD-4 should reduce the likelihood of, but not eliminate, this possibility. Thus, functional developers should take care to avoid overfitting of functionals to particular thermochemical data and to be cognisant of Goodhart's law that "when a measure becomes a target, it ceases to be a good measure."[15,16] This is particularly important for artificial intelligence (AI) applications of DFT, as well as AI-guided DFA development. It is more sound to consider multiple statistical measures for a data set, rather than focusing on a single number as a target metric.

Acknowledgements

KRB and ERJ thank the Natural Sciences and Engineering Research Council (NSERC) of Canada for financial support and the Atlantic Computing Excellence Network (ACENET) for computational resources. ERJ additionally thanks the Royal Society for a Wolfson Visiting Fellowship, while KRB thanks the Killam Trust, the Government of Nova Scotia, and the Mary Margaret Werner Graduate Scholarship Fund.

Data Availability Statement

The data that support the findings of this study are available in the supplementary information.

Conflicts of Interest

There are no conflicts to report.

References

1 L. A. Burns, M. S. Marshall and C. D. Sherrill, *J. Chem. Phys.*, 2014, **141**, 234111.

2 M. Kodrycka and K. Patkowski, *J. Chem. Phys.*, 2019, **151**, 070901.

3 L. Goerigk, A. Hansen, C. Bauer, S. Ehrlich, A. Najibi and S. Grimme, *Phys. Chem. Chem. Phys.*, 2017, **19**, 32184–32215.

4 J. Liang and M. Head-Gordon, *arXiv preprint 2508.13468v2*, 2025.

5 N. Mardirossian and M. Head-Gordon, *Mol. Phys.*, 2017, **115**, 2315–2372.

6 N. Mehta, M. Casanova-Páez and L. Goerigk, *Phys. Chem. Chem. Phys.*, 2018, **20**, 23175–23194.

7 G. Santra, M. Cho and J. M. Martin, *J. Phys. Chem. A*, 2021, **125**, 4614–4627.

8 L. Wittmann, H. Neugebauer, S. Grimme and M. Bursch, *J. Chem. Phys.*, 2023, **159**, 224103.

9 A. D. Becke, *J. Chem. Phys.*, 2024, **160**, 204118.

10 A. M. Teale, T. Helgaker, A. Savin, C. Adamo, B. Aradi, A. V. Arbuznikov, P. W. Ayers, E. J. Baerends, V. Barone, P. Calaminici, E. Cancès, E. A. Carter, P. K. Chattaraj, H. Chermette, I. Ciofini, T. D. Crawford, F. D. Proft, J. F. Dobson, C. Draxl, T. Frauenheim, E. Fromager, P. Fuentealba, L. Gagliardi, G. Galli, J. Gao, P. Geerlings, N. Gidopoulos, P. M. W. Gill, P. Gori-Giorgi, A. Görling, T. Gould, S. Grimme, O. Gritsenko, H. J. A. Jensen, E. R. Johnson, R. O. Jones, M. Kaupp, A. M. Köster, L. Kronik, A. I. Krylov, S. Kvaal, A. Laestadius, M. Levy, M. Lewin, S. Liu, P.-F. Loos, N. T. Maitra, F. Neese, J. P. Perdew, K. Pernal, P. Pernot, P. Piecuch, E. Rebolini, L. Reining, P. Romaniello, A. Ruzsinszky, D. R. Salahub, M. Scheffler, P. Schwerdtfeger, V. N. Staroverov, J. Sun, E. Tellgren, D. J. Tozer, S. B. Trickey, C. A. Ullrich, A. Vela, G. Vignale, T. A. Wesolowski, X. Xu and W. Yang, *Phys. Chem. Chem. Phys.*, 2022, **24**, 28700–28781.

11 M. Müller and S. Ehlert, *gmtkn55*, 2021, https://github.com/grimme-lab/GMTKN55, https://github.com/grimme-lab/GMTKN55.

12 G. Santra, N. Sylvetsky and J. M. Martin, *J. Phys. Chem. A*, 2019, **123**, 5129–5143.

13 G. Santra, E. Semidalas and J. M. L. Martin, *J. Phys. Chem. Lett.*, 2021, **12**, 9368–9376.

14 A. D. Becke, *J. Chem. Phys.*, 2023, **159**, 241101.

15 C. A. Goodhart, in *Monetary theory and practice: The UK experience*, Springer, 1984, pp. 91–121.

16 M. Strathern, *Eur. Rev.*, 1997, **5**, 305–321.