The Journal of Chemical Physics

Probabilistic performance estimators for computational chemistry methods: Systematic improvement probability and ranking probability matrix. II. Applications

Cite as: J. Chem. Phys. **152**, 164109 (2020); https://doi.org/10.1063/5.0006204
Submitted: 28 February 2020 . Accepted: 07 April 2020 . Published Online: 28 April 2020

Pascal Pernot, and Andreas Savin

The Journal of Chemical Physics | ARTICLE | scitation.org/journal/jcp

Probabilistic performance estimators for computational chemistry methods: Systematic improvement probability and ranking probability matrix. II. Applications

Cite as: J. Chem. Phys. **152**, 164109 (2020); doi: 10.1063/5.0006204
Submitted: 28 February 2020 • Accepted: 7 April 2020 •
Published Online: 28 April 2020

View Online Export Citation CrossMark

Pascal Pernot[1,a)] and Andreas Savin[2,b)]

AFFILIATIONS

[1] Institut de Chimie Physique, UMR8000, CNRS, Université Paris-Saclay, 91405 Orsay, France
[2] Laboratoire de Chimie Théorique, CNRS and UPMC Université Paris 06, Sorbonne Universités, 75252 Paris, France

[a)] **Author to whom correspondence should be addressed:** Pascal.Pernot@universite-paris-saclay.fr
[b)] **Electronic mail:** Andreas.Savin@lct.jussieu.fr

ABSTRACT

In Paper I [P. Pernot and A. Savin, J. Chem. Phys. **152**, 164108 (2020)], we introduced the systematic improvement probability as a tool to assess the level of improvement on absolute errors to be expected when switching between two computational chemistry methods. We also developed two indicators based on robust statistics to address the uncertainty of ranking in computational chemistry benchmarks: P_{inv}, the inversion probability between two values of a statistic, and $\mathbf{P}_r$, the ranking probability matrix. In this second part, these indicators are applied to nine data sets extracted from the recent benchmarking literature. We also illustrate how the correlation between the error sets might contain useful information on the benchmark dataset quality, notably when experimental data are used as reference.

I. INTRODUCTION

In Paper I,[1] we considered the uncertainty sources impacting the values of benchmarking statistics (scores), and we presented tools to estimate the uncertainty on statistics and to compare them. We briefly summarize them here.

First, one compares system-by-system the absolute errors of two methods M_i and M_j. The systematic improvement probability (SIP$_{i,j}$) is defined as the fraction of systems for which M_i has smaller absolute errors than M_j. A SIP matrix can be built for a set of methods, enabling to detect the methods with the best performances in terms of absolute errors. A mean gain (MG$_{i,j}$; a negative value) is estimated, providing the expected decrease in absolute errors when using M_i instead of M_j. The mean loss (ML$_{i,j}$) is defined accordingly. The MUE difference between both methods can be expressed as a combination of SIP, MG, and ML, illustrating the balance between gains and losses when switching between two methods.

Then, one compares statistics, taking into account their uncertainty and correlation. For comparison of pairs of values, one uses P_{inv}, which gives the probability that the sign of the observed difference is the opposite of the true one, considering the use of limited size datasets. We have shown in Paper I[1] that $P_{inv} \simeq p_g/2$, where p_g is the p-value for the test of the equality of the two values. This is tested in the first example below, as well as the comparison to a p-value of the test ignoring correlations, p_{unc}. To compare the statistics for a set of several methods, we use the ranking probability matrix $\mathbf{P}_r$, which gives the probability for each method to have any rank, considering the limited size of the data set.

To avoid hypotheses on the errors distributions, bootstrap-based sampling methods were used for the estimation of statistics uncertainty, p-values, P_{inv} and $\mathbf{P}_r$. The algorithms are detailed in Paper I,[1] and some specific choices have been made regarding the statistics: based on the recommendations of Wilcox and Erceg-Hurn,[2] quantiles are estimated by the Harrell and Davis method,[3] and correlation coefficients are estimated by the Spearman method (rank correlation), unless stated otherwise.

In the following, these methods are illustrated and validated on nine datasets taken from the recent benchmarking literature and covering a wide range of dataset sizes and properties. In Sec. II, the datasets are introduced and treated sequentially with a common framework. *A reader not interested in the detailed treatment of the example datasets can skip directly to* Sec. III*, where the major findings are reported*. This is a global discussion covering the topics of both papers.

II. APPLICATIONS

Nine datasets have been extracted from the recent benchmarking literature. Our selection is mostly based on the coverage of a representative range of properties, dataset sizes (between a few tens to a few thousands), and reference type (experimental or calculated) (Table I). Besides such selection criteria, a major quality of the datasets is their *availability*, and their authors have to be praised for that.

Through these various examples, *our intent is not to validate or invalidate the original studies*, but only to illustrate the properties and interest of our proposed tools.

The cases are treated with a common framework: an introduction; the analysis of the correlation matrices for error sets and statistics (MUE and Q_{95}); the analysis of the MUE and Q_{95} statistics and their inversion probabilities; the analysis of the SIP statistics; and finally, the ranking probability matrices.

A. PER2018

We consider here the intensive atomization energies[13] estimated with nine DFAs on the G3/99 dataset[14] and extracted from a recent article by Pernot and Savin.[4,15] This medium-sized dataset ($N = 222$) presents several non-normal error distributions and was used to illustrate the interest for benchmarks of using Q_{95} as a complement to the MUE and to illustrate our former definition of P_{inv}. Here, we focus on the correlations and their impact on the comparison of statistics.

1. Correlations

The correlation matrices between the error sets and their statistics are represented in Fig. 1, along with histograms of their non-diagonal elements. The error sets are positively correlated, with a wide distribution of correlation coefficients, except for pairs involving BH&HLYP, which presents negative correlations with four other methods. When considering the scores, all correlations are positive or null. Globally, the correlations are weaker for Q_{95} than for the MUE, except for a few pairs. The maximum of the histograms shifts from 0.6 for MUE to 0 for Q_{95}, but large correlation values are, nevertheless, still observed for Q_{95}. These observations confirm the main trends from the numerical study of correlation transfer in Paper I.[1]

2. Statistics

The statistics are reported in Table II. Note that due to the use of a different quantile estimation algorithm, the values of Q_{95} have changed slightly from the values reported in the original article.[4]

There is a group of three methods (B97-1, CAM-B3LYP, and PBE0) with small MUE values. Considering the p_g values, one cannot reject the hypothesis that the observed differences are due to the limited size of the datasets. Note that the same conclusion would have been reached when ignoring correlation (p_{unc}), as the neglect of correlation increases the p-values, but no other one reaches the 0.05 threshold. However, the p_{unc} value for LC-ωPBE reaches 0.03, not far from the threshold. Consistently, the MUE inversion probability P_{inv} computed in the reference article[15] included LC-ωPBE in the group of methods with a sizable risk of inversion. As demonstrated in Eq. (31) of Paper I,[1] the revised version of P_{inv} accounting for correlations is now practically equal to $p_g/2$, which rejects LC-ωPBE as a contender for the head group. When picking B97-1 instead of CAM-B3LYP based on the MUE, there is a 29% chance to be wrong, i.e., that the MUE of CAM-B3LYP is indeed smaller than B97-1 due to the dataset size. This risk falls to 12% for PBE0.

The situation is different for Q_{95}, where the neglect of correlation would lead to the conclusion that PBE0 [3.3(5) kcal/mol] is not significantly distinct from B97-1 [2.7(4) kcal/mol; $p_{unc} = 0.33$], whereas the correct value is given by $p_g = 0.02$. In this example, Q_{95} can help us to rank the three best methods, for which the MUE is not discriminant. This is linked to the presence of different tails in the absolute errors distributions [cf. Fig. 3(a)].

This example illustrates and confirms the relations between p_{unc}, p_g, and P_{inv} expressed in Paper I,[1] Sec. II D 3. In the following examples, only P_{inv} is reported to alleviate the results tables.

TABLE I. Case studies: N is the number of systems in the dataset and K is the number of compared methods.[1] Nature of the reference data: experimental (expt.) or calculated (cal.).

Case	Property	N	K	Reference[1]	Source
PER2018	Intensive atomization energies	222	9	Expt.	4
BOR2019	Bandgaps	471	15	Expt.	5
NAR2019	Enthalpies of formation	469	4	Cal.	6
CAL2019	London dispersion corrections	41	10 × 3	Cal.	7
JEN2018	Non-covalent interaction energies	66	6	Cal.	8
DAS2019	Dielectric constants	23	6	Expt.	9
THA2015	Polarizability	135	7	Expt.	10
WU2015	Polarizability	145	7	Cal.	11
ZAS2019	Effective atomization energies	6211	3	Cal.	12

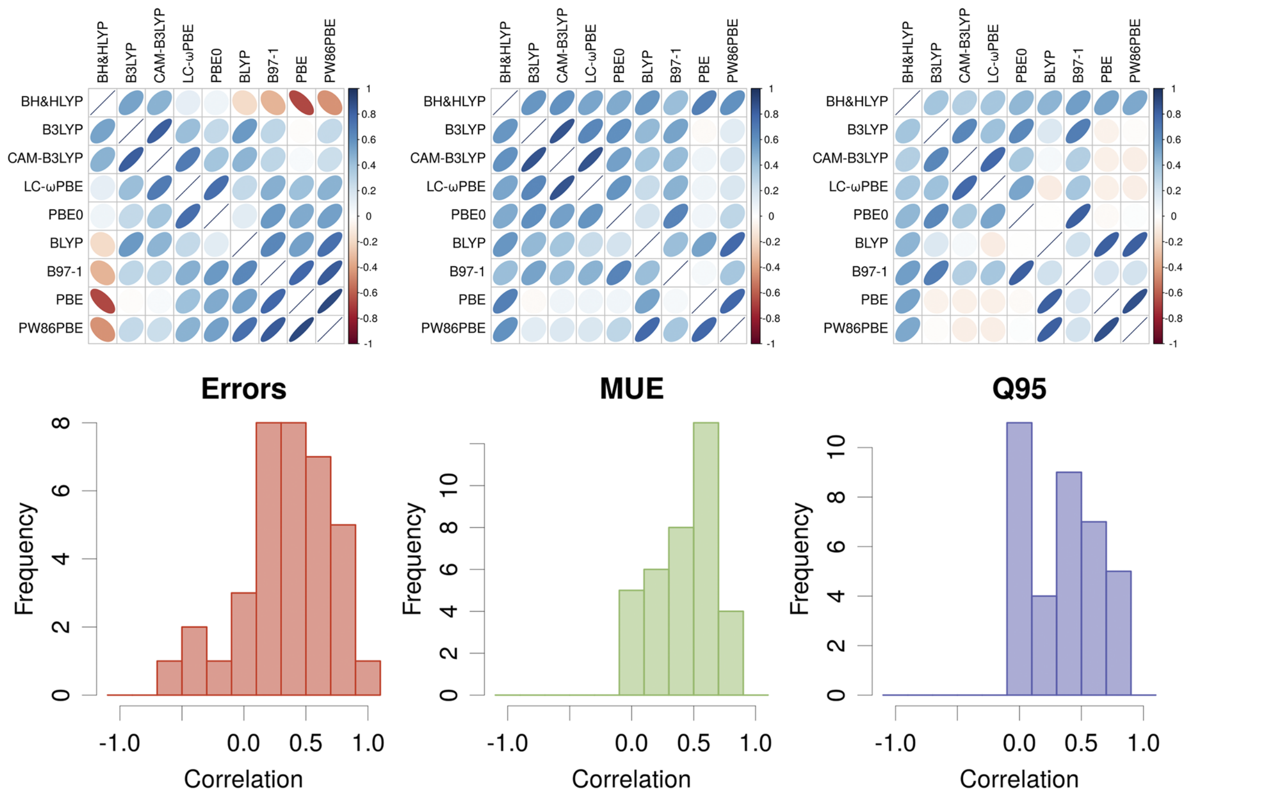

FIG. 1. Case PER2018—correlations: (top) rank correlation matrices between error sets, MUE, and Q_{95}, and (bottom) histogram of non-diagonal elements of the corresponding correlation matrices. The methods are ordered by a clustering of the errors correlation matrix by the complete linkage method[16] implemented in the `R` function `hclust`.[17]

3. SIP analysis

The SIP analysis brings another view on the head trio (B97-1, CAM-B3LYP and PBE0), as the method with the highest MSIP is CAM-B3LYP. One can see on the SIP matrix in Fig. 2 that indeed, the row for CAM-B3LYP is fully reddish, when those for B97-1 and PBE0 present also blue and white patches. We note also that B97-1 provides a nearly full improvement over BH&HLYP [SIP = 0.95(2)].

The ECDF of the difference of absolute errors for CAM-B3LYP and B97-1 helps to understand the contradiction between

TABLE II. Case PER2018—absolute error statistics: *p*-values, inversion probabilities and SIP statistics for comparison with the DFA of smallest MUE (B97-1). The best scores and the values for which $p_g > 0.05$ are in boldface. The SIP, MG and ML columns correspond to the B97-1 row of the corresponding matrices. Uncertainty is presented in parenthesis notation.

Methods	MUE (kcal/mol)	p_{unc}	p_g	P_{inv}	Q_{95} (kcal/mol)	p_{unc}	p_g	P_{inv}	MSIP	SIP	MG (kcal/mol)	ML (kcal/mol)
B3LYP	1.18(9)	0.00	0.00	0.00	4.5(5)	0.00	0.00	0.00	0.57(3)	0.53(3)	−1.05(10)	0.48(5)
B97-1	**0.85(5)**	...	...	...	**2.7(4)**	...	...	...	0.61(3)	...	...	...
BH&HLYP	4.8(2)	0.00	0.00	0.00	11.7(6)	0.00	0.00	0.00	0.06(1)	0.95(2)	−4.3(2)	0.8(2)
BLYP	1.6(1)	0.00	0.00	0.00	5.3(6)	0.00	0.00	0.00	0.43(3)	0.77(3)	−1.2(1)	0.6(1)
CAM-B3LYP	**0.90(9)**	0.64	0.57	0.29	4.1(4)	0.00	0.00	0.00	**0.74(3)**	0.33(3)	−1.3(2)	0.59(4)
LC-ωPBE	1.09(10)	0.03	0.00	0.00	4.3(5)	0.01	0.00	0.00	0.65(3)	0.43(3)	−1.1(1)	0.44(3)
PBE	2.8(2)	0.00	0.00	0.00	8.1(8)	0.00	0.00	0.00	0.30(2)	0.81(3)	−2.6(2)	0.8(1)
PBE0	**0.92(7)**	0.44	0.24	0.12	3.3(5)	0.33	0.02	0.01	0.66(3)	0.50(3)	−0.74(7)	0.61(4)
PW86PBE	1.6(1)	0.00	0.00	0.00	6.1(9)	0.00	0.00	0.00	0.49(3)	0.59(3)	−1.6(2)	0.43(6)

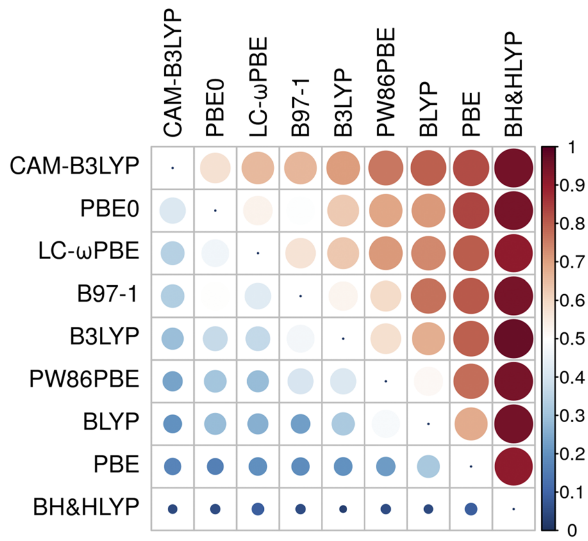

FIG. 2. Case PER2018: SIP matrix. A line with a majority of red patches signals a method with good SIP performances. The SIP value is color-coded, and the area of a disk is proportional to the corresponding value. The methods are ordered by decreasing the value of MSIP.

the MUE and MSIP ranks [Fig. 3(b)]. The MUE difference for this pair is statistically not significant (p_g = 0.57), the SIP value for CAM-B3LYP over B97-1 is 0.67 (1–0.33), the mean gain is −0.6 kcal/mol, and the mean loss is 1.3 kcal/mol, due to the heavy tail in the CAM-B3LYP error distribution (these numbers correspond to the reciprocal comparison of the one presented in Table II). So, by switching from B97-1 to CAM-B3LYP, one would have to accept a 33% risk to degrade the intensive atomization energies by 1.3 kcal/mol in average and up to 4 kcal/mol, but one would improve the estimations in 67% of the cases by 0.6 kcal/mol in average. The same comparison between CAM-B3LYP and PBE0 [Fig. 3(c)] shows that there is no strong basis to favor one of these methods.

4. Ranking

The ranking probability matrices (Fig. 4) confirm the previous analysis. The group of three methods (B97-1, CAM-B3LYP, and PBE0) at the top of the MUE ranking presents a blurred image (no clear diagonal), whereas the first Q_{95} rank of B97-1 is not ambiguous. As expected, the MSIP ranking favors solidly CAM-B3LYP. Globally, B97-1 should be preferred to minimize the risk of large errors, where CAM-B3LYP would provide, overall, smaller absolute errors.

B. BOR2019

Bandgap estimations for a set of 471 systems[18] by 15 DFAs were extracted from the supplementary material of a recent article by Borlido *et al.*[5] For a full description of the dataset, we refer the reader to the original article.

The reference authors reported and analyzed relative errors, but as there is a large range of bandgaps in this set, this causes a dispersion of relative errors over six orders of magnitude and an unsuitable distortion of the errors distributions, with large relative errors for small bandgaps and small relative errors for large bandgaps. It is true that for some methods (e.g., LDA), the errors increase with the value of the bandgap, but this is mostly due to a systematic deviation (trend), not to an increase in the dispersion of the errors. In consequence, we chose to treat here the raw errors.

Borlido *et al.*[5] discuss the uncertainties on the reference band gaps in their dataset and estimate it to a few tenths of eV. Without more detailed information, we assume that this represents a uniform uncertainty for the dataset.

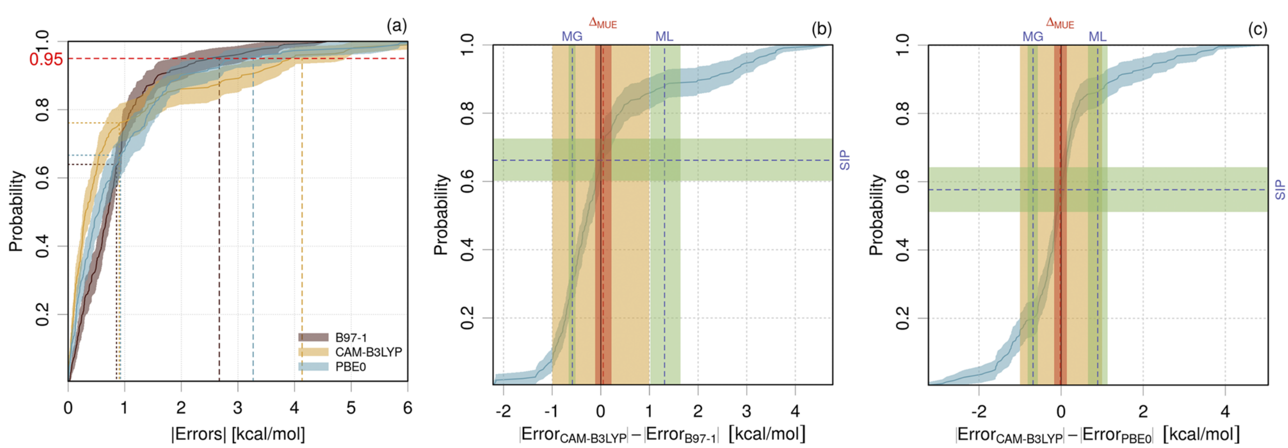

FIG. 3. Case PER2018—absolute errors statistics: (a) ECDF and statistics of absolute errors. The MUE values are depicted by vertical dotted lines, and the Q_{95} values are depicted by vertical dashed lines; [(b) and (c)] ECDF and statistics of the difference of absolute errors. The green- and red-shaded bands represent 95% confidence intervals (CIs) for the reported statistics (SIP: systematic improvement probability; MG: mean gain; ML: mean loss, and Δ_{MUE}: MUE difference). The orange bar depicts the chemical accuracy (1 kcal/mol). It is a visual aid to evaluate the pertinence of the observed differences.

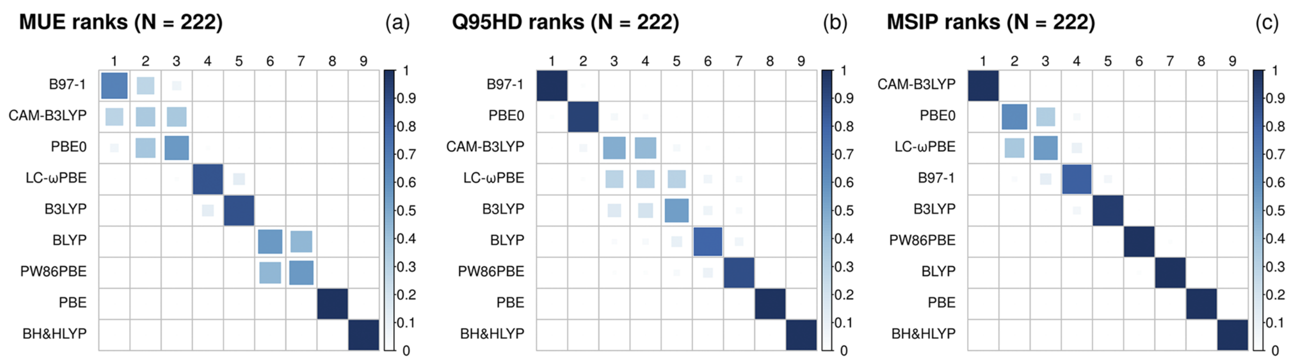

FIG. 4. Case PER2018: ranking probability matrix for (a) MUE, (b) Q_{95}, and (c) MSIP.

1. Correlations

One sees in Fig. 5 that, across the spectrum of methods, all error sets' correlation coefficients are positive and can reach very large values, up to 0.998. Only about 30% of the dataset pairs have correlation coefficients below 0.6, involving notably PBE0_mix and HSE_mix. If the error sets are dominated by method errors (i.e., there are no large reference data errors, nor outliers), the correlation matrix can be used to infer a clustering of methods, describing the relationships of the methods for the current property/dataset. Error sets with large correlation coefficients are related by a linear or monotonous transformation, and the corresponding methods are clustered together. The presence of well delimited clusters indicates that the error sets are not dominated by reference data errors. From the correlation matrix, the clusters would be (HLE16, HLE16 + SOC), (BJ, SCAN, LDA, PBE, PBE_SOL, LDA + SOC, PBE + SOC), (HSE_mix, PBE0_mix) and (HSE06,PBE0). mBJ and HSE14 stay alone. This clustering seems to produce blocks that correspond to physical intuition: LDA, PBE, SCAN, ... have all an electron-gas background. This is relaxed for HLE16 that differs from HLE16 + SOC only by taking into account spin–orbit coupling. These methods are further decoupled from hybrid methods (PBE0 and HSE06).

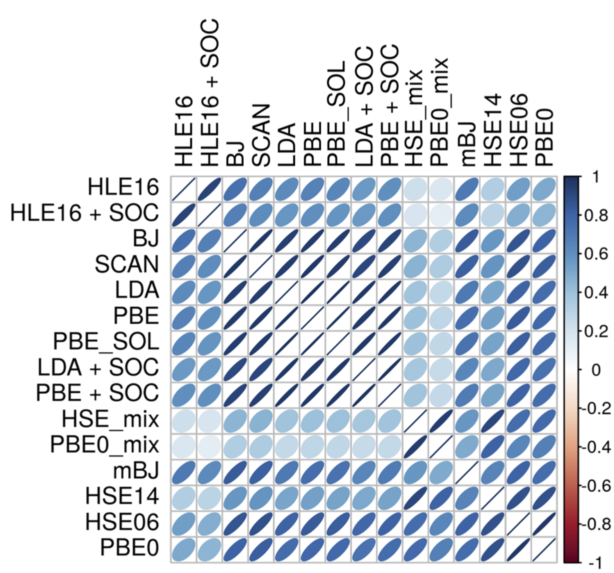

FIG. 5. Case BOR2019—rank correlation between error sets. The methods are ordered by a clustering algorithm using the complete linkage method[16] implemented in the R function `hclust`.[17]

2. Statistics

The values are reported in Table III. Although mBJ presents the smallest MUE [0.50(2) eV], the value for HSE06 is very close [0.53(5) eV] and one cannot exclude that the difference is due to a mere sampling effect ($p_g \simeq 2P_{inv} = 0.16$). Besides, HSE06 is the only method with a notably non-zero P_{inv} value with mBJ for the MUE. mBJ is also the method with the smallest Q_{95}, and no other method is able to challenge this rank. mBJ has the largest MSIP, but its value is moderate (0.7), indicating that mBJ does not provide a full systematic improvement over (some of) the other methods.

3. SIP analysis

The SIP values for mBJ lie between 0.49 and 0.86. The latter value is against LDA + SOC, which means that for 14% of the systems, LDA + SOC achieves smaller absolute errors than mBJ, despite its poor scores. Interestingly, small values, close to 0.5, are also observed against HLS16, HLSE16 + SOC, and HSE06, indicating a notable risk of performance loss for individual systems when switching from one of these methods to mBJ.

As seen in Paper I[1] (Fig. 3), when going from LDA to mBJ, one has about 15% chance to perform better using LDA, and the mean gain more than doubles the mean loss. By contrast, the comparison of mBJ to HSE06 [Fig. 6(b)] is an example of undecidability: the Δ_{MUE} is not significantly different from zero, and one has as much to lose as to gain by switching between both methods.

The SIP matrix (Fig. 7) provides a convenient summary of these observations. The mBJ line is mostly reddish with white spots indicating neutral comparisons. In contrast, the LDA + SOC line is fully blueish, indicating that it is dominated by all other methods.

TABLE III. Case BOR2019—absolute error statistics: inversion probabilities and SIP statistics for comparison with the DFA of smallest MUE (mBJ). The best scores and the values for which (p_g = $2P_{inv}$) > 0.05 are in boldface.

Methods	MUE (eV)	P_{inv}	Q_{95} (eV)	P_{inv}	MSIP	SIP	MG (eV)	ML (eV)
LDA	1.17(5)	0.00	3.2(2)	0.00	0.25(2)	0.84(2)	−0.87(4)	0.41(4)
LDA + SOC	1.24(5)	0.00	3.3(2)	0.00	0.16(2)	0.86(2)	−0.92(4)	0.38(4)
PBE	1.05(5)	0.00	3.0(2)	0.00	0.41(2)	0.82(2)	−0.76(4)	0.40(3)
PBE + SOC	1.12(5)	0.00	3.0(2)	0.00	0.30(2)	0.83(2)	−0.82(4)	0.37(4)
PBE_SOL	1.12(5)	0.00	3.1(2)	0.00	0.30(2)	0.83(2)	−0.82(4)	0.42(4)
HLE16	0.60(4)	0.00	1.9(2)	0.00	0.66(2)	0.49(2)	−0.44(4)	0.23(2)
HLE16 + SOC	0.61(4)	0.00	2.0(2)	0.00	0.65(2)	0.49(2)	−0.48(4)	0.25(2)
BJ	0.79(4)	0.00	2.3(2)	0.00	0.55(2)	0.75(2)	−0.49(3)	0.31(2)
mBJ	**0.50(2)**	...	**1.41(7)**	...	**0.69(2)**	...	...	...
SCAN	0.81(4)	0.00	2.4(2)	0.00	0.55(2)	0.74(2)	−0.53(3)	0.30(2)
HSE06	**0.53(3)**	0.09	1.7(2)	0.00	0.68(2)	0.52(2)	−0.28(3)	0.25(2)
HSE14	0.57(3)	0.00	1.8(1)	0.00	0.63(2)	0.56(2)	−0.38(2)	0.33(2)
HSE06_mix	0.64(3)	0.00	2.0(1)	0.00	0.60(2)	0.58(2)	−0.51(3)	0.36(3)
PBE0	0.78(3)	0.00	1.8(1)	0.00	0.44(2)	0.72(2)	−0.57(2)	0.46(4)
PBE0_mix	0.82(4)	0.00	2.4(2)	0.00	0.47(2)	0.66(2)	−0.67(4)	0.37(3)

4. Ranking

Ranking probability matrices for the MUE, Q_{95} and MSIP are presented in Figs. 8(a)–8(c). They illustrate the previous results and show that ranking by MUE beyond the second place becomes uncertain. This is even more notable for Q_{95}. The MSIP ranking selects the same group of five methods as the MUE ranking, with some inversions. At the opposite, an end-group of five methods is rather well ascertained for all three statistics.

These matrices are a convenient tool to visualize the impact of dataset size on the ranking quality. We estimated them for reduced error sets (N = 235 and N = 100), sampled randomly from the original one. The impact is clearly visible in Figs. 8(d)–8(i), as the diagonal contributions get weaker when N decreases. For the MUE, the block of ranks 1 and 2 is quite robust, but the situation deteriorates for the upper ranks. For Q_{95}, the first place of mBJ is very stable, but the upper ranks become very uncertain, up to the last ranks for N = 100. As for the MUE, the MSIP ranking suffers from the reduced datasets, but a head group of five methods is well preserved.

C. NAR2019

The dataset contains the calculated enthalpies of formation by G4MP2 for 469 molecules having experimental values with small uncertainty (Pedley test set).[6] The G4MP2 values are compared with those of B3LYP, M06–2X, and ωB97X-D.

1. Correlations

The most remarkable feature of the correlation matrices in Fig. 9 is the decorrelation of G4MP2 errors from the other error sets. For the MUE and Q_{95}, weak positive correlations appear, more notably for Q_{95}.

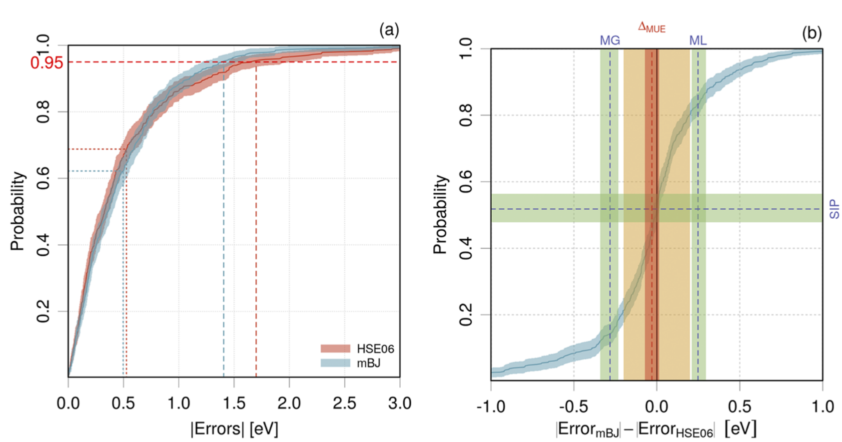

FIG. 6. Case BOR2019—absolute errors statistics: (a) ECDF of the absolute errors; (b) ECDF of the difference of absolute errors for mBJ and HSE06. See Fig. 3 for details. The orange band depicts a reasonable level of uncertainty in the dataset (0.2 eV).

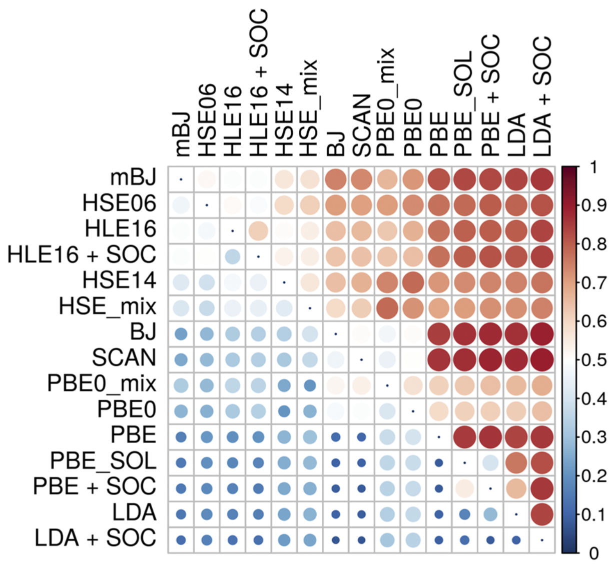

FIG. 7. Case BOR2019—SIP matrix.

2. Statistics

The statistics reported in Table IV show the supremacy of G4MP2 over the three DFAs for all statistics. Narayanan *et al.*[6] claim an "accuracy"[19] (MUE) of 0.79 kcal/mol with G4MP2. However, a look at the absolute errors CDFs [Fig. 10(a)] shows that for G4MP2, there is still a probability of about 20% that the absolute errors exceed 1 kcal/mol and 5% to exceed 2.2 kcal/mol.

3. SIP analysis

G4MP2 presents a high degree of systematic improvement over the three DFAs (MSIP = 0.81). Nonetheless, there is about 27% probability (1–0.73) that ωB97X-D performs better, but with a rather small value of ML (0.62 kcal/mol), when compared to the chemical accuracy [Fig. 10(d)]. In contrast, the mean gain when using G4MP2 instead of ωB97X-D is about −1.7 kcal/mol for 73% of the systems. The advantage of G4MP2 over B3LYP is more spectacular [Fig. 10(c)].

D. CAL2019

The impact of an atomic-charge dependent London dispersion correction (D4 model) has been evaluated by Caldeweyher *et al.*[7] on a large series of datasets. From those, we selected one of the largest ones, i.e., the reference energies for the MOR41 transition metal reaction benchmark set,[20] available as Tables 14–18 in the supplementary material of the reference article.[21] The reference data are calculated values, with *a priori* no significant numerical uncertainty. The London dispersion corrections have been tested on a series of 10 DFAs. Note that the nomenclature used here for the corrections is the one provided in the supplementary material, table, which differs somewhat from the one used in the reference article.

It is important to note that we picked here for illustration a single dataset among the dozen used in the original study,[7] and that our conclusions for MOR41 are not generalizable to the other datasets.

1. Statistics

The results are reported in Table V, where DFT-D3 has been taken as reference throughout for P_{inv} estimation. The aim here is to check if DFT-D4 brings significant differences. It is notable that with a set of size 41, the sampling uncertainty is rather large for both statistics (typically on the second or first digit). Nevertheless, significant MUE improvements are observed when passing from DFT-D3 to DFT-D4, except for revPBE and PW6B95. In the latter case, the better MUE of the D3 calculations, noted by the reference authors, might be due to a random effect of dataset selection. Based on Q_{95} the improvements due to D4 are not significant, except for DOD-PBE, DSD-PBE, and RPBE. So, for most of the studied DFAs, DFT-D4 improves the MUE, but does not reduce the Q_{95} values for the MOR41 dataset. Note that comparisons of Q_{95} values have to be performed with care, considering the small size of the dataset.

2. SIP analysis

Let us consider several examples with the SIP approach:

- **PBE0-Dn**. Inspection of Fig. 11(a) shows that the 95% confidence interval (CI) for the SIP value of 0.61 for PBE0-D4-ATM over PBE0-D3 does not exclude the neutral value (0.5), with a tiny advantage of the mean gain over the mean loss. One can note also that, despite the large uncertainty on the MUE values 2.3(3) and 2.6(4), the small difference Δ_{MUE} = 0.3 between these two methods is significantly different from 0 (its 95% confidence interval excludes 0), an effect of the strong positive correlation between the error sets (0.98) as discussed in Paper I.[1]
- **PW6B95-Dn**. This case is an inversion of the previous one, where the confidence interval on the SIP value of nearly 0.4 (disadvantaging D4) does not exclude the neutral value, and the CI on the MUE difference Δ_{MUE} does not exclude 0. One cannot firmly conclude that the D3 version performs better than the D4 ones for this DFA.
- **RPBE-Dn**. For this case, one has a rare instance where D4 improves almost systematically over D3, with a SIP of 0.95(3), and a mean gain overwhelming the mean loss.

Except for RPBE-Dn, where the SIP value of D4 over D3 is about 0.95 and DOD-PBE (SIP = 0.83), all the estimated SIP values lie near or below 0.75, down to 0.45, meaning that there is no systematic improvement when passing from D3 to D4 for this dataset. In several cases, the uncertainty due to the limited set size does not allow to conclude clearly.

3. Ranking

Considering that both DFT-D4 options are mostly indiscernible over the MOR41 dataset, we built global ranking probability matrices for the DFT-D3 and DFT-D4-ATM data. The results are reported in Fig. 12 (top). Although the rankings of the Dn options for each DFA are mostly unambiguous, a global ranking is clearly very uncertain. Based on the MUE, DOD-PBE-D4-ATM and PBE0-D4-ATM would share the leading places. Beyond that, the situation

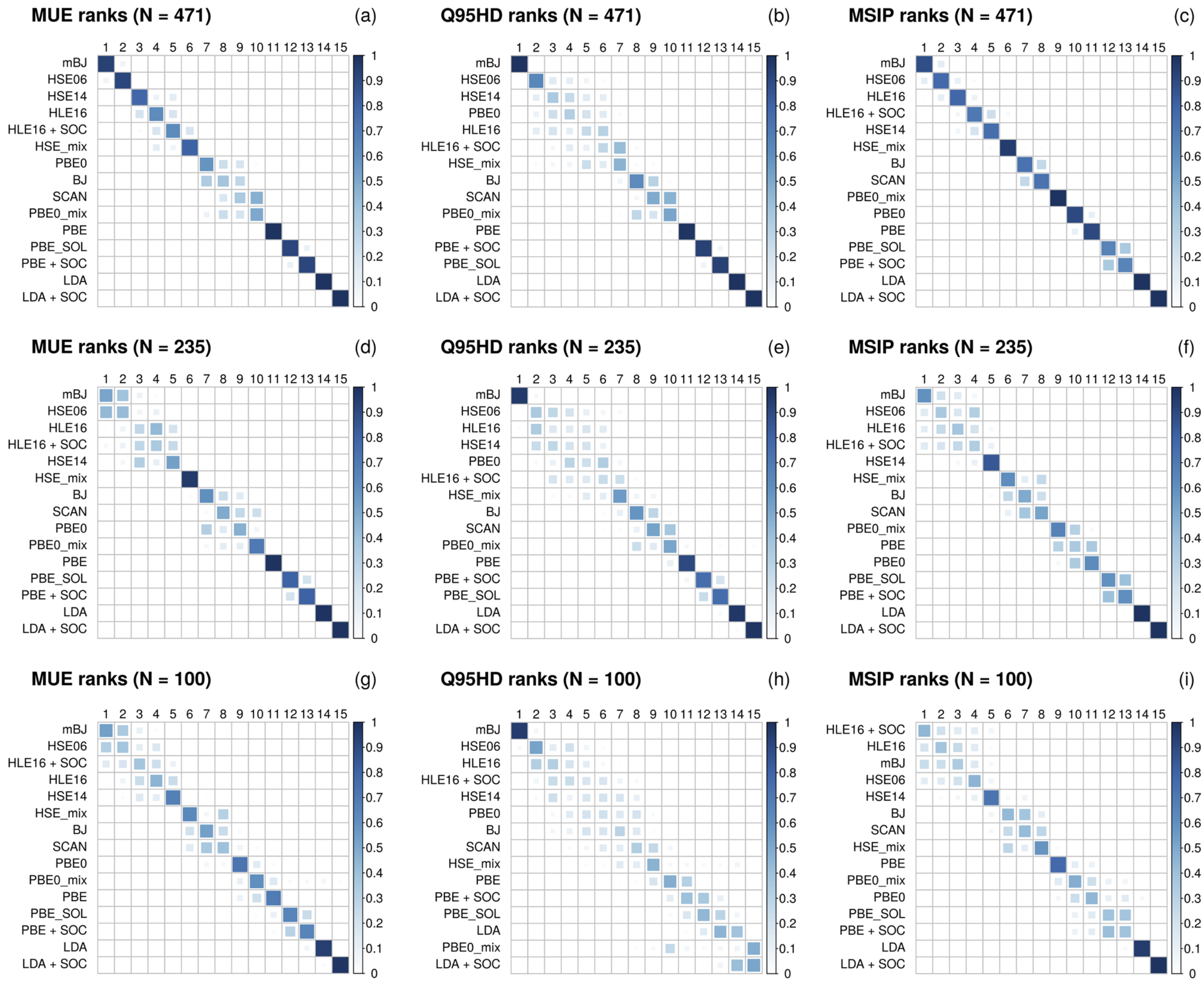

FIG. 8. Case BOR2019: ranking probability matrices for the full dataset, N = 471 [(a)–(c)], and for random reduced sets N = 235 [(d)–(f)] and N = 100 [(g)–(i)].

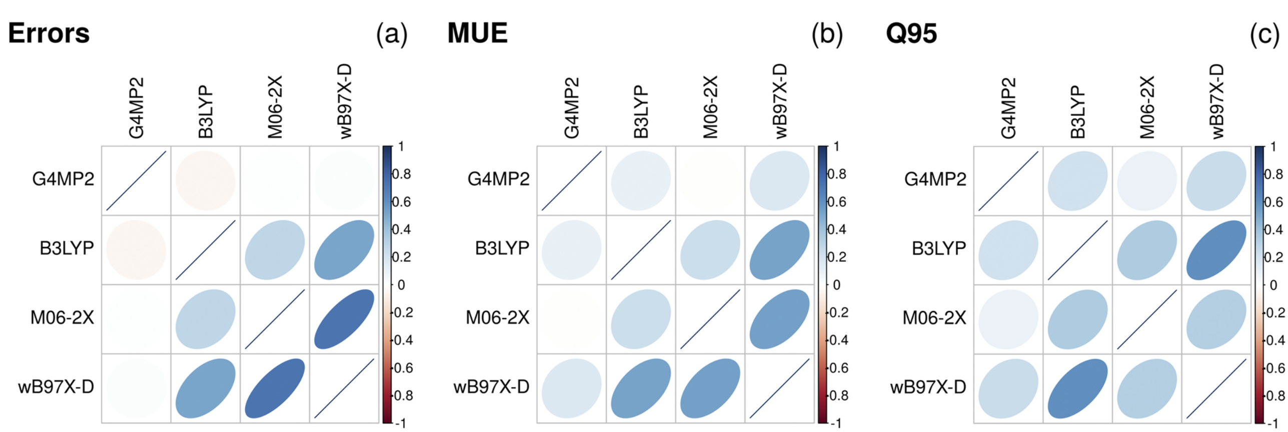

FIG. 9. Case NAR2019—rank correlation matrices: (a) errors; (b) MUE; and (c) Q_{95}.

TABLE IV. Case NAR2019—absolute error statistics: inversion probabilities and SIP statistics for comparison with the DFA of smallest MUE (G4MP2). The best scores are in boldface.

Methods	MUE (kcal/mol)	P_{inv}	Q_{95} (kcal/mol)	P_{inv}	MSIP	SIP	MG (kcal/mol)	ML (kcal/mol)
G4MP2	**0.79(3)**	…	**2.21(9)**	…	**0.81(2)**	…	…	…
B3LYP	4.0(2)	0.0	9.3(6)	0.0	0.22(2)	0.89(1)	−3.7(2)	0.52(7)
M06–2X	2.71(10)	0.0	6.1(5)	0.0	0.37(2)	0.83(2)	−2.5(1)	0.82(7)
ωB97X-D	1.85(9)	0.0	5.2(4)	0.0	0.59(2)	0.73(2)	−1.7(1)	0.62(5)

is utterly scrambled, the only clear point being the last ranks for M06-L-D3 and RPBE-D3. The picture based on Q_{95} is even less well defined, with no clear leading method within a head group of five. The MSIP ranking is akin to the MUE ranking.

If one restricts the methods to DFT-D4-ATM (Fig. 12, bottom), the situation is slightly better defined for the leading and tailing places for the three scores, but remains very undecidable in intermediate ranks. This illustrates how, for a given dataset, the uncertainty in ranking is also affected by the number of methods to be ranked.

E. JEN2018

This dataset contains non-covalent interaction energies estimated by M06-L with six different basis sets for 66 systems in the S66 dataset.[22,23] This is a part of the results reported in Table 8 of a recent article by Jensen[8] and available as supplementary material to this article. This dataset was used by Jensen to study the impact of error cancellations when using standard or optimized medium-sized basis sets. Six basis sets are considered (pop2 = 6-31G(d,p), pop3 = 6-311G(2df,2pd), pcseg-1, pcseg-4, pop2-opt, and pcseg1-opt), where

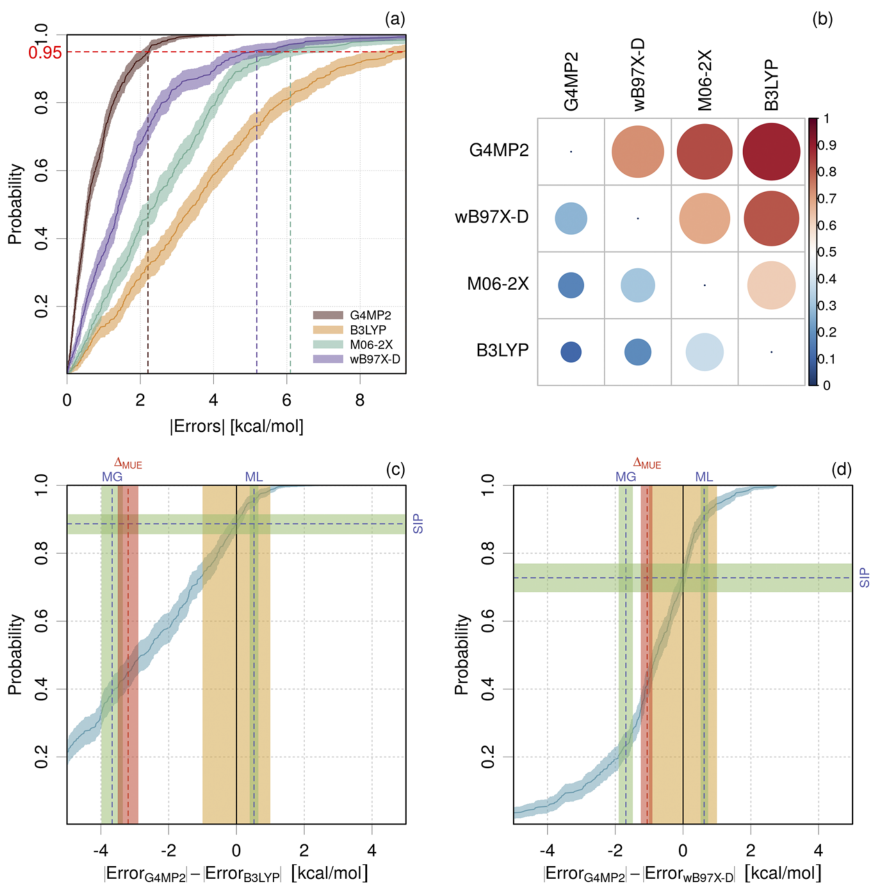

FIG. 10. Case NAR2019: (a) ECDF of the absolute errors; (b) SIP matrix; and [(c) and (d)] ECDF of the difference of absolute errors of B3LYP and ωB97X-D with respect to G4MP2 (see Fig. 3 for details).

TABLE V. Case CAL2019—absolute error statistics: inversion probabilities are calculated for comparison with DFT-D3, for each DFT. The SIP statistics are calculated for comparison with the smallest MUE within each DFT. The best scores and the values for which $p_g > 0.05$ are in boldface.

Methods	MUE (kcal/mol)	P_{inv}	Q_{95} (kcal/mol)	P_{inv}	MSIP	SIP	MG (kcal/mol)	ML (kcal/mol)
DOD-PBE-D4-ATM	**2.1(4)**	0.00	**7(2)**	0.00	0.63(6)	...	...	...
DOD-PBE-D4-MBD	**2.1(4)**	0.00	**8(2)**	0.00	0.65(6)	0.44(8)	−0.28(4)	0.24(5)
DOD-PBE-D3	3.5(4)	...	10(2)	...	0.13(4)	0.83(6)	−1.8(3)	0.8(2)
DSD-PBE-D4-ATM	**2.9(5)**	0.00	**11(3)**	0.00	0.35(4)	...	...	...
DSD-PBE-D4-MBD	**2.9(5)**	0.00	**11(3)**	0.00	0.35(4)	0	0	0
DSD-PBE-D3	3.7(5)	...	12(2)	...	0.29(6)	0.71(7)	−1.5(2)	0.7(1)
B3LYP-D4-ATM	**4.2(5)**	0.00	**11(3)**	0.07	0.56(6)	0.41(8)	−0.22(4)	0.21(3)
B3LYP-D4-MBD	**4.2(5)**	0.00	**11(3)**	0.11	0.56(6)	...	...	...
B3LYP-D3	4.8(6)	...	**13(3)**	...	0.26(6)	0.71(7)	−1.1(2)	0.8(2)
PBE0-D4-ATM	**2.3(3)**	0.01	**8(1)**	0.08	0.30(4)	...	...	...
PBE0-D4-MBD	**2.3(3)**	0.01	**8(1)**	0.08	0.30(5)	0	0	0
PBE0-D3	2.6(4)	...	**8(1)**	...	0.29(6)	0.61(8)	−0.7(1)	0.4(1)
PW6B95-D4-ATM	**3.2(4)**	0.02	**7.9(9)**	0.30	0.35(6)	0.56(8)	−1.6(2)	1.0(2)
PW6B95-D4-MBD	**3.0(4)**	0.08	**7.8(8)**	0.31	0.48(6)	0.54(8)	−1.3(2)	1.0(2)
PW6B95-D3	**2.7(4)**	...	**7.4(9)**	...	0.55(6)	...	...	...
CAM-B3LYP-D4-ATM	**3.7(4)**	0.00	**9(1)**	0.04	0.38(4)	...	...	...
CAM-B3LYP-D4-MBD	**3.7(4)**	0.00	**9(1)**	0.04	0.38(4)	0	0	0
CAM-B3LYP-D3	4.3(4)	...	**10(1)**	...	0.20(5)	0.76(7)	−0.8(1)	0.5(1)
revPBE-D4-ATM	**3.3(5)**	0.11	**12(2)**	0.33	0.43(6)	0.54(8)	−0.27(6)	0.28(6)
revPBE-D4-MBD	**3.3(6)**	0.08	**12(2)**	0.39	0.54(7)	...	...	...
revPBE-D3	**3.8(6)**	...	**12(1)**	...	0.46(7)	0.54(8)	−2.0(4)	1.3(3)
M06L-D4-ATM	**5.1(6)**	0.00	**13(1)**	0.08	0.35(4)	...	...	...
M06L-D4-MBD	**5.1(6)**	0.00	**13(1)**	0.08	0.35(4)	0	0	0
M06L-D3	5.5(6)	...	**14(1)**	...	0.22(5)	0.71(7)	−0.7(1)	0.5(2)
PBE-D4-ATM	**3.5(5)**	0.00	**12(2)**	0.34	0.45(6)	0.51(8)	−0.20(5)	0.16(2)
PBE-D4-MBD	**3.4(5)**	0.00	**12(2)**	0.48	0.60(6)	...	...	...
PBE-D3	3.9(5)	...	**12(2)**	...	0.30(6)	0.68(7)	−1.0(1)	0.5(2)
RPBE-D4-ATM	**3.4(6)**	0.00	**12(2)**	0.00	0.48(2)	...	...	...
RPBE-D4-MBD	**3.4(6)**	0.00	**12(2)**	0.00	0.48(2)	0	0	0
RPBE-D3	8.3(9)	...	20(5)	...	0.05(3)	0.95(3)	−5.3(7)	2(1)

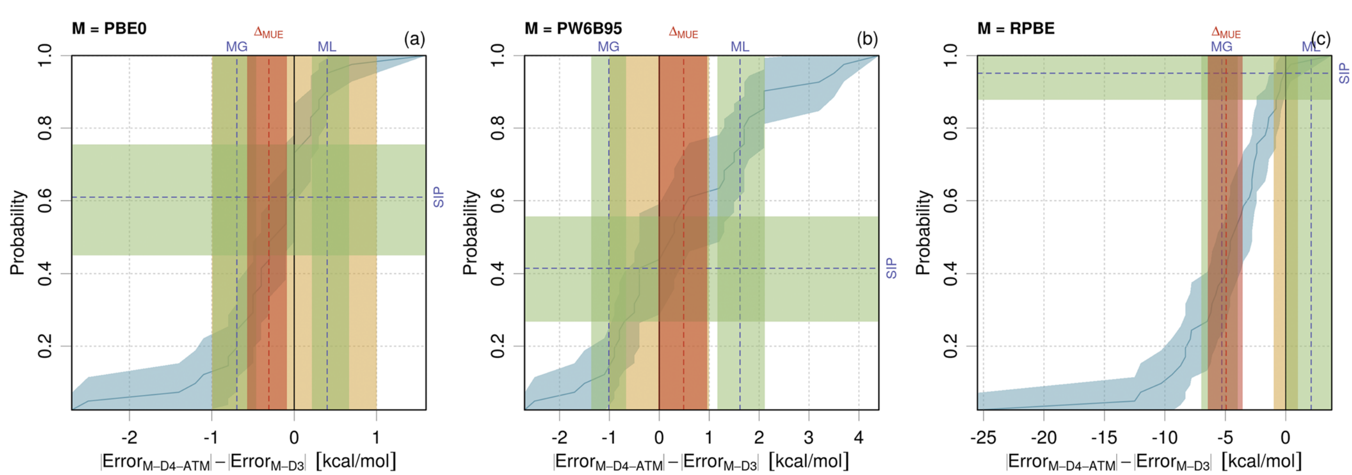

FIG. 11. Case CAL2019—selected SIP plots. The orange band depicts the chemical accuracy (1 kcal/mol).

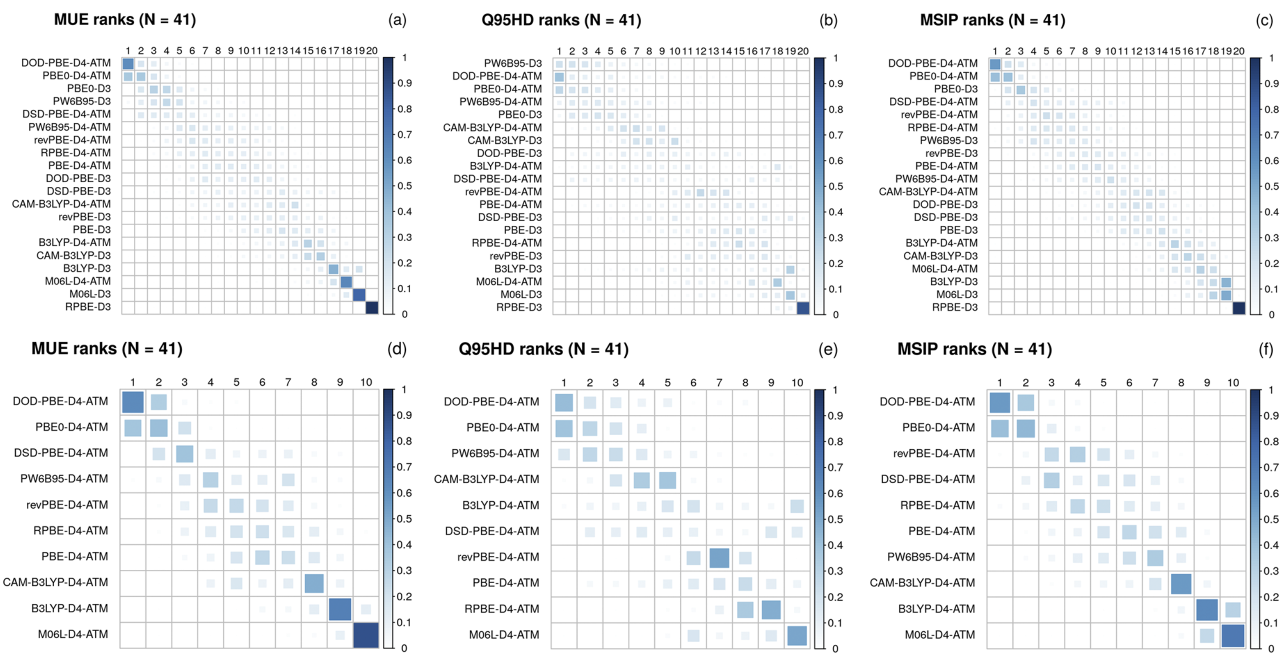

FIG. 12. Case CAL2019: ranking probability matrices for [(a)–(c)] DFT-D3 and DFT-D4-ATM methods and [(d)–(f)] DFT-D4-ATM methods only.

the "-opt" ones have optimized contraction coefficients with respect to the reference data.

1. Correlations

The error sets of the "-opt" methods are practically uncorrelated to the other sets [Fig. 13(a)], and in the remaining methods, pcseg4 errors are anti-correlated with the other ones. A striking feature of this dataset is that this negative correlation persists for the MUE, contradicting the trends observed in Appendix B of Paper I.[1] Otherwise, the correlations globally weaken for Q_{95}, except for the pop2/pop3 and pcseg1/pcseg1-opt cases, for which the correlation is stronger as the one between the error sets.

2. Statistics

The statistics in Table VI show the strong impact of basis-set optimization, and both optimized basis sets provide comparable results for the MUE and Q_{95}. All statistics show that the ranking between both "-opt" methods is not strict.

3. SIP analysis

They both also stand out by their MSIP, with a slight advantage for pcseg1-opt. Once again, the importance of error cancellations stands out through the medium values of the SIP of pcseg1-opt over the other cases. The strongest improvement is 0.9 over pcseg4, and

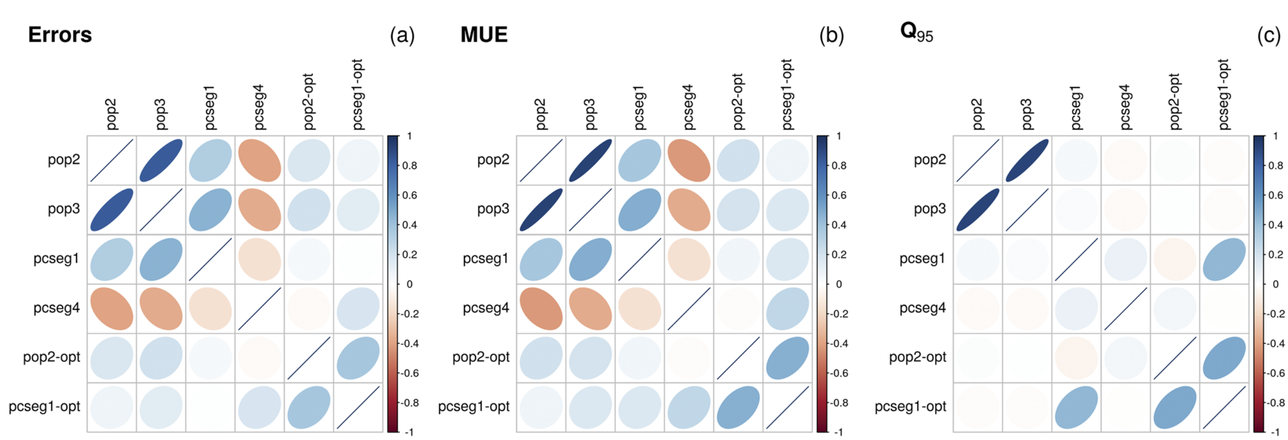

FIG. 13. Case JEN2018—rank correlation matrices: (a) errors; (b) MUE; and (c) Q_{95}.

TABLE VI. Case JEN2018—absolute error statistics: inversion probabilities and SIP statistics for comparison with the method of smallest MUE (pcseg1-opt). The best scores and the values for which ($p_g = 2P_{inv}$) > 0.05 are in boldface.

Methods	MUE (kJ/mol)	P_{inv}	Q_{95} (kcal/mol)	P_{inv}	MSIP	SIP	MG (kJ/mol)	ML (kJ/mol)
pop2	2.9(3)	0.00	7.2(7)	0.00	0.35(5)	0.77(5)	−2.9(3)	0.8(2)
pop3	2.4(3)	0.00	6.4(7)	0.00	0.47(5)	0.74(5)	−2.3(3)	0.8(1)
pcseg1	2.5(2)	0.00	5.6(4)	0.00	0.42(5)	0.76(5)	−2.3(2)	0.9(2)
pcseg4	2.5(1)	0.00	4.8(4)	0.00	0.33(5)	0.89(4)	−1.8(1)	0.6(2)
pop2-opt	**1.06(10)**	0.05	**2.6(2)**	0.24	0.67(5)	0.62(6)	−0.66(8)	0.65(9)
pcseg1-opt	**0.90(9)**	...	**2.5(3)**	...	**0.76(5)**	...	...	...

the smallest is 0.6 over pop2-opt. The plots in Fig. 14 illustrate these features. The SIP matrix shows clearly that the optimized basis sets provide a partial improvement and a slight advantage of pcseg1-opt over pop2-opt. The major gain when going from pop2 to pop2-opt is visible in Fig. 14(c) where the medium SIP (~0.7) is compensated by the very small mean loss (0.6 kJ/mol). In contrast, Fig. 14(d) shows that the improvement of pcseg1-opt over pop2-opt is marginal, with SIP values close to the neutral value (0.5) and symmetrical MG and ML values.

4. Ranking

The leading position of the "-opt" methods is solid and confirmed by our three scores (Fig. 15).

F. DAS2019

A set of 24 dielectric constants for 3D metal oxides has been reported by Das *et al.*[9] in their work (Table 3). One of the experimental values being unknown, the dataset is limited to 23 values.

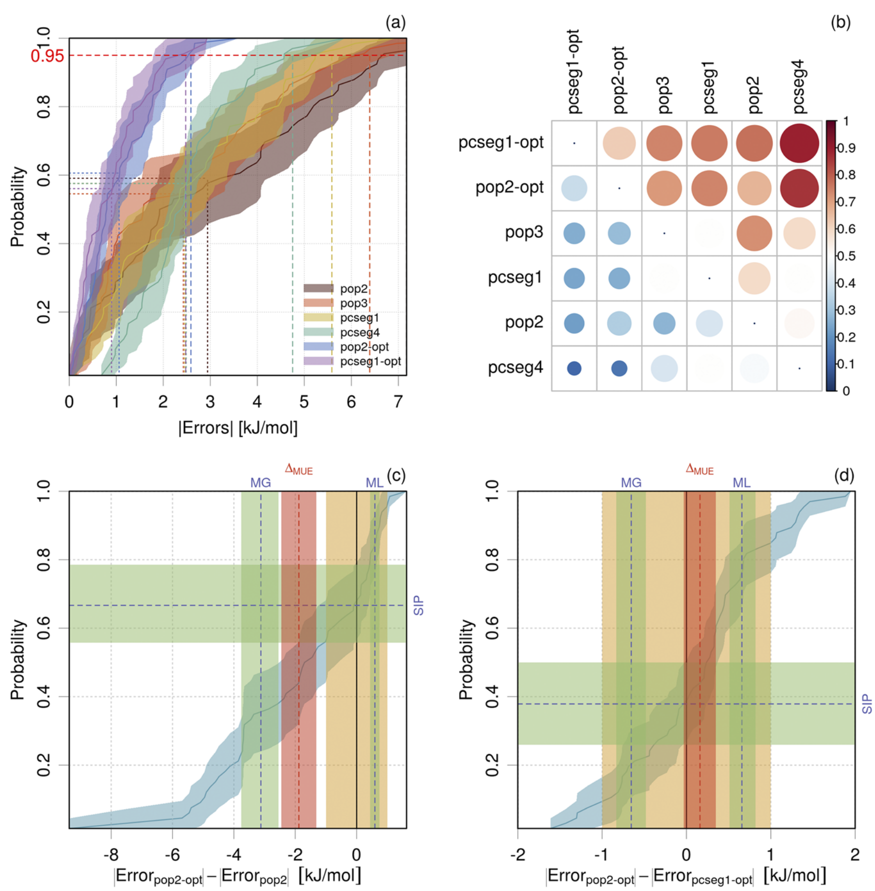

FIG. 14. Case JEN2018: (a) ECDF of the absolute errors; (b) SIP matrix; and [(c) and (d)] ECDF of the difference of absolute errors of pop2 and pcseg1-opt with respect to pop2-opt (see Fig. 3 for details). The orange bar represents a chemical accuracy of 1 kJ/mol.

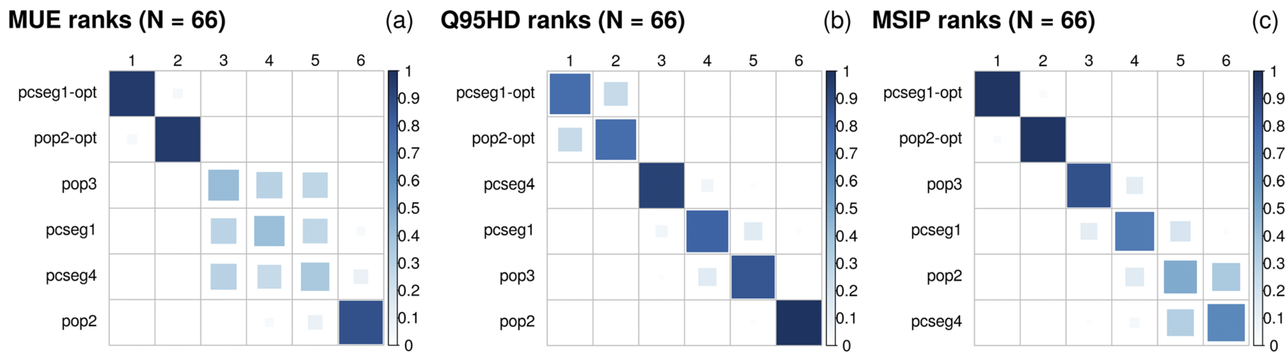

FIG. 15. Case JEN2018: ranking probability matrices.

Experimental uncertainties are not specified. The predictions by six DFAs are reported, three global hybrids (PBE0, B3LYP, and DD-B3LYP) and three range-separated hybrids (SC-BLYP, DD-SCBLYP, and DD-CAM-B3LYP). This is a small dataset, below the standards, required for low type I errors (false positive) in the comparison of MUE ($N > 30$) and Q_{95} ($N > 60$) (Paper I[1]-Appendix C).

1. Correlations

The correlation matrices of the errors, MUE, and Q_{95} have uniformly strongly positive elements (Fig. 16, top). This is an unusual situation when compared to the previous cases. Knowing that correlation coefficients are sensitive to outliers (even if rank correlation

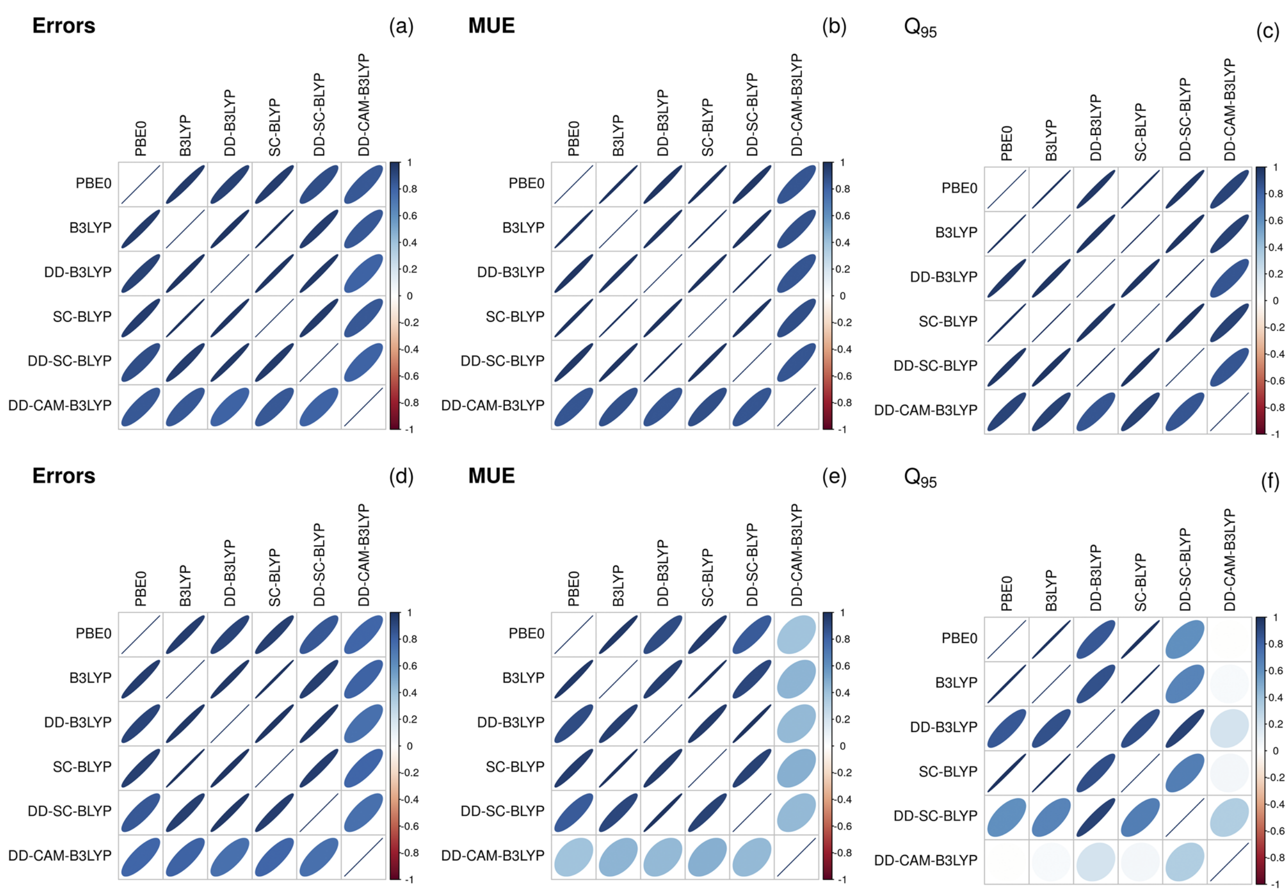

FIG. 16. Case DAS2019—rank correlation matrices: [(a)–(c)] original data set ($N = 23$); [(d)–(f)] after removal of two outliers ($N = 21$).

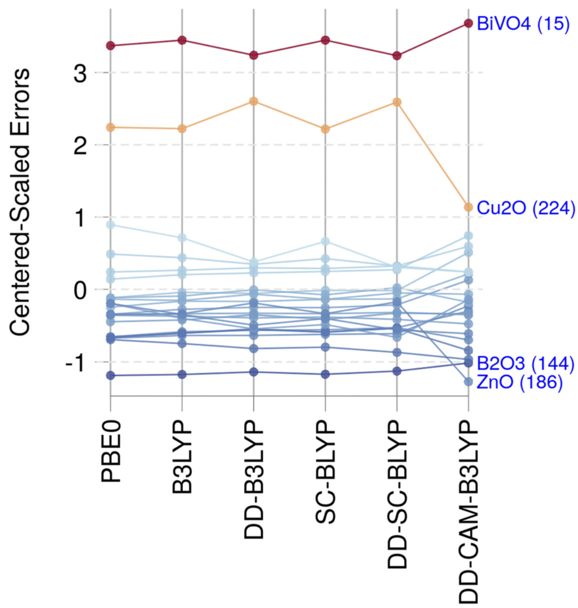

FIG. 17. Case DAS2019: parallel plot of scaled and centered error sets, used to identify global outliers.

is a little more robust), we explored the dataset for outliers. A parallel plot (Fig. 17) of the scaled and centered error sets enables to identify systems that deviate significantly from the core distribution for all methods (global outliers). Two such systems exist for all methods: $BiVO_4$ and Cu_2O. After removal of these two points, the correlation matrix for the errors is slightly relaxed (the smallest correlation coefficient decreases from 0.81 to 0.74), but that for MUE and Q_{95} are visibly more affected [(Fig. 16, bottom)]. In fact, the parallel plot reflects the strong correlations between all error sets (many quasi-parallel horizontal lines), except for DD-CAM-B3LYP. The pruned dataset (N = 21) is used in the following.

2. Statistics

Considering the small size of the sample, few clear-cut conclusions are possible. Only DD-CAM-B3LYP stands out significantly, either by its MUE, Q95, or MSIP values (Table VII). On the contrary, although its MUE and Q_{95} values are not distinguishable from those of PBE0, B3LYP, SC-BLYP, and DD-SC-BLYP, DD-B3LYP is the worst performer of the group based on the SIP statistics.

3. SIP analysis

The best and worse methods are clearly identifiable in the SIP matrix [Fig. 18(a)], with a full reddish line for DD-CAMB3LYP and a full bluish line for DD-B3LYP. The impact of the small set size on this conclusion is illustrated in Figs. 18(b) and 18(c), where the ECDFs of the differences of absolute errors are plotted for DD-CAM-B3LYP vs B3LYP and DD-B3LYP vs B3LYP. Despite being very large, the error bars on the statistics enable to validate these conclusions.

4. Ranking

All ranking matrices confirm a solid leading place for DD-CAM-B3LYP (Fig. 19). The MUE and MSIP rankings would then favor SC-BLYP and B3LYP, in disagreement with the Q_{95} ranking, for which the three DD-X methods have leading ranks. An example of an N'-out of-N bootstrap ($N' = N/3$) is shown on the bottom row. The uncertainty is slightly enhanced, notably for the Q_{95} ranks above the first, but the main features are mostly preserved.

G. THA2015 AND WU2015

Thakkar and Wu[10] compiled a database of polarizabilities for 135 molecules, from triatomics to 26-atom systems. The experimental data are given with their uncertainty, and computational results are provided for seven methods. Dataset THA2015 for our study was extracted from Tables II–IV of the reference article. The raw errors present a dispersion increasing with the polarizability; hence, relative errors are used in the reference article and this study.

The relative uncertainties for the reference experimental data cover a large range, from 0.09% to 12.4%, and the median value is 1.7%. The authors identified eight outliers and a total of 32 systems in need of further experimental study. The outliers do not contain the points with the extreme uncertainties, so that even after removal of the 32 problematic systems, the range of relative uncertainties stays the same. The dispersion of uncertainties would certainly justify the use of weighted statistics. This was not the choice of Thakkar *et al.*, and we proceed with unweighted statistics, keeping in mind that the results might be influenced by reference data errors instead of model errors.

TABLE VII. Case DAS2019—absolute error statistics for the pruned dataset (N = 21): inversion probabilities and SIP statistics for comparison with the DFA of smallest MUE (DD-CAM-B3LYP). The best scores are in boldface.

Methods	MUE (a.u.)	P_{inv}	Q_{95} (a.u.)	P_{inv}	MSIP	SIP	MG (a.u.)	ML (a.u.)
PBE0	0.66(9)	0.00	1.6(2)	0.00	0.47(9)	0.76(9)	−0.44(8)	0.19(4)
B3LYP	0.61(8)	0.00	1.4(2)	0.00	0.49(8)	0.76(10)	−0.38(6)	0.21(6)
DD-B3LYP	0.70(7)	0.00	1.30(7)	0.00	0.19(8)	0.90(6)	−0.41(6)	0.4(1)
SC-BLYP	0.58(8)	0.00	1.3(1)	0.00	0.62(8)	0.76(9)	−0.36(6)	0.22(7)
DD-SC-BLYP	0.68(7)	0.00	1.23(5)	0.00	0.29(8)	0.90(6)	−0.39(6)	0.4(1)
DD-CAM-B3LYP	**0.36(6)**	…	**0.83(7)**	…	**0.82(8)**	…	…	…

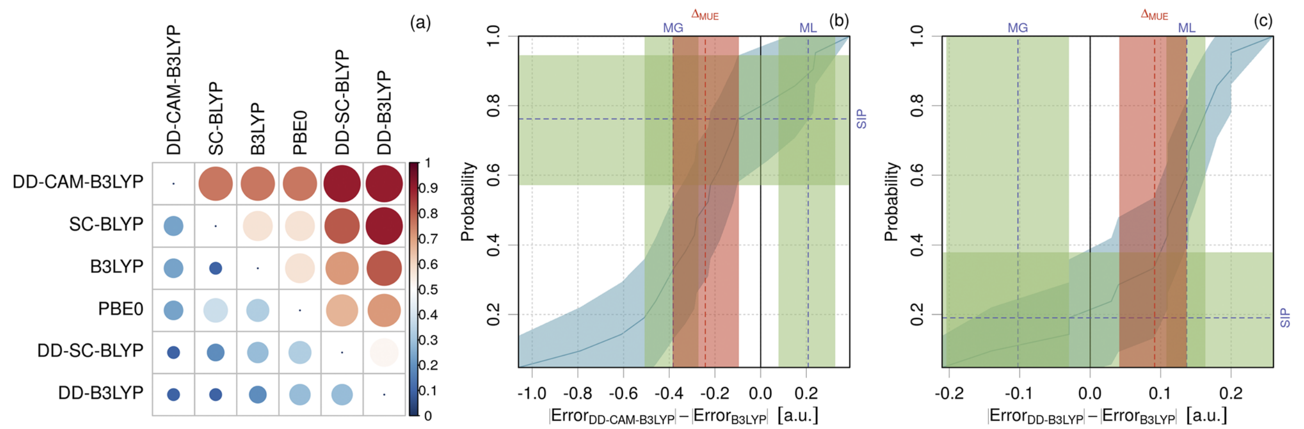

FIG. 18. Case DAS2019: (a) SIP matrix; (b) ECDF of the difference of absolute errors of methods DD-CAMB3LYP and B3LYP; (c) idem for DD-B3LYP and B3LYP (see Fig. 3 for details).

In a complementary study, Wu *et al.*[11] calculated polarizabilities for a set of 145 molecules with HF, MP2, CCSD(T), and 34 DFAs. In this study, CCSD(T) was used as reference to evaluate the other methods. In the following, we select the subset of seven methods common to both datasets (WU2015). This enables us to study the impact of the reference data (experimental vs calculated) on the correlation and ranking matrices.

1. Correlations

The Pearson correlation matrix of the error sets [Fig. 20(a)] is uniformly strongly positive. The smallest CC value is 0.8. To appreciate the role of data points with large deviations (outliers) in these strong correlations, we removed a set of eight outliers identified by Thakkar *et al.*[10] [Fig. 20(b)]. Most of the correlations weaken

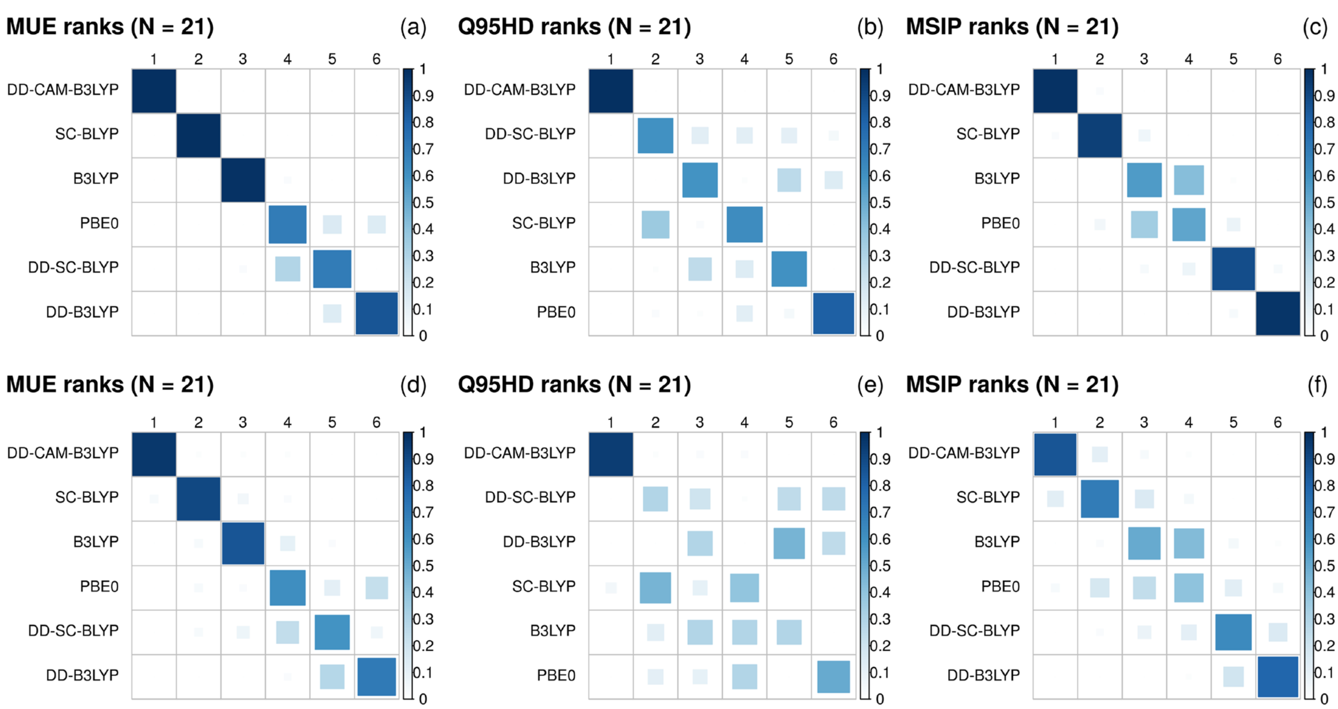

FIG. 19. Case DAS2019—ranking probability matrices: [(a)–(c)] *N*-out of-*N* bootstrap; [(d)–(f)] *N*/3-out of-*N* bootstrap.

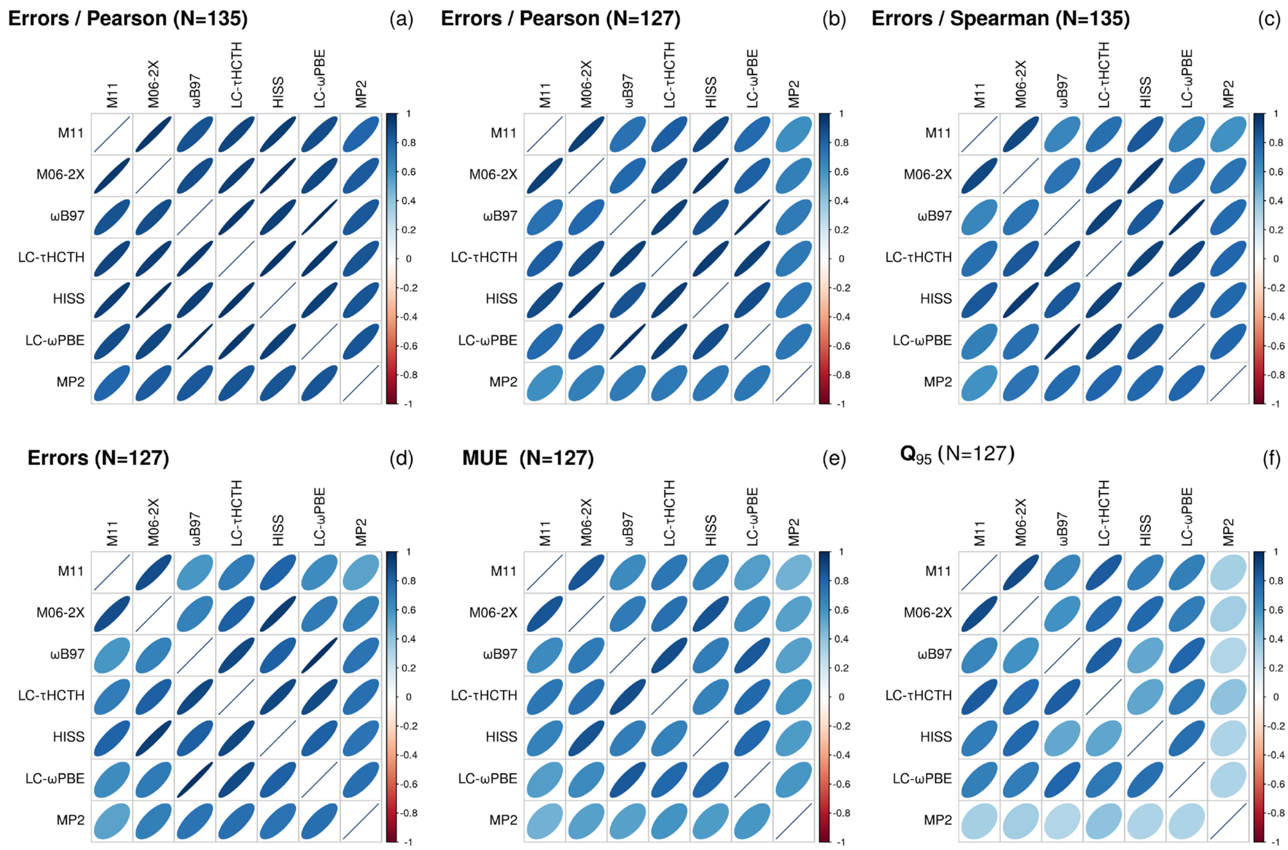

FIG. 20. Case THA2015—correlation matrix: (a) Pearson correlation of the full data set (N = 135); (b) Pearson correlation of the pruned dataset (N = 127); (c) Spearman/rank correlation of the full data set; (d): Errors rank correlation; (e): MUE rank correlation; and (f) Q_{95} rank correlation.

notably. For comparison, the rank correlation matrix was calculated for the full dataset [Fig. 20(c)]. This matrix is very similar to the one with outliers removed, illustrating the better resilience of rank correlations to outliers. Finally, the errors, MUE ,and Q_{95} rank correlation matrices were estimated on the pruned (N = 127) dataset [Figs. 20(d)–20(f)]. Globally, the structure of the errors correlation matrix seems to be transferred to the statistics, with attenuated correlation intensities.

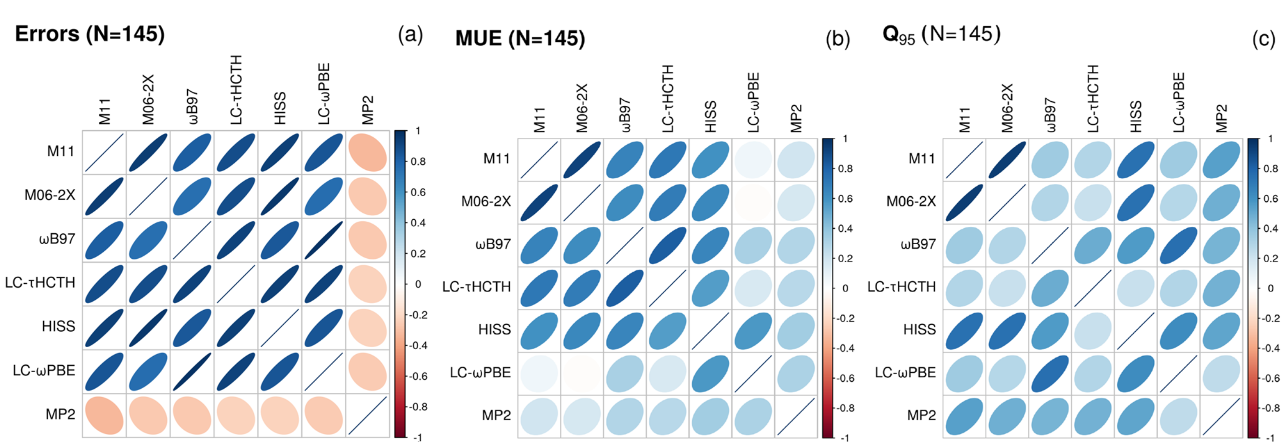

FIG. 21. Case WU2015—rank correlation matrix: (a) errors; (b) MUE; and (c) Q_{95}.

TABLE VIII. Case THA2015—absolute error statistics for the full dataset (N = 145): inversion probabilities and SIP statistics for comparison with the DFA of smallest MUE (LC-τHCTH), except for Q_{95} inversion probability, where the reference is the DFA with smallest Q_{95}. The best scores and the values for which ($p_g = 2P_{inv}$) > 0.05 are in boldface.

Methods	MUE (%)	P_{inv}	Q_{95} (%)	P_{inv}	MSIP	SIP	MG (%)	ML (%)
M11	**3.1(3)**	0.34	**10(1)**	...	0.58(4)	0.47(4)	−1.4(1)	1.16(10)
M06–2X	**3.2(3)**	0.09	**10(2)**	0.50	0.57(4)	0.53(4)	−1.2(1)	1.0(1)
ωB97	3.3(3)	0.00	**11(2)**	0.21	0.53(4)	0.59(4)	−0.94(7)	0.72(7)
LC-τHCTH	**3.0(3)**	...	**10(2)**	0.30	**0.59(4)**	...	...	...
HISS	3.8(3)	0.00	**10(2)**	0.38	0.34(4)	0.72(4)	−1.62(10)	1.5(1)
LC-ωPBE	3.9(3)	0.00	**11(1)**	0.25	0.31(4)	0.78(3)	−1.39(8)	1.2(1)
MP2	**3.2(3)**	0.22	**11(2)**	0.34	0.56(4)	0.45(4)	−1.3(3)	0.8(1)

The error, MUE, and Q_{95} rank correlation matrices were also calculated for the WU2015 dataset (Fig. 21). In the absence of reference data uncertainties, MP2 errors are now weakly anticorrelated to the other error sets, while all DFAs remain positively correlated.

The differences between both sets of correlation matrices, notably when MP2 is concerned, might be due in a large part to the presence of large experimental errors in the THA2015 dataset.

2. Statistics

The values of MUE and Q_{95} for the full THA2015 dataset are reported in Table VIII. The MUE values agree with those of the reference article, but the uncertainty bears on the second digit, showing that a third digit is essentially irrelevant. The analysis of P_{inv} for the MUE leads us to conclude that there is a group of four methods (M11, M06–2X, LC-τHCTH, and MP2) with similar

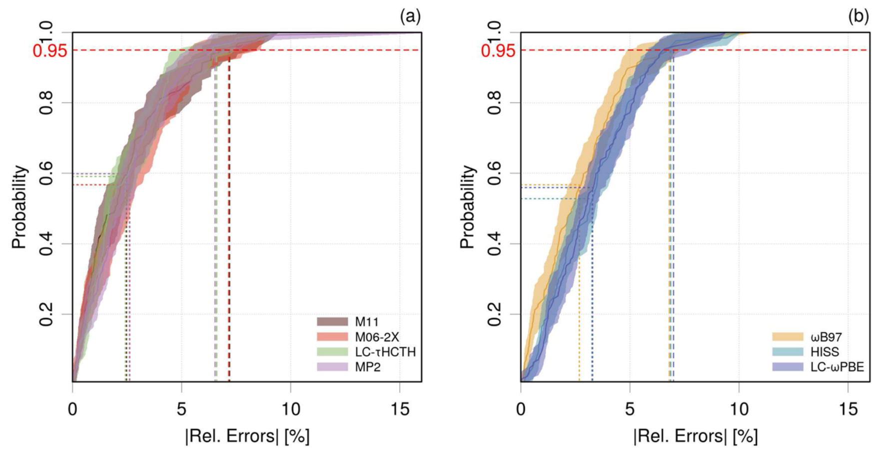

FIG. 22. Case THA2015—ECDFs of absolute relative errors: (a) methods with smallest, indiscernible, MUE values, and (b) other methods.

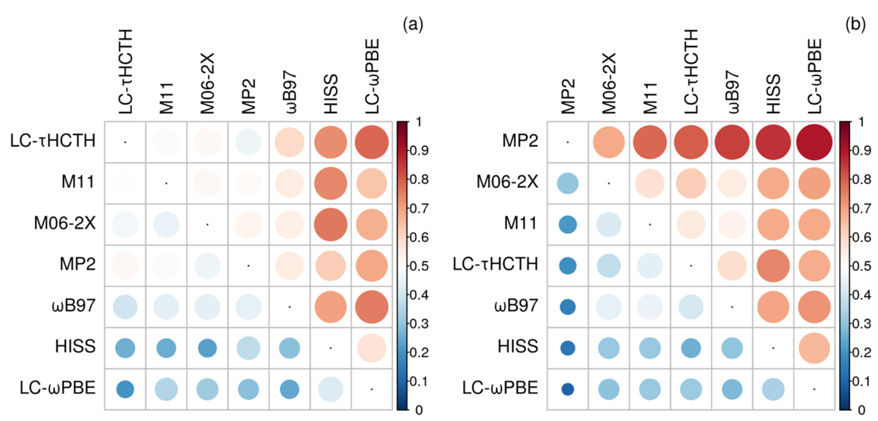

FIG. 23. SIP matrix: (a) case THA2015 (N = 127); (b) case WU2015. The methods are sorted by decreasing MSIP value.

performances, which is confirmed by the comparison of their empirical cumulated distribution functions[4] (Fig. 22). These ECDFs overlap over the whole error range. Besides, these methods cannot be discriminated on the basis of their Q_{95} values, as it appears that all values are indiscernible. These conclusions are unchanged when one removes the eight outliers identified by Thakkar *et al.* (not shown).

3. SIP analysis

The SIP matrix [Fig. 23(a)] for the THA2015 dataset reveals a leading group of four methods identical to those identified above. When passing to WU2015 [Fig. 23(b)], there is a better discrimination between methods, and MP2 presents SIP values over all the other methods.

4. Ranking

The ranking matrices are plotted in Fig. 24. The top row concerns dataset THA2015. The ranking probability matrices for the MUE confirm the problem seen above for the four best methods. It shows also that the rank of MP2 is quite ill-defined. For Q_{95}, as expected, any ranking seems illusory. The same matrices have been estimated after the removal of eight outliers defined above (Fig. 24, middle row). This has a negligible impact on the MUE ranking,

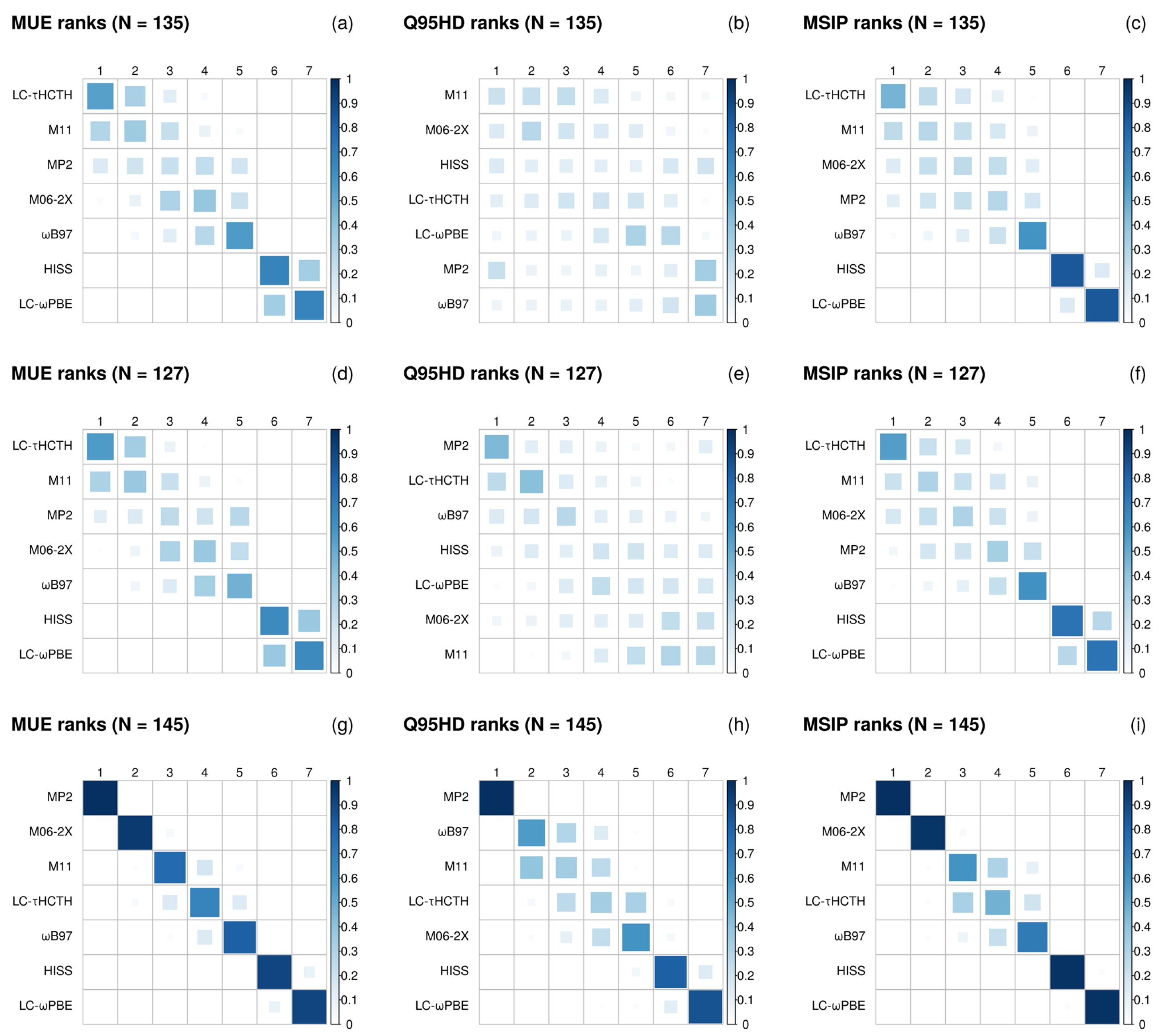

FIG. 24. Ranking probability matrices: [(a)–(c)] case THA2015 full dataset (N = 135); [(d)–(f)] case THA2015 dataset pruned from eight outliers (N = 127); and [(g)–(i)] case WU2015 (N = 145).

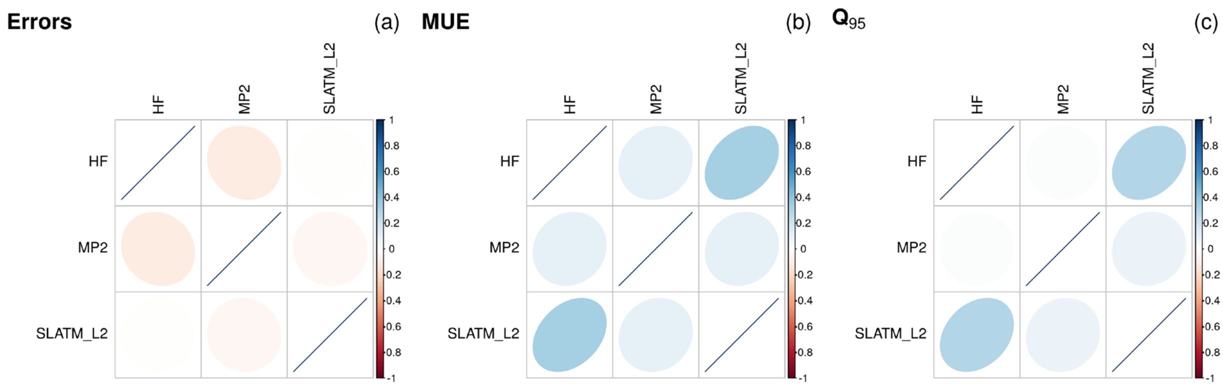

FIG. 25. Case ZAS2019—rank correlation matrices: (a) errors; (b) MUE; and (c) Q_{95}.

but fully scrambles the Q_{95} one, M11 passing from the first to the last place, MP2 from the 8th to the first, and so on. In fact, ill-defined ranking matrices can be expected to be very sensitive to any alteration of the dataset.

When considering the WU2015 dataset, the ranking matrices (Fig. 24, bottom row) show much less dispersion, underlining the deleterious role of experimental errors on ranking. Note that there remains a notable uncertainty to rank ωB97, M11, M06–2X, and LC-τHCTH using Q_{95}.

Depending on the reference dataset [experimental or CCSD(T)], one obtains different rankings: LC-τHCTH seems a better option to predict experimental values (possibly an artifact due to some large experimental reference data errors), whereas MP2 is a better proxy for CCSD(T) calculations.

H. ZAS2019

The effective atomization energies (E^*) for the QM7b dataset,[24] for 7211 molecules up to seven heavy atoms (C, N, O, S, or Cl), are available for several basis sets (STO-3g, 6-31g, and cc-pvdz), three quantum chemistry methods [HF, MP2, and CCSD(T)], and four machine learning algorithms (CM-L1, CM-L2, SLATM-L1, and SLATM-L2). The data have been provided on request by the authors of Zaspel *et al.*[12] The machine learning methods have been trained over a random sample of 1000 CCSD(T) energies (learning set), and the test set contains the prediction errors for the 6211 remaining systems.[12] We retain here only HF, MP2, and SLATM-L2 and compare their ability to predict CCSD(T) values.

1. Correlations

The error sets are essentially uncorrelated (Fig. 25), whereas small positive correlations can be noted for the MUE and Q_{95}.

2. Statistics

The values are reported in Table IX. There is a contrast between the MUE and Q_{95}. SLATM-L2 and MP2 have close MUE values, with an above-threshold p-value ($p_g \simeq 2P_{inv} = 0.06$) and a slight advantage for SLATM-L2. However, MP2 has a significantly smaller Q_{95}. As seen on the absolute errors ECDFs [Fig. 26(a)], SLATM-L2 has indeed a pronounced tail of large errors

This case emphasizes the fact that similar values of the MUE can result by chance from very distinct error distributions and that no conclusion should be taken on the basis of MUE alone.

3. SIP analysis

The SIP matrix [Fig. 26(b)] shows that SLATM-L2 presents a notable improvement probability (~0.75) over HF and a moderate one over MP2 (~0.61). Even if SLATM-L2 has significantly better statistics than HF [Fig. 26(c)], there remains a 25% chance that the latter provides smaller absolute errors. In most case studies presented above, the mean gain was larger in absolute value than the mean loss. In the comparison between SLATM-L2 and MP2, one observes the opposite by choosing SLATM-L2 over MP2 [Fig. 26(d)]

TABLE IX. Case ZAS2019—absolute error statistics: inversion probabilities and SIP statistics for comparison with the DFA of smallest MUE (SLATM-L2), except for Q_{95} inversion probability, where the reference is the DFA with smallest Q_{95} (MP2). The best scores and the values for which ($p_g = 2P_{inv}$) > 0.05 are in boldface.

Methods	MUE (kcal/mol)	P_{inv}	Q_{95} (kcal/mol)	P_{inv}	MSIP	SIP	MG (kcal/mol)	ML (kcal/mol)
HF	2.38(3)	0.00	6.1(1)	0.00	0.283(5)	0.743(6)	−2.03(2)	1.50(5)
MP2	**1.31(1)**	0.03	**3.35(5)**	…	0.538(5)	0.613(6)	−1.08(2)	1.58(5)
SLATM-L2	**1.26(3)**	…	4.7(1)	0.00	**0.678(5)**	…	…	…

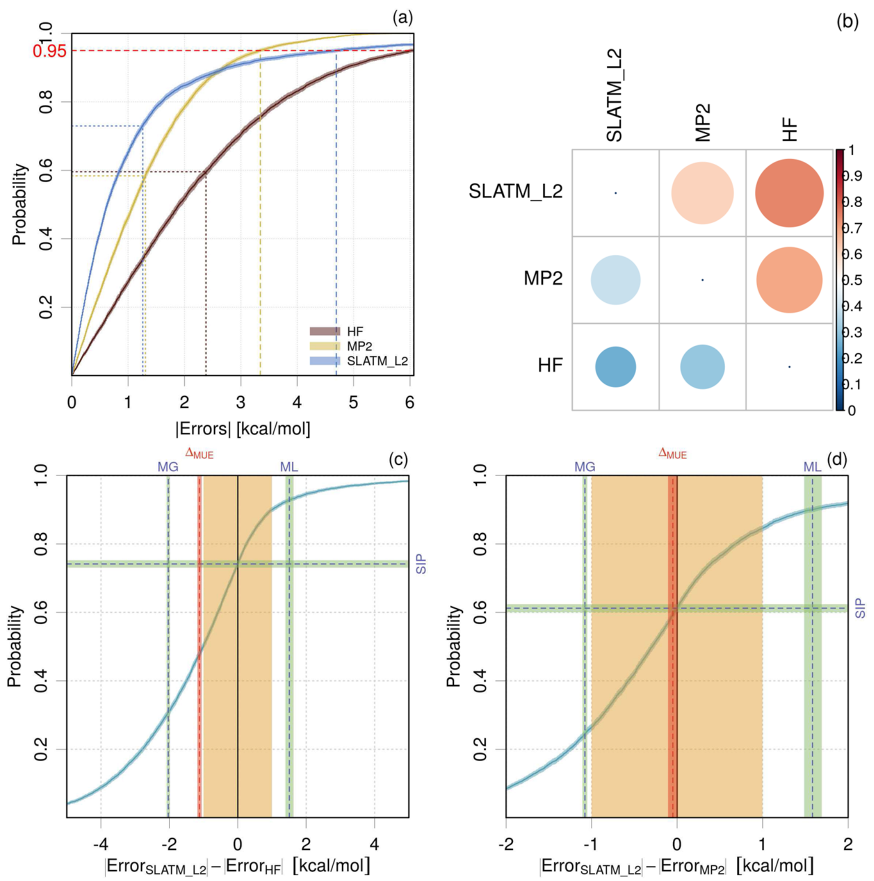

FIG. 26. Case ZAS2019: (a) ECDF of the absolute errors; (b) SIP matrix; and [(c) and (d)] ECDF of the difference of absolute errors of HF (c) and MP2 (d) with respect to SLATM-L2 (see Fig. 3 for details). The orange band represents the chemical accuracy (1 kcal/mol).

and one has 61% chance to get better results, with a mean gain MG $\simeq -1.1$ kcal/mol, and 39% chance to deteriorate the MP2 values with a mean loss ML $\simeq 1.6$ kcal/mol. In agreement with the Q_{95} analysis, this is due to the notable tail of large errors of SLATM-L2.

III. DISCUSSION

A. Extracting data from articles and supplementary material

The raw data of benchmark studies are important assets for the community, and their accessibility and reusability are essential for intercomparison studies or the development of alternative statistical analyses, as performed in this study. When gathering the data, we found that many benchmarking studies have practically inaccessible data, failing the FAIR principle of Open Data.[25] Besides the trivial case of non-available data, we have stumbled on data stored in complex databases and requiring non-trivial coding for their extraction or data stored in inappropriate formats, such as PDF (a Page Description Format), instead of recognized machine-readable data storage formats, such as CSV tables.

Note that for some of the cases we gathered here, we were able to extract data from PDF articles or supplementary material files, but not without some difficulty, involving several steps of manual operations. Typical problems for the data extraction from tables in PDF documents are excessive numerical truncation, empty cells or complex table mapping, typographical (−) instead of numerical (-) minus sign, rotated tables, compact notations for uncertainty [either 123(4) or 123 ± 4], and bibliographical references attached to the data (generally processed by extraction tools as spurious decimals)... Most of these features preclude fully automated data extraction and require error-prone human processing.

So, unless the structure of the data is complex, and this should not be the case for most benchmark studies, it is warmly recommended to use "flat" numerical tables stored in an open format, such as CSV, and to avoid putting more than one information per table cell. "Think Open, think FAIR !"

B. Impact of dataset size

The examples mentioned above have shown that dataset size impacts considerably the ability to rank methods or to assert the impact of an improved method. Size effect on the uncertainty of statistics is well known for the mean value, and similar formulae can be derived for other statistics under normality hypotheses. However, the non-normality of error sets requires the use of numerical methods, typically bootstrap sampling. This enables to show how the usual benchmark statistics are affected by sample size. We have seen, for instance, that there is a notable probability to conclude erroneously that two Q_{95} values are different when they are not (type I errors or false positive) if $N < 60$ (Paper I[1]-Appendix C). For the MUE, this limit is smaller ($N = 30$). Moreover, for small datasets (a few tens of points), even the first digit of the statistics is often affected by the uncertainty.

It is practically impossible to predict the dataset size required for a stable and robust ranking. Many factors other than set size are involved, notably the number and nature of methods to be ranked. When a lot of DFAs are compared, a hierarchical ranking is often performed, for instance, by first choosing the best method at each rung of the Jacob's ladder and then comparing these methods together.[11] This is one way to reduce the ranking uncertainty that is likely to result from the direct comparison of a large number of methods, as illustrated, for instance, in case CAL2019 (Sec. II D, Fig. 12).

C. The correlation matrix as a sanity check

When we started this study, the correlation matrices were mainly intended to illustrate the importance to consider correlation when comparing statistics. When cumulating the case studies, we realized that errors correlation matrices may contain pertinent information on the quality of the benchmark dataset. Considering that model errors in computational chemistry are mostly systematic, one expects that error patterns over a dataset are characteristic of each method or a family of methods. This seems to be a basic requirement for sound benchmarking studies. One should, thus, expect that closely related methods produce similar error patterns and have strongly correlated error sets, the correlation level decreasing with a "distance" between methods. This is clearly illustrated in case BOR2019 (Fig. 5), where the correlation matrix clusters nicely into relevant DFA groups. There seems also to be a genuine decorrelation between MP2 or MP2-based methods and DFAs (NAR2019, Fig. 9; WU2015, Fig. 21). Similarly, one observes no correlation between HF, MP2, and a machine-learning method calibrated on CCSD(T) in case ZAS2019, Fig. 25.

As a consequence, when the method set contains unrelated methods, a uniform strongly positive correlation matrix should raise an alert. We have seen in cases DAS2019 and THA2015/WU2015 that outliers and/or large reference data errors could dominate the correlation matrix and influence the benchmark statistics. Outliers common to all error sets (global outliers) can be efficiently identified on a parallel plot, as shown in case DAS2019 (Fig. 17). If the ranking study is to reflect the methods performances, the curation and possible pruning of the dataset from such global outliers is a necessary preliminary step. Otherwise, more complex statistical models have to be used to alleviate the impact of those points (see Paper I[1]-Appendix A and Refs. 26–28).

Note that strongly correlated error sets do not imply similar performances. For instance, a set of linearly scaled harmonic vibrational frequencies typically has better statistics than the unscaled set,[29] whereas their correlation coefficient is one because of the linear transformation between both error sets. One should also remember that the correlation coefficient between calculated and reference values that is still presented in some benchmarks is not a reliable performance statistic.[30] At most, it reveals a linear (Pearson) or monotonic (Spearman and Kendall) association between datasets, but their proximity to the identity line.

D. Impact of error sets correlation on ranking

The correlation between error sets is partially or totally transferred to benchmark statistics. Except for linear transformations of the errors, where the transfer is trivial, one has to use Monte Carlo methods to estimate it. In many cases, such as for normal, Student's-*t*, or g-and-h error distributions,[31] one observes that the correlation intensity mainly decreases when passing from errors to MUE to Q_{95}. The case studies mentioned above show, however, that there are exceptions to this ideal trend. We cannot presently rationalize the observed exceptions. In a vast majority of the cases studied above, the correlation matrices for MUE and Q_{95} have positive coefficients. These contribute to a reduction of the uncertainty on statistics differences, with better discernibility between uncertain statistics. Globally, positive correlations increase the robustness of rankings.

However, unlike for the error correlations, the visualization and analysis of correlations between statistics might be of secondary interest for benchmarks. In fact, the paired samples bootstrap algorithms used in this study enable to account directly for these correlations, without having to estimate intermediate correlation matrices.

E. Systematic improvement analysis

We introduced a new criterion, the systematic improvement probability (SIP), which has the major advantage to be independent of the usual descriptive statistics. It is based on a sign statistic of the differences of absolute error pairs. It is a useful complement to the MUE, as it enables to analyze MUE differences. All the cases studied above show that a decrease of MUE results from a balance between gains and losses. Only two methods pairs were found, in cases PER2018 (Sec. II A) and CAL2019 (Sec. II D), with SIP values reaching 0.95, close to the full systematic improvement. We did not find a "best method" that fully improves the results of all lower rank methods. Because of the well known error compensations in computational chemistry methods,[32] even physics-based improvements in DFAs do not lead to systematic improvements for all systems. Of course, this balance is not a discovery, but the SIP enables to quantify it and provides a basis for the user to estimate the risk taken when switching from an old, faithful method to a new one. We have seen for, instance, that for bandgaps, mBJ degrades LDA predictions for 16% of the systems (BOR2019, Table III). In fact, there is often a non-negligible percentage of systems for which a "bad" method is better than a "good" one, all across Jacob's ladder.

We have also introduced the mean SIP as a possible ranking statistic. The main advantage of the MSIP is its independence from the usual summary statistics; its main drawback is that it depends on the set of methods being compared, and it is not transferable to comparisons out of its definition set. Conflicts of the MSIP with the MUE reveal disparities in the errors distribution.

F. Ranking probability matrix

The ranking probability matrix $\mathbf{P}_r$ provides a diagnostic on the robustness of the ranking by any statistic. Our tests of MUE, Q_{95}, and

MSIP rankings show that the dataset size and the number of methods influence notably the ranking uncertainty. Without any surprise, the closer the performances of a group of methods, the more uncertain their ranking. Depending on the datasets, the MUE and Q_{95} rankings might conflict and present different levels of robustness (cf. case THA2015, Fig. 24). We would advise to publish systematically both of them, as they provide complementary information.

In the various cases treated above, the rankings provided by the MSIP most often conform to the MUE rankings and are as sensitive as the other rankings to sampling uncertainty. When ranking conflicts for the first places occur with the MUE, as was observed in case PER2018 (Fig. 4), one gets alerted that the method with the lowest MUE is not the one providing the largest proportion of small absolute errors. Due to the non-normality of error distributions, such scenarios are to be expected, as for inversions in MUE and Q_{95} rankings.

G. Extension to composite datasets

We considered here only datasets based on a single property. Many modern benchmarks are based on composite datasets, involving weighting schemes to incorporate data with different units.[33] The applicability of the SIP to such datasets is straightforward, but the mean gain and mean loss statistics, having dimensions, should become multivariate.

The estimation of P_{inv} and ranking probability matrices for composite statistics (e.g., WTMAD[33]) can use directly the pair-based bootstrap sampling algorithms described in the present article, although care should be taken to avoid imbalance between the various components of a dataset by using the so-called *stratified* bootstrap,[34] preserving the cardinal number of each component in the generated sample.

IV. CONCLUSION

In Paper I,[1] we proposed several tools to test the robustness of rankings or comparisons of methods based on error statistics for non-exhaustive, limited size datasets. In order to avoid hypotheses on the errors distributions, bootstrap-based methods were used for the estimation of statistics uncertainty, p-values, and ranking uncertainty. In this paper, we illustrated and validated these methods on nine datasets covering a representative panel of properties and sizes.

Most of these tools take into account the correlation between error sets or their statistics, and we illustrated repeatedly that large correlations occur that cannot be neglected. Moreover, we have seen that the error sets correlation matrix can be useful to appreciate the quality of a benchmark dataset, notably when experimental reference data are used. To our knowledge, this topic has not previously been discussed, and benchmarking studies do not presently make use nor report such correlation matrices.

The systematic improvement probability (SIP) is based on the system-wise difference of absolute errors between two methods, and in conjunction with the mean gain (MG) and mean loss (ML) statistics, it quantifies the risk taken by a user when passing from a method to another. We have seen in the applications that choosing a method with a lower MUE might imply a non-negligible risk to produce large errors. Moreover, only two of the showcased examples revealed a method that provides a (nearly) full systematic improvement over one of its concurrents. Even when comparing an elaborate composite method such as G4MP2 to DFAs one observes partial SIP values (case NAR2019, Table IV). A pedagogical virtue of the SIP is to clearly show that computational chemistry is a science of compromises.

We based the comparison between values of a statistic for two methods on the inversion probability P_{inv}, which is simply linked to the p-value for the test of the equality of those statistics ($p_g \simeq 2P_{inv}$). It is, thus, an important tool to assess if a difference between two values is a real effect or if it might be due to the choice of dataset. For ranking statistics, we suggest to report P_{inv} with respect to the method with the smallest value in results table.

The ranking probability matrix $\mathbf{P}_r$ for a chosen statistic provides a clear diagnostic on the robustness of the corresponding ranking. The impact of dataset size and number of compared methods can be thoroughly tested, as shown in the examples above. It appeared in these examples that the intermediate ranks are often weakly defined. The robustness of the ranking might also depend on the ranking statistic, and the statistic providing the most robust ranking depends on the dataset. As we suggested earlier, one should, therefore, not rely on the MUE alone to rank methods. We encourage benchmark authors to provide ranking probability matrices for several statistics (at least the MUE and Q_{95}), which can be obtained with a negligible overcharge in computer time.

We considered here for simplicity raw error sets, from which no care has been taken to remove systematic trends. When this is possible, such trend corrections, often simply linear, will provide much better generalizability of the summary statistics derived from these error sets. Besides, this is a necessary step if one wishes to estimate the prediction uncertainty of any method,[26–28] notably when dealing with non-uniform reference data uncertainties.

ACKNOWLEDGMENTS

The authors are grateful to Professor O. A. von Lilienfeld for providing the datasets of case ZAS2019 and to Professor S. Grimme for providing a corrected copy of the supplementary material for case CAL2019.

DATA AVAILABILITY

Data that support the findings of this study that are openly available in Zenodo at http://doi.org/10.5281/zenodo.3678481.[35] Application `ErrView` implementing the methods described in this article is archived in Zenodo at http://doi.org/10.5281/zenodo.3628489); a test web interface is freely accessible at http://upsa.shinyapps.io/ErrView.

REFERENCES

[1] P. Pernot and A. Savin, "Probabilistic performance estimators for computational chemistry methods: Systematic improvement probability and ranking probability matrix. I. Theory," J. Chem. Phys. **152**, 164108 (2020); arXiv:2003.00987.

[2] R. R. Wilcox and D. M. Erceg-Hurn, "Comparing two dependent groups via quantiles," J. Appl. Stat. **39**, 2655–2664 (2012).

[3] F. E. Harrell and C. Davis, "A new distribution-free quantile estimator," Biometrika **69**, 635–640 (1982).

[4]P. Pernot and A. Savin, "Probabilistic performance estimators for computational chemistry methods: The empirical cumulative distribution function of absolute errors," J. Chem. Phys. **148**, 241707 (2018).

[5]P. Borlido, T. Aull, A. W. Huran, F. Tran, M. A. Marques, and S. Botti, "Large-scale benchmark of exchange–correlation functionals for the determination of electronic band gaps of solids," J. Chem. Theory Comput. **15**, 5069–5079 (2019).

[6]B. Narayanan, P. C. Redfern, R. S. Assary, and L. A. Curtiss, "Accurate quantum chemical energies for 133000 organic molecules," Chem. Sci. **10**, 7449–7455 (2019).

[7]E. Caldeweyher, S. Ehlert, A. Hansen, H. Neugebauer, S. Spicher, C. Bannwarth, and S. Grimme, "A generally applicable atomic-charge dependent London dispersion correction," J. Chem. Phys. **150**, 154122 (2019).

[8]F. Jensen, "Method calibration or data fitting?," J. Chem. Theory Comput. **14**, 4651–4661 (2018).

[9]T. Das, G. Di Liberto, S. Tosoni, and G. Pacchioni, "Band gap of 3D metal oxides and quasi-2D materials from hybrid density functional theory: Are dielectric-dependent functionals superior?," J. Chem. Theory Comput. **15**, 6294–6312 (2019).

[10]A. J. Thakkar and T. Wu, "How well do static electronic dipole polarizabilities from gas-phase experiments compare with density functional and MP2 computations?," J. Chem. Phys. **143**, 144302 (2015).

[11]T. Wu, Y. N. Kalugina, and A. J. Thakkar, "Choosing a density functional for static molecular polarizabilities," Chem. Phys. Lett. **635**, 257–261 (2015).

[12]P. Zaspel, B. Huang, H. Harbrecht, and O. A. von Lilienfeld, "Boosting quantum machine learning models with a multilevel combination technique: Pople diagrams revisited," J. Chem. Theory Comput. **15**, 1546–1559 (2019).

[13]J. P. Perdew, J. Sun, A. J. Garza, and G. E. Scuseria, "Intensive atomization energy: Re-thinking a metric for electronic structure theory methods," Z. Phys. Chem. **230**, 737–742 (2016).

[14]L. A. Curtiss, K. Raghavachari, P. C. Redfern, and J. A. Pople, "Assessment of Gaussian-3 and density functional theories for a larger experimental test set," J. Chem. Phys. **112**, 7374–7383 (2000).

[15]P. Pernot and A. Savin, "Erratum: "Probabilistic performance estimators for computational chemistry methods: The empirical cumulative distribution function of absolute errors" [J. Chem. Phys. 148, 241707 (2018)]," J. Chem. Phys. **150**, 219906 (2019).

[16]D. Defays, "An efficient algorithm for a complete link method," Comput. J. **20**, 364–366 (1977).

[17]R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2019, Version 3.6.1, URL: https://www.R-project.org/.

[18]The original dataset contains 472 systems, but several values are missing for $NaYbP_2S_6$, which was excluded.

[19]The MUE is sometimes abusively used to characterize the accuracy of a method, which cannot be the case when error distributions are not zero-centered normal.[4,27]

[20]S. Dohm, A. Hansen, M. Steinmetz, S. Grimme, and M. P. Checinski, "Comprehensive thermochemical benchmark set of realistic closed-shell metal organic reactions," J. Chem. Theory Comput. **14**, 2596–2608 (2018).

[21]Reproducibility note: These data are inconsistent with the results reported in Fig. 9 of the reference article and the subsequent discussion. We contacted the corresponding author (S. Grimme) who kindly sent us a corrected version of the supplementary material.

[22]J. Rezác, K. E. Riley, and P. Hobza, "S66: A well-balanced database of benchmark interaction energies relevant to biomolecular structures," J. Chem. Theory Comput. **7**, 2427–2438 (2011).

[23]J. Rezác, K. E. Riley, and P. Hobza, "Erratum to "S66: A well-balanced database of benchmark interaction energies relevant to biomolecular structures"," J. Chem. Theory Comput. **10**, 1359–1360 (2014).

[24]G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller, and O. Anatole von Lilienfeld, "Machine learning of molecular electronic properties in chemical compound space," New J. Phys. **15**, 095003 (2013).

[25]M. D. Wilkinson *et al.*, "The FAIR guiding principles for scientific data management and stewardship," Sci. Data **3**, 160018 (2016).

[26]K. Lejaeghere, J. Jaeken, V. V. Speybroeck, and S. Cottenier, "*Ab initio* based thermal property predictions at a low cost: An error analysis," Phys. Rev. B **89**, 014304 (2014).

[27]P. Pernot, B. Civalleri, D. Presti, and A. Savin, "Prediction uncertainty of density functional approximations for properties of crystals with cubic symmetry," J. Phys. Chem. A **119**, 5288–5304 (2015).

[28]J. Proppe and M. Reiher, "Reliable estimation of prediction uncertainty for physicochemical property models," J. Chem. Theory Comput. **13**, 3297–3317 (2017).

[29]A. P. Scott and L. Radom, "Harmonic vibrational frequencies: An evaluation of Hartree-Fock, Möller-Plesset, quadratic configuration interaction, density functional theory, and semiempirical scale factors," J. Phys. Chem. **100**, 16502–16513 (1996).

[30]J. Bland and D. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," Lancet **327**, 307–310 (1986).

[31]D. C. Hoaglin, "Summarizing shape numerically: The G-and-H distributions," in *Exploring Data Tables, Trends, and Shapes* (Wiley, New York, 1985), pp. 461–513.

[32]T. H. Dunning, "A road map for the calculation of molecular binding energies," J. Phys. Chem. A **104**, 9062–9080 (2000).

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

[34]T. C. Hesterberg, "What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum," Am. Stat. **69**, 371–386 (2015).

[35]P. Pernot and A. Savin (2020). "Codes and data that support the findings of this study," Zenodo. http://doi.org/10.5281/zenodo.3678481