On the comparison of error sets and their statistics in benchmarking studies

Pascal PERNOT (Pascal.Pernot@u-psud.fr)
Institut de Chimie Physique, UMR8000, CNRS,
Université Paris-Saclay, 91405, Orsay, France
email: pascal.pernot@universite-paris-saclay.fr

Andreas SAVIN (Andreas.Savin@lct.jussieu.fr)
Laboratoire de Chimie Théorique, CNRS and UPMC Université Paris 06,
Sorbonne Universités, 75252 Paris, France

Tuesday 14th January, 2020

Abstract

The comparison of benchmark error sets is an essential tool for the evaluation of theories in computational chemistry. The standard ranking of methods by their Mean Absolute Error is unsatisfactory for several reasons linked to the non-normality of the error distributions and the underlying trends. Complementary statistics have recently been proposed to palliate such deficiencies, such as quantiles of the absolute errors distribution or the mean prediction uncertainty. We introduce here a new score, the systematic improvement probability (SIP), based on the direct pair-wise comparison of absolute errors, bypassing the need of other descriptive statistics. Independently of the chosen scoring rule, the uncertainty of the statistics due to the incompleteness of the benchmark data sets is also generally overlooked. However, this uncertainty is essential to appreciate the robustness of score-based rankings. In the present article, we develop two methods based on robust statistics to address this problem: P_{inv}, the inversion probability between two statistics, and $\mathbf{P}_r$, the ranking uncertainty matrix. We demonstrate also the essential contribution of the correlations between error sets in these scores comparisons. These methods are validated and compared on a set of diverse benchmark data extracted from the recent literature.

Contents

1 Introduction

Benchmarks are a central tool for the evaluation of new theories/methods in quantum chemistry [1]. Amongst many possible metrics [2], the most common benchmarking statistics are the mean unsigned error (MUE/MAD/MAE...), mean signed error (MSE), root mean squared error (RMSE) and root mean squared deviation (RMSD). The explicit definition of these scores is given in Ref. [3]. In a vast majority of benchmark studies, the MUE, or some variant of it, is used to compare methods performance. If the ranking of methods is precious for developers who want to assess the impact of their latest methods, it might be of less interest for final users. In particular, it does not generally offer the choice and criteria for picking another method than the 'best one'. Recently, we proposed a more informative probabilistic score, the 95th percentile of the absolute errors distribution (Q_{95}) [3].[1]

Whichever the statistic used, the question remains of the robustness of such scores and rankings with respect to the choice of the reference dataset. One easily conceives that the values of these statistics change unpredictably when one adds or removes a few points in the dataset. Benchmarks implicitly assume that the error sets are representative samples of unknown distributions characterizing model errors for each method – the more systems in the dataset, the best the approximation of the underlying distributions. The quest for large datasets incurs heavy computer charges to perform benchmarks, and there is a trend to reduce this burden by looking for small, optimally representative, datasets [7,8]. Besides, there are several properties for which the reference data are rather sparse, leading to rather small datasets. Another trend, fueled by the development of machine learning if to replace experimental values values by gold standard calculations, with limitations on the size of accessible systems. As the estimated values of the statistics and their uncertainties depend on the size of the dataset, it is important to assess this size effect and its impact on statistics comparison and ranking.

This question has been considered recently by Proppe and Reiher [9], who used bootstrapping to assess the impact of dataset size and reference data uncertainty on the first rank in an intercomparison of Mössbauer isomer shifts estimated by a dozen of DFAs. They concluded that for their dataset of $N = 39$ values, at least three methods were competing for the first rank, with a slight probabilistic advantage for PBE0. This is a very interesting contribution to the quality assessment of benchmarking tools. We considered another approach to this problem by defining an inversion probability P_{inv} for the ranking of two methods [3]. Our definition was based on the assumption of a normal distribution of statistics differences and neglected error sets correlations, and merits a

[1] We argued that Q_{95} is more informative than the MUE, because the latter does not provide probabilistic information if the errors distribution is not zero-centered normal, a rather unlikely occurrence. In contrast, Q_{95} gives us the error level that one has only 5 % chance to exceed in a new calculation (provided that the reference dataset is representative of the systems for which predictions are sought). The end-users can easily check if this threshold meets their expectations. We recently realized that the 90th percentile (noted P_{90}) has been used by Thakkar and colleagues [4,5]. We think Q_{95} is more appropriate because of its natural link (for symmetric zero-centered error distributions) to the enlarged uncertainty u_{95} recommended in the thermochemistry literature [6,3].

more general setup.

In this study, we revisit the ranking uncertainty problem along several lines:

1. we consider the statistical significance difference between two statistics: it depends both on the uncertainty on the difference, which is influenced by the dataset size, and on the correlation between the estimated values, which is due in a large part to their use of a common reference dataset [10]. There are a few specific points to be considered: the non-normality of the error sets distributions, the small size of some datasets, and some properties of quantiles estimators. The impact of reference data uncertainty has also to be considered.

2. we define a ranking probability matrix P_r, generalizing the proposition of Proppe and Reiher [9], which enables us to propose efficient visual assessments of the robustness of rankings.

3. we introduce a new statistics (the systematic improvement probability, SIP) that convey the proportion of systems in the benchmark data set for which one method is better than the other, and the expected gain or loss when switching between methods.

In the next section, we consider the uncertainty sources impacting the values of benchmarking statistics (scores) and we present the tools best adapted to estimate the uncertainty on statistics and to compare them. These methods are then validated on several datasets taken from the recent benchmarking literature and covering a wide range of dataset sizes and of reference data uncertainty. The discussion considers the impact of these observations on the benchmarking practice and proposes several suggestions on their reporting, as well as on the best practice to share benchmark data.

2 Statistical methods

2.1 Error sets, their uncertainty and correlation

Benchmarking is based on the statistical analysis of errors sets for a method M ($E_M = \{e_i(M)\}_{i=1}^N$) for a set of N calculated ($C_M = \{c_i(M)\}_{i=1}^N$) and reference data ($R = \{r_i\}_{i=1}^N$), where

$$e_i(M) = r_i - c_i(M) \tag{1}$$

Uncertainty. As the reference data or even the calculated values can be uncertain, one should consider that the error sets E_M contain uncertain values when estimating and comparing statistics. Experimental or computational uncertainties being typically estimated by standard deviations, one can use the method of combination of variances to get the uncertainty on the errors [11],

$$u(e_i) = \sqrt{u(r_i)^2 + u(c_i)^2} \tag{2}$$

where $u(x)$ is the uncertainty on x. This formula assumes safely that the individual errors on the reference data and calculated values are uncorrelated. For a reference datum r_i, $u(r_i)$ would typically be a measurement uncertainty for experimental data. For a computed reference datum r_i and for a calculated datum c_i, uncertainty might come from numerical uncertainty due to the use of finite precision arithmetics and discretization errors [12, 13], statistical uncertainty (*e.g.*, for Monte Carlo methods [14]), parametric uncertainty (*e.g.*, for calibrated parametric methods [15, 14, 16]), or careful calibration of computational protocols [17, 18].

We consider here deterministic computational chemistry methods with assumed low and controlled arithmetic uncertainty. The uncertainty on errors is then equal to the reference data uncertainty $u(e_i) \equiv u(r_i)$. For the sake of generality, the $u(e_i)$ notation is preserved in the following.

Correlation. Let us consider a set of K methods $\{M_i\}_{i=1}^{K}$. The covariance of the error sets for two method can be decomposed as

$$\begin{aligned} \mathrm{cov}(E_i, E_j) &= \mathrm{cov}(R - C_i, R - C_j) && (3) \\ &= \mathrm{var}(R) + \mathrm{cov}(C_i, C_j) - \mathrm{cov}(R, C_i) - \mathrm{cov}(R, C_j) && (4) \end{aligned}$$

where, for brevity, we use shortened notations such as $E_i \equiv E_{M_i}$. It is not possible to predict the values of the individual terms and of their sum, but a few considerations might be helpful:

- even if the predictions of two methods are statistically independent ($\mathrm{cov}(C_1, C_2) = 0$), an unlikely occurrence), their errors are not;
- if reference data uncertainties are larger than prediction errors, the covariance will be dominated by $\mathrm{var}(R)$, and all error sets will be strongly correlated.

In the following case studies (Section 3), we report and analyze the correlation coefficients between error sets (normalized covariances)

$$\mathrm{corr}(E_i, E_j) = \frac{\mathrm{cov}(E_i, E_j)}{\sigma_{E_i} \sigma_{E_j}} \quad (5)$$

where $\sigma_{E_i}$is the standard deviation of the error set E_i, assumed finite. We will show through case studies that the correlation matrix contains relevant information on the quality of datasets and the proximity of methods.

2.2 Statistics, their uncertainty and correlation

Uncertainty. The value s of a statistic S (MSE, MUE, Q_{95}...) estimated on an error set is generally uncertain, with uncertainty estimated by its standard error $u(s)$. Two main uncertainty sources should be considered: (1) the limited size N of the reference data sample, and (2) the error

set uncertainties, $u(e_i)$. Unless the dataset is exhaustive (*e.g.*, a dataset containing a property for a complete class of systems), the first source is always present. For experimental reference data, the second source is also always present, but experimental uncertainties are rarely available for large datasets, and a common practice seems to be to ignore them in the statistical analysis (although they are often discussed to assess the quality of the dataset). Some studies considered the effect of representative uncertainty levels on benchmarking conclusions [19, 20, 9].

In Appendix A, the impact of both uncertainty sources is illustrated on the mean value (MSE), for which analytical formulae are available. The strategy to handle reference data uncertainty depends on their distribution. If the reference data uncertainties are uniform over the dataset, the hypothesis of *i.i.d.* errors holds, and standard statistical procedures can be applied (unless one is interested in quantifying specifically model errors [19, 9]). Otherwise, weighted statistics have to be used [19, 9], which will no be considered here. We will consider here that datasets should not include data with extreme uncertainty values.

Simple formulae for standard errors, such as those for the mean, are not available for non-linear statistics such as the MUE or Q_{95}. Moreover, in order to avoid some of the limitations implied by such formulae (*e.g.*, normality hypothesis), one can use a general method to estimate the standard error of any statistic: the bootstrap [21–23]. The bootstrap is a Monte Carlo sampling method which consists in random draws with replacement of M values from a dataset of size N. In the standard bootstrap, one uses $M = N$, *i.e.*, the generated samples have the same size than the original set. The bootstrap has been shown to provide reliable estimations of uncertainties, but the mean values unavoidably reflect the bias due to the reference data set [23]. In consequence, we estimate in the following the mean values from the original sample and the uncertainties from the bootstrap samples. The main limitation of the bootstrap is its hypothesis of *i.i.d.* data, but it is consistent with our choice to avoid weighted statistics and to avoid reference datasets with a large uncertainty range.

Correlation. The statistics covariance $\mathrm{cov}(S_1, S_2)$ derives from the mathematical expression of S and from the variances and covariance of the error sets, $\mathrm{cov}(E_1, E_2)$. To estimate $\mathrm{cov}(s_1, s_2)$ in the case of linear statistics, one can directly apply the generalization of the combination of variances to several model outputs [24]. For the MSE, it is easy to demonstrate that the covariance is transferred in totality: $\mathrm{cov}(\overline{e}_1, \overline{e}_2) = \mathrm{cov}(E_1, E_2)$. More generally, for linear statistics, $\mathrm{cov}(E_1, E_2) = 0 \Longrightarrow \mathrm{cov}(s_1, s_2) = 0$. For non-linear statistics, as the MUE or Q_{95}, the combination of covariances is unsuitable, and one has to refer to sampling strategies.

To illustrate the transfer of correlation from error sets to non-linear statistics, we performed a Monte Carlo study, detailed in Appendix B, with scenarii implying diverse distribution shapes. Few global trends can be derived from this study, notably that $\mathrm{cor}(S_1, S_2)$ is a convex, positive function of $\mathrm{cor}(E_1, E_2)$. Moreover, for a given value of $\mathrm{cor}(E_1, E_2)$ one has $\mathrm{cor}(MUE_1, MUE_2) \geq \mathrm{cor}(Q_{95,1}, Q_{95,2})$.

As we explored only a fraction of the possible scenarios for the errors distributions, these trends cannot be considered as robust. The main point is that the correlation of error sets is at least partly transferred to the derived statistics, a fact to be considered when comparing the values of these statistics.

2.3 Pair-wise comparison of errors

The systematic improvement probability (SIP) between two methods M_i and M_j is the proportion of systems in the reference set for which the absolute error decreases when using M_i instead of M_j. It is estimated as

$$SIP_{i,j} = \frac{D_{i,j}}{N} \tag{6}$$

$$D_{i,j} = \sum_{k=1}^{N} \mathbf{1}_{\delta_k(M_i,M_j)<0} \tag{7}$$

where $\mathbf{1}_X$ is the indicator function, taking for value 1 if X is true and 0 otherwise, and

$$\delta_k(M_i, M_j) = |e_k(M_i)| - |e_k(M_j)| \tag{8}$$

Note that, because of the possible presence of ties, one has

$$SIP_{i,j} + SIP_{j,i} \lesssim 1 \tag{9}$$

Interpretation. A line of the SIP matrix, provides the SIP values for the corresponding method over all the other ones. If a new method M_1 provides systematic improvement over M_2, in the sense that it has smaller absolute errors for all systems in the reference set, one should have $SIP_{1,2} = 1$. Values smaller than 0.5 indicate a degradation. Note however that M_1 can achieve small values of the SIP and still have better scores (MUE, Q_{95}), as a few large improvements might overwhelm many small degradations. The interest of the SIP indicator is mainly to alert the user that using a "better method" M_1 can lead to a degradation of results with respect to M_2, with a probability approximately $(1 - SIP_{1,2})$.

Mean SIP. In order to compare and rank a set of K methods, one defines the Mean SIP ($MSIP$) as the mean value of a line of the SIP matrix (excluding the diagonal)

$$MSIP(M_i) = \frac{1}{K} \sum_{j=1}^{K} SIP_{i,j} \, (1 - \delta_{ij}) \tag{10}$$

The largest MSIP value points to a method which in average provides the best level of improvement over the other methods in the set. Note that the MSIP is not transferable for comparisons with

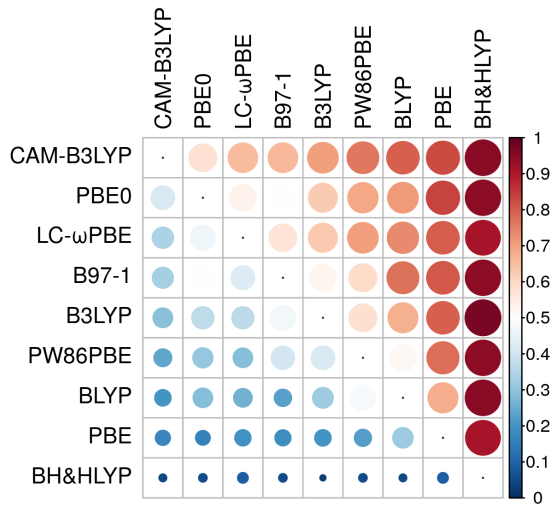

Figure 1: SIP matrix for a set of 9 methods compared with the G99 set of enthalpies (case PER2018, *cf.* Section 3.1). The methods are ordered by decreasing value of MSIP (Eq. 10).

new methods out of its definition set.

Representation. The SIP matrix can be represented by a levels image, where the color scale goes from blue (0.0) to red (1.0) with a white midpoint (0.5). The diagonal is left undefined to alleviate the graph. The methods are ordered by decreasing value of MSIP. Fig. 1 provides an example extracted from Section 3.1. It shows clearly that in this case BH&HLYP is problematic and is systematically and strongly outperformed by all other methods. At the opposite, the line for CAM-B3LYP is the only one to contain exclusively reddish patches (values above 0.5), albeit CAM-B3LYP does not achieve the best MUE or Q_{95} scores within this set of methods (*cf.* Table 2).

Mean gain and loss. In order to appreciate the amplitude of the possible losses or gains, we define the mean gain (MG) as the mean of the positive values of $\delta_k(M_i, M_j)$

$$MG_{i,j} = \frac{1}{D_{i,j}} \sum_{k=1}^{N} \mathbf{1}_{\delta_k(M_i,M_j)<0}\, \delta_k(M_i, M_j) \tag{11}$$

$$ML_{i,j} = -MG_{j,i} \tag{12}$$

where the lean loss (ML) is the opposite of the mean gain for the reciprocal comparison. These statistics are intended to convey an amplitude of the improvement of M_i over M_j: MG is therefore a negative value (corresponding to a decrease of absolute errors), and ML a positive value. Moreover,

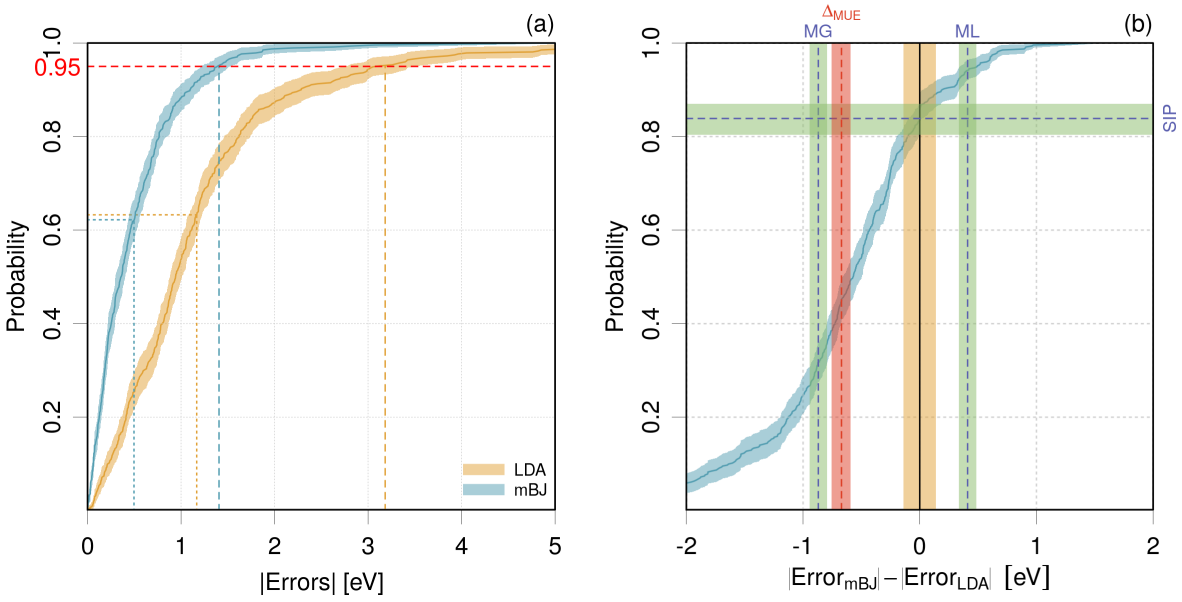

Figure 2: Statistics of absolute errors and their pair-wise differences: (a) ECDF of two error sets to be compared. The MUE values are depicted by vertical dotted lines, and the Q_{95} values by vertical dashed lines; (b) ECDF of the difference of absolute errors on band gaps for methods mBJ and LDA (case BOR2019, *cf.* Section 3.2). The green- and red-shaded bands represent 95 % confidence intervals for the reported statistics (SIP: systematic improvement probability; MG: mean gain; ML: mean loss, Δ_{MUE}: MUE difference). The orange band represents the chemical accuracy (0.14 eV).

the SIP and MG provide a decomposition of the MUE difference between two methods:

$$\Delta_{MUE_{i,j}} = MUE(M_i) - MUE(M_j) \tag{13}$$

$$= SIP_{i,j} * MG_{i,j} - SIP_{j,i} * MG_{j,i} \tag{14}$$

This shows that, except for methods pairs with an extreme SIP value, any MUE difference is the balance between losses and gains distributed over the systems. One should not expect that a method with a smaller MUE will systematically provide better results.

ECDF of $\delta_k(M_i, M_j)$. The scores (SIP, MG and ML) can be visualized on a single graph of the Empirical Cumulated Density Function of the differences of absolute errors between two methods, as shown in Fig. 2(b). This example is extracted from Section 3.2, on the prediction of band gaps. It compares mBJ (MUE = 0.50 eV) and LDA (MUE = 1.17 eV), showing that the large MUE difference (Δ_{MUE}) between these methods is the balance of a mean gain $MG = -0.86$ eV for 85 % of the systems (SIP), and a mean loss $ML = 0.37$ eV for 15 % of the systems. In the hypothesis of a representative dataset, a user switching from LDA to mBJ has to accept a 15 % risk to see his results be degraded in average by 0.37 eV, up to 1 eV.

Note that this information is not available when considering the ECDFs of the absolute errors (Fig. 2(a)). For the chosen example, the comparison of these ECDFs might leave the false impression

that mBJ has consistently smaller absolute errors than LDA, which is an artifact due to the ignorance of data pairing in this representation.

2.4 Pair-wise comparison of statistics

2.4.1 The testing framework

Using the error sets for two methods M_1 and M_2, one calculates the values $s_1 = S(E_1)$ and $s_2 = S(E_2)$ of a statistic S. The usual procedure to compare these two values is to test if their difference is significantly larger than their combined uncertainty, *i.e.*

$$|s_1 - s_2| > \kappa u(s_1 - s_2) \tag{15}$$

where $u(s_1 - s_2)$ is the uncertainty on the difference of scores, and κ is an enlargement factor typically taken as $\kappa = 2$ (or 1.96) in metrology [25]. In the hypothesis of a normal distribution for the statistics difference, $\kappa = 1.96$ corresponds to a confidence level of 95 % for a two-sided test, implied by the absolute value in Eq. 15. If one has evidence that the distribution of differences is not normal, κ has to be chosen accordingly. If the test is positive, there is less than 5 % probability that $s_1 = s_2$.

Assuming that $u(s_1 - s_2)$ cannot be null nor infinite, it is convenient to recast the test by using a discrepancy factor

$$\xi(s_1, s_2) = \frac{|s_1 - s_2|}{u(s_1 - s_2)} \tag{16}$$

to be compared to the threshold κ. A probability value (p-value) corresponding to ξ is derived from the cumulated density function of the absolute statistics difference. In the normal case

$$p_t = 1 - \Phi_H(\xi) \tag{17}$$

$$= 2 * (1 - \Phi(\xi)) \tag{18}$$

where $\Phi_H(.)$ is the cumulative distribution function (CDF) of the standard half-normal distribution, and $\Phi(.)$ is the CDF of the standard normal distribution. The half-normal distribution is used to account for the absolute value in Eq. 16. The t index of p_t refers here to the analogy with the two-sample t-test for equal means [26]. For testing, the probability threshold corresponding to $P(\xi > \kappa = 1.96)$ is 0.05. For p_t above this value, one cannot reject the hypothesis that the observed difference between two values is due to statistical noise.

In order to be able to estimate p_t, one needs to evaluate the uncertainty on the difference of s_1 and s_2. Formally, it can be obtained by the combination of variances [11]

$$u(s_1 - s_2) = \sqrt{u^2(s_1) + u^2(s_2) - 2\mathrm{cov}(s_1, s_2)} \tag{19}$$

where $\mathrm{cov}(s_1, s_2)$ is the covariance between both statistics. The usefulness of this formula depends on several assumptions (theoretical limits of the statistics not within a high probability interval around their values, non-symmetric error intervals for non-linear statistics [27, 10],...). Nevertheless, it shows that the covariance between statistics can have a major effect on the amplitude of $u(s_1 - s_2)$. In the limit of very strong positive correlation, the uncertainty on the difference can become very small, impacting $\xi(s_1, s_2)$ and p_t.

To estimate of the effect of correlation on the comparison of scores, we introduce a variant p_{unc} of p_t, based on a version of the discrepancy ignoring correlation

$$\xi_{unc}(s_1, s_2) = \frac{|s_1 - s_2|}{\sqrt{u(s_1)^2 + u(s_2)^2}} \tag{20}$$

$$p_{unc} = 2 * (1 - \Phi(\xi_{unc})) \tag{21}$$

In the hypothesis of mostly positive covariances for the ranking statistics of interest, p_{unc} is expected to overestimate the p-value.

2.4.2 Bootstrap-based comparison of statistics

Several strategies can be considered to compare pairs of statistics (s_1, s_2) through a p-value.

Estimate $u(s_1)$, $u(s_2)$ and $\mathrm{cov}(s_1, s_2)$. The uncertainty on the statistics of interest (except for the MSE and RMSD) and their covariance are not, to our knowledge, available in analytical form. In consequence, one has to use a numerical procedure, such as the bootstrap to estimate them [21, 23]. The application of the bootstrap to individual terms of Eq. 19 will result in an accumulation of sampling uncertainties. Besides, the estimation of covariances is very sensitive to outliers. This approach is clearly suboptimal and is not recommended.

Estimate directly $u(s_1 - s_2)$. A better approach in the present context is to estimate directly (by bootstrap) the uncertainty on the difference of scores. This relieves some distributional hypotheses in Eq. 19, and enables the explicit correlation of samples of s_1 and s_2 through paired data sampling. However, estimating a discrepancy factor leads us to use Eq.18 to estimate the p-value, with the associated normality hypothesis.

Generalized p-value. The use of the generalized p-value (p_g), as proposed by Wilcox and Erceg-Hurn [29, 28] (method M; *cf.* Algorithm 1), conveniently avoids to estimate $u(s_1 - s_2)$, and the incurring normality hypothesis of p_t. It is based on a simple counting of null and negative bootstrapped differences of statistics with paired samples. Note that the use of paired samples is essential to capture inter-statistics correlations. Wilcox and Erceg-Hurn [28] have shown that their 'method M' provides a well controlled level of type I errors for the comparison of quantiles at the

Algorithm 1 Method M [28]: testing the equality of a statistic S for two paired samples by bootstrap and a generalized p-value (p_g).

Input: Two paired error sets E_1, E_2 of size N, and a statistic estimator S

1. Bootstrap the statistics difference
   (a) For $j = 1 : B$
      i. Generate a N-sample of paired data with replacement $\longrightarrow (E_1^*, E_2^*)$
      ii. Estimate $d_j = S(E_1^*) - S(E_2^*)$
2. Calculate a generalized p-value to test $S(E_1) = S(E_2)$

   $p_g = 2\min(p^*, 1 - p^*)$,where
   $p^* = (A + 0.5C)/B$
   $A = \sum_{i=1}^{B} 1_{d_i<0}$
   $C = \sum_{i=1}^{B} 1_{d_i=0}$

0.05 level. They estimated that dataset sizes of $N \geq 30$ are necessary when comparing quantiles up to 0.9. Using the same protocol, we estimated that for the comparison of Q_{95} values at the same 0.05 level, $N \geq 60$ is requested. Details are presented in Appendix C.

2.4.3 Rank inversion probability P_{inv}

In a previous article [3,30], we defined a probability for ranking inversion $P_{inv} = P(S_1 < S_2 | s_1 > s_2)$, based on the normal distribution. Our goal was to estimate the probability that a given ranking results from a limited benchmark dataset. Using the present notations, P_{inv} can be reformulated as

$$P_{inv} = \Phi(0, \mu = s_1 - s_2, \sigma = \sqrt{u^2(s_1) + u^2(s_2)}) \tag{22}$$
$$= \Phi(0, \mu = \xi_{unc}) \tag{23}$$
$$= \Phi(-\xi_{unc}) \tag{24}$$
$$= 1 - \Phi(\xi_{unc}) \tag{25}$$
$$= p_{unc} \,/\, 2 \tag{26}$$

This shows the limitations of this previous definition of P_{inv}, *i.e.*, the normality hypothesis and the neglect of error sets correlations. Using the same difference statistics used for p_g (Algo. 1), one can generalize P_{inv} as the probability to have differences in the bootstrap sample with a sign opposite to the reference one ($\text{sign}(s_1 - s_2)$), *i.e.*,

$$P_{inv} = \frac{1}{B}\left(\sum_{i=1}^{B} 1_{\text{sign}(d_i) \neq \text{sign}(s_1 - s_2)} - \sum_{i=1}^{B} 1_{d_i=0}\right) \tag{27}$$

where the null differences, with sign 0, are excluded from the count.

Algorithm 2 Estimating the rank probabilities for a set of methods.

Input: K paired error sets, $E_1, \ldots, E_K$ of size N, and a statistic estimator S

1. Bootstrap the ranks
   (a) For $j = 1 : B$
      i. Generate a N-sample of paired data with replacement $\longrightarrow (E_1^*, \ldots, E_K^*)$
      ii. Estimate the statistics vector $S^* = (S(E_1^*), \ldots, S(E_K^*))$
      iii. Estimate the ranks by increasing order of S^*: $O_j^* = \text{order}(S^*)$, where O_j^* is a K-vector of integer values.
2. Estimate for each method its probability to have any rank

$$P_{r,jk} = \frac{1}{B} \sum_{i=1}^{B} 1_{O_{ij}^*=k}$$

If one takes as reference s_2 the smallest value ŝ of a statistic within a set of methods, one gets

$$P_{inv} = \frac{1}{B} \left(\sum_{i=1}^{B} 1_{\text{sign}(d_i) \neq \text{sign}(s_1 - \hat{s})} - \sum_{i=1}^{B} 1_{d_i=0} \right) \tag{28}$$

$$= \frac{1}{B} \left(\sum_{i=1}^{B} 1_{d_i<0} - \sum_{i=1}^{B} 1_{d_i=0} \right) \tag{29}$$

$$\simeq p_g \,/\, 2 \tag{30}$$

where the relation to p_g assumes a negligible probability to have null statistics differences and exploits the fact that that $\sum_{i=1}^{B} 1_{d_i<0} < \sum_{i=1}^{B} 1_{d_i>0}$ for our choice of reference $\hat{s}$.

2.4.4 Ranking probability matrix $\mathbf{P}_r$

A measure of the reliability of a statistic-based ranking can be estimated by bootstrap [31]. This approach has notably been used by Proppe and Reiher [9] to study how the sample size affects the probability for a DFA to be ranked first on the basis of its prediction uncertainty. We apply it here to compute, for a set of K methods scored by a statistic S, a ranking probability matrix $\mathbf{P}_r$ giving, for each method, its probability to have any rank

$$P_{r,jk} = P(\text{rank}(S_j) = k);\ j, k = 1, \ldots, K \tag{31}$$

The algorithm to generate this matrix is described in Algo. 2.

Representations. Two representations for this matrix are used by Hall and Miller [31], either a levels image (Fig. 3(a)), or a summary by probability intervals (Fig. 3(b)). In the following, we will

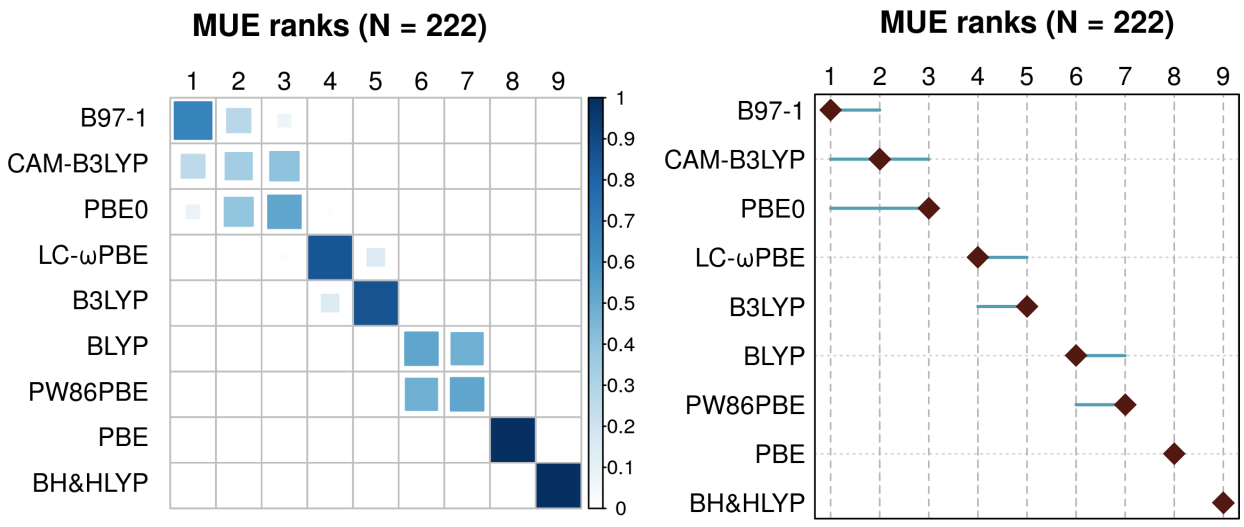

Figure 3: Graphical representations of sample-size effect on ranking: (left) levels image of the ranking probability matrix; (right) summary of the ranking probability matrix by the modes (diamonds) and 90 % probability intervals. The data are taken from the case Pernot2018 (Section 3.1). Both representations indicate a possible ranking inversion between B97-1, CAM-B3LYP and PBE0, *i.e.*, the reference ranking based on the MUE is not certain for this trio. Similar problems occur within two other groups, notably PW86PBE and BLYP. The ranks of PBE (8) and BH&HLYP (9) seem well established.

use mostly the levels image representation which we find easier to read and interpret.

A summary in results tables can also be considered, by reporting for each method its mode in ranking probability and the corresponding probability, which indicates the strength of this rank. These tools enable to appreciate easily the robustness of the reference ranking.

Remarks.

- As discussed by Hall and Miller [31], the standard bootstrap used in the present article (N-out of-N sampling) tends to underestimate the dispersion of the ranks. Better estimates would be obtained by a M-out of-N sampling ($M < N$), but the best choice of M is left to the appreciation of the analyst. For the sake of simplicity, we consider here that the standard method provides a reasonable qualitative appreciation of ranking deficiencies.

- As a general trend, one expects that ranking uncertainty will increase for smaller error sets, but one should also consider that ranking uncertainty might increase with the number of compared methods K (*cf.* Section 3.4). We did not explore this, but there should be an optimal balance between the minimal dataset size and the number of methods that can be effectively ranked.

Code	Property	N	K	Source
PER2018	Intensive atomization energies	222	9	[3, 30]
BOR2019	Band gaps	471	15	[37]
NAR2019	Enthalpies of formation	469	4	[38]
CAL2019	London Dispersion Corrections	41	3*10	[39]
JEN2018	Non-covalent interaction energies	66	6	[40]
DAS2019	Dielectric Constants	23	6	[41]
THA2015	Polarizability	135	7	[4]
WU2015	Polarizability	145	34	[5]
ZAS2019	Effective atomization energies	6211	3	[42]

Table 1: Case studies: N is the number of systems in the dataset and K is the number of compared methods.

2.5 Implementation

Calculations have been made in the `R` language [32], using several packages, notably for bootstrap (`boot` [33]). Bootstrap estimates are based on 1000 replicates.

Quantiles. Wilcox and Erceg-Hurn [28] recommend the use of Harrell and Davis method for quantiles estimation [34], which provides a better stability for the bootstrap sampling of quantiles. The impact of this choice is illustrated in Appendix D. In the case studies, all quantiles were estimated by the Harrell and Davis method [34], as implemented in package WSR2 [28, 35, 36].

Correlation. The estimation of correlation coefficients by the standard Pearson method is reputed to be very sensitive to the presence of outliers [35]. As the presence of a small amount of outliers is a frequent feature of the benchmarking data sets, we used systematically the more robust rank-correlation (Spearman) method.

3 Case studies

In this section, we validate and illustrate the methods presented in the previous section on several representative cases. The choice of dataset is mostly based on their availability and on the coverage of a representative range of dataset sizes – between a few tens to a few thousands – and uncertainty (Table 1).

3.1 PER2018

We consider here the intensive atomization energies [43] estimated with 9 DFAs on the G3/99 dataset [44], and extracted from Ref. [3, 30]. 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}

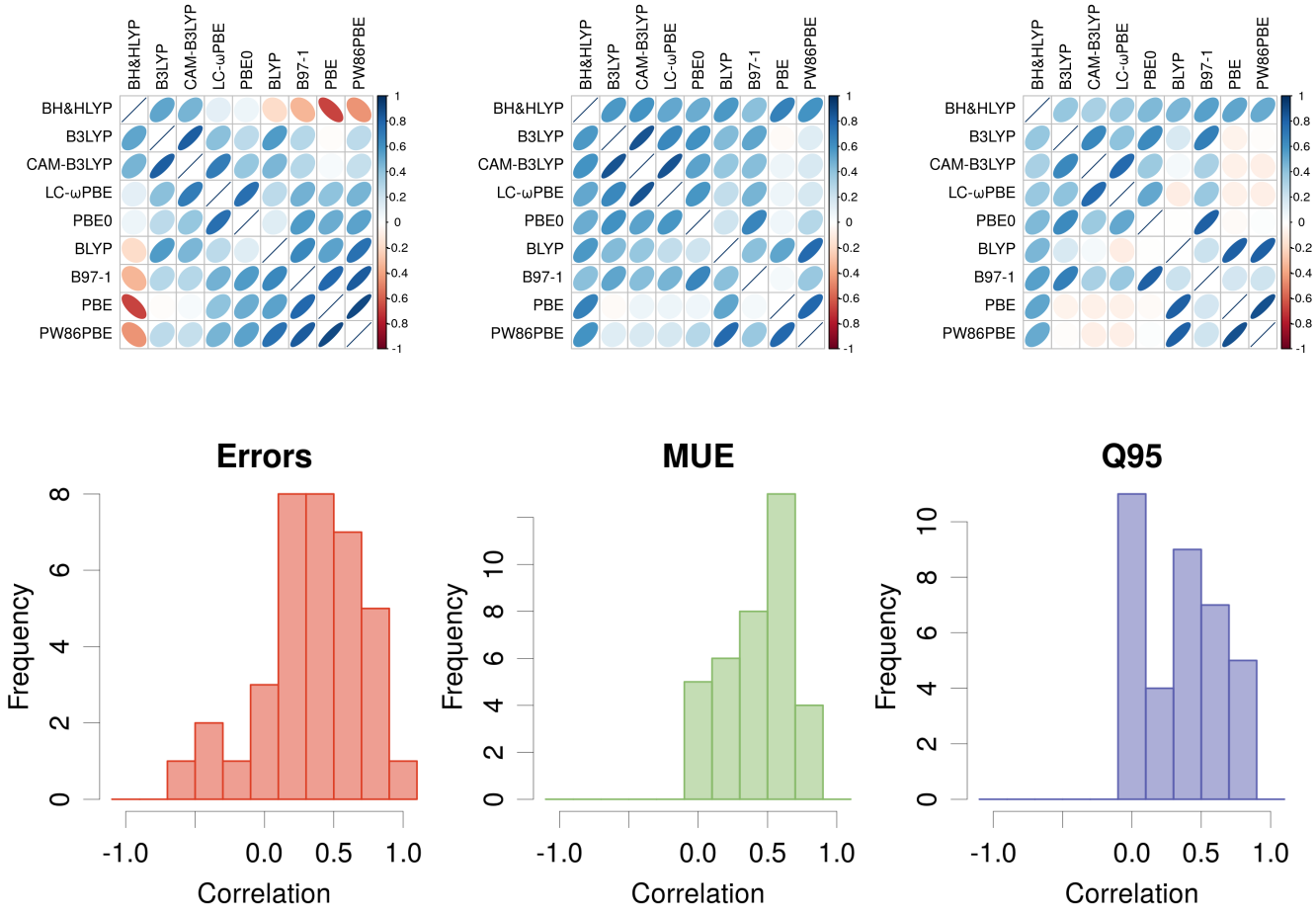

Figure 4: Case PER2018 - correlations: (top) rank correlation matrices between Errors sets, MUE and Q_{95} ; (bottom) histogram of non-diagonal elements of the corresponding correlation matrices. The methods are ordered by a clustering algorithm of the errors correlation matrix (`hclust` [32])

as a complement to the MUE, and to illustrate the former definition of P_{inv}. Here we focus on the correlations and their impact on the comparison of statistics.

Correlations. The correlation matrices between the error sets and their statistics are represented in Fig. 4, along with histograms of correlation coefficients of their non-diagonal elements. The errors sets are all positively correlated, with a wide distribution, except for pairs involving BH&HLYP which presents negative correlations with several methods. When considering the scores, all correlations are positive, and the checkerboard pattern is rather well preserved. Globally, the correlations are weaker for Q_{95} than for the MUE. 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}..

Statistics. The statistics are reported in Table 2. Note that, due to the change in quantile estimation algorithm, the values of Q_{95} have changed slightly from the values reported in Ref. [3].

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)

Table 2: Case PER2018: Absolute error statistics, p-values and inversion probabilities and SIP statistics for comparisons with respect to the DFA with the smallest MUE (B97-1), the reported SIP values correspond to the B97-1 line of the SIP matrix.

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 chance. 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 [30], included LC-ωPBE in the group of methods with a sizable risk of inversion. As demonstrated in Eq. 30, 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 30 % chance to be wrong, *i.e.*, that the MUE of CAM-B3LYP is indeed smaller than B97-1 due to the restricted sample size. This risks falls to 10% when PBE0 is concerned.

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, where the MUE is not discriminant.

This example enabled to illustrate and confirm the relations between p_{unc}, p_g and P_{inv} expressed in Section 2.4.3. In the following examples, only P_{inv} will be reported.

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. 1, that indeed, the row for CAM-B3LYP is fully reddish, when those for B97-1 and PBE0 present also blue and white patches. At the opposite, the deficiency of BH&HLYP for this dataset is clearly visible as a full deep blue line.

The ECDF of the difference of absolute errors for CAM-B3LYP and B97-1 helps to understand the contradiction between the MUE and MSIP ranks (Fig. 5(a)). The MUE difference for this pair is statistically not significant ($p_{g=0.57}$), the SIP value is about 0.65 – a small improvement of CAM-B3LYP over B97-1 – the mean gain -0.6 kcal/mol and the mean loss -1.3 kcal/mol, due to the heavy tail in the CAM-B3LYP error distribution. So by switching from B97-1 to CAM-B3LYP, one has to accept a risk of 35 % to degrade the intensive atomization energies by 1.3 kcal/mol in average and up to 4 kcal/mol. The same comparison between CAM-B3LYP and PBE0 (Fig. 5(b)) shows that there is no strong basis to favor either method.

Ranking. The ranking probability matrices (Figs 3 and 6) confirm the 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.

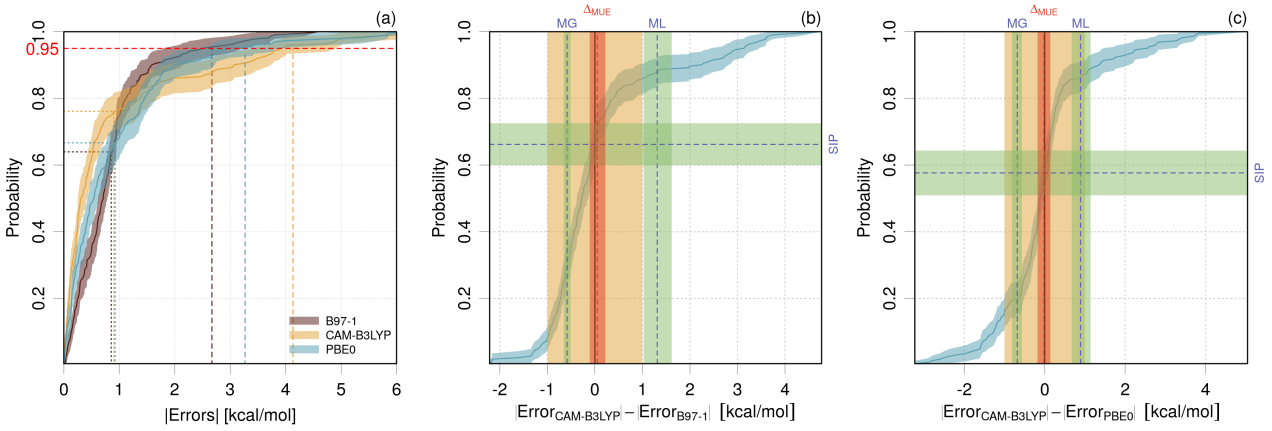

Figure 5: Case PER2018 - absolute errors statistics: (a) ECDF and statistics of absolute errors; (b-c) ECDF and statistics of the difference of absolute errors. See Fig. 2 for details. The orange band depicts the chemical accuracy (1 kcal/mol).

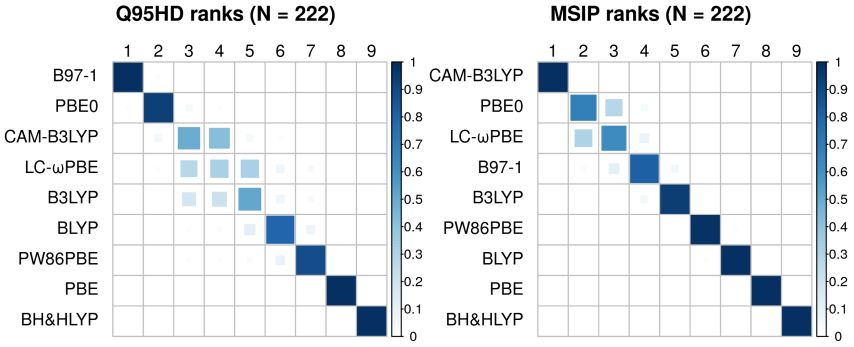

Figure 6: Case PER2018: levels representation of the ranking probability matrix for Q_{95} and MSIP.

3.2 BOR2019

Band gap estimations for a set of 471 systems[2] by 15 DFAs were extracted from the Supplementary Information of a recent article by Borlido *et al.* [37]. For a full description of the dataset, see the original article. The reference authors reported and analyzed relative errors, but this causes an unsuitable distortion of the errors distributions, with large relative errors for small band gaps, and small errors for large band gaps. It is true that for some methods (*e.g.*, LDA) the errors increase with the value of the band gap, but this is due mostly to a systematic deviation, not an increase in the dispersion of the values. In consequence, we treat here the 'absolute' errors, as defined in Eq. 1.[3]

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

[3]The reference authors report in their Table 1 Kendall and Pearson correlation coefficients between the calculated and reference values. First, one should remind that correlation coefficients are not reliable performance indicators [45]. At most, they reveal a linear (Pearson) or monotonic (Kendall) association between two variables X and Y, but not their proximity to the identity ($X = Y$) line. Nevertheless, the notable difference (between 0.1 and 0.2), for each method, between both estimators point to the presence of outliers and/or a non-linear relationship.

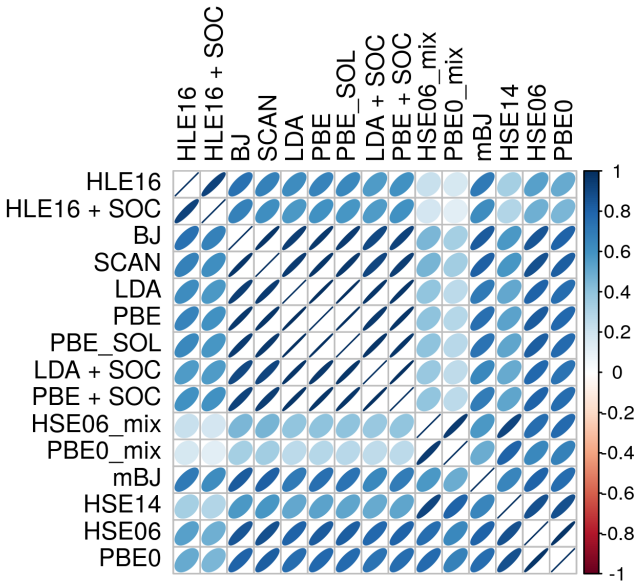

Figure 7: Case BOR2019: rank correlation between errors sets. The methods are ordered by a clustering algorithm (`hclust` [32]).

Correlations. One sees in Fig. 7 that across the spectrum of methods, all 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 HSE06_mix. If the error sets are dominated by method errors (*i.e.*, there are no large reference data uncertainty, 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 (Fig. 7). Error sets with large correlation coefficients are related by a linear or monotonous transformation and the corresponding methods are assigned to the same class. Andreas, tu peux un peu développer ?

Statistics. The values are reported in Table 3. 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 ($p_g = 0.15$). Besides HSE06 is the only method with a notably non-zero p_g value, either for MUE or Q_{95}.

SIP analysis. mBJ has also the largest MSIP, but its value is moderate (0.7), indicating that mBJ does not provide a full systematic improvement over some of the methods. The SIP for mBJ with respect to the other methods values range 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 of 0.52-0.53 are also observed against HLS16, HLSE16+SOC and HSE06, indicating a notable risk of performance loss when switching from one of these methods to mBJ.

As seen previously, when going from LDA to mBJ (Fig. 2), one has less than 15 % chance to

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)**	0.08	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)

Table 3: Case BOR2019: Absolute error statistics and p-values for the comparison with respect to the DFA with the smallest MUE (B97-1).

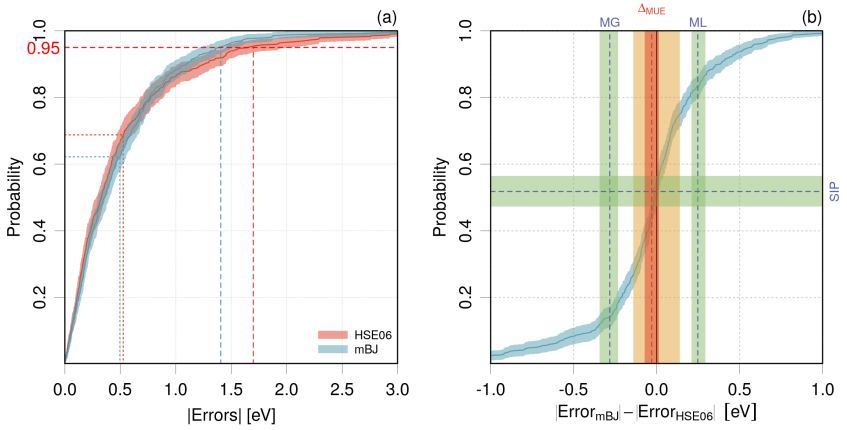

Figure 8: Case BOR2019 - absolute errors statistics: (a) ECDF of the absolute errors; (b) ECDF of the difference of absolute errors for mBJ and HSE06. See Fig. 2 for details. The orange band depicts the chemical accuracy (0.14 eV).

perform better by LDA than mBJ and the mean gain more than doubles the mean loss. By contrast, the comparison of mBJ to HSE06 (Fig. 8) is an example of undecidability: the MUE difference is not significantly different from zero, and one has as much to loose as to gain by switching between both methods.

The SIP matrix (Fig. 9) provides a convenient summary of these observations. The mBJ line is mostly reddish with white spots indicating neutral comparisons. In contrast,the LDA+SOC is fully blueish, indicating that it is dominated by all other methods.

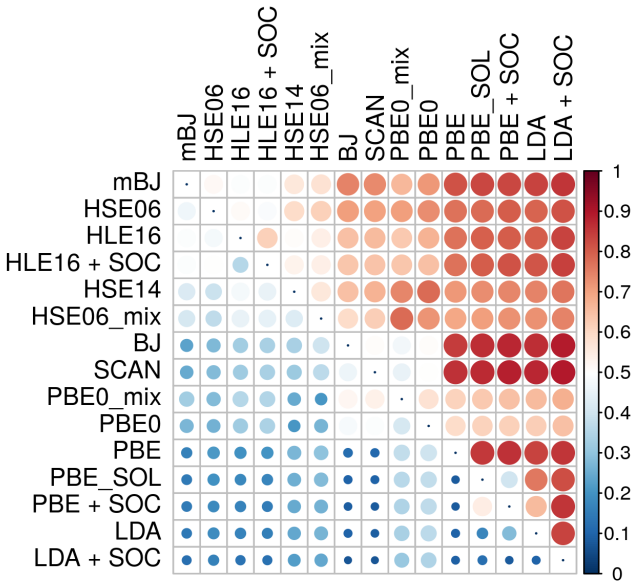

Figure 9: Case BOR2019: SIP matrix.

Ranking. Ranking probability matrices for the MUE and Q_{95} are presented in Fig. 10. They illustrate the previous results and show that ranking by MUE beyond the second place is adventurous. 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.

These matrices are convenient 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, 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 are very uncertain, up to the last ranks for $N = 100$. As for the MUE, the MSIP ranking suffers from the reduced datasets, but the head group of five methods is preserved.

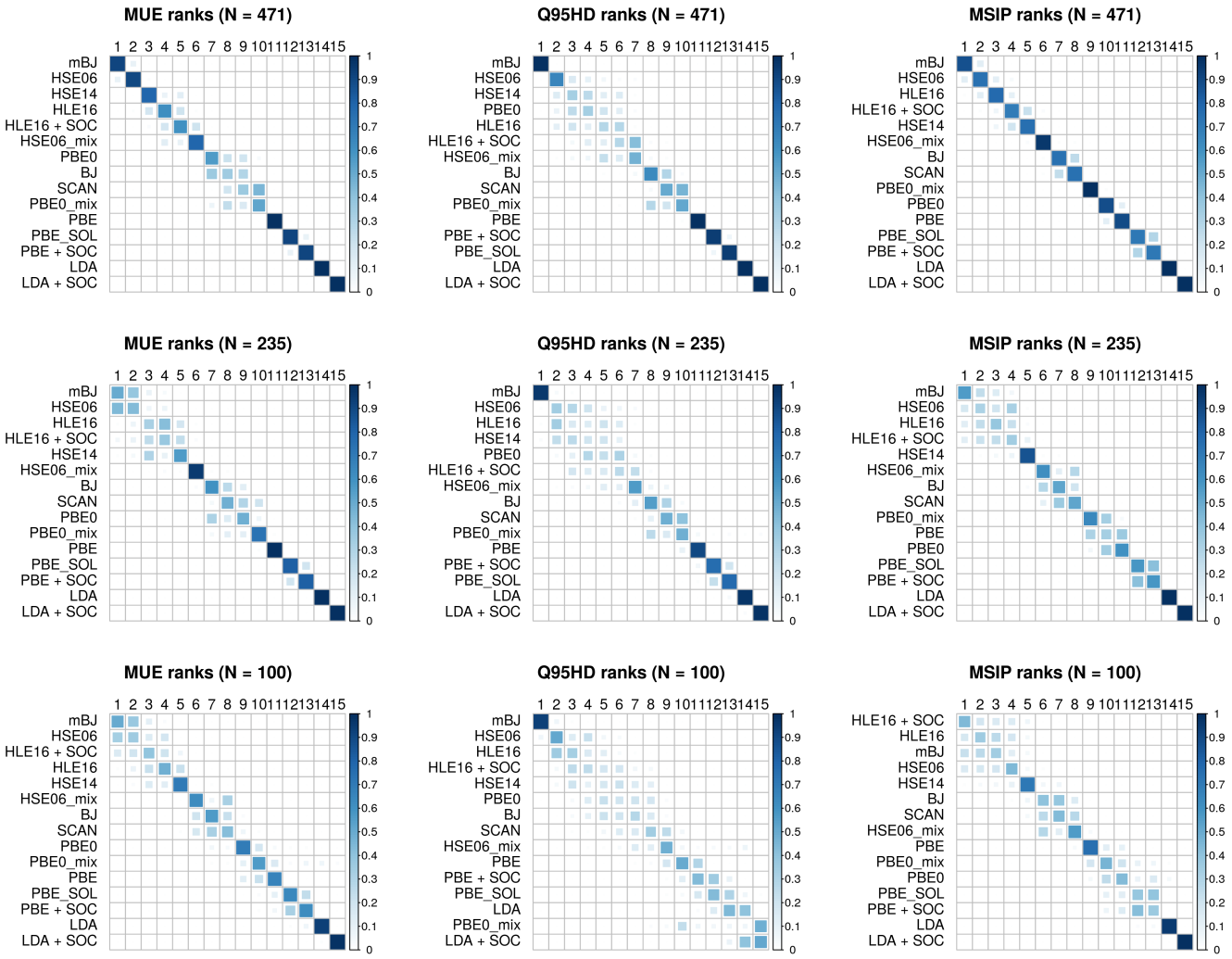

Figure 10: Case BOR2019: ranking probability matrices for the full dataset (top row, $N = 471$), and for reduced sets ($N = 235$ and 100).

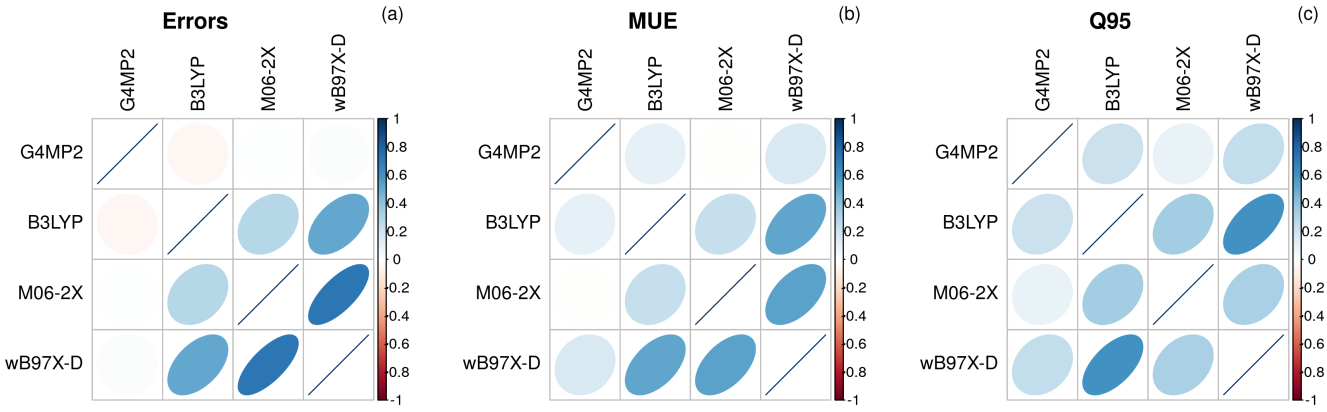

Figure 11: Case NAR2019 - rank correlation matrices: (a) Errors; (b) MUE; (c) Q_{95}.

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)

Table 4: Case NAR2019: same as Table 2 for case Narayanan2019.

3.3 NAR2019

The dataset (Pedley test set) contains the calculated enthalpies of formation by G4MP2 for 469 molecules having experimental values with small uncertainty [38]. The G4MP2 values are compared with those of B3LYP, M06-2X and ωB97X-D. The authors claim an "accuracy"[4] (MUE) of 0.79 kcal/mol with G4MP2 and that the best of the DFAs (ωB97X-D) has a MUE larger than twice the G4MP2 value.

Correlations. The most remarkable feature in the correlation matrices in Fig. 11 is the decorrelation of G4MP2 errors from those of the other methods. Weak positive correlations appear, notably for Q_{95}.

Statistics. The values of Q_{95} confirms the superiority of G4MP2 over the 3 DFAs (Table 4) noted for the MUE. A look at the absolute errors CDFs (Fig. 12(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. The method does not therefore guarantee predictions with chemical accuracy.

SIP analysis. G4MP2 presents a high degree of systematic improvement over the three DFAs (MSIP = 0.81). Nonetheless, there is about 25 % probability that ωB97X-D performs better, but with

[4]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 [19, 3].

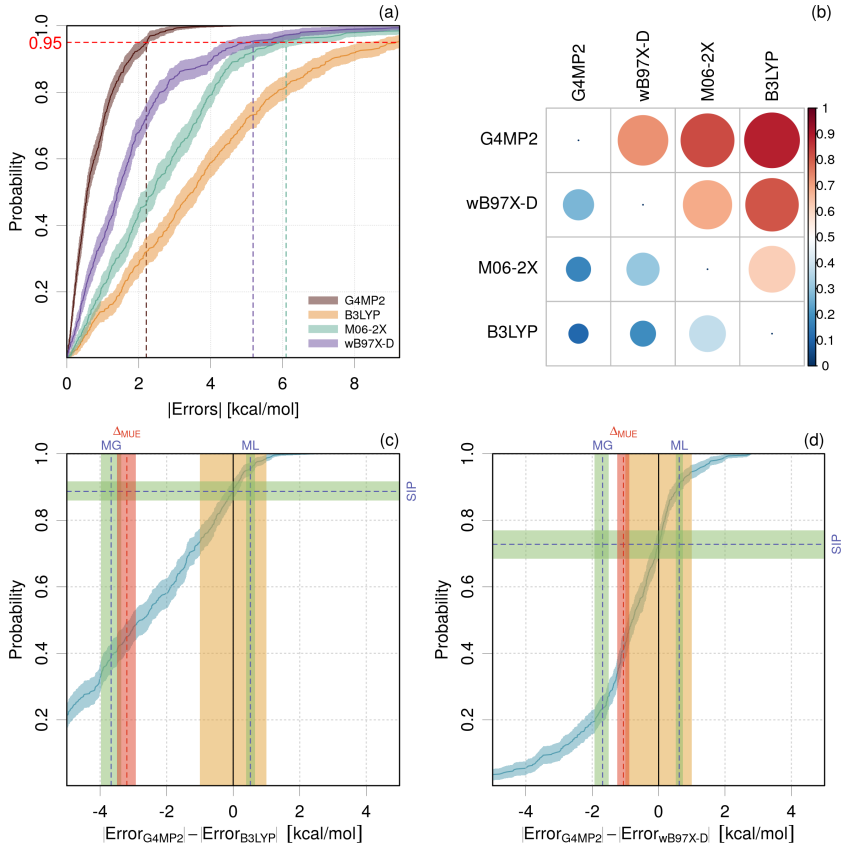

Figure 12: Case NAR2019: (a) ECDF of the absolute errors; (b) SIP matrix; (c,d) ECDF of the difference of absolute errors of B3LYP and ωB97X-D with respect to G4MP2.

a rather small value of ML (0.62 kcal/mol), when compared to the chemical accuracy (Fig. 12(d)). In contrast, the mean gain when using G4MP2 instead of ωB97X-D is about -1.7 kcal/mol for about 75 % of the systems. The advantage of G4MP2 over B2LYP is more spectacular (Fig. 12(c)).

Ranking. All the ranking probability matrices are diagonal (not shown). There is no risk of inversion, a conjunction of the use of a large dataset and few characteristic methods.

3.4 CAL2019

The impact of an atomic-charge dependent London dispersion correction (D4 model) has been evaluated by Caldeweyher *at al.* [39] on a large series of datasets. From those, we selected one of the largest one, *i.e.*, the reference energies for the MOR41 transition metal reaction benchmark set [46], available as Tables 14-18 in the Supplementary Information of the reference article.[5] The London dispersion corrections 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 SI table, which differs from the one used in the reference article.

Statistics. The results are reported in Table 5, where DFT-D3 has been taken as reference throughout. 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). Based on MUE, significant improvements are observed when passing from DFT-D3 do 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 sample effect. Based on Q_{95} the improvements due to D4 are not significant, except for DOD-PBE, DSD-PBE and RPBE. Globally, DFT-D4 does not reduce the risk of large errors.

SIP analysis. Let us consider several examples with the SIP approach.

- **PBE0-Dn**. There is a small MUE decrease using D4 (from 2.6 to 2.3), no effect on Q_{95}, and the best MSIP is for PBE0-D4-ATM, with a neutral value of 0.5. Inspection of Fig. 13(a) shows that the 95 % confidence interval for the SIP value of 0.6 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 their large error bars, the small MUE difference between these two methods is significantly different from 0 (its 95 % confidence interval excludes 0), an effect of the correlation between error sets.

- **PW6B95-Dn**. This case is an inversion of the previous one, where the SIP value of 0.4 (disadvantaging D4) does not exclude the neutral value. The MUE difference does not reject 0. One cannot conclude that the D3 version performs better than the D4 ones.

- **RPBE-Dn**. For this case, one has a rare instance where D4 improves 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

[5] 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 Information.

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)**	0.08	0.56(6)	0.41(8)	-0.22(4)	0.21(3)
B3LYP-D4-MBD	**4.2(5)**	0.01	**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)**	0.00	**8(1)**	0.08	0.30(4)	-	-	-
PBE0-D4-MBD	**2.3(3)**	0.00	**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)**	0.03	**7.9(9)**	0.33	0.35(6)	0.56(8)	-1.6(2)	1.0(2)
PW6B95-D4-MBD	**3.0(4)**	0.11	**7.8(8)**	0.34	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)**	0.05	0.38(4)	-	-	-
CAM-B3LYP-D4-MBD	**3.7(4)**	0.00	**9(1)**	0.05	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)**	0.10	**12(2)**	0.33	0.43(6)	0.54(8)	-0.27(6)	0.28(6)
revPBE-D4-MBD	**3.3(6)**	0.07	**12(2)**	0.38	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)**	0.10	0.35(4)	-	-	-
M06L-D4-MBD	**5.1(6)**	0.00	**13(1)**	0.10	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)**	0.32	0.45(6)	0.51(8)	-0.20(5)	0.16(2)
PBE-D4-MBD	**3.4(5)**	0.00	**12(2)**	0.46	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)

Table 5: Case CAL2019: same as Table 2. DFT-D3 has been taken as reference throughout.

improvement when passing from D3 to D4. In several cases, the uncertainty due to the limited set size does not allow to conclude clearly.

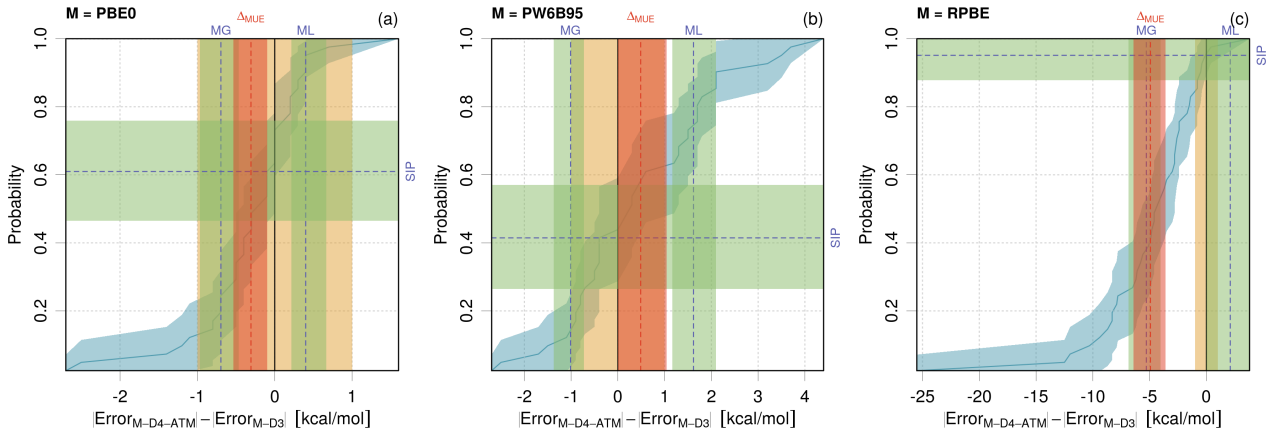

Figure 13: Case CAL2019: SIP plots

Ranking. Considering that both DFT-D4 options are mostly indiscernible, we built global ranking probability matrices for the DFT-D3 and DFT-D4-ATM data. The results are reported in Fig. 14(top). Although the rankings for each DFA were mostly unambiguous, a global ranking is clearly very uncertain. Based on the MUE, DOD-PBE-D4 and PBE0-D4 would share the leading places. Beyond that, the situation is very 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 group of five. The tail of the ranking comforts the MUE analysis. If one restricts the methods to DFT-D3 (Fig. 14 bottom), the situation is slightly better for the leading and tailing places for both MUE and Q_{95}, but remains very undecidable in the middle. This illustrates how, for a given dataset, the uncertainty in ranking is also affected by the number of methods to be ranked.

3.5 JEN2018

This dataset contains non-covalent interaction energies estimated by M06-L with six different basis sets for 66 systems in the S66 dataset [47, 48]. This is a part of the results reported in Table 8 of Ref. [40], and available as Supplementary Information 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 the '-opt' ones have optimized contraction coefficients with respect to the reference data.

Correlations. The error sets of the '-opt' methods are uncorrelated to the other sets (Fig. 15(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. Otherwise, the correlations globally weaken for Q_{95}, except for the pop2/pop3, where the correlation remains as strong as the one between the error sets.

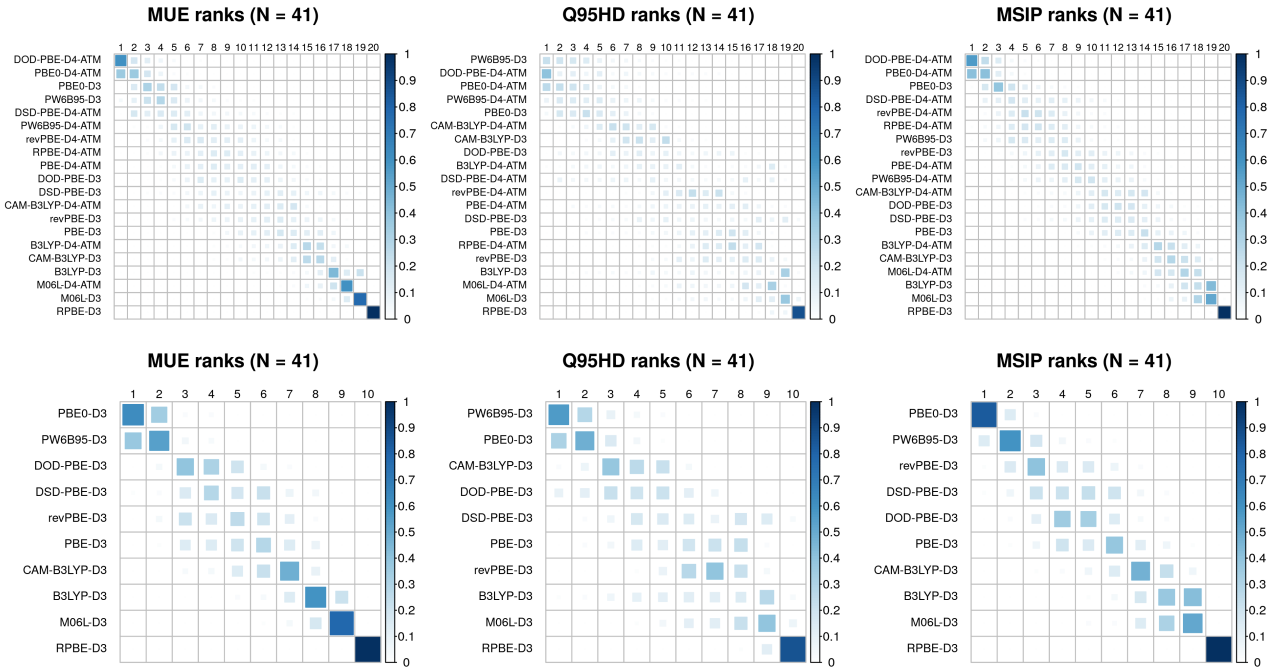

Figure 14: Case CAL2019: ranking probability matrices for (top) DFT-D3 and DFT-D4-ATM methods, and (bottom) DFT-D3 methods only.

Statistics. The statistics in Table 6 show the strong impact of basis-set optimization, both optimized basis sets provide comparable results for the MUE and Q_{95}.

SIP analysis. They both also stand out bey their MSIP, with a light advantage for pcseg1-opt. One more, 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, the smallest 0.6 over pop2-opt. The plots in Fig. 16 illustrate these features. The SIP matrix shows clearly the medium supremacy of the optimized basis sets, and a light advantage of pcseg1-opt over pop2-opt. The major gain when going to pop2 to pop2-opt is visible in Fig. 16(c) where the medium SIP

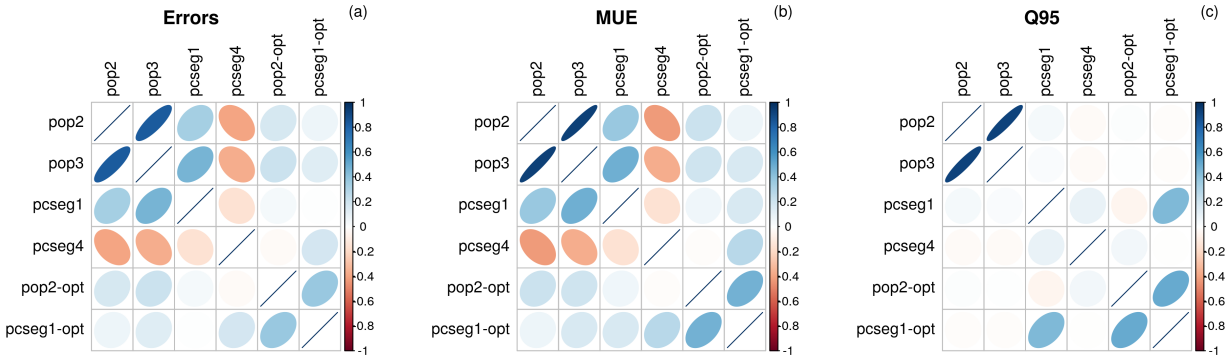

Figure 15: Case JEN2018: rank correlation matrices.

Methods	MUE kJ/mol	P_{inv}	Q_{95} kJ/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)**	-	-	-

Table 6: Case JEN2018: same as Table 2.

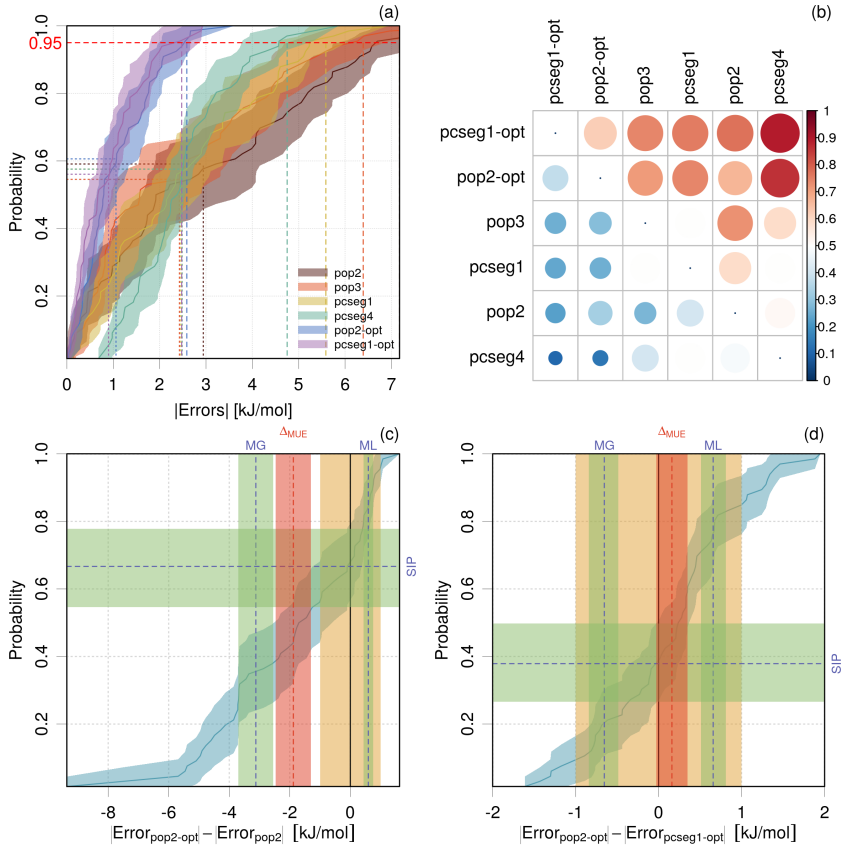

Figure 16: Case JEN2018: (a) ECDF of the absolute errors; (b) SIP matrix; (c,d) ECDF of the difference of absolute errors of pop2 and pcseg1-opt with respect to pop2-opt.

(~0.7) is compensated by the very small mean loss (0.6 kJ/mol). In contrast, Fig. 16(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.

Ranking. The leading position of the '-opt' methods is solid and confirmed by our three scores.

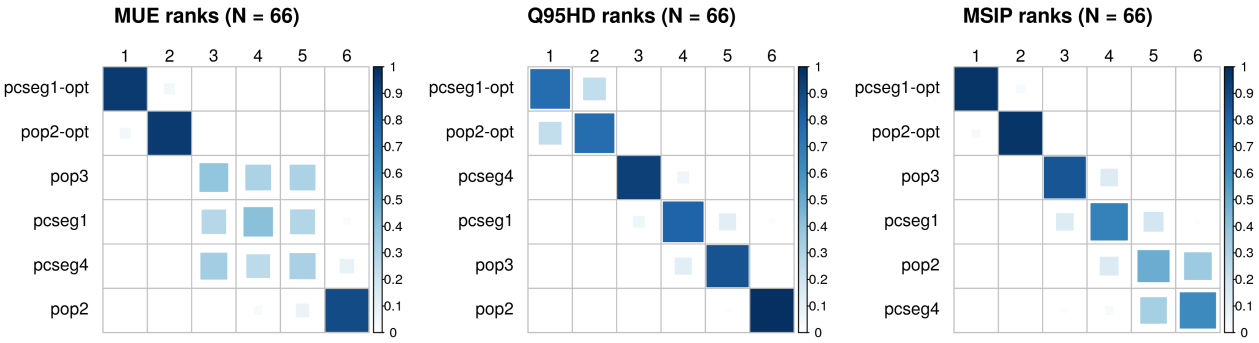

Figure 17: Case JEN2018: ranking probability matrices.

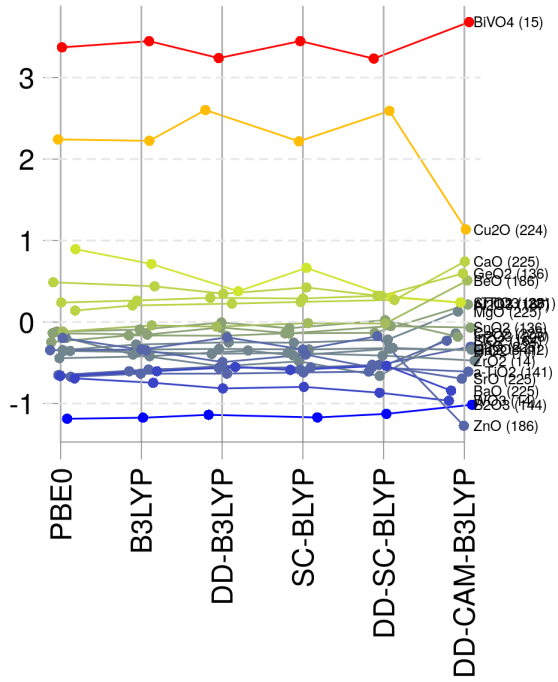

Figure 18: Case DAS2019: parallel plot of scaled and centered error sets to identify outliers.

3.6 DAS2019

A set of 24 dielectric constants for 3D metal oxides has been reported by Das *et al.* [41] in their Table 3. One of the experimental values being unknown, this limits the dataset to 23 values. 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).

Correlations. The correlation matrices of the errors, MUE and Q_{95} have uniformly strongly positive elements (Fig. 19-top). This is an unusual situation when compared to the previous cases. Knowing that correlation coefficients are sensitive to outliers (even if rank correlations are a little more robust), we explored the dataset for outliers. A parallel plot of the scaled and centered error sets enables to identify systems which deviate significantly from the core distribution. Fig. 18 shows that 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 those for MUE and Q_{95} are visibly more affected ((Fig. 19-bottom)). The parallel plot reflect the strong correlations between the errors sets (many parallel lines), except for DD-CAM-B3LYP. The pruned dataset ($N = 21$) is used in the following.

Statistics. Considering the reduced size of the sample, few clear-cut conclusions are possible. One is notably below the estimated limit of 60 points to get reliable estimates of Q_{95} uncertainty. Only DD-CAM-B3LYP stands out significantly, either by its MUE, Q95 and MSIP values (Table 7).

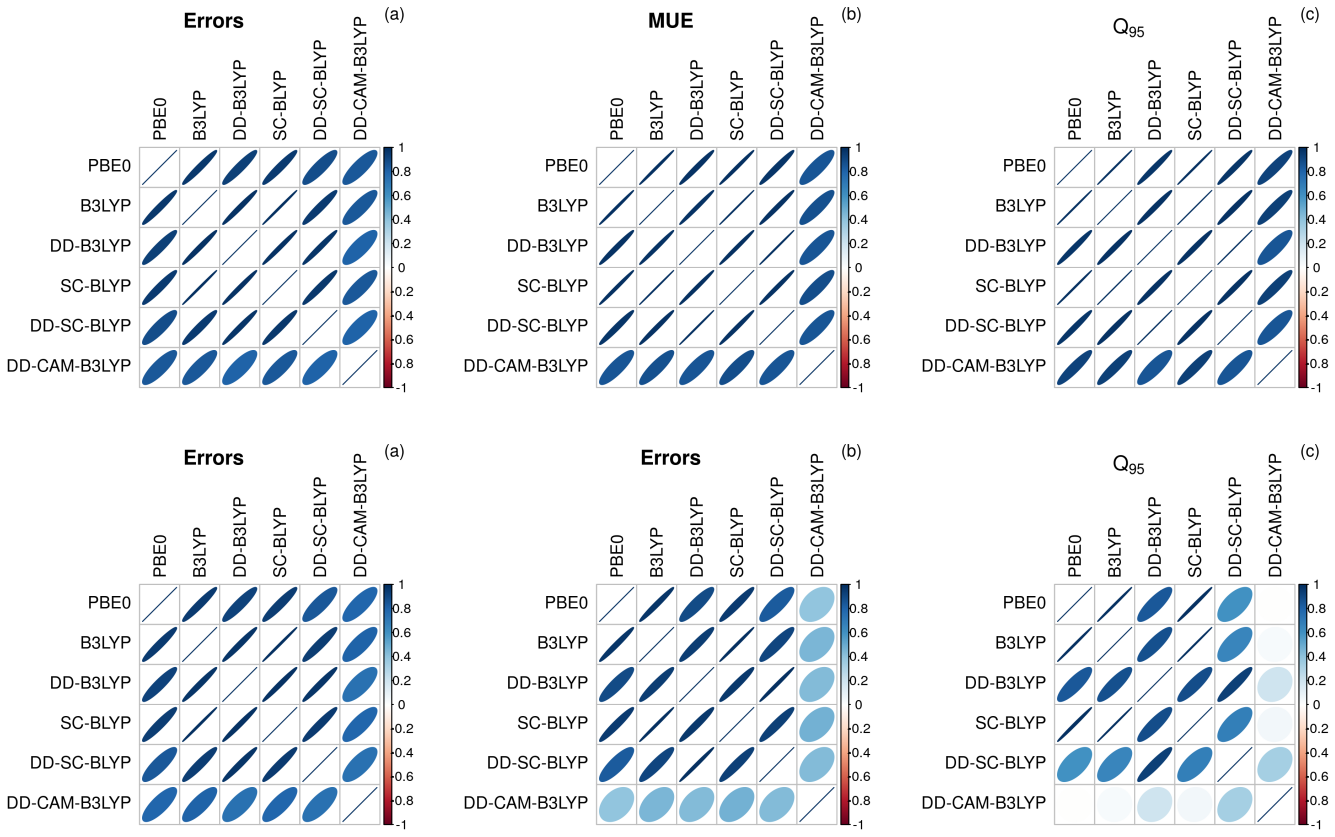

Figure 19: Case DAS2019: rank correlation matrices. (top) original data set ($N = 23$); (bottom) after removal of two outliers ($N = 21$).

Methods	MUE	P_{inv}	Q_{95}	P_{inv}	MSIP	SIP	MG	ML
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)**	-	-	-

Table 7: Same as Table 2 for case DAS2019.

At the opposite, 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.

SIP analysis. These two methods are clearly identifiable in the SIP matrix (Fig. 20), with a full reddish line for DD-CAMB3LYP, and a full blueish line for DD-B3LYP. The impact on the small set size on this conclusion is illustrated in Fig. 20(c-d), 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. Any ranking

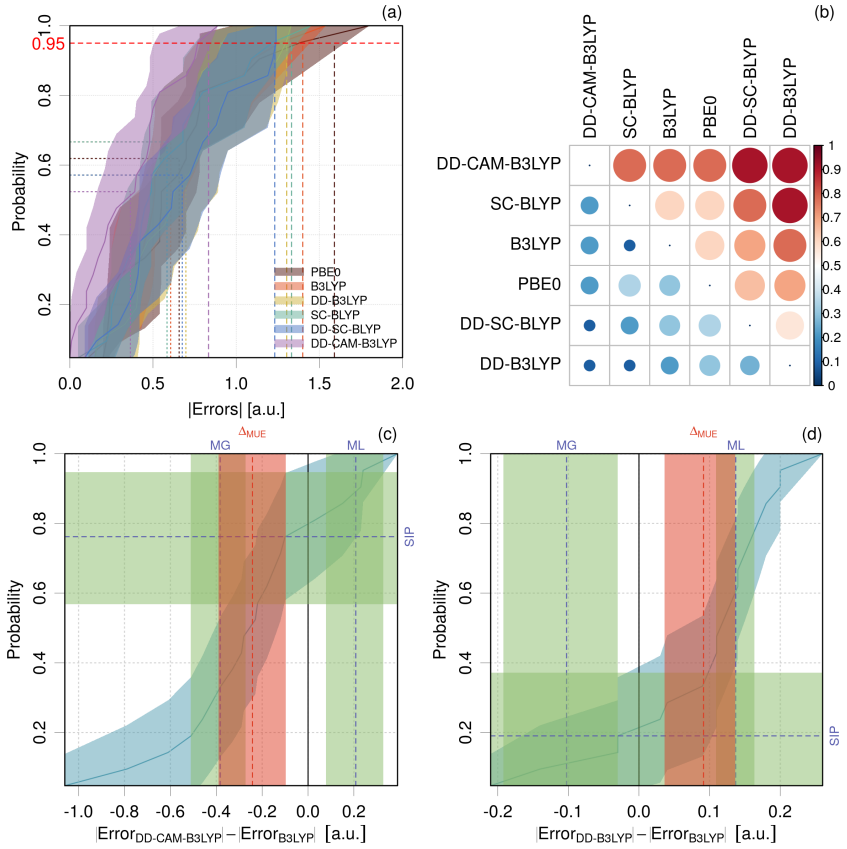

Figure 20: Case DAS2019: ECDF of the absolute errors.

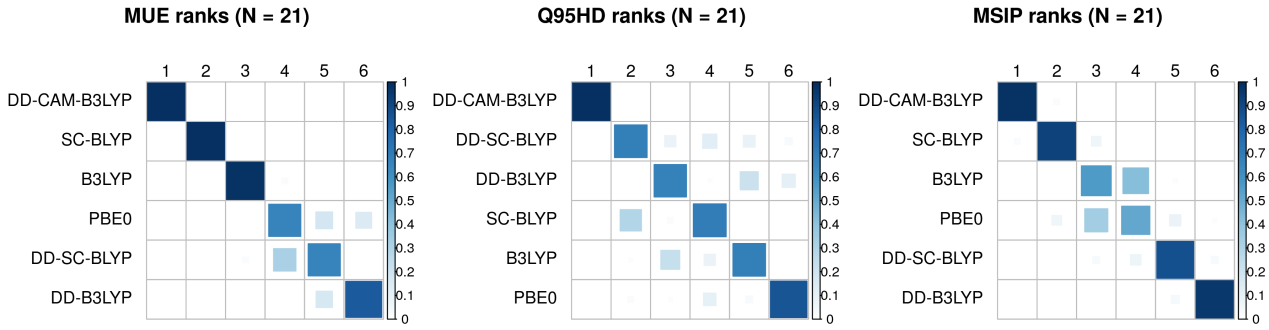

Figure 21: Case DAS2019: ranking probability matrices.

of the remaining four DFAs would be unreliable.

Ranking. All ranking matrices confirm a solid leading place for DD-CAM-B3LYP (Fig. 21). 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.

3.7 THA2015 / WU2015

Thakkar *et al.* [4] compiled a database of polarizabilities for 135 molecules, from triatomics to 26-atoms systems. The experimental data are given with their uncertainty, and computational results are provided for 7 methods. Data for our study were extracted from Tables II-IV in the reference article. The same year, Wu *et al.* [5] 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 7 methods common to both datasets. This enables us to study the impact of the reference data (experimental *vs.* calculated) on the correlation and ranking matrices. The raw errors present a dispersion increasing with the polarizability, hence relative errors are used in the reference article and this study.

Correlations. The Pearson correlation matrix of the error sets (Fig. 22(a)) is uniformly strongly positive. The smallest CC value is 0.8. This is in contrast with what we observed in previous examples, where only "related" methods tend to show strong correlations. To appreciate the role of data points with large deviations (outliers) in these strong correlations, we removed a set of 8 outliers identified by Thakkar *et al.* [4] ((Fig. 22(b)). Most of the correlations weaken notably. For comparison, the rank correlation matrix was calculated for the full dataset ((Fig. 22(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 (Fig. 22(d-f)). The structure of the errors correlation matrix is transferred to the statistics, with attenuated correlation intensities.

The error, MUE and Q_{95} rank correlation matrices were also calculated for the full WU2015 dataset (Fig. 23). In absence of experimental errors, MP2 errors are weakly anticorrelated to the other error sets, while all DFAs remain positively correlated. Again, one observes a weakening of MUE and Q_{95} correlations, except for the M11/M06-2X pair.

Statistics. The values of MUE and Q_{95} are reported in Table 8. The MUE values agree with those of the reference articles, 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 performances, which is confirmed by the comparison of their empirical cumulated distribution functions [3] (Fig. 24). These ECDFs overlap over the whole 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 8 outliers identified by Thakkar *et al.* (not shown).

SIP analysis. The SIP matrix (Fig. 25) for the THA2015 dataset reveals a leading group of four methods identical to those identified above. When passing to WU2015, there is a better

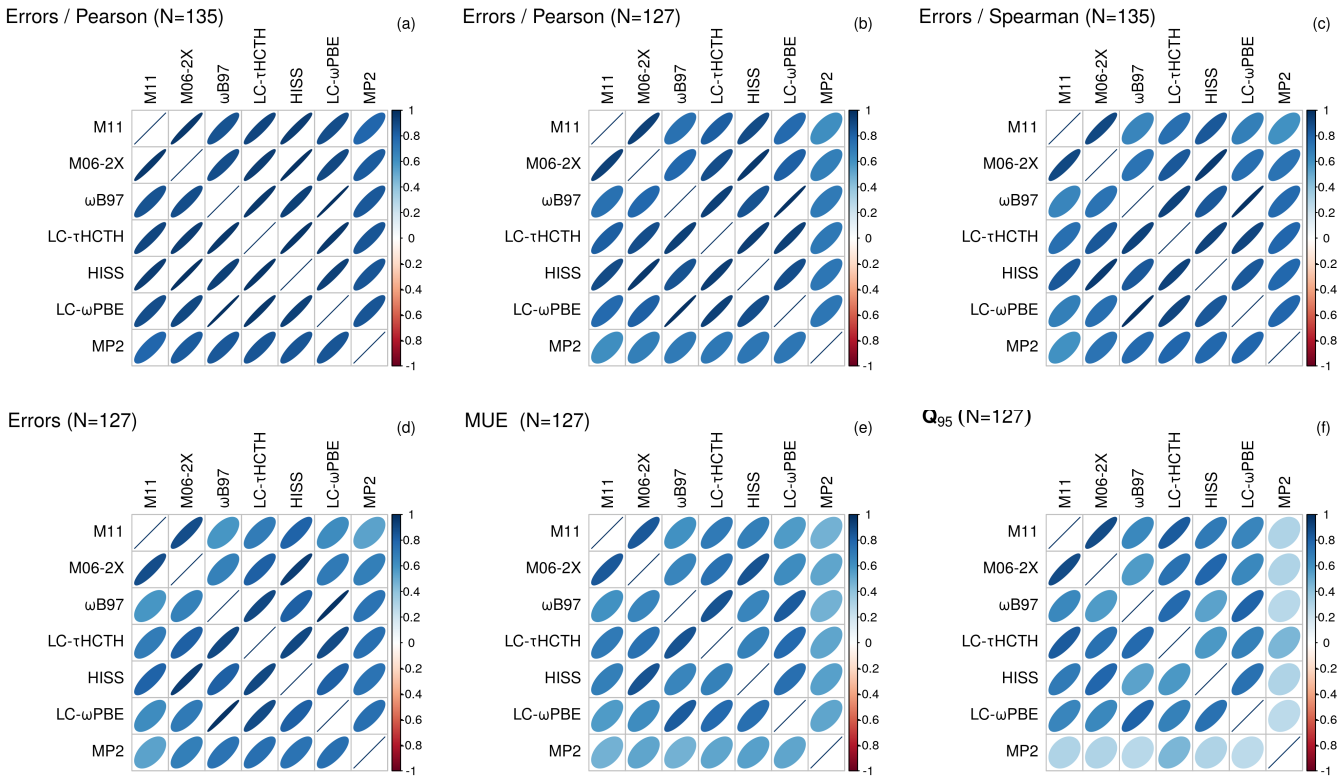

Figure 22: Case THA2015 - correlation matrix: (a) Pearson correlation of the full data set ($N = 135$); (b) Pearson correlation of the pruned dataset ($N = 127$); (c) Spearman/rank correlation of the full data set; (d): Errors rank correlation; (e): MUE rank correlations; (f) Q_{95} rank correlations.

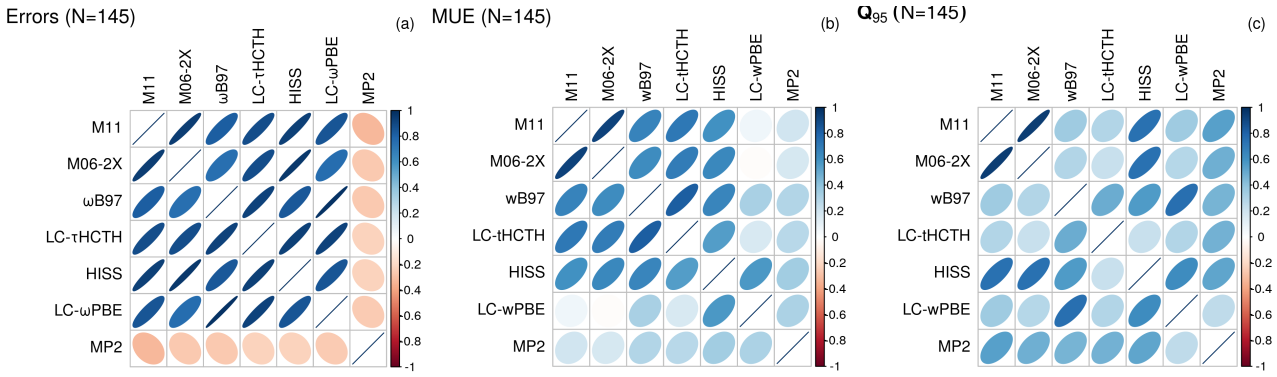

Figure 23: Case WU2015 - rank correlation matrix: (a) errors; (b) MUE; (c) Q_{95}.

discrimination between methods, and MP2 presents strong improvement probabilities over all the other methods.

Ranking. The ranking matrices are plotted in Fig. 26. The ranking probability matrices for the MUE confirm the problem seen above for the three 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

Methods	MUE %	P_{inv}	Q_{95} %	P_{inv}	MSIP	SIP	MG %	ML %
M11	**3.1(3)**	0.36	**10(1)**	-	0.58(4)	0.47(4)	-1.4(1)	1.16(10)
M06-2X	**3.2(3)**	0.08	**10(2)**	0.48	0.57(4)	0.53(4)	-1.2(1)	1.0(1)
ωB97	3.3(3)	0.00	**11(2)**	0.23	0.53(4)	0.59(4)	-0.94(7)	0.72(7)
LC-τHCTH	**3.0(3)**	-	**10(2)**	0.32	**0.59(4)**	-	-	-
HISS	3.8(3)	0.00	**10(2)**	0.35	0.34(4)	0.72(4)	-1.62(10)	1.5(1)
LC-ωPBE	3.9(3)	0.00	**11(1)**	0.26	0.31(4)	0.78(3)	-1.39(8)	1.2(1)
MP2	**3.2(3)**	0.21	**11(2)**	0.35	0.56(4)	0.45(4)	-1.3(3)	0.8(1)

Table 8: Same as Table 2 for the case THA2015.

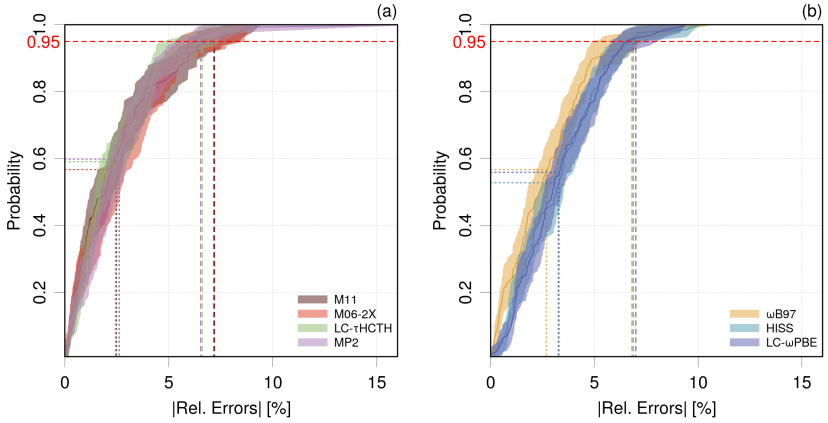

Figure 24: Thakkar2015: ECDFs of absolute relative errors: (a) methods with indiscernible MUE; (b) other methods.

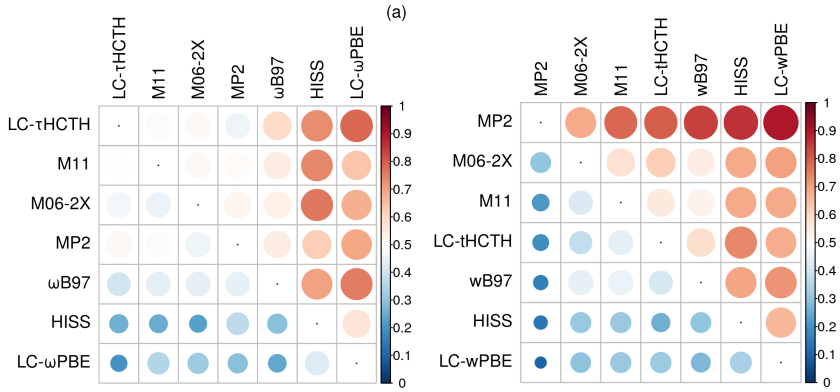

Figure 25: SIP matrix: (a) case THA2015 ($N = 127$); (b) case WU2015.

have been estimated after the removal of 8 outliers defined above. This has a negligible impact on the MUE ranking, 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. 26-bottom) show much less dispersion, underlining the deleterious role of

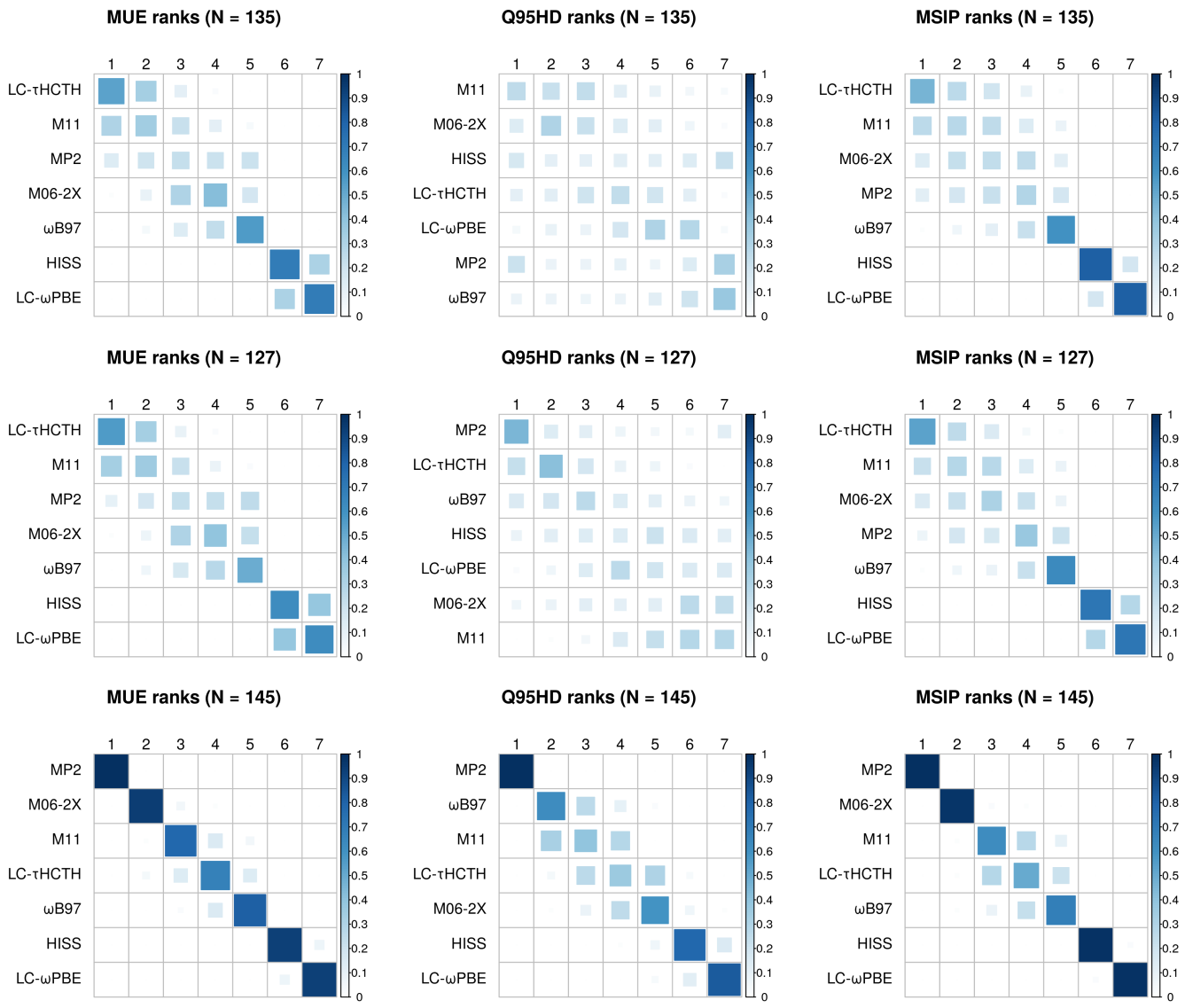

Figure 26: Ranking probability matrices: (top) case THA2015 full dataset ($N = 135$); (middle) case THA2015 dataset pruned from 8 outliers ($N = 127$); (bottom) case WU2015 ($N = 145$).

experimental errors. Note that there remains a notable uncertainty to rank ωB97, M11, M06-2X and LC-τHCTH using Q_{95}.

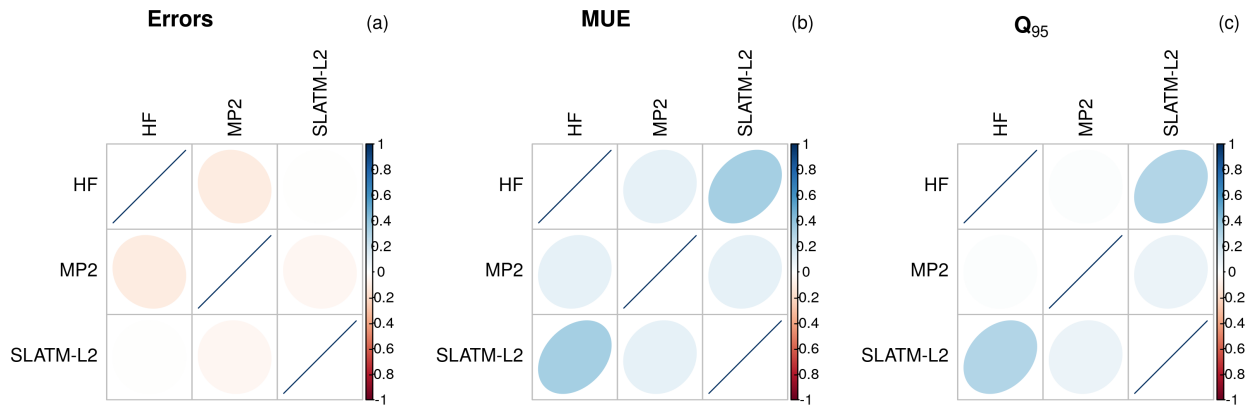

Figure 27: Case ZAS2019: rank correlation matrices.

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)**	-	-	-

Table 9: Same as Table 2 for case ZAS2019.

3.8 ZAS2019

The data have been provided by the authors of Ref. [42], to which the data are linked. The effective atomization energies (E^*) for the QM7b dataset [49], for 7211 molecules up to 7 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 ML 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 [42].

Correlations. The error sets are essentially uncorrelated (Fig. 27), whereas small positive correlations can be noted for the MUE and Q_{95}. In this problem, it would therefore be possible to ignore correlations when computing P_{inv}.

Statistics. The values are reported in Table 9. There is a contrast between the MUE and Q_{95}. SLATM-L2 and MP2 have close MUE values, with an above -threshold p-value ($p_g = 2P_{inv} = 0.06$), and a slight advantage for SLATM-L2. However, MP2 has a significantly smaller Q_{95}. As seen on the absolute errors ECDFs (Fig.28(a)), SLATM-L2 has a pronounced tail of large errors.

SIP analysis. The SIP matrix (Fig. 28(b)) shows that SLATM-L2 presents a notable improvement probability ($\sim$0.75) over HF and a more moderate one aver MP2 ($\sim$ 0.61). Even if SLATM-L2 has

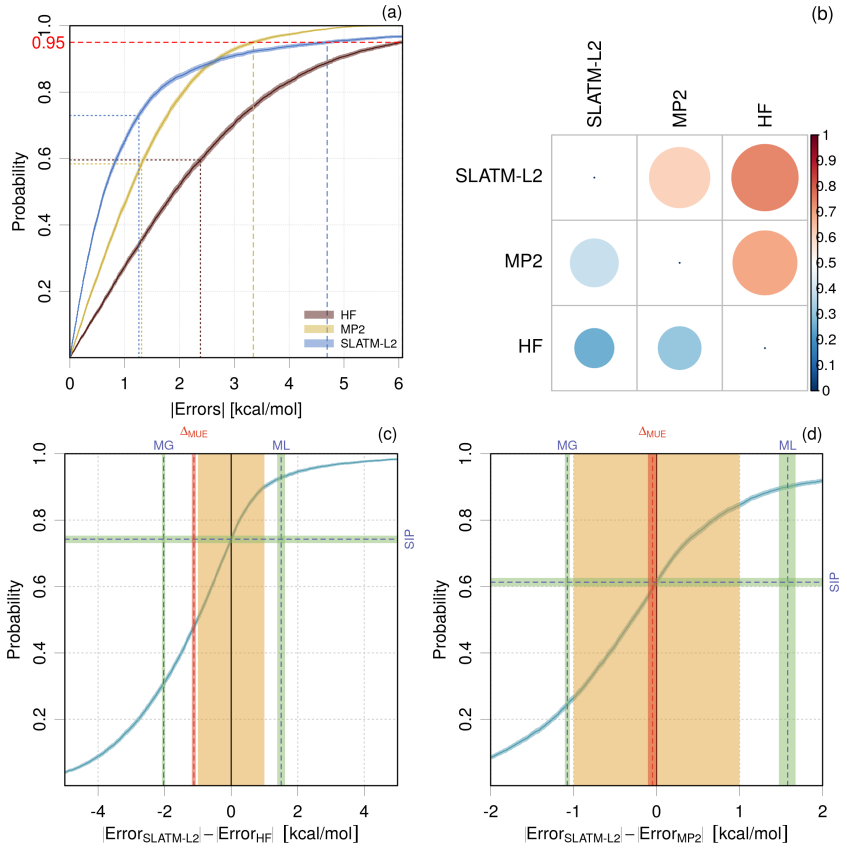

Figure 28: Case ZAS2019: ECDF of the absolute errors.

significantly better statistics than HF, 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, one has thus 61 % chance to get better results, with a mean gain $MG \simeq -1.1$, and 39 % chance to deteriorate the MP2 values with a mean loss $ML \simeq 1.6$.

4 Discussion

4.1 Extracting data from articles and supplementary material

The raw data of benchmark studies are strong assets for the community, and their accessibility and reusability are very important for intercomparison studies or the development of alternative statistical analyses, as performed in this study. We want to emphasize that the datasets used in the previous sections were not chosen to point out problems in the original studies. Besides other choice criteria such as dataset size and presence of reference data uncertainty, they were, above all other considerations, *available*, and their authors have to be praised for that.

We found many benchmarking studies with practically inaccessible data, *i.e.*, failing the 'A' (Accessible) and/or the 'R' (Reusable) of the FAIR principle of Open Data [50]. 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 SI 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), bibliographical references attached to the data (generally processed as spurious decimals)... Most of these features preclude automated data extraction and require error-prone human intervention.

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 to put more than one information in each cell. "Think Open, think FAIR !"

4.2 The correlation matrix as a sanity check

When we started this study, the correlation matrices were mainly intended to illustrate the importance to consider correlations when comparing statistics. When cumulating the case studies, it appeared that they contain a lot of information relevant to the quality of the benchmark dataset. To our knowledge, this subject has not previously been discussed, and benchmarking studies do not report such correlation matrices.

Considering that model errors in computational chemistry are mostly systematic, one expects that error patterns over a dataset are characteristic of the methods. This seems to be a basic requirement for sound benchmarking studies. Considering the errors correlation matrix, the guiding line is thus that closely related methods should produce similar error patterns and have strongly

correlated error sets, the correlation level decreasing with a "distance" between methods. This is clearly illustrated in the cased BOR2019, where the correlation matrix clusters nicely into relevant DFA groups. There seems also to be a clean decorrelation between MP2 or MP2-based methods and DFAs (NAR2019, WU2015). Similarly, one observes no correlation between HF, MP2 and a machine-learning method calibrated on CCSD(T) in case ZAS2019.

A contrario, when the methods set is diverse, an uniform strongly positive correlation matrix should raise an alert. We have seen in cases DAS2019 and THA2019 that outliers and/or large reference data errors could dominate the correlation matrix and influence the benchmark statistics. If the ranking study is to reflect the methods performances, the curation and possible pruning of the dataset has to be performed. Otherwise, more complex statistical models have to be used to alleviate the impact of those points (see Appendix A and Refs [51, 19, 9]).

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 that the unscaled set, whereas their correlation coefficient is 1 because of the linear transformation.

4.3 Impact of error sets correlation on ranking

The correlation between error sets is partially or totally transferred to statistics. Except for linear transformations of the errors, where the transfer is total, including the sign, one has to use Monte Carlo methods to estimate this transfer. In many cases, such as for normal, student's or g-and-h error distributions, one observes that the correlation intensity mainly decreases when passing from errors to MUE to Q_{95}. The case studies above show however that there are exceptions to this basic trend. We cannot presently rationalize the observed exceptions, but the main conclusion is that in most cases, one should not ignore correlations when comparing statistics.

However, unlike as shown above for the error correlations, the visualization of correlations between statistics might be of secondary interest. In fact, the paired samples bootstrap algorithms used in this study enable to account directly for these correlations without having to estimate them.

In a vast majority of the case 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.

4.4 Impact of dataset size

The examples above have shown that data set size impact considerably the ability to rank methods or to assert the impact of an improved method. Data set 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 the comparison of Q_{95} values might have an unacceptable level of type I errors if $N < 60$. Moreover, for small datasets (a few tens of points), the first digit of the statistics is often affected by the uncertainty. In such cases, ranking with the second or third digit of a statistics has no sense, unless correlations are taken into account.

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. We can only encourage benchmark authors to provide adequate uncertainty estimations and ranking probability matrices, which can be obtained with a negligible overcharge in computer time.

4.5 Systematic improvement analysis

We introduced a new criterion, the systematic improvement probability, which has the major advantage to be independent of the usual descriptive statistics. It simply counts the signs of the absolute energy differences. It is a useful complement to the MUE, as it enables to analyze a MUE difference. All the case studies show that a decrease of MUE results from the balance between gains and losses. At the exception of one methods pair in CAL2019, we did not find a 'best method' which improves the results of lower rank methods for the full benchmark dataset. A point we wish to make here is that *even physics-based improvements in DFAs do not lead to systematic improvements for all systems*. We have seen for instance that for band gaps, mBJ degrades LDA predictions for 16 % of the systems (BOR2019). In fact, there is often a non-negligible percentage of systems for which the 'bad' method is better than the 'good' one, all across Jacob's ladder. As long as the performances of computational chemistry methods rely on error cancellations, physics-based improvements of DFAs can be seen as a kind of statistical correction.

5 Conclusion

In this article, we proposed several tools to test the robustness of rankings or comparisons of methods based on error statistics for non-exhaustive, limited sized data samples. In order to avoid normality hypotheses on the errors distributions, bootstrap-based methods were adopted, as suggested by Proppe and Reiher [9] for the estimation of prediction uncertainty of DFT methods. Our target statistics were the MUE and Q_{95}, but these tools are straightforwardly applicable to other statistics.

Before any ranking, we have seen that the correlation matrix of error sets can be useful to appreciate the quality of a benchmark dataset. Then, 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 the number of compared methods can be thoroughly tested.

When considering pairs of methods, we generalized our previous definition of the inversion

probability P_{inv} to account for correlations between statistics. We also introduced the systematic improvement probability (SIP) which is independent of other descriptive statistics. We have seen that the use of MUE for ranking of methods hides a complex interplay between the genuine method improvement and the error cancellations inherent to most computational chemistry methods. In particular, we have shown how a difference in MUE is a balance between gains and losses in absolute errors. Estimation of the systematic improvement probability (SIP) and the mean gain (MG) and loss (ML) statistics can help understand this balance, and to assess the risks of switching between two methods. None of the showcased examples provided a method which provides a full systematic improvement over its concurrents, even when comparing an elaborate composite method such as G4MP2 to simple DFAs. The pedagogical virtue of the SIP is to clearly show that computational chemistry is a science of compromises.

We considered here only homogeneous datasets. Many modern benchmarks are based on composite datasets, involving weighting schemes to incorporate data with different units [52]. The applicability of the SIP to such datasets is direct, but the mean gain and mean loss statistics should become multivariate. For the estimation of P_{inv} and ranking probability matrices for composite statistics (e.g. WTMAD [52]), adaptation of the paired-sample bootstrap is straightforward, although care should be taken to avoid imbalance between the various components of the dataset.

Finally, 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 linear, will provide much better generalizability of the summary statistics derived from error sets. This is a necessary step if one wishes to estimate the prediction uncertainty of any method [19,9], notably when dealing with non-uniform reference data uncertainties.

Supplementary Information

All datasets + code to reproduce results of paper + online app

References

[1] R. A. Mata and M. A. Suhm. Benchmarking quantum chemical methods: Are we heading in the right direction? *Angew. Chem. Int. Ed.*, 56(37):11011–11018, 2017. doi:10.1002/anie.201611308.

[2] B. Civalleri, D. Presti, R. Dovesi, and A. Savin. On choosing the best density functional approximation. In *Chemical Modelling: Applications and Theory*, volume 9, pages 168–185. Royal Soc. Chem., 2012. doi:10.1039/9781849734790-00168.

[3] 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. doi:10.1063/1.5016248.

[4] A. J. Thakkar and T. Wu. How well do static electronic dipole polarizabilities from gas-phase experiments compare with density functional and MP2 computations? *The Journal of chemical physics*, 143(14):144302, 2015. doi:10.1063/1.4932594.

[5] T. Wu, Y. N. Kalugina, and A. J. Thakkar. Choosing a density functional for static molecular polarizabilities. *Chemical Physics Letters*, 635:257–261, 2015. doi:10.1016/j.cplett.2015.07.003.

[6] B. Ruscic. Uncertainty quantification in thermochemistry, benchmarking electronic structure computations, and active thermochemical tables. *Int. J. Quantum Chem.*, 114:1097–1101, 2014. doi:10.1002/qua.24605.

[7] T. Gould. 'diet gmtkn55' offers accelerated benchmarking through a representative subset approach. *Phys. Chem. Chem. Phys.*, 20:27735–27739, 2018. doi:10.1039/C8CP05554H.

[8] P. Morgante and R. Peverati. Statistically representative databases for density functional theory via data science. *Phys. Chem. Chem. Phys.*, 21:19092–19103, 2019. doi:10.1039/C9CP03211H.

[9] J. Proppe and M. Reiher. Reliable estimation of prediction uncertainty for physicochemical property models. *J. Chem. Theory Comput.*, 13:3297–3317, 2017. doi:10.1021/acs.jctc.7b00235.

[10] A. Nicholls. Confidence limits, error bars and method comparison in molecular modeling. Part 2: comparing methods. *J. Comput.-Aided Mol. Des.*, 30:103–126, 2016. doi:10.1007/s10822-016-9904-5.

[11] BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, and OIML. Evaluation of measurement data - Guide to the expression of uncertainty in measurement (GUM). Technical Report 100:2008, Joint Committee for Guides in Metrology, JCGM, 2008. URL: http://www.bipm.org/utils/common/documents/jcgm/JCGM_100_2008_F.pdf.

[12] P. P. Janes and A. P. Rendell. Placing rigorous bounds on numerical errors in Hartree–Fock energy computations. *J. Chem. Theory Comput.*, 7:1631–1639, 2011. doi:10.1021/ct200026t.

[13] Cancès, Eric and Dusson, Geneviève. Discretization error cancellation in electronic structure calculation: toward a quantitative study. *ESAIM: M2AN*, 51:1617–1636, 2017. doi:10.1051/m2an/2017035.

[14] F. Cailliez and P. Pernot. Statistical approaches to forcefield calibration and prediction uncertainty of molecular simulations. *J. Chem. Phys.*, 134:054124, 2011. doi:10.1063/1.3545069.

[15] J. J. Mortensen, K. Kaasbjerg, S. L. Frederiksen, J. K. Nørskov, J. P. Sethna, and K. W. Jacobsen. Bayesian error estimation in density-functional theory. *Phys. Rev. Lett.*, 95:216401, Nov 2005. doi:10.1103/PhysRevLett.95.216401.

[16] P. Pernot. The parameter uncertainty inflation fallacy. *J. Chem. Phys.*, 147(10):104102, September 2017. doi:10.1063/1.4994654.

[17] D. Bakowies. Estimating systematic error and uncertainty in ab initio thermochemistry. I. Atomization energies of hydrocarbons in the ATOMIC(hc) protocol. *J. Chem. Theory Comput.*, 15:5230–5251, 2019. doi:10.1021/acs.jctc.9b00343.

[18] D. Bakowies. Estimating systematic error and uncertainty in ab initio thermochemistry: II. ATOMIC(hc) enthalpies of formation for a large set of hydrocarbons. *J. Chem. Theory Comput.*, 2020. doi:10.1021/acs.jctc.9b00974.

[19] 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. doi:10.1021/jp509980w.

[20] S. De Waele, K. Lejaeghere, M. Sluydts, and S. Cottenier. Error estimates for density-functional theory predictions of surface energy and work function. *Phys. Rev. B*, 94:235418, 2016. doi:10.1103/PhysRevB.94.235418.

[21] B. Efron. Bootstrap Methods: Another Look at the Jackknife. *Ann. Stat.*, 7(1):1–26, January 1979. doi:10.1214/aos/1176344552.

[22] B. Efron and R. Tibshirani. Statistical data analysis in the computer age. *Science*, 253:390–395, 1991. doi:10.1126/science.253.5018.390.

[23] T. C. Hesterberg. What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum. *The American Statistician*, 69:371–386, 2015. doi:10.1080/00031305.2015.1089789.

[24] I. BIPM, I. IFCC, I. ISO, and O. IUPAP. Evaluation of measurement data – supplement 2 to the 'guide to the expression of uncertainty in measurement' – extension to any number of output quantities. *JCGM*, 102, 2011.

[25] R. N. Kacker, R. Kessel, and K.-D. Sommer. Assessing differences between results determined according to the guide to the expression of uncertainty in measurement. *J. Res. Nat. Inst. Stand. Technol.*, 115(6):453, 2010. doi:10.6028/jres.115.031.

[26] G. W. Snedecor and W. G. Cochran. *Statistical Methods, Eighth edition*. Iowa State University Press, 1989.

[27] A. Nicholls. Confidence limits, error bars and method comparison in molecular modeling. Part 1: The calculation of confidence intervals. *J. Comput.-Aided Mol. Des.*, 28:887–918, 2014. doi:10.1007/s10822-014-9753-z.

[28] R. R. Wilcox and D. M. Erceg-Hurn. Comparing two dependent groups via quantiles. *J. App. Stat.*, 39:2655–2664, 2012. doi:10.1080/02664763.2012.724665.

[29] R. Y. Liu and K. Singh. Notions of limiting p-values based on data depth and bootstrap. *Journal of the American Statistical Association*, 92:266–277, 1997. URL: http://www.jstor.org/stable/2291471, doi:10.2307/2291471.

[30] 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)]. *The Journal of Chemical Physics*, 150:219906, 2019. doi:10.1063/1.5110025.

[31] P. Hall, H. Miller, et al. Using the bootstrap to quantify the authority of an empirical ranking. *The Annals of Statistics*, 37:3929–3959, 2009. doi:10.1214/09-AOS699.

[32] 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/.

[33] A. Canty and B. Ripley. *boot: Bootstrap Functions (Originally by Angelo Canty for S)*, 2019. R package version 1.3-22. URL: https://CRAN.R-project.org/package=boot.

[34] F. E. Harrell and C. Davis. A new distribution-free quantile estimator. *Biometrika*, 69:635–640, 1982. doi:10.2307/2335999.

[35] R. R. Wilcox and G. A. Rousselet. A guide to robust statistical methods in neuroscience. *Curr. Prot. Neuroscience*, 82:8.42.1–8.42.30, 2018. doi:10.1002/cpns.41.

[36] P. Mair and R. Wilcox. *WRS2: A Collection of Robust Statistical Methods*, 2019. R package version 1.0-0. URL: https://CRAN.R-project.org/package=WRS2.

[37] 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. doi:10.1021/acs.jctc.9b00322.

[38] B. Narayanan, P. C. Redfern, R. S. Assary, and L. A. Curtiss. Accurate quantum chemical energies for 133000 organic molecules. *Chem. Sci.*, 2019. doi:10.1039/c9sc02834j.

[39] 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. doi:10.1063/1.5090222.

[40] F. Jensen. Method calibration or data fitting? *J. Chem. Theory Comput.*, 14(9):4651–4661, 2018. doi:10.1021/acs.jctc.8b00477.

[41] 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. doi:10.1021/acs.jctc.9b00545.

[42] 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(3):1546–1559, 2019. doi:10.1021/acs.jctc.8b00832.

[43] J. P. Perdew, J. Sun, A. J. Garza, and G. E. Scuseria. Intensive atomization energy: Rethinking a metric for electronic structure theory methods. *Z. Phys. Chem.*, 230:737–742, 2016. doi:10.1515/zpch-2015-0713.

[44] 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(17):7374–7383, 2000. doi:10.1063/1.481336.

[45] J. Bland and D. Altman. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, i:307–310, 1986. URL: http://www-users.york.ac.uk/~mb55/meas/ba.htm.

[46] 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(5):2596–2608, 2018. doi:10.1021/acs.jctc.7b01183.

[47] J. Rezác, K. E. Riley, and P. Hobza. S66: A well-balanced database of benchmark interaction energies relevant to biomolecular structures. *Journal of chemical theory and computation*, 7:2427–2438, 2011. doi:10.1021/ct2002946.

[48] 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. doi:10.1021/ct5000692.

[49] 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. doi:10.1088/1367-2630/15/9/095003.

[50] M. D. Wilkinson et al. The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3:160018, 2016. doi:10.1038/sdata.2016.18.

[51] 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, jan 2014. doi:10.1103/physrevb.89.014304.

[52] 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. doi:10.1039/C7CP04913G.

[53] R. Kacker and A. Jones. On use of bayesian statistics to make the Guide to the Expression of Uncertainty in Measurement consistent. *Metrologia*, 40:235–248, 2003. doi:10.1088/0026-1394/40/5/305.

[54] M. Evans, N. Hastings, and B. Peacock. *Statistical Distributions*. Wiley-Interscience, 3rd edition, 2000.

[55] P. R. Bevington and D. K. Robinson. *Data Reduction and Error Analysis for the Physical Sciences*. McGraw-Hill, New York, 1992.

[56] R. N. Kacker. Combining information from interlaboratory evaluations using a random effects model. *Metrologia*, 41:132–136, 2004. doi:10.1088/0026-1394/41/3/004.

[57] A. L. Rukhin. Weighted means statistics in interlaboratory studies. *Metrologia*, 46:323, 2009. doi:10.1088/0026-1394/46/3/021.

[58] C. Rivier, M. Désenfant, M. Crozet, C. Rigaux, D. Roudil, B. Tufféry, and A. Ruas. Use of an excess variance approach for the certification of reference materials by interlaboratory comparison. *Accreditation and Quality Assurance*, 19:269–274, 2014. doi:10.1007/s00769-014-1066-3.

[59] K. Lejaeghere, V. Van Speybroeck, G. Van Oost, and S. Cottenier. Error estimates for solid-state density-functional theory predictions: An overview by means of the ground-state elemental crystals. *Crit. Rev. Solid State Mater. Sci.*, 39:1–24, 2014. doi:10.1080/10408436.2013.772503.

[60] D. C. Hoaglin. *Exploring data tables, trends, and shapes*, chapter Summarizing shape numerically: The g-and-h distributions, pages 461–513. Wiley, New York, 1985.

[61] J. V. Bradley. Robustness? *Br. J. Math. Stat. Psychol.*, 31(2):144–152, 1978. doi:10.1111/j.2044-8317.1978.tb00581.x.

[62] R. R. Wilcox. *WRS: A Package of R.R. Wilcox' Robust Statistics Functions*, 2019. R package version 0.36.

[63] R. J. Hyndman and Y. Fan. Sample quantiles in statistical packages. *The American Statistician*, 50:361–365, 1996. doi:10.2307/2684934.

Appendices

A Estimation of the mean value and its uncertainty

Let us consider the mean (signed) value of the errors (MSE). In absence of uncertainty, it is defined as

$$\overline{e} = \frac{1}{N}\sum_{i=1}^{N} e_i \tag{32}$$

and its uncertainty (standard error) is estimated as

$$u(\overline{e}) = \sqrt{\frac{s_e^2}{N}} \tag{33}$$

where s_e^2 is a sample-based estimator of the population variance

$$s_e^2 = \frac{1}{N-1}\sum_{i=1}^{N}(e_i - \overline{e})^2 \tag{34}$$

Eq. 33 gives the well-known dependence of the MSE uncertainty with the dataset size for independent and identically distributed (*i.i.d.*) errors, assuming a finite variance, which might exclude error sets with heavy-tailed distributions, *e.g.*, Cauchy.

Note that $u(\overline{e})$ in Eq. 33 does not account for the uncertainty on s_e. Taking this factor into account leads to a larger uncertainty, which can be estimated as [53]

$$u(\overline{e}) = \sqrt{\frac{N-1}{N-3}\frac{s_e^2}{N}} \tag{35}$$

This formula is based on the properties of the Student-t distribution [54]. The impact of the correction factor is notable only for very small datasets (smaller than 3 % for $N \geq 30$), and we will consider the standard formula .

If uncertainty on errors $u(e_i)$ is negligible, s_e is an estimation of the standard deviation of the errors distribution σ, which represents the dispersion of model errors. If the reference data are uncertain, s_e quantifies a dispersion due to both model errors and reference data uncertainty. In consequence, it overestimates the dispersion of model errors, and specific models have to be

designed if one wishes to estimate this specific contribution [19, 9]. This points to the necessity of using accurate reference data if the benchmark based on standard statistics is to reflect the properties of the studied methods.

To be more specific, in the presence of uncertainty on errors, the weighted mean is the maximum likelihood estimator of the distribution mean under normality assumptions [55]

$$\overline{e} = \sum_{i=1}^{N} w_i e_i \tag{36}$$

$$w_i = \frac{u(e_i)^{-2}}{\sum_{j=1}^{N} u(e_j)^{-2}} \tag{37}$$

giving less weight to the more uncertain data. Direct application of the combination of variances to this expression leads to [55]

$$u(\overline{e})^2 = \frac{1}{\sum_{j=1}^{N} u(e_j)^{-2}} \tag{38}$$

Note that in the case of identical uncertainty for all data, one recovers the expression for the unweighted case (Eq. 33).

The validity of this estimations has to be tested by computing the weighted chi-squared

$$\chi^2_w = \sum_i \frac{(e_i - \overline{e})^2}{u(e_i)^2} \tag{39}$$

If the errors on the reference data are assumed to be normally distributed, χ^2_w has a chi-squared distribution with $N-1$ degrees of freedom (χ^2_{N-1}). χ^2_w should be close to the mean of this distribution, $N-1$, and lie within its 95 % high probability interval. If χ^2_w is too small, the $u(e_i)$ are over-estimated and should be reconsidered, or the benchmarked method is over-fitting the data, which is unlikely, unless the method is parametric and has been calibrated on this same dataset. If χ^2_w is too large there is an excess of variance in the E_M error set [56–58]. In the typical benchmarking of computational chemistry methods, this is generally the case because of the extraneous dispersion due to model errors. To ensure the statistical validity of the weighted mean and its uncertainty, one has therefore to define a more complex error model, considering explicitly the two sources of dispersion, and to redefine the weights, accounting for the excess of variance and possible biases in the error sets [51, 59, 19, 20, 9].

Considering σ as the dispersion of model errors, one can stipulate that the dispersion of the errors is the combined effect of model error and reference data uncertainty and redefine the weights as [57]

$$w_i = \frac{(\sigma^2 + u(e_i)^2)^{-1}}{\sum_{j=1}^{N} (\sigma^2 + u(e_j)^2)^{-1}} \tag{40}$$

with which

$$u(\overline{e})^2 = \frac{1}{\sum_{j=1}^{N} \left(\sigma^2 + u(e_j)^2\right)^{-1}} \tag{41}$$

converges properly to the standard limit when the reference data errors become negligible before the model errors. The model error variance σ^2 can be estimated by decomposing the total variance of the errors into the variance of model errors plus the mean variance of the data (known as Cochran's ANOVA estimate [56, 58])

$$\mathrm{var}(e) = \sigma^2 + \frac{1}{N}\sum_{j=1}^{N} u(e_j)^2 \tag{42}$$

This variance analysis ensures that χ^2_w is correct. Note that other reweighting schemes exist [56, 58], but Cochran's is the simplest. Besides, all reweighting methods are iterative: σ depends on $\overline{e}$, which itself depends on σ.

If the dispersion of reference data uncertainties is small, *i.e.*, smaller than the model errors contribution, one can reasonably consider that the weights are identical and that the unweighted mean can be used. Formally, its uncertainty (Eq. 41) depends on σ, which can be directly estimated through Eq. 42, but by construction, one will recover results given by Eq. 33.

One will therefore consider that, unless a large dispersion of data reference uncertainty is observed, these uncertainties can be ignored in the estimation of the mean and its standard error. Otherwise, one should use the weighted mean with the standard uncertainty estimate.[6] This advanced modeling of uncertainty sources is crucial if one wishes a reliable estimate of the MSE, and of the various uncertainty contributions [19]. In standard benchmarking, the aim is mostly to compare methods, knowing that the reference datasets are incomplete. If reference data uncertainty plays a significant role – that would be the case if data with very different uncertainty levels were aggregated in the dataset – one might assume that its impact will be the same for all methods to be compared. The values of the dispersion statistics will be consistently overestimated for all methods. As long as one is not interested in the accurate estimation of the underlying properties of the error distributions, such as the model prediction uncertainty [19, 9], it is simpler to rely on unweighted schemes and properly curated datasets.

B Numerical study of the covariance of nonlinear statistics

To illustrate the transfer of covariance from the errors sets to their statistics, one generates random samples E_1 and E_2 from a bivariate distribution with prescribed correlation coefficient ρ. For sample point, the statistics values S_1 and S_2 are calculated, and $\mathrm{cov}(S_1, S_2)$ is estimated from the statistics samples. The error sets correlation coefficient ρ has been varied between -1 and 1, and the corresponding correlation coefficients have been estimated for the MSE, MUE and Q_{95} statistics.

[6] Note that the dispersion of model errors σ is related to the model prediction uncertainty and is a score of interest for the ranking of models [19, 3].

The dataset size is $N = 100$ and the correlation coefficient statistics are based on 10^3 Monte Carlo samples.

The results for four cases of the g-and-h distribution of error sets are reported in Fig. 29(a-d). The g-and-h distribution's shape is defined by two parameters and contains the normal distribution as a special case ($g = h = 0$). Besides the normal, three typical cases are [28]: heavy-tailed symmetric ($g = 0$; $h = 0.2$), light-tailed asymmetric ($g = 0.2$; $h = 0$), and heavy-tailed asymmetric ($g = h = 0.2$). In this example, both error sets E_1 and E_2 have the same distribution with unit variance, only their correlation varies.

These simulations confirms the identity for the MSE, independently of the underlying distribution. The correlation coefficients for the other statistics are positive (within numerical uncertainty) and systematically smaller than $|\rho|$. They are symmetrical with respect to $\rho = 0$ for symmetrical distributions. The values for the MUE are consistently larger than or or equal to the values for Q_{95}. In all cases, the correlation coefficient for the MUE is very close to ρ^2. For negative values of ρ, the correlation coefficient of Q_{95} is sensitive to the asymmetry or the errors distribution.

The same procedure has been applied to shifted means ($\overline{e}_1 = -0.2$, $\overline{e}_2 = -0.5$) for normal and Student's distribution with 5 degrees of freedom (Fig. 29(e,f)). For the normal distribution the symmetry observed above is broken, as well as the pure quadratic trend for the MUE. For the Student's distribution, the correlations lie above a positive threshold.

Simulation of correlated error samples enabled us to illustrate properties of correlation transfer to statistics: identical correlation for the MSE, and smaller positive correlations for the MUE and Q_{95}. As we covered only a limited set of scenarii, these features cannot be considered as universal. Indeed, the case studies in Section 3 reveal some exceptions.

C Probabilities of Type I errors for the comparison of MUE and Q_{95} pairs

We applied the procedure followed by Wilcox and Erceg-Hurn [28] to estimate the probability of type I errors for the comparison of quantiles with their method M (Algo. 1).

A false positive (type I error) is obtained when a true null hypothesis is rejected by the test. In the present context, one draws two samples E_1 and E_2 of size N from the same distribution and compute p_g for the comparison of the values of a statistic S, s_1 and s_2, respectively. A value of $p_g < 0.05$ leads to a false rejection of the (null) hypothesis $s_1 = s_2$. The process is repeated M times ($M = 2000$), and the proportion of false rejections provides an estimation of the probability $\hat{\alpha}$ of type I errors.

For their tests, Wilcox and Erceg-Hurn use the g-and-h distribution [60] to generate the data samples. The g-and-h distribution's shape is defined by two parameters and contains the normal distribution as a special case ($g = h = 0$). Besides the normal, the cases considered by Wilcox

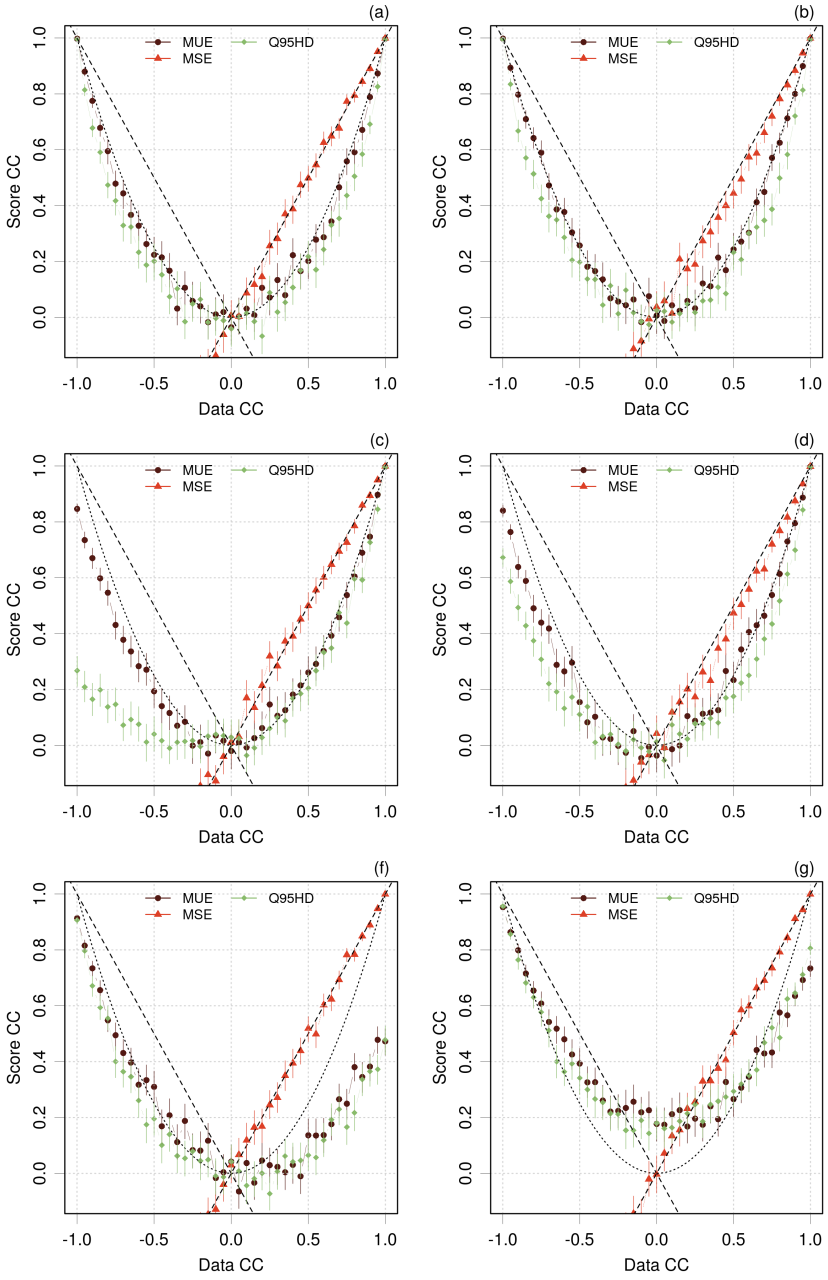

Figure 29: Correlation coefficients (CC) of several statistics/scores as a function of the CC of the samples used to estimate them. The error bars represent 95 % intervals for sampling errors. Four cases of the g-and-h distribution are considered for the error sets: (a) normal ($g = h = 0$); (b) heavy-tailed symmetric ($g = 0$; $h = 0.2$); (c) light-tailed asymmetric ($g = 0.2$; $h = 0$); (d) heavy-tailed asymmetric ($g = h = 0.2$). Additional cases with shifted distributions, μ= (-0.2,0.5) : (e) normal ; (f) Student's-t (ν= 5).

and Erceg-Hurn [28] were: heavy-tailed symmetric ($g = 0$; $h = 0.2$), light-tailed asymmetric ($g = 0.2$; $h = 0$), and heavy-tailed asymmetric ($g = h = 0.2$). Several levels of correlation between E_1 and E_2 were also considered ($\rho = 0$ or 0.7) in the original study. Through these tests, one wishes to determine the sample size N required to reach a probability of type I errors $\hat{\alpha}$ close to the

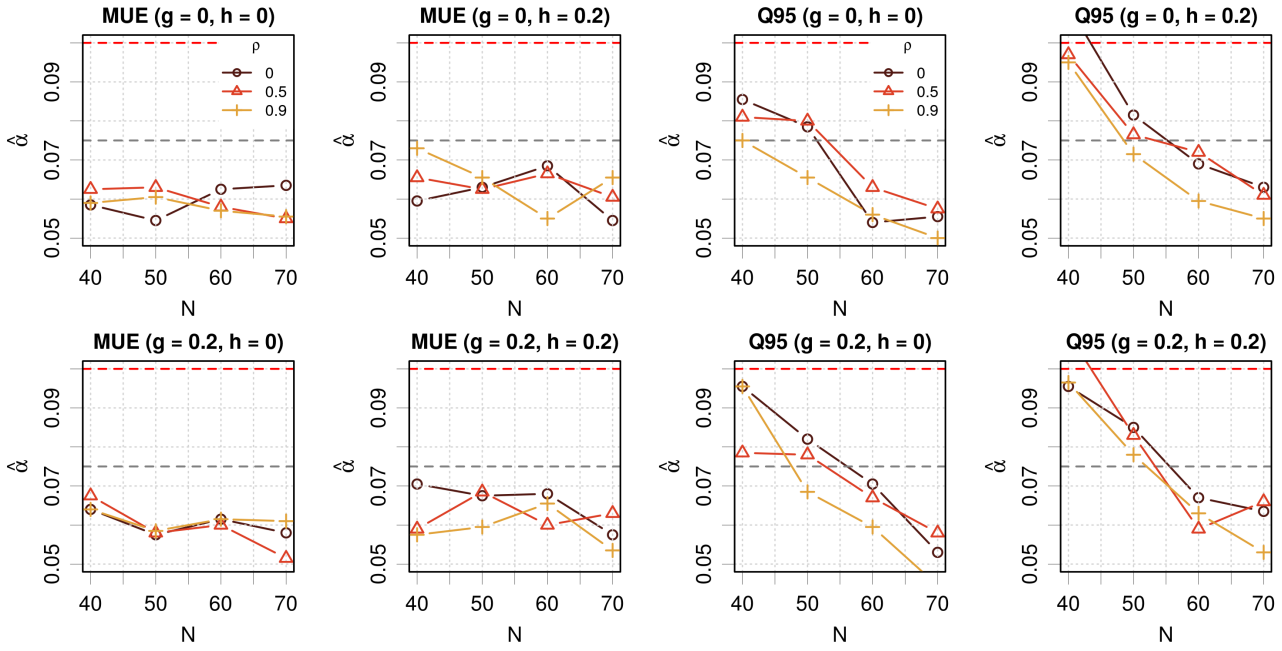

Figure 30: Probability of type I errors $\hat{\alpha}$ for the MUE (left) and Q_{95} (right), as a function of dataset size N. Each graph corresponds to a type of g-and-h distribution for the data samples (see text for details). The lines correspond to a value of the datasets correlation coefficient ρ.

statistical testing threshold. The recommendation in the original study is to stay below $\hat{\alpha} = 0.075$ for testing at the 0.05 level [61].

These test cases did not include our conditions of interest, and we performed new simulations, using functions provided in `R` packages `WRS` [62] and `WRS2` [36], after assessing the reproducibility of the original results. We kept the same generative distribution and scenarii for g and h parameters, and we extended the exploration for dataset size to $N = 40$, 50, 60, 70 and correlation coefficient $\rho = 0$, 0.5, 0.9, more representative of the conditions of interest in the present study. For compatibility with the original study, the number of replications was kept to $M = 2000$, and the number of bootstrap samples to $B = 1000$. The results are summarized in Fig. 30.

For the MUE, the safety region ($\hat{\alpha} \leq 0.075$) is reached in all cases, and all values are close to the nominal value (0.05). There is no remarkable trend with respect to the type of g-and-h distribution, nor the correlation coefficient. We have shown previously [3] that the MUE is typically located between the 0.5 and 0.75 quantiles, for which Wilcox and Erceg-Hurn [28] have conclude that the minimal request is $N \geq 30$, which we confirm.

For Q_{95}, one sees that for $N = 40$, the situation is more favorable for the normal distribution, but in all cases, the recommended limit is reached globally for $N = 60$. Strong correlation coefficients ($\rho = 0.9$) seem also to be almost systematically more favorable, and there is a slight deleterious effect below $N = 50$ for heavy-tailed distributions ($h = 0.2$). Nevertheless, even for $N = 40$, $\hat{\alpha}$ does not exceed notably the 10 % type I error probability.

Set	MSE	RMSD	MUE	Q_{95}
E_1	0	1.1	0.88	2.16
E_2	0.1	1.0	0.80	1.97

Table 10: Reference values for the univariate statistics of datasets E_1 and E_2 described by Eq. 43, for $\mu_1 = 0$, $\mu_2 = 0.1$, $\sigma_1 = 1.1$ and $\sigma_2 = 1.0$.

Remark. Establishing the power of the test $(1 - \beta)$, where β is the probability of type II errors (false negative, or the non-rejection of a false null hypothesis) is practically impossible unless one defines a specific alternative hypothesis. In the present case, there is a infinity of ways to realize the $s_1 \neq s_2$ alternative, so the power estimation is intractable.

D Numerical study of the Harrell and Davis algorithm

This example is intended to outline the advantages of Harrell and Davis (HD) algorithm for quantiles estimation, notably when associated with bootstrap sampling. One considers the values s_1 and s_2 of a statistic S on two datasets E_1 and E_2, which are drawn from a bivariate normal distribution

$$(E_1, E_2) \sim \mathcal{N}\left(\boldsymbol{\mu} = (\mu_1, \mu_2), \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right) \tag{43}$$

where the error samples have different means (μ_1, μ_2) and variances (σ_1^2, σ_2^2), and $\mathrm{cov}(E_1, E_2) = \rho\sigma_1\sigma_2$. The values of the parameters for the simulations and the corresponding statistics are given in Table 10. The values for the MUE and Q_{95} are obtained as described in Ref. [3]. Those values are fairly representative of the problems treated in the case studies.

D.1 Comparison of quantiles estimated with $\hat{Q}_7$

The numerical case is based on the distribution for E_2, as presented above. In a first test, E_2 sets of increasing sizes between 20 and 500 were generated by random sampling from the corresponding normal distribution, and Q_{95} was estimated by two algorithms: the HD algorithm and the $\hat{Q}_7$ method of Hyndman and Fan [63], which is the default algorithms in the `quantile()` function of `R` [32]. The procedure is repeated 10000 times, and the distributions of Q_{95} values are summarized by a set of five quantiles (0.05, 0.25, 0.5, 0.75, 0.95). The results are presented in Fig. 31. This simulation shows that the HD quantiles converge faster than the reference one, with less bias for small samples ($N < 100$).

In the second test, a unique E_2 sample of size $N = 500$ is generated, and subsets of increasing size are taken as initial data for a bootstrap procedure (10000 repeats). The bootstrap samples are analyzed as above and plotted in Fig. 31(b). The difference between both quantile algorithms is less striking, but the reference algorithm seems to produce quite asymmetric distributions, where

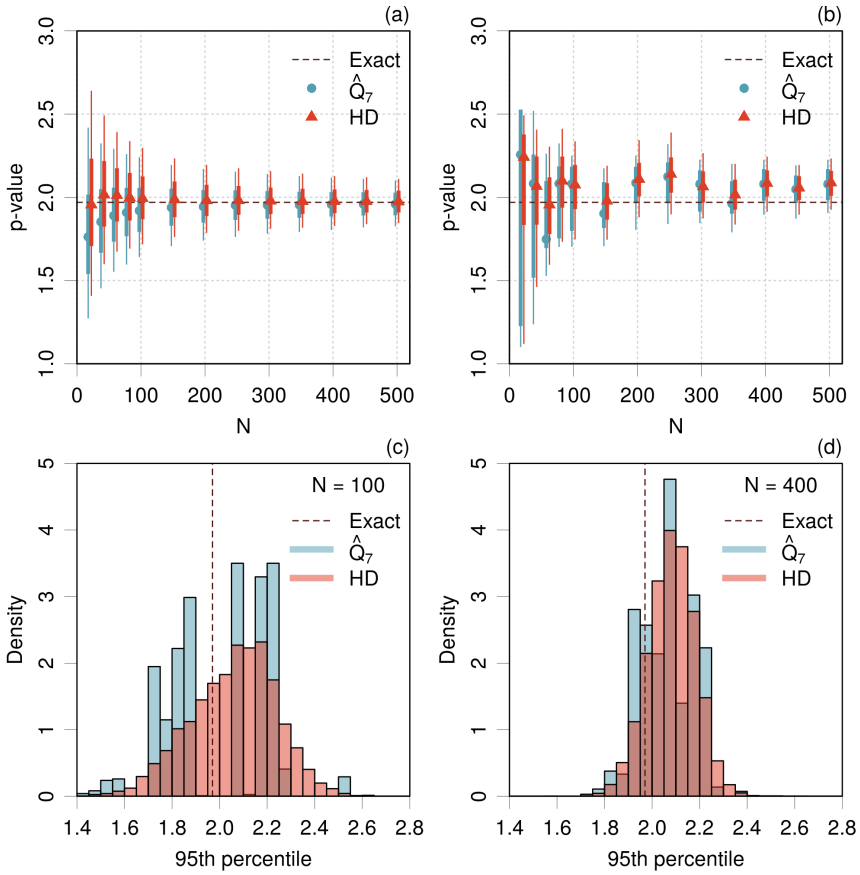

Figure 31: Comparison of Q_{95} estimation algorithms, $\hat{Q}_7$ and HD: (a) Monte Carlo sampling; (b) bootstrap sampling; (c) bootstrap sample histogram for $N = 100$; (d) idem for $N = 400$. The thicker bars in (a,b) represent 25-75 % probability intervals and the finer bars 5-95 % probability intervals.

the median is close to one of the quartiles. If one looks at the histograms of sampled values for $N = 100$ (Fig. 31(c)), one sees that the HD algorithms produces a much smoother distribution, where $\hat{Q}_7$ produces rugged histograms. The same is still visible, to a lesser extent, for $N = 400$ (Fig. 31(d)). This feature of the HD method explains its good performances for small samples, when used in conjunction with the bootstrap [28].

D.2 Estimation of p-values

The estimation of p-values is obtained by Monte Carlo sampling of E_1 and E_2 sets of size N varying between 20 and 500 ($\rho = 0.9$). One first checks that the generalized p-value p_g (Algo. 1) is identical to the analytical value of p_t for the comparison of mean values (Fig. 32(a)). Then, the interest of the Harrell-Davis algorithms for the estimation of p_g values for the comparison of quantiles is shown in Fig. 32(b): reaching the 0.05 threshold requires about 250 points for the HD method, whereas the reference quantile algorithms requires about 380 points. Besides, the HD curve is smoother than the reference one, due to the smoothness properties of the HD estimator shown above.

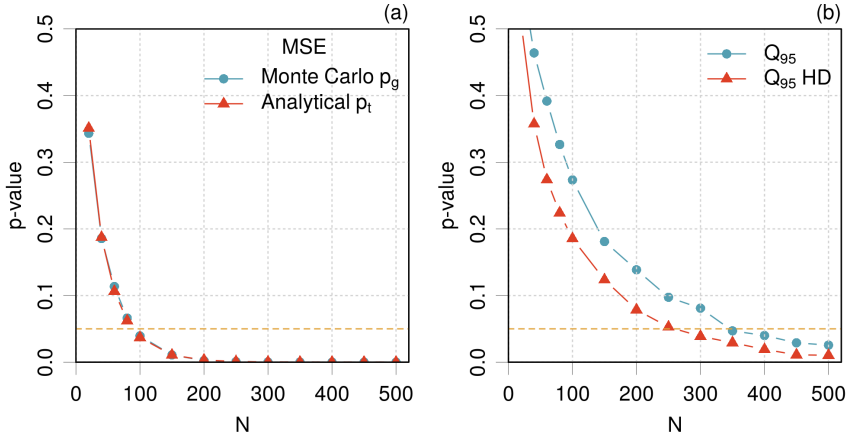

Figure 32: Validation of methodological choices for p-value estimation: (a) generalized p-value p_g for the comparison of means (MSE) compared to the analytical result p_t ; (b) impact of the quantile estimation algorithms on p_g for the comparison of Q_{95}.