Theoretical Chemistry Accounts (2021) 140:24
https://doi.org/10.1007/s00214-021-02725-0

REGULAR ARTICLE

Using the Gini coefficient to characterize the shape of computational chemistry error distributions

Pascal Pernot[1] · **Andreas Savin**[2]

Received: 17 December 2020 / Accepted: 29 January 2021

Abstract
The distribution of errors is a central object in the assessment and benchmarking of computational chemistry methods. The popular and often blind use of the mean unsigned error as a benchmarking statistic leads to ignore distributions features that impact the reliability of the tested methods. We explore how the Gini coefficient offers a global representation of the errors distribution, but, except for extreme values, does not enable an unambiguous diagnostic. We propose to relieve the ambiguity by applying the Gini coefficient to mode-centered error distributions. This version can usefully complement benchmarking statistics and alert on error sets with potentially problematic shapes.

Keywords Benchmarking · Statistics · Large errors · Errors distribution

1 Introduction

The reliability of a computational chemistry method is conditioned by the distribution of its prediction errors. Distributions with heavy tails elicit a risk of large prediction errors. As a benchmarking statistic, the popular mean unsigned error (MUE) bears no information on such a risk [1–4]. We have recently reported a case where two unbiased error distributions with identical values of the MUE present widely different risks of large errors because of heavy tails in one of them [4, 5]. It would therefore be very useful to complement the MUE with a statistic indicating or quantifying the risk of large errors.

We recently proposed alternative statistics such as Q_{95} [2], P_η, [2] and systematic improvement probability (SIP) [3]. In terms of risk, these statistics offer the following interpretations:

- There is a 5 % risk for absolute errors to exceed Q_{95}.
- There is a probability P_η that absolute errors are larger than a chosen threshold η. P_η provides a direct quantification of the risk of large errors, but η has to be defined *a priori* and might be user dependent, which complicates its reporting in benchmarking studies.
- For two methods M_1 and M_2, the SIP provides the system-wise probability that the absolute errors of M_1 are smaller than the absolute errors of M_2, informing on the risk incurred by switching between two methods. Interestingly, the SIP analysis provides a decomposition of the MUE difference between two methods in terms of gain and loss probabilities [3].

The Q_{95} and P_η partly answer the question, but they are point estimates on the cumulated density function of the absolute errors, and a statistic accounting for the whole distribution might be of interest. Besides, it is well established in econometrics that measures of dispersions such as the variance perform poorly at risk estimation and that higher moments of the distributions have to be considered [6]. This would lead

This work is dedicated to Ramon Carbó-Dorca for his 80th birthday. It reflects his interest for cross-disciplinary aspects of science, especially the role of mathematics in chemistry.

Published as part of the special collection of articles "Festschrift in honour of Prof. Ramon Carbó-Dorca".

✉ Pascal Pernot
pascal.pernot@universite-paris-saclay.fr

Andreas Savin
andreas.savin@lct.jussieu.fr

[1] Institut de Chimie Physique, UMR8000 CNRS, Université Paris-Saclay, 91405 Orsay, France

[2] Laboratoire de Chimie Théorique, CNRS and UPMC Université Paris 06, Sorbonne Universités, 75252 Paris, France

Published online: 15 February 2021

us to such measures as skewness and kurtosis, but none of these alone would be able to cover all the scenarios. The risk of large errors through heavy tails of the errors distribution might be associated with large skewness or large kurtosis or a combination of them.

The Lorenz curve [7] is widely used in econometrics to represent the distribution of wealth in human populations. Its summary statistics, notably the Gini coefficient (noted G) [8–10], are used to evaluate the degree of inequality within these populations. The Gini coefficient is also used, for instance, in ecology, to estimate the inequality of properties within plant populations [9, 11], in astronomy, to characterize the morphology of galaxies [12], or in information theory, to characterize the sparsity of datasets [13].

The Lorenz curve has many mathematical representations, the most interesting one, for us, being its formulation as an integral of the quantile function, a direct link with our study of probabilistic metrics [2–4]. More precisely, we explore here the interest of the Gini coefficient as a complement to the MUE in benchmarking studies.

We introduce the statistical tools and their implementation in Sect. 2, and apply them to a series of datasets to illustrate the interest and limitations of the Gini coefficient in Sect. 4. An adaptation of the Gini coefficient is proposed to relieve its main drawbacks when applied to error datasets.

2 Statistical methods

2.1 Definitions

Let us consider a distribution of errors e with probability density function (PDF) $f(e)$. The absolute errors $\varepsilon = |e|$ have a *folded* PDF $f_F(\varepsilon)$. To avoid ambiguity, statistics based on absolute errors are indexed by F.

2.1.1 CDF and quantile function

The cumulative distribution function (CDF) of the absolute errors is noted

$$C_F(z) = \int_0^z f_F(\varepsilon)\, d\varepsilon \tag{1}$$

from which the quantile function is the inverse

$$q_F(p) = C_F^{-1}(p) \tag{2}$$

2.1.2 Mean unsigned error

The mean unsigned error (MUE) is defined as

$$\mu_F = \int_0^\infty \varepsilon f_F(\varepsilon)\, d\varepsilon \tag{3}$$

Using the change of variable $\varepsilon = C_F^{-1}(p)$, $p = C_F(\varepsilon)$ and $dp = f_F(\varepsilon)\, d\varepsilon$, the MUE also can be shown to be the integral of the quantile function

$$\mu_F = \int_0^1 C_F^{-1}(p)\, dp \tag{4}$$

$$= \int_0^1 q_F(p)\, dp \tag{5}$$

2.1.3 The Lorenz curve

The Lorenz curve gives the percentage of cumulated absolute errors due to the $100 \times p$ % smallest values or, equivalently, the portion of the MUE due to the $100 \times p$ % smallest absolute errors:

$$L_F(p) = \frac{1}{\mu_F} \int_0^p q_F(t)\, dt \tag{6}$$

As shown in Fig. 1a, it is the ratio between the slanted shaded area and the total slanted area. Its value for p' is reported on the corresponding Lorenz curve graph (Fig. 1b).

The Lorenz curve provides a *scale-invariant* representation of the CDF $C_F(z)$ [14], with the following properties: $L_F(p)$ is concave, increasing with p, such as $0 \le L_F(p) \le p \le 1$, $L_F(0) = 0$ and $L_F(1) = 1$. $L_F(p)$ lies one the identity line ($L_F(p) = p$) when all the errors are equal, *i.e.*, $f_F(\varepsilon) = \delta(\varepsilon - c)$. Note that this case corresponds to a discontinuous CDF, with a jump at $\varepsilon = c$. The deviation of an error distribution from this extreme case can be measured by the Gini coefficient.

2.1.4 The Gini coefficient

It is related to the area between $L_F(p)$ and the identity line (Fig. 1b)

$$G_F = 2 \int_0^1 \{p - L(p)\}\, dp \tag{7}$$

where the factor two is used to scale G_F between 0 and 1. The smaller G_F, the closer the Lorenz curve to the identity line. The Gini coefficient, usually noted G, is generally used for distributions with positive support. Our notation G_F is a reminder that we are working here with distributions of absolute errors $f_F(\varepsilon)$.

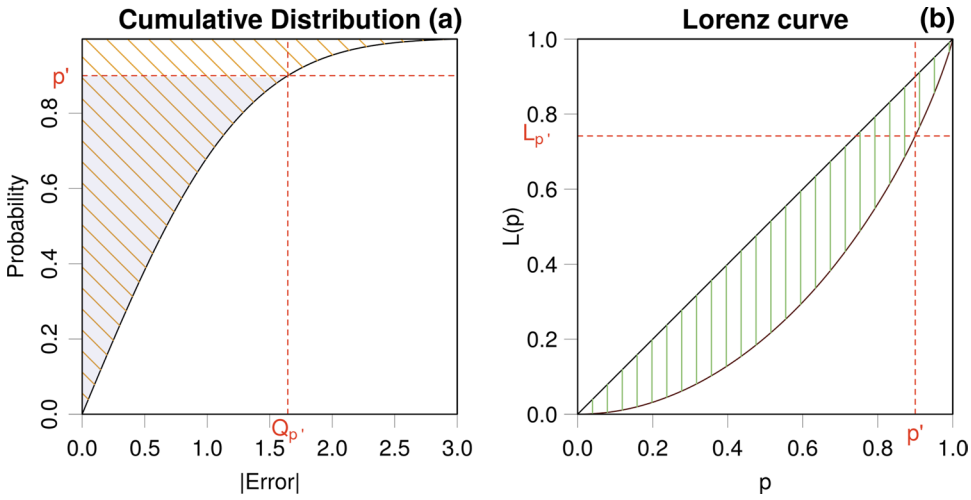

Fig. 1 Schematic cumulative density function **a** and Lorenz curve **b** for a folded standard normal density function of absolute errors. The area above the CDF (slanted) is the mean unsigned error (MUE). For a given probability p', the ratio of the shaded area to the total slanted area gives the value of the Lorenz curve $L_{p'} = L(p')$. $Q_{p'}$ is the quantile for probability p'. The area between the Lorenz curve and the identity axis (vertical bars) is half the Gini coefficient.

For sets of *absolute* errors with a normal distribution $N(\varepsilon;\mu_F,\sigma_F)$, G_F is proportional to the coefficient of variation $c_v = \sigma_F/\mu_F$ [11], where σ_F the standard deviation of the absolute errors

$$G_F \sim c_v/\sqrt{\pi} \tag{8}$$

Note that this relationship does hold only when all errors are of the same sign ($\mu_F \gg \sigma_F$), therefore with small c_v values.

Two typical values of G_F will be useful in the following:

- for any zero-centered normal distribution of errors $N(0,\sigma)$, the distribution of absolute errors is the half-normal distribution, with value $G_{FN} = \sqrt{2} - 1 \simeq 0.41$ [15];
- for any zero-centered uniform distribution, $U(-a,a)$, or any uniform distribution with a bound at zero, $U(-a,0)$ or $U(0,a)$, folding produces a uniform distribution with the minimal bound at zero, with value $G_{FU} = 1/3$ [15].

2.1.5 Skewness and kurtosis

Skewness measures the asymmetry of a distribution, while kurtosis is used as a measure either of its "peakedness" or "tailedness" [16] The moment-based formulae for skewness and kurtosis are not robust to outliers, and more robust quantile-based formulae have been proposed by several authors [6, 16–18].

For the skewness, we use a measure using the difference between the mean and median

$$\beta_{GM} = \frac{\mu - q(0.5)}{< |e - q(0.5)| >} \tag{9}$$

where the brackets indicate the mean value, $q(0.5)$ is the median of *signed* error, e, and the *GM* subscript refers to the authors of this definition, Goeneveld and Meeden [17]. β_{GM} takes its values between -1 and 1, and is 0 for symmetric distributions.

For kurtosis, an estimate based on quantiles is used [6] (originating from a similar form proposed by Crown and Siddiqui [19], hence the *CS* subscript)

$$\kappa_{CS} = \frac{q(0.975) - q(0.025)}{q(0.75) - q(0.25)} - 2.91 \tag{10}$$

where $q(.)$ is the quantile function for *signed* errors. The correction factor for the normal distribution (2.91) makes that κ_{CS} measures an *excess* kurtosis. Datasets with $\kappa_{CS} > 0$ have heavier tails than a normal distribution, and the opposite for negative values.

Specific notations will be introduced below when these statistics are applied to sets of absolute errors.

2.2 Estimation

We consider in this section the application of the previous statistics to finite error samples, and the corresponding formulae. Let us consider a set of errors ($E = \{e_i\}_{i=1}^N$), derived from a set of N calculated values ($C = \{c_i\}_{i=1}^N$) and reference data ($R = \{r_i\}_{i=1}^N$), by $e_i = r_i - c_i$. The absolute errors are noted $\Xi = \{\varepsilon_i = |e_i|\}_{i=1}^N$.

MSE, MUE and mode The mean signed error (MSE) is estimated as $\mu = \frac{1}{N}\sum_{i=1}^N e_i$, and the mean unsigned error (MUE) as $\mu_F = \frac{1}{N}\sum_{i=1}^N \varepsilon_i$.

As one is not dealing with necessarily symmetric distributions, the mode is an interesting location statistic, notably in correspondence to the tails of a distribution. The mode locates the part of the population with the highest density,

which is expected to bring a large contribution to μ_F and therefore influence the Lorenz curve and Gini coefficient. As one cannot assume the unimodality of the underlying distributions, one will consider the main mode. A nonparametric robust method, Bickel's half-range mode (HRM) estimator [20, 21], is used to estimate the location of the error samples main mode. This methods proceeds by iterative bipartition of modal intervals (intervals with highest density).

$L_F(p)$ and G_F

Let us introduce the cumulated sum of the $n \leq N$ *smallest* absolute errors

$$S_n = \sum_{i=1}^{n} \varepsilon_{[i]} \tag{11}$$

where $\varepsilon_{[i]}$ is the ith order statistic (*i.e.*, the value with rank i) of the sample of absolute errors. For consistency, one sets $S_0 = 0$.

The Lorenz curve is estimated as

$$L_F(p) = \frac{S_{p \times N}}{S_N} \tag{12}$$

where $0 \leq p \leq 1$. Note that, due to the use of finite samples, p takes its values in $\{i/N\}_{i=0}^{N}$.

Using a fast sorting of the sample of absolute errors, an efficient estimation of G_F uses the formula [9, 15, 22, 23]

$$G_F = \frac{\sum_{i=1}^{N} (2i - N - 1)\varepsilon_{[i]}}{N \sum_{i=1}^{N} \varepsilon_{[i]}} \tag{13}$$

A slower, but equivalent expression in terms of mean values is [10]

$$G_F = \frac{1}{MUE} < \max\{0, \varepsilon_1 - \varepsilon_2\} > \tag{14}$$

where ε_1 and ε_2 are two elements of Ξ and the mean is taken on all pairs.

β_{GM} and κ_{CS}

For skewness and kurtosis, Eqs. 9 and 10 are applied directly, with the robust method to estimate quantiles due to Harrell and Davis [3, 24, 25].

Uncertainty

Uncertainty on any statistic X, noted $u(X)$, is estimated by bootstrapping [26] with 1000 samples. Note that there is a known risk of underestimation of G_F for small datasets ($N < 100$) [15].

2.3 Implementation

All calculations have been made in the `R` language [27], using several packages, notably for the Gini coefficient (package `ineq` [23]), the HRM estimator (package `genefilter` [28]) and bootstrapping (package `boot` [29]).

The Gini coefficient, Lorenz curves, G_{MCF}, β_{GM}, κ_{CS} and mode estimator have been implemented in the freely available R [27] package `ErrViewLib (v1.3,` https://doi.org/10.5281/zenodo.3628475). The datasets can be analyzed with the `ErrView` graphical interface (source: https://doi.org/10.5281/zenodo.3628489; web interface: http://upsa.shinyapps.io/ErrView).

3 Datasets

3.1 Model datasets

Before applying the Gini coefficient to literature datasets, one explores its properties on error sets generated from distributions with controlled properties: uniform, normal, Student's-t, lognormal [30] and g-and-h [3, 25, 31].

It is important to note that we explore only distributions with a single dominant, more or less structured peak, such as the ones encountered in most computational chemistry error datasets. In the list of analytical distributions above, the uniform is an exception because of its undefined mode. We use it as an extreme case of single peaked, continuous distribution, with negative excess kurtosis.

Besides, it is easy to design multi-peaked distributions for which our conclusions on the Gini coefficient would not be valid. In fact, none of the usual summary statistics (MUE, MSE, skewness, kurtosis...) would describe properly such distributions.

3.2 Literature datasets

The statistical tools described above are applied to datasets gathered in the computational chemistry literature. These

Table 1 Literature datasets: N is the number of systems in the dataset and K is the number of methods.

Case	Property	N	K	Source
BOR2019	Band gaps (eV)	471	15	[32]
NAR2019	Enthalpies of formation (kcal/mol)	469	4	[33]
PER2018	Intensive atomization energies (kcal/mol)	222	9	[2]
SCH2018	Adsorption energies (eV)	195	7	[34]
THA2015	Polarizability (relative errors, in %)	135	7	[35]
WU2015	Polarizability (relative errors, in %)	145	36	[36]
ZAS2019	Effective atomization energies (kcal/mol)	6211	3	[37]
ZHA2018	Solid formation enthalpies (kcal/mol)	196	2	[38]

are summarized in Table 1 and strongly overlap with those studied in more details in a previous article [4], from which we removed small datasets ($N < 100$). The statistics, empirical cumulative density functions and Lorenz curves corresponding to these datasets are provided as Supplementary Information.

4 Applications

In its usual application fields, the Gini coefficient is applied to distributions with positive support. Our application to computational chemistry error sets involves the intermediate folding operation, which is not reversible. In a first part, we show how this limits the information on the errors distribution that can be inferred from the Gini coefficient. To relieve this difficulty, we propose a mode-centering operation before folding, which better preserves some of the tail properties of the original distributions. The Gini coefficient is then compared to other tails statistics, notably at the level of statistical uncertainty.

4.1 Gini coefficient versus bias

The link between the Gini coefficient and the coefficient of variation (Eq. 8) tells us that, for a normal distribution of given standard deviation, a decreasing bias should result in increasing values of G_F. At some point, this relation is broken by the folding operation: as noted earlier, for a centered normal distribution, one has $G_F = 0.41$, for which Eq. 8 does not hold. The dependence between a distribution shift and G_F is plotted in Fig. 2a for uniform $U(-1, 1)$, normal $N(0, 1)$, Student's-$t(\nu = 2)$, lognormal $LN(1, 0.5)$ and g-and-h $GH(g = 1, h = 0)$ distributions.

The zero-centered unimodal symmetric distributions (normal and Student's-t) have their maximal G_F value when the bias is null. The G_F curve for the uniform distribution reaches $G_{FU} = 1/3$ when the distribution is centered or shifted by ± 1. For intermediate absolute values of the bias, the folded distribution is not uniform and presents higher values of G_F. The value decreases when the bias is large enough to exclude zero from the range of non-null densities. Having heavier tails, the Student's-t distribution has a larger Gini coefficient than the normal. Any added bias leads to a decrease in G_F.

The decay curves are symmetric with respect to the sign of the bias. This is no longer the case for asymmetric distributions (lognormal and g-and-h), for which the peak is reached for non-null values of the bias and the decay curves are non-symmetric.

In Fig. 2b we plot similar curves for distributions centered on their mode before adding a bias (the symmetric distributions have been left unchanged). This shows that the maximal value of G_F is reached when the mode of the distribution is at, or near, the origin. This assertion is validated in the next section (Sect. 4.2).

Another important point illustrated by these curves is that the level of information that can be recovered from G_F is not uniform over the range of G_F values. For instance, $G_F = 0.41$ might as well occur for a centered normal distribution as for positively or negatively biased Student's-t, lognormal of g-and-h distributions, whereas values above 0.41 exclude normal and uniform distributions. Within the restrictions on the distributions shapes we considered above, one might infer that a value smaller than $G_{FU} = 1/3$ is likely to reveal a biased error distribution, while a value above $G_{FN} = 0.41$ is likely to signal distributions with one or two extended tail(s) (compared to normal tails), and with a possible bias. Between these bounds lies a blind zone where compensation

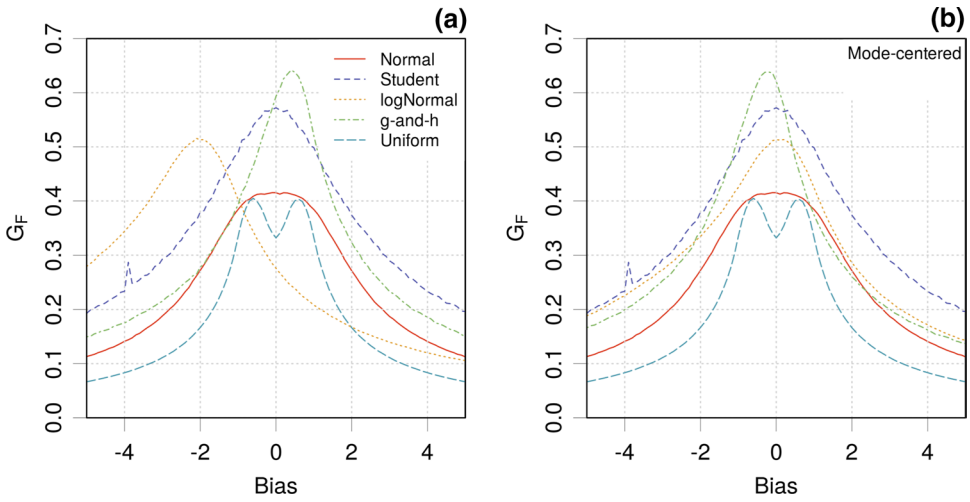

Fig. 2 Variation of G_F with a bias value added to several model distributions: **a** no centering applied; **b** mode centering applied before adding bias (Uniform, Student and Normal are unchanged).

between bias and shape factors prevent any inference on either of them.

4.2 Does mode centering maximize G_F?

In order to avoid the blind zone effect observed above and be able to characterize the shape of a distribution from its G_F value, mode centering seems to offer an interesting way to relieve the bias/shape compensation. Mode centering a distribution results in a folded distribution where both tails overlap and mix, but it ensures that the contribution of the most extended tail will prevail. In the absence of mode centering, when a biased distribution has a large tail encompassing zero, the folding around zero might considerably reduce this tail.

The assertion that mode centering maximizes G_F is tested here by comparing the results for mode centering with those obtained by explicitly maximizing the Gini coefficient with respect to a bias value. We define b_{max} as the value of the bias which maximizes G_F

$$b_{max} = \max_b G_F(|E - b|) \tag{15}$$

and note $G_{Fmax} = G_F(|E - b_{max}|)$. This equation is solved numerically by the Nelder and Mead optimizer [39].

The values of b_{max} and G_{Fmax} were computed for the literature datasets and compared to the mode $m(E)$ and G_{MCF} respectively, through z-scores

$$z_b = \frac{m(E) - b_{max}}{\sqrt{u(m(E))^2 + u(b_{max})^2}} \tag{16}$$

and

$$z_G = \frac{G_{MCF} - G_{Fmax}}{\sqrt{u(G_{MCF})^2 + u(G_{Fmax})^2}} \tag{17}$$

where the uncertainties are estimated by bootstrap.

In the hypothesis of a normal distribution of z-scores, a test threshold of 2 is generally chosen for the absolute value of the z-score [40]. For absolute values above 2, there is less than 5 percent of probability that the difference is due to sampling effects. For values below, one does not reject the hypothesis that the compared values are equal [3].

Histograms for the z-scores are shown in Fig. 3. At the exception of one point, the absolute value of all z-score values are smaller than 2, and we have therefore no reason to reject the hypothesis that these values are equal considering their uncertainty. The outlying point, with $z_b = -3.8$ and $z_G = -2.5$ corresponds to the MP2 method in dataset ZAS2019, which has a practically normal distribution [5]. One has $G_{MCF} = 0.418(6)$ and $G_{Fmax} = 0.436(3)$ for a distance of 1.07 between the mode and b_{max}, to be compared to the standard deviation of the distribution, 1.7. As shown in Fig. 2a, there is a flat area near the top of the G_F curve as a function of bias for a normal distribution: very small deviations from a perfect normal distribution (as hinted to by the value of G_{Fmax} being larger than 0.41) can deviate the optimal point over a wide range.

For all practical purposes in the present study, one can therefore estimate that mode centering maximizes the value of the Gini coefficient, at a fraction of the computing cost for the search for b_{max}. We further note that for distributions having distinct mean and mode, centering on the mean would not maximize G_F and therefore preserve some of the ambiguity due to bias/shape compensation.

4.3 G_{MCF} versus G_F

We note the Gini coefficient of mode-centered folded distributions G_{MCF}. Figure 4a displays G_{MCF} versus G_F for the literature datasets. It is clear that mode centering increases all G values, *i.e.*, $G_{MCF} \geq G_F$ for all datasets, within the

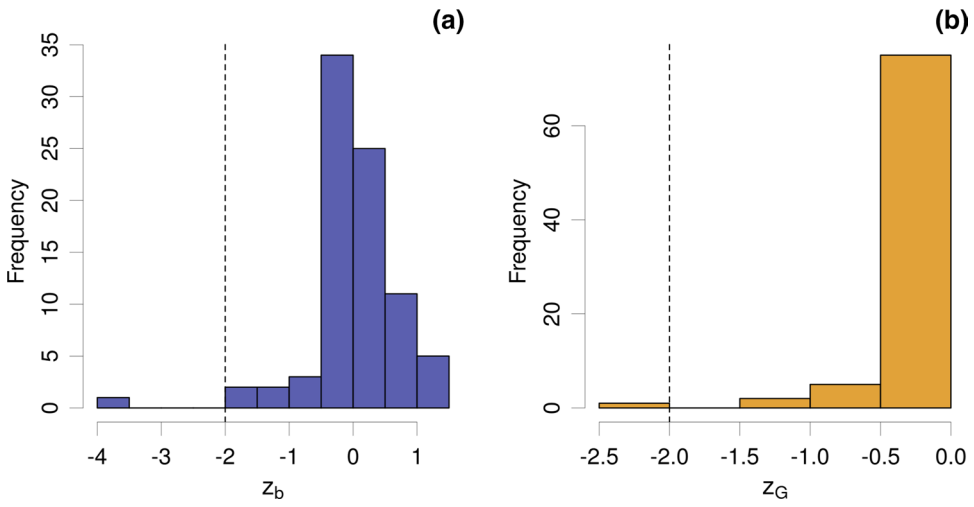

Fig. 3 Histograms of z-scores for **a** the position of the mode versus the G_F maximizer and **b** the values of the corresponding Gini coefficients.

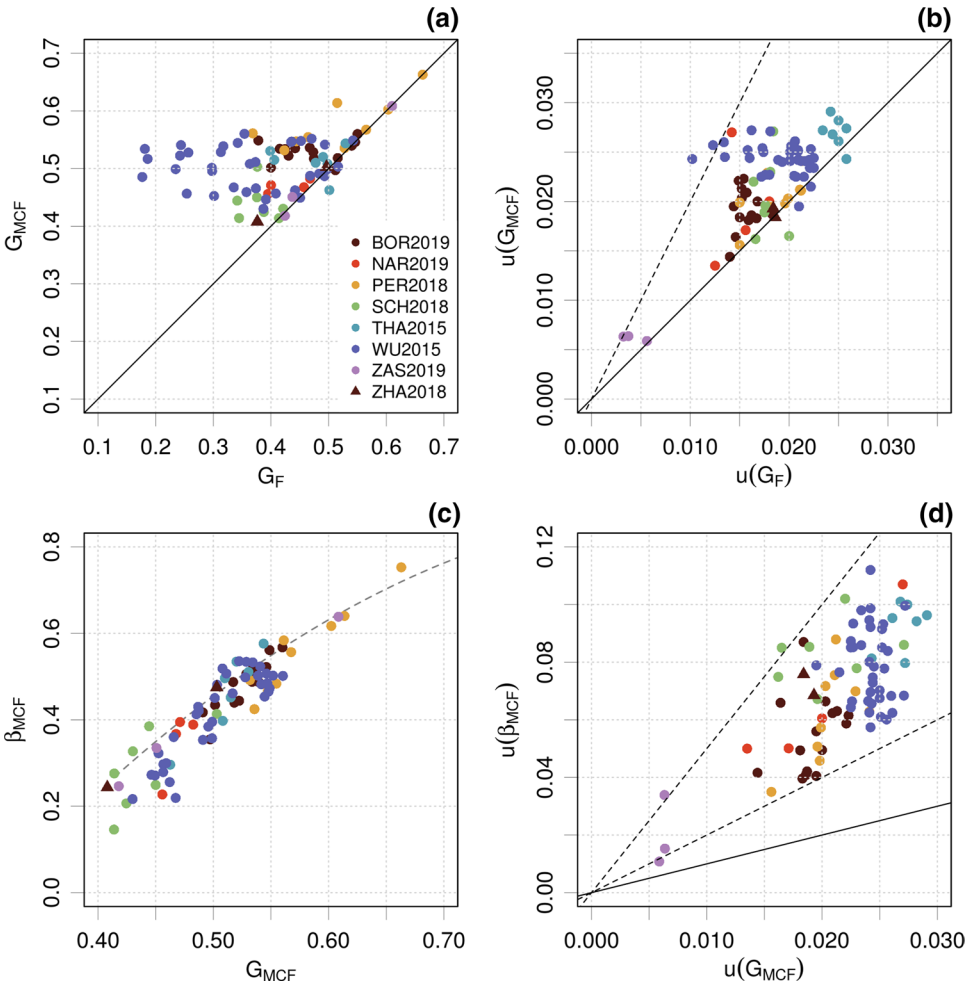

Fig. 4 Comparison of G_{MCF} with other statistics: **a** correlation of G_F and G_{MCF} for the literature datasets and **b** comparison of their uncertainties (the dashed line has a slope of 2); **c** correlation of G_{MCF} and β_{MCF} for the literature datasets (points) and a series of large samples ($N = 10^6$) of Student's-t and g-and-h distributions with a range of shape parameters (dashed line) and **d** comparison of their uncertainties (the points are bracketed by lines of slope 2 and 5).

estimation uncertainties. The G-scale is now reduced to values above 0.4, in conformity with our interpretation that all values below $G_{FU} = 1/3$ were due to bias.

The uncertainties are reported in Fig. 4(b), showing that for some datasets the uncertainty on G_{MCF} is larger than the uncertainty on G_F, up to a factor two. This extra uncertainty is due to the uncertainty on the mode value. We note also a size effect, the smallest datasets (THA2015, $N = 145$) having the largest uncertainty, and the largest dataset (ZAS2019, $N = 6211$), the smallest one.

4.4 G_{MCF} as a shape statistic

In parallel with the Gini coefficient, the skewness of the distribution has also been considered as an estimator of inequality [11]. One is interested here in comparing G_{MCF} with β_{MCF}, which is the skewness β_{GM} (Eq. 9) of the mode-centered folded distribution.

The values for our selection of literature datasets are shown in Fig. 4c. There is an excellent correlation between those statistics, considering the uncertainties reported in Fig. 4d. Using model datasets of large size ($N = 10^6$) for Student's-t and g-and-h distributions with a range of shape parameters, one observes a nearly perfect nonlinear correlation (dashed line, resulting from a quadratic fit of the sampled values). The dispersion of the points for the literature datasets about this curve is mostly due to statistical uncertainty (the points for the largest dataset (ZAS2019, $N = 6211$) are very close to the curve). It is important to note that the uncertainty on G_{MCF} is a factor two to five smaller than the uncertainty on β_{MCF} and therefore performs better for smaller datasets.

We can therefore conclude that G_{MCF} is apt at estimating the heaviness of the errors distribution tail after mode centering and folding. In order to appreciate the information about the *signed* error distribution that can be extracted from G_{MCF}, we plotted it against the skewness β_{GM} and excess kurtosis κ_{CS} (Fig. 5).

Considering skewness (Fig. 5a), all the points seem to lie within an angular sector, indicating that distributions with large skewness have necessarily large G_{MCF} values. For instance, if the absolute value of the skewness is above

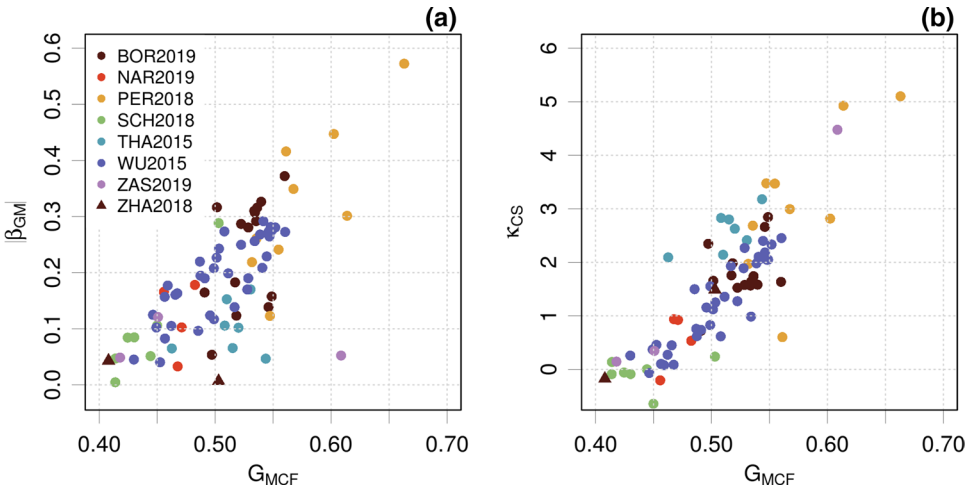

Fig. 5 Comparison of G_{MCF} with shape statistics of error distributions: **a** absolute value of skewness $|\beta_{GM}|$; **b** excess kurtosis κ_{CS}.

Table 2 The ten methods with the largest G_{MCF} values and the corresponding skewness and kurtosis.

Dataset	Methods	G_{MCF}	β_{GM}	κ_{CS}
PER2018	CAM-B3LYP	0.663(16)	0.572(57)	5.1(1.3)
PER2018	B3LYP	0.614(20)	0.301(74)	4.93(96)
ZAS2019	SLATM_L2	0.6086(59)	0.052(22)	4.48(25)
PER2018	LC-ωPBE	0.602(20)	0.447(64)	2.82(82)
PER2018	PBE0	0.568(24)	0.349(63)	2.99(86)
PER2018	BH&HLYP	0.561(20)	0.416(52)	0.60(48)
WU2015	τHCTHhyb	0.560(24)	−0.273(84)	2.45(83)
BOR2019	HLE16 + SOC	0.560(19)	0.372(43)	1.64(47)
PER2018	BLYP	0.555(20)	−0.241(67)	3.47(78)
WU2015	B97-2	0.552(22)	−0.281(80)	2.33(77)

0.3, G_{MCF} is larger than 0.5. Reciprocally, G_{MCF} provides only an upper limit to $|\beta_{GM}|$ (for $G_{MCF} = 0.55$, the absolute value of the skewness cannot be above 0.4). For kurtosis (Fig. 5b), there is a lax positive trend between both statistics, and, globally, large values of G_{MCF} are associated with large excess kurtosis, which might be due to heavy tails or outliers. In both graphs, values of G_{MCF} below 0.45 are associated with low skewness *and* excess kurtosis. Although information is lost because of folding, G_{MCF} can still provide some information about the shape of the distribution of signed errors, and notably about the kurtosis.

Let us consider a few examples to illustrate this point. We see in Fig. 5 that most points fall between 0.4 and 0.55, but a few methods reach higher values. The largest G_{MCF} value in this study is 0.66 (orange dot) for method CAM-B3LYP in the PER2018 set [2] (*cf.* Table 2). This corresponds to large values of both β_{GM} and κ_{CS}. The authors discussed how this DFA is in the head group of two methods with similar MUE values, but does not minimize the risk of large errors (Sect. II.A [2]). From the same set, B3LYP has the second largest G_{MCF} value (0.61) and presents the same tail features than CAM-B3LYP. The third largest G_{MCF} value (0.61, violet dot) belongs to the ZAS2019 dataset, and it presents a null skewness and a large kurtosis. An in-depth study has been published for this case [5], where the errors distribution for the SLATM-L2 method was shown to have large tails, despite having the smallest MUE among the compared methods. In the same set, the MP2 method has a practically normal errors distribution and can be found in the lower part of the Gini scale (0.42). This analysis can be repeated for the ten methods with largest G_{MCF} values (Table 2), showing that large G_{MCF} values point indeed toward error sets with high kurtosis and/or skewness.

4.5 Application of G_{MCF} to ranking

To evaluate the interest of G_{MCF} in a practical scenario, one might consider it as an alert mechanism to complement a MUE ranking. But a question remains: "What is the threshold for G_{MCF} one should use to detect problematic error distributions ?" We have seen above that the ten largest G_{MCF} values, above 0.55, point to distributions with notable tails. We propose for the present study to adopt an alert value based on the median of the G_{MCF} values for our full dataset (0.51) and round it to 0.5. This might be reevaluated when more data are analyzed. Using this threshold, one might flag distributions suspected of having tails unsuitable for reliable predictions.

Figure 6a shows, for each dataset, the flagging of the methods with the best ranking. If one considers the first rank, five methods are flagged, but it is striking that three datasets have all their 10 lowest MUE-ranked methods flagged (BOR2019, PER2018 and THA2015). For BOR2019, all methods present some excess kurtosis and variable levels of skewness. We can relate this to an increasing trend of the errors with the band gap value [4]. In the case of PER2018,

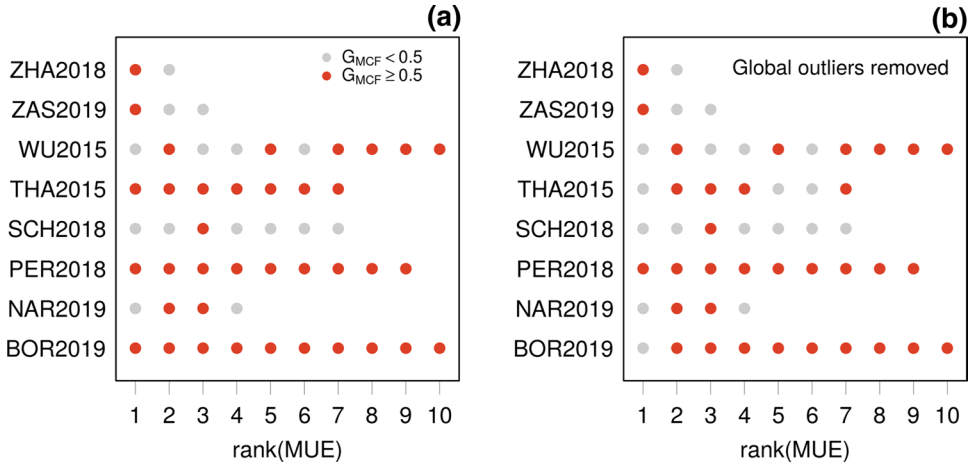

Fig. 6 Flagging of large G_{MCF} values after mode centering of at most 10 lowest MUE-ranked methods in each dataset.

Table 3 Statistics for the methods of the ZHA2018 dataset, before and after linear correction ("lc-" prefix).

Methods	MUE (kcal/mol)	MSE (kcal/mol)	Q_{95} (kcal/mol)	β_{GM}	κ_{CS}	G_{MCF}
PBE	0.2106 (98)	−0.205(10)	0.467(34)	−0.043(64)	−0.17(28)	0.408(18)
SCAN	0.1024 (69)	−0.0165(97)	0.291(21)	−0.007(67)	1.49(53)	0.503(19)
lc-PBE	0.0923 (61)	0.0	0.287(38)	−0.138(71)	1.02(43)	0.479(24)
lc-SCAN	0.0917 (63)	0.0	0.276(26)	−0.082(67)	1.27(41)	0.475(20)

the error distributions present also large skewness and kurtosis, which can be associated with the chemical heterogeneity of the dataset [2]. For THA2015, it was noted previously [4, 35] that some experimental reference data with large measurement uncertainty could not be reproduced by any method in the studied set. These outliers contribute to the tails of all the error distributions (so-called *global* outliers) and affect G_{MCF} values. Note that, more generally, reference data are not necessarily the origin of global outliers, as a missing physical contribution in the tested methods could produce similar effects.

To explore the role of global outliers, Fig. 6b reports the same analysis after search and removal of global outliers, defined as systems lying out of the $[q(0.025), q(0.975)]$ interval for *all methods* of a dataset. The removal of 6 systems affects strongly the case THA2015, confirming the previous analysis. The results for BOR2019 are mostly unchanged, except for the best MUE-ranked method (mBJ) which benefits from the removal of a single global outlier. No effect is observed for the PER2018 dataset, confirming the intrinsic heavy-tailed shape of these heterogeneous atomization energy error sets [2].

The other datasets with leading G_{MCF}-flagged methods are ZAS2019 and ZHA2018. The former has already been discussed (Sect. 4.4), and the removal of several global outliers has no impact. In the case ZHA2018, the first MUE-ranked method is SCAN, which has a G_{MCF} value just above the threshold (0.503) and presents no skewness and a slight level of kurtosis. Removal of 4 global outliers does not improve the shape of the errors distribution.

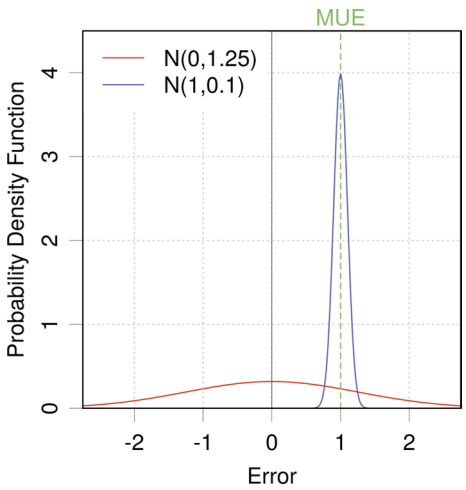

Fig. 7 Example of different error distributions having the same MUE (1.0) and offering contradictory results for some tail statistics. The probability to have *absolute* errors larger than 1.0 is $P_1 = 0.50$ for the blue curve and 0.42 for the red curve, hiding the fact that the red distribution contains much worse results that the blue one. In this case, the problem is solved by the values of Q_{95}, giving 1.16 for the blue curve, and 2.46 for the red one. Shape statistics, such as the kurtosis, would not enable to discriminate between both normal distributions.

However, the errors on the formation enthalpies present a linear trend as a function of the calculated values. Correcting this trend [1] improves slightly the performance and the shape of the SCAN distribution, but most notably of the PBE distribution, which performances get indistinguishable from those of SCAN (Table 3).

4.6 Limits of the G_{MCF} coefficient

We have shown above that G_{MCF} might be a useful complement to the usual ranking statistics, in order to detect error distributions with shapes that might reveal problem in prediction reliability. However, there remain cases where the G_{MCF} index is insufficient to reveal underlying problems. Figure 7 proposes a scenario of two normal distributions ($G_{MCF} = 0.41$) with the same value of the MUE (1.0) and yet very different risks of large prediction errors. This is clearly a case showing that a quantile-based statistic, such as Q_{95}, is an essential complement to the MUE.

5 Conclusion

The Gini coefficient presents an interesting addition to the computational chemistry benchmarking statistical toolbox. We focused here on its properties in relation to features of the error distributions, such as bias and shape (skewness and kurtosis). The interest of the Gini coefficient is that it correlates with these features and offers a one-number summary. This is also one of its weaknesses, as there is no unique mapping from the Gini coefficient to these features.

To unscramble this situation, we propose to use the Gini coefficient of the mode-centered distributions, G_{MCF}, which offers a simpler to interpret, shape-based, measure of tailedness. Large G_{MCF} values, *e.g.*, above 0.5, alert us about large tails that might be due to large skewness, kurtosis and/or the presence of outliers. For high ranking methods, this is an incentive to inspect closely the error distributions and check whether the selected methods might have problems of reliability in their predictions. It might then be worth to investigate whether the distorted shape of the distribution is due to systematic trends in the errors, as they can often be corrected by simple linear transformations [1, 41–44]. The impact of such corrections on the shape of error distributions is the prospect of further studies.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00214-021-02725-0.

Data availability statement The data and codes that support the findings of this study are openly available at the following URL: https://doi.org/10.5281/zenodo.4333217

References

1. Pernot P, Civalleri B, Presti D, Savin A (2015) Prediction uncertainty of density functional approximations for properties of crystals with cubic symmetry. J Phys Chem A 119:5288–5304. https://doi.org/10.1021/jp509980w
2. Pernot P, Savin A (2018) Probabilistic performance estimators for computational chemistry methods: the empirical cumulative distribution function of absolute errors. J Chem Phys 148:241707. https://doi.org/10.1063/1.5016248
3. Pernot P, Savin A (2020) Probabilistic performance estimators for computational chemistry methods: systematic improvement probability and ranking probability matrix. I. Theory J Chem Phys 152:164108. https://doi.org/10.1063/5.0006202
4. Pernot P, Savin A (2020) Probabilistic performance estimators for computational chemistry methods: Systematic improvement probability and ranking probability matrix. II. Appl J Chem Phys 152:164109. https://doi.org/10.1063/5.0006204
5. Pernot P, Huang B, Savin A (2020) Impact of non-normal error distributions on the benchmarking and ranking of Quantum Machine Learning models. Mach Learn Sci Technol 1:035011. https://doi.org/10.1088/2632-2153/aba184
6. Bonato M (2011) Robust estimation of skewness and kurtosis in distributions with infinite higher moments. Finance Res Lett 8:77–87. https://doi.org/10.1016/j.frl.2010.12.001
7. Lorenz MO (1905) Methods of measuring the concentration of wealth. Publ Am Stat Assoc 9:209–219. https://doi.org/10.2307/2276207
8. Gini C (1912) Variabilità e mutabilità
9. Damgaard C, Weiner J (2000) Describing inequality in plant size or fecundity. Ecology 81:1139–1142. https://doi.org/10.2307/177185
10. Eliazar II, Sokolov IM (2010) Measuring statistical heterogeneity: the Pietra index. Phys A 389:117–125. https://doi.org/10.1016/j.physa.2009.08.006
11. Bendel RB, Higgins SS, Teberg JE, Pyke DA (1989) Comparison of skewness coefficient, coefficient of variation, and Gini coefficient as inequality measures within populations. Oecologia 78:394–400. https://doi.org/10.1007/BF00379115
12. Florian MK, Li N, Gladders MD (2016) The Gini coefficient as a morphological measurement of strongly lensed galaxies in the image plane. Astrophys J 832:168. https://doi.org/10.3847/0004-637X/832/2/168
13. Hurley N, Rickard S (2009) Comparing measures of sparsity. IEEE Trans Inf Theory 55:4723–4741. https://doi.org/10.1109/TIT.2009.2027527
14. Kleiber C (2005) The Lorenz curve in economics and econometrics. techreport, TU Dortmund, March. https://doi.org/10.17877/DE290R-14481
15. Dixon PM, Weiner J, Mitchell-Olds T, Woodley R (1987) Bootstrapping the Gini coefficient of inequality. Ecology 68:1548–1551. https://doi.org/10.2307/1939238
16. Ruppert D (1987) What is kurtosis? An influence function approach. Am Stat 41:1. https://doi.org/10.2307/2684309
17. Groeneveld RA, Meeden G (1984) Measuring skewness and kurtosis. Stat 33:391–399. http://www.jstor.org/stable/2987742, https://doi.org/10.2307/2987742
18. Suaray K (2015) On the asymptotic distribution of an alternative measure of kurtosis. Int J Adv Stat Proba 3:161–168. https://doi.org/10.14419/ijasp.v3i2.5007
19. Crow EL, Siddiqui MM (1967) Robust estimation of location. J Am Stat Assoc 62:353–389. https://doi.org/10.2307/2283968
20. Bickel DR (2002) Robust estimators of the mode and skewness of continuous data. Comput Stat Data Anal 39:153–163. https://doi.org/10.1016/S0167-9473(01)00057-3

21. Hedges SB, Shah P (2003) Comparison of mode estimation methods and application in molecular clock analysis. BMC Bioinform 4:31. https://doi.org/10.1186/1471-2105-4-31
22. Glasser GJ (1962) Variance formulas for the mean difference and coefficient of concentration. J Am Stat Assoc 57:648–654. https://doi.org/10.1080/01621459.1962.10500553
23. Zeileis A (2014) ineq: measuring inequality, concentration, and poverty. R package version 0.2-13. URL: https://CRAN.R-project.org/package=ineq
24. Harrell FE, Davis C (1982) A new distribution-free quantile estimator. Biometrika 69:635–640. https://doi.org/10.2307/2335999
25. Wilcox RR, Erceg-Hurn DM (2012) Comparing two dependent groups via quantiles. J Appl Stat 39:2655–2664. https://doi.org/10.1080/02664763.2012.724665
26. Efron B (1979) Bootstrap methods: another look at the jackknife. Ann Stat 7(1):1–26. https://doi.org/10.1214/aos/1176344552
27. R Core Team (2019) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org/
28. Gentleman R, Carey V, Huber W, Hahne F (2019) genefilter: methods for filtering genes from high-throughput experiments. R package version 1(68)
29. Canty A, Ripley BD (2019) boot: bootstrap R (S-Plus) Functions. R package version 1.3-22
30. Evans M, Hastings N, Peacock B (2000) Statistical distributions. Wiley-Interscience, 3rd edition
31. Hoaglin DC (1985) Exploring data tables, trends, and shapes, chapter Summarizing shape numerically: the g-and-h distributions, pp 461–513. Wiley, New York
32. Borlido P, Aull T, Huran AW, Tran F, Marques MA, Botti S (2019) Large-scale benchmark of exchange-correlation functionals for the determination of electronic band gaps of solids. J Chem Theory Comput 15:5069–5079. https://doi.org/10.1021/acs.jctc.9b00322
33. Narayanan B, Redfern PC, Assary RS, Curtiss LA (2019) Accurate quantum chemical energies for 133000 organic molecules. Chem Sci 10:7449–7455. https://doi.org/10.1039/c9sc02834j
34. Schmidt PS, Thygesen KS (2018) Benchmark database of transition metal surface and adsorption energies from many-body perturbation theory. J Phys Chem C 122:4381–4390. https://doi.org/10.1021/acs.jpcc.7b12258
35. Thakkar AJ, Wu T (2015) How well do static electronic dipole polarizabilities from gas-phase experiments compare with density functional and MP2 computations? J Chem Phys 143:144302. https://doi.org/10.1063/1.4932594
36. Wu T, Kalugina YN, Thakkar AJ (2015) Choosing a density functional for static molecular polarizabilities. Chem Phys Lett 635:257–261. https://doi.org/10.1016/j.cplett.2015.07.003
37. Zaspel P, Huang B, Harbrecht H, von Lilienfeld OA (2019) Boosting quantum machine learning models with a multilevel combination technique: people diagrams revisited. J Chem Theory Comput 15(3):1546–1559. https://doi.org/10.1021/acs.jctc.8b00832
38. Zhang Y, Kitchaev DA, Yang J, Chen T, Dacek ST, Sarmiento-Perez RA, Marques MAL, Peng H, Ceder G, Perdew JP, Sun J (2018) Efficient first-principles prediction of solid stability: towards chemical accuracy. npj Comput Mater 4:9. https://doi.org/10.1038/s41524-018-0065-z
39. Nelder JA, Mead R (1965) A simplex method for function minimization. Comput J 7:308–313. https://doi.org/10.1093/comjnl/7.4.308
40. Kacker RN, Kessel R, Sommer K-D (2010) 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. https://doi.org/10.6028/jres.115.031
41. Lejaeghere K, Jaeken J, Speybroeck VV, Cottenier S (2014) Ab initio based thermal property predictions at a low cost: an error analysis. Phys Rev B 89:014304. https://doi.org/10.1103/physrevb.89.014304
42. Lejaeghere K, Vanduyfhuys L, Verstraelen T, Speybroeck VV, Cottenier S (2016) Is the error on first-principles volume predictions absolute or relative? Comput Mater Sci 117:390–396. https://doi.org/10.1016/j.commatsci.2016.01.039
43. Proppe J, Husch T, Simm GN, Reiher M (2016) Uncertainty quantification for quantum chemical models of complex reaction networks. Faraday Discuss 195:497–520. https://doi.org/10.1039/c6fd00144k
44. Proppe J, Reiher M (2017) Reliable estimation of prediction uncertainty for physicochemical property models. J Chem Theory Comput 13:3297–3317. https://doi.org/10.1021/acs.jctc.7b00235