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

Pascal PERNOT (pascal.pernot@universite-paris-saclay.fr)
Institut de Chimie Physique, CNRS, Univ. Paris-Sud,
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, F-75252 Paris, France
email:andreas.savin@lct.jussieu.fr

Monday 23rd November, 2020

**Abstract**

The distribution of errors being a central object in the assesment and benchmarking of quantum chemistry methods, we show how the Lorenz curve and the associated statistics can be used to characterize it. TO BE COMPLETED...

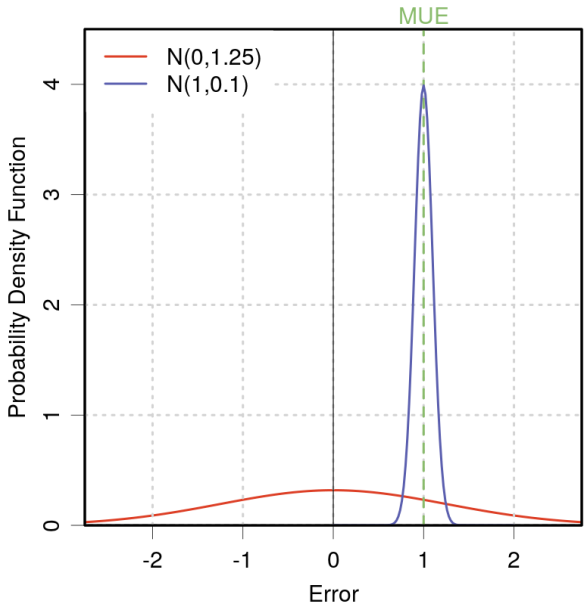


Figure 1: Example of different error distributions having the same MUE (1.0) and offering contradictory results for some tail statistics. The probability to have *absolute* errors larger than 1.0 is $P_1 = 0.50$ for the blue curve and 0.43 for the red curve, hiding the fact that the red distribution 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. Statistical correction of the blue curve by centering would reconciliate the ranking by the MUE, $P_1$ and $Q_{95}$ statistics.

# 1 Introduction

The reliability of a computational chemistry method is strongly conditioned by the distribution of its prediction errors. Distributions with heavy tails entail a risk of large prediction errors. As a benchmarking statistic, the popular mean unsigned error (MUE) bears no information on such a risk [1, 2, 3, 4]. We have recently reported a case where two error distributions with identical values of the MUE have widely different risks of large errors [5, 4]. This was mainly due to heavy tails in one of the two practically unbiased distributions, but cases might occur where bias adds even more complexity to the MUE comparison. As summarized in Fig. 1, the duality of the MUE as a location and dispersion statistic prevents its use to predict the risk of large errors. 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], $P_\eta$ [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 $P_\eta$ that absolute errors are larger than a chosen threshold $\eta$. $P_\eta$ provides a direct quantification of the risk of large errors, but $\eta$ has to be defined *a priori* and might

be usage-dependent.

- 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 $P_\eta$ partly answer to the question, but they are point estimates on the cumulated density function of the absolute errors, and a more global 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 us to such measures as skewness and kurtosis, but none of these alone would be able to cover all the scenarii. 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, 9, 10], are used to evaluate the degree of inequality within these populations. It is also used in biology to estimate the inequality of properties within plant populations. The Lorenz curve has many mathematical representations, the most interesting one for us being its formulation as an integral of the quantile function, making the link with our study of probabilistic metrics [2, 3, 4].

In this short article, we explore the interest of $G$ as a complement to the MUE. We introduce the statistical tools and their implementation in Section 2, and apply them to a series of datasets to illustrate the pertinence of the Gini coefficient in Section 3.

# 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

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

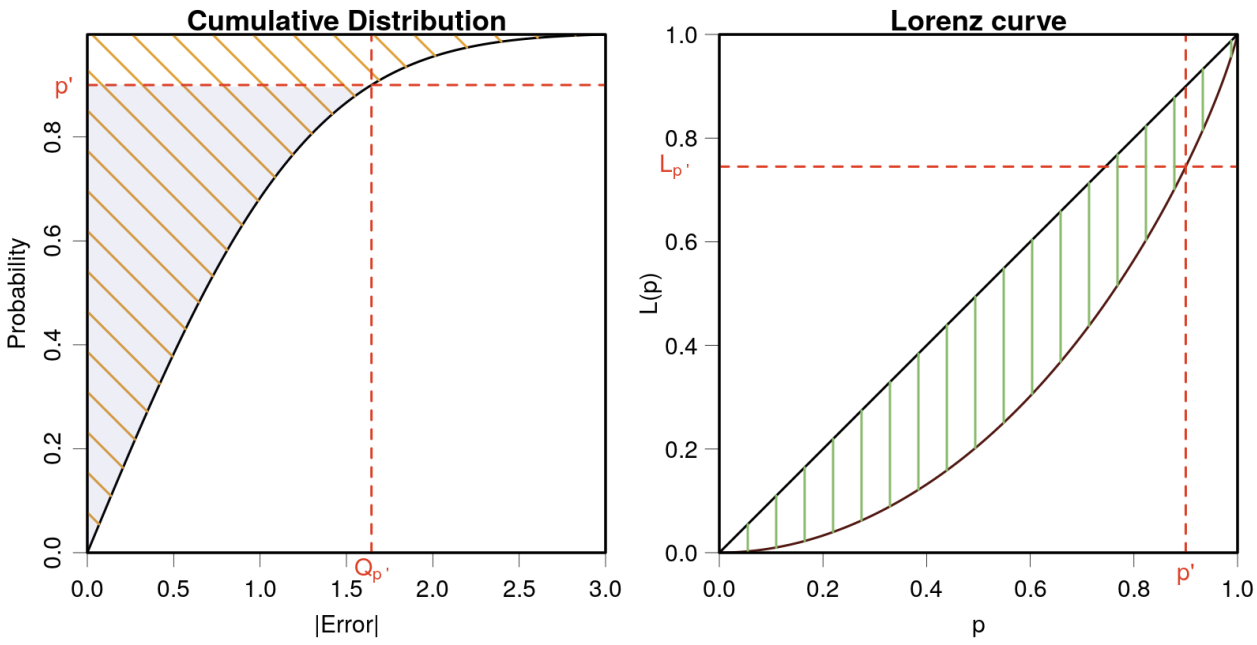


Figure 2: Schematic cumulative density function (CDF; left) and Lorenz curve (right) for a folded standard normal density function of absolute errors. The area above the CDF (slanted) is the mean unsigned error (MUE). For a given probability $p'$, the ratio of the unshaded area to the total area gives the value of the Lorenz curve $L_{p'} = L(p')$. The area between the Lorenz curve and the identity axis is half the Gini coefficient.

from which the quantile function is the inverse

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

In Fig. 2, the quantile for probability $p'$ is noted $Q_{p'}$.

### 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:

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

As shown in Fig. 2(left), 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. 2(right)).

The Lorenz curve provides a scale-invariant representation of the CDF [11], with the following properties: $L(p)$ is concave, increasing with $p$, such as $0 \leq L(p) \leq p \leq 1$, $L(0) = 0$ and $L(1) = 1$. $L(p)$ lies one the identity line ($L(p) = p$) when all the errors are equal, *i.e.* $f_F(\varepsilon) = \delta(\varepsilon - c)$. The deviation of the error distributions from this case can be estimated by the Gini coefficient.

### 2.1.4 The Gini coefficient

It is related to the area between $L(p)$ and the identity line (Fig. 2(right))

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

where the factor two is used to scale $G$ between 0 and 1. The smaller $G$, the closer the Lorenz curve to the identity line.

For sets of errors with normal distributions, $G$ is proportional to the coefficient of variation $c_v = \sigma/\mu$, where $\mu$ is the mean, and $\sigma$ the standard deviation of the signed errors; $G = c_v/\sqrt{\pi}$ [12]. Usually seen as a measure of inequality in a distribution, $G$ is also related to the relative precision. Note that this relationship does not hold for absolute errors, except if all errors are of the same sign, therefore with a small absolute $c_v$ value.

### 2.1.5 Kurtosis

Kurtosis is used as a measure either of "peakedness" or "tailedness" of a distribution [13], compared to a normal one. The moments-based formula for kurtosis is not robust to outliers, and a more robust quantile-based formula has been proposed by several authors [13, 6, 14] (originating from a similar form proposed by Crown and Siddiqui [15], hence the 'CS' subscript)

$$\kappa_{CS} = \frac{q(0.975) - q(0.025)}{q(0.75) - q(0.25)} - 2.905 \tag{8}$$

where $q(.)$ is the quantile function for *signed* errors. The correction factor for the normal distribution (2.905) makes that $\kappa_{CS}$ measures an excess kurtosis.

## 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 $\varepsilon = \{\epsilon_i = |e_i|\}_{i=1}^N$.

The Mean Signed Error (MSE), Mean Unsigned Error (MUE) and Root Mean Squared Error (RMSD) are estimated as

$$MSE = \frac{1}{N}\sum_{i=1}^{N} e_i \tag{9}$$

$$MUE = \frac{1}{N}\sum_{i=1}^{N} \epsilon_i \tag{10}$$

$$RMSD = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(e_i - MSE)^2} \tag{11}$$

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

$$S_n = \sum_{i=1}^{n} \epsilon_{[i]} \tag{12}$$

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

### 2.2.1 The Lorenz curve

The Lorenz curve is estimated as

$$L(p) = \frac{S_{p\times N}}{S_N} \tag{13}$$

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

### 2.2.2 The Gini coefficient

Using a fast sorting of the sample of absolute errors, an efficient estimation of $G$ uses the formula [9, 16]

$$G = \frac{\sum_{i=1}^{N}(2i - N - 1)\epsilon_{[i]}}{N\sum_{i=1}^{N}\epsilon_{[i]}} \tag{14}$$

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

$$G = \frac{1}{<\varepsilon>} < \max\{0, \varepsilon_1 - \varepsilon_2\} > \tag{15}$$

where $\varepsilon_1$ and $\varepsilon_2$ are two elements of $\varepsilon$ and the mean is taken on all pairs.

| Case | Property |
|---|---|
| BiasNorm | Normal distributions with the mean $\mu$ varying from 0.0 to 5.0 by 0.5 |
| Student | Student's-$t$ distributions with degrees of freedom in {2-10,20,50,100} |
| BiasStudent1 | Same as above, with a shift of +1 |
| BiasStudent2 | Same as above, with a shift of +2 |
| GandH | $g$-and-$h$ distributions with asymmetry parameter $g$ varying from 0.1 to 1.0 by 0.1 |
| BiasGandH1 | Same as above, with a shift of +1 |
| BiasGandH2 | Same as above, with a shift of +2 |

Table 1: Reference datasets. All sets were generated with $N = 1000$ samples.

Uncertainty on $G$ is generally estimated by bootstrapping [17], with a known risk of underestimation for small samples ($N < 100$) [18].

#### 2.2.3 Kurtosis

The direct application of Eq. 8 is used, with the robust method for quantiles due to Harrel and Davis [19, 20, 3].

### 2.3 Codes

All calculations have been made in the `R` language [21], using several packages, notably for the Gini coefficient from package `ineq` [16]. Statistics and Lorenz curves are implemented in ErrViewLib (`https://doi.org/10.5281/zenodo.3628475`). The datasets can also be analyzed with the ErrView code (`https://doi.org/10.5281/zenodo.3628489`), or its web interface (`http://upsa.shinyapps.io/ErrView`).

## 3 Applications

Before applying the Gini coefficient to literature datasets, one explores its properties on error sets generated from distributions with controlled properties (reference dataset).

### 3.1 Reference datasets

As underlined above, the Gini coefficient is influenced both by the position and shape of the distribution. In order to map the correlations of the Gini coefficient with other statistics, we generated an ensemble of datasets from standard distributions with varying degrees of bias (BiasNorm), asymmetry (GandH), kurtosis (Student) and some combinations of these (BiasStudent, BiasGandH). These seven datasets are described in Table 1. In all cases, the distributions refer to signed errors. The statistics, probability density functions and Lorenz curves corresponding to these datasets are provided in Supplementary Information.

The Gini coefficient was plotted against the inverse coefficient of variation ($1/c_v$, see Sect.2.1.4), estimated as |MSE|/RMSD, to avoid the singularity of $c_v$ for unbiased distribution (Fig. 3(a)). The dashed line in this figure represents the $1/c_v = 1/(\sqrt{\pi}G)$ relationship. It is closely followed by strongly biased normal distribution (black squares) at low $G$ values ($G \lesssim 0.3$). At small bias values, the properties of the folded normal distribution take precedence, and the symbols deviate from the inverse law to converge to the value $G \simeq 0.41$ for a centered normal distribution.

The Student case (red squares) gives us milestones for centered symmetric distributions with different kurtosis ($\kappa_{CS}$) values. The red squares lie on the $1/c_v = 0$ axis, the rightmost value corresponding to a Student's-$t$ distribution with 2 degree of freedom, at $G \simeq 0.55$. As expected, when the number of degrees of freedom increases, the points shift towards the centered normal case. For the GandH case (light blue squares), the asymmetry/skewness is explored, in a series starting from the normal distribution ($g = 0$), and acquiring a small bias as the asymmetry parameter increases up to $g = 1$. As for the Student case, $G$ increases when the shape of the normal distribution is altered, up to $G \simeq 0.62$ in this case.

Combining bias with kurtosis and asymmetry is explored in sets BiasStudent1 (red dots), BiasStudent2 (red triangles), BiasGandH1 (blue dots) and BiasGandH2 (blue triangles). All series converge to the corresponding biased normal case, the higher $G$ values corresponding to the higher values of asymmetry or kurtosis. For a bias value of 1, the series follow distinct trajectories, but when the bias increases to 2, they strongly overlap. For large bias, the distribution shape becomes unimportant.

A similar analysis is presented in Fig. 3(b) for the correlation between $G$ and kurtosis. The dashed line is a linear fit to the points of the GandH set. The most striking feature in this plot, is the exclusion zone on the bottom-right of the graph, implying that one cannot have large $G$ values without a large kurtosis or asymmetry (heavy tails). On the left of the graph, one sees again that the shape of distributions becomes less and less relevant when the bias increases (overlap of the red and blue triangles tracks).

Based on these observations, one might propose an interpretation for a $G$ scale, with three regions:

- $G \leq 0.35$: highly biased datasets; practically no information on distribution shape;
- $0.35 \leq G \leq 0.5$: no information due to possible cancellation between bias and shape; this could be considered as a "blind zone" of the $G$ scale;
- $0.5 \leq G$: heavy tails, either due to a strong asymmetry or to a large kurtosis; this might reveal a dominant role of outliers.

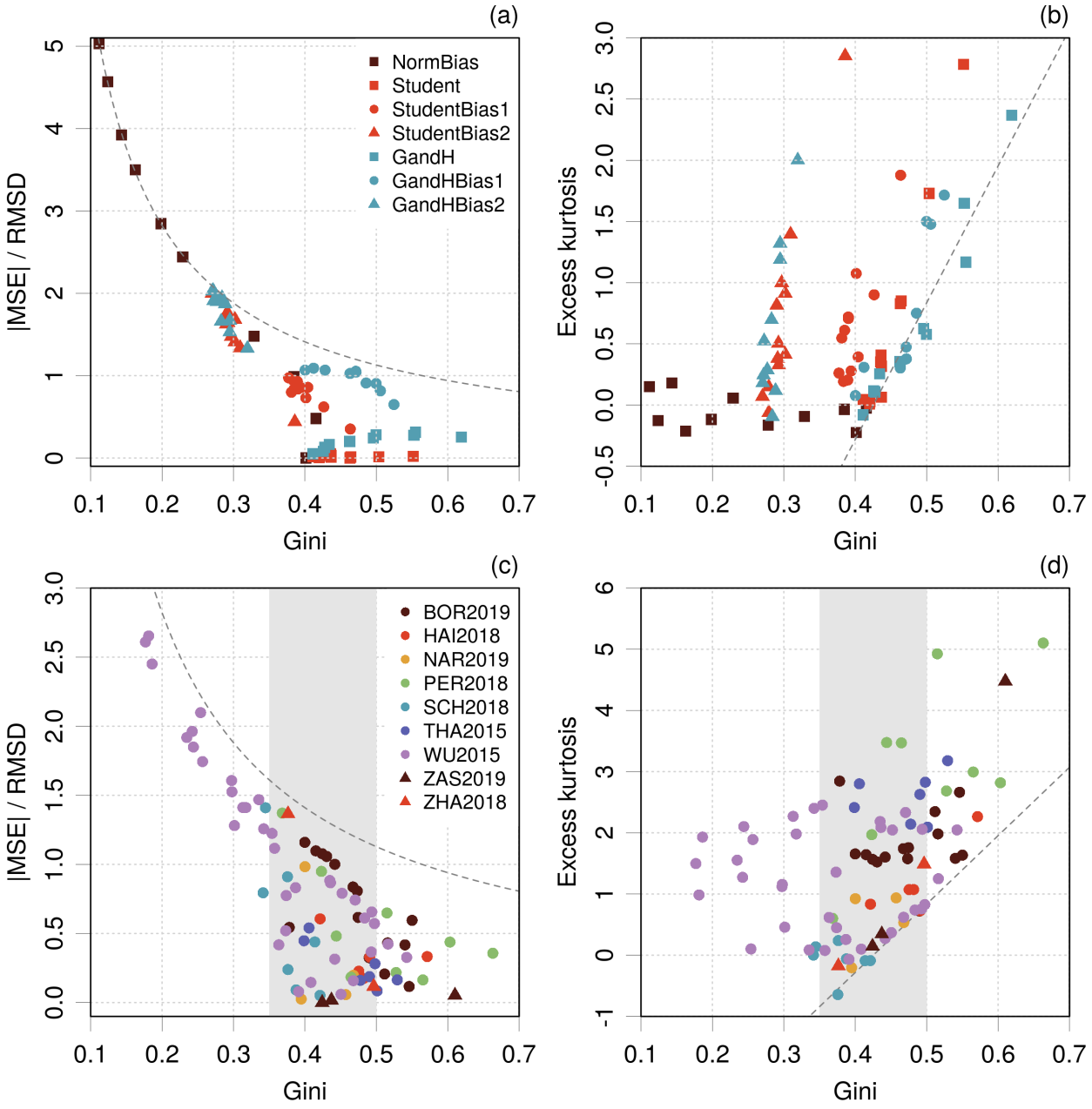


Figure 3: Scatterplots between the Gini coefficient and the ratio |MSE|/RMSD (a,c), and the excess kurtosis (b,d), for reference error sets (a,b) and computational chemistry literature error sets (c,d).

## 3.2 Literature datasets

In order to test the proposed $G$-scale interpretation, an analysis similar to the one for the reference datasets is performed on datasets gathered in the computational chemistry literature, in order to cover a wide spectrum of properties. These are described in Table 2, and strongly overlap with those studied in more details in Ref. [4]. The statistics, probability density functions and Lorenz curves corresponding to these datasets are provided in Supplementary Information.

Fig. 3(c,d) presents the scatterplots between $G$, $1/c_v$ and $\kappa_{CS}$ for theses datasets. One notes first

| Case | Property | $N$ | $K$ | Source |
|---|---|---|---|---|
| BOR2019 | Band gaps | 471 | 15 | [22] |
| HAI2018 | Dipole moments | 149 | 5 | [23] |
| NAR2019 | Enthalpies of formation | 469 | 4 | [24] |
| PER2018 | Intensive atomization energies | 222 | 9 | [2] |
| SCH2018 | Chemisorption energies | 195 | 7 | [25] |
| THA2015 | Polarizability | 135 | 7 | [26] |
| WU2015 | Polarizability | 145 | 34 | [27] |
| ZAS2019 | Effective atomization energies | 6211 | 3 | [28] |
| ZHA2018 | Solid formation enthalpies | 196 | 2 | [29] |

Table 2: Case studies: $N$ is the number of systems in the dataset and $K$ is the number of methods.

that the space covered by the data cloud in Fig. 3(c) is very similar to the one in Fig. 3(a), albeit with less extreme bias values. The $1/(\sqrt{\pi}G)$ curve seems to provide a reliable upper boundary in the ($G$,$1/c_v$) plane.

Some datasets fall fully within the "gray" $G$ zone, between 0.35 and 0.5 (NAR2019, ZHA2018), but most have some elements outside, that would require some attention. For instance, errors on polarizabilities in the WU2015 dataset (purple dots) present many methods with large bias, and a few with no bias and heavy tails. As discussed in the source paper [27], a non-negligible part of the reference dataset of experimental static polarizabilities presents large measurement uncertainties, which can be an important source of outliers. As one can see in Fig. 3(d), most of the error sets in this case present values of $\kappa_{CS}$ above 1.0. It therefore is probable that points in the gray area benefit from bias/outliers cancellation.

Cases BOR2019, HAI2018, PER2018, THA2015 and ZAS2018 present cases with large $G$ values. Predictions with the corresponding methods should probably be handled with circumspection, as some errors might stray far away from the MUE, at least more than one might expect from a normal distribution. An in-depth study has been published for case ZAS2019 [5], where the errors distribution for one of the methods was shown to have problematic large tails, despite having the smallest MUE amongst the three studied methods.

On Fig. 3(d), one can see that many datasets have large excess kurtosis, independently of their $G$ value. However, as seen for the reference datasets, there seems to be a limit in the ( $G, \kappa_{CS}$) plane showing that large $G$ values imply large excess kurtosis ($G > 0.5 \Rightarrow \kappa_{CS} > 1$).

**Ranking.** In order to evaluate the interest of the $G$-scale in practical scenarios, one might consider it as an alert mechanism to complement a MUE ranking. Fig. 4 shows how, for each dataset, the methods with best ranking are tagged by their $G$ value. Using the uncertainty on $G$ estimated by bootstrapping, the methods are flagged if either $P(G \leq 0.35) \geq 0.05$ or $P(G \geq 0.5) \geq 0.05$, based on a normal distribution hypothesis for $G$.

If one considers the first rank, about half of the methods are in the blind zone, but five methods

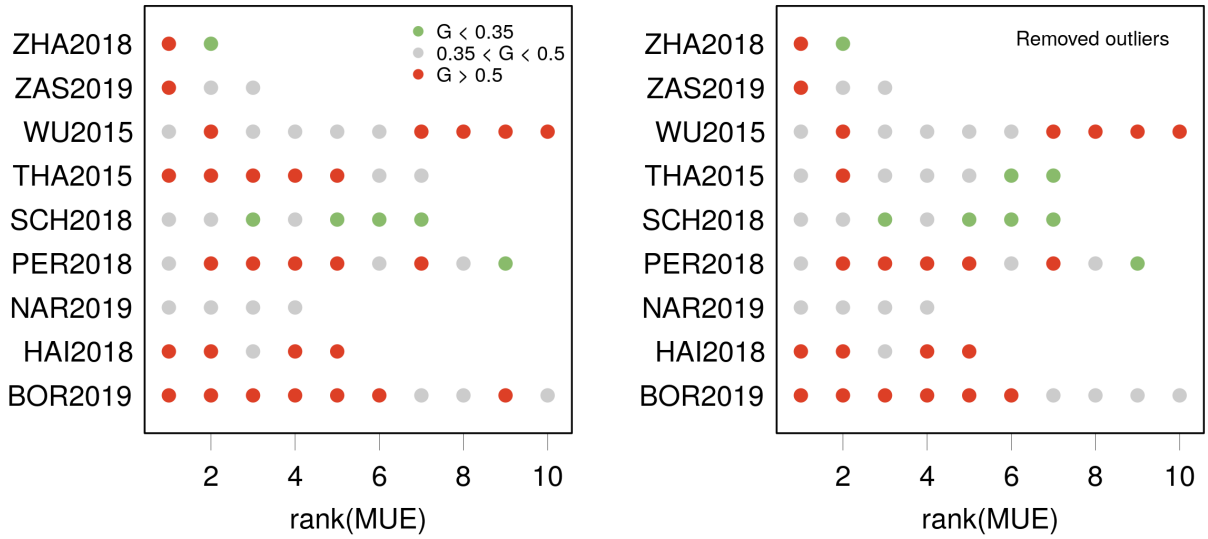


Figure 4: Gini-scale classification of 10 lowest MUE-ranked methods in each case. Red dots depict methods with large $G$ values, green dots methods with small $G$ values, and gray dots methods in the "blind zone".

are flagged as heavy-tailed. In case ZAS2019, the flagged method has very heavy tails, as has been reviewed in previous studies [4, 5]. For case THA2015, the five lowest ranked methods have $G$ values around 0.5, which might point to the existence of global outliers in the reference dataset, *i.e.* data that no method in the panel is able to predict correctly. This can be checked in Fig. 4(b), where the same analysis has been performed after the removal of global outliers, if any. Removal of global outliers (6 systems) for THA2015 changes remarkably the situation, with much less red-flagged methods. In the other cases, no effect is observed, which might be due to either the absence of global outliers (HAI2018, NAR2019, PER2018, SCH2018, WU2015), or the intrinsic shape of the distributions (BOR2019, SAZ2019, WU2015). In all these cases, it might be worth to investigate if the distorted shape of the distribution is due to systematic trends in the errors, as they could be profitably corrected [30, 1, 31, 32, 33].

# 4 Conclusion

The Lorenz curve and Gini coefficients present an interesting addition to the computational chemistry benchmarking statistical toolbox. We focused here on the Gini statistic to study its properties in relation with features of error distributions, such as bias and shape (kurtosis, asymmetry). We have shown that one can establish a scale of $G$ values, with tree distinct zones. Values of the Gini coefficient outside the interval $0.35-0.5$ should raise an alert, as the corresponding datasets present features that might imped the unambiguous interpretation of popular benchmarking statistics such

as the MUE.

# References


[1] 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.

[2] 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.

[3] P. Pernot and A. Savin. Probabilistic performance estimators for computational chemistry methods: Systematic improvement probability and ranking probability matrix. I. Theory. *J. Chem. Phys.*, 152:164108, 2020. doi:10.1063/5.0006202.

[4] P. Pernot and A. Savin. Probabilistic performance estimators for computational chemistry methods: Systematic improvement probability and ranking probability matrix. II. Applications. *J. Chem. Phys.*, 152:164109, 2020. doi:10.1063/5.0006204.

[5] P. Pernot, B. Huang, and A. Savin. Impact of non-normal error distributions on the benchmarking and ranking of Quantum Machine Learning models. *Mach. Learn.: Sci. Technol.*, 1:035011, 2020. doi:10.1088/2632-2153/aba184.

[6] M. Bonato. Robust estimation of skewness and kurtosis in distributions with infinite higher moments. *Finance Research Letters*, 8:77–87, 2011. doi:10.1016/j.frl.2010.12.001.

[7] M. O. Lorenz. Methods of measuring the concentration of wealth. *Publications of the American Statistical Association*, 9:209–219, 1905. doi:10.1080/15225437.1905.10503443.

[8] C. Gini. *Variabilità e mutabilità*. 1912.

[9] C. Damgaard and J. Weiner. Describing inequality in plant size or fecundity. *Ecology*, 81:1139–1142, 2000. doi:10.2307/177185.

[10] I. I. Eliazar and I. M. Sokolov. Measuring statistical heterogeneity: The pietra index. *Physica A*, 389:117–125, 2010. doi:10.1016/j.physa.2009.08.006.

[11] C. Kleiber. The Lorenz curve in economics and econometrics. techreport, TU Dortmund, March 2005. doi:10.17877/DE290R-14481.

[12] R. B. Bendel, S. S. Higgins, J. E. Teberg, and D. A. Pyke. Comparison of skewness coefficient, coefficient of variation, and Gini coefficient as inequality measures within populations. *Oecologia*, 78:394–400, 1989. doi:10.1007/BF00379115.

[13] D. Ruppert. What is kurtosis?: An influence function approach. *The American Statistician*, 41:1, 1987. doi:10.2307/2684309.

[14] K. Suaray. On the asymptotic distribution of an alternative measure of kurtosis. *Int. J. Adv. Stat. Proba.*, 3:161–168, 2015. doi:10.14419/ijasp.v3i2.5007.

[15] E. L. Crow and M. M. Siddiqui. Robust estimation of location. *Journal of the American Statistical Association*, 62:353–389, 1967. doi:10.2307/2283968.

[16] A. Zeileis. *ineq: Measuring Inequality, Concentration, and Poverty*, 2014. R package version 0.2-13. URL: https://CRAN.R-project.org/package=ineq.

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

[18] P. M. Dixon, J. Weiner, T. Mitchell-Olds, and R. Woodley. Bootstrapping the gini coefficient of inequality. *Ecology*, 68:1548–1551, 1987. doi:10.2307/1939238.

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

[20] 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.

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

[22] 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.

[23] D. Hait and M. Head-Gordon. How accurate is density functional theory at predicting dipole moments? an assessment using a new database of 200 benchmark values. *J. Chem. Theory Comput.*, 14:1969–1981, 2018. doi:10.1021/acs.jctc.7b01252.

[24] B. Narayanan, P. C. Redfern, R. S. Assary, and L. A. Curtiss. Accurate quantum chemical energies for 133000 organic molecules. *Chem. Sci.*, 10:7449–7455, 2019. doi:10.1039/c9sc02834j.

[25] P. S. Schmidt and K. S. Thygesen. Benchmark database of transition metal surface and adsorption energies from many-body perturbation theory. *J. Phys. Chem. C*, 122:4381–4390, 2018. doi:10.1021/acs.jpcc.7b12258.

[26] A. J. Thakkar and T. Wu. How well do static electronic dipole polarizabilities from gas-phase experiments compare with density functional and MP2 computations? *J. Chem. Phys.*, 143:144302, 2015. doi:10.1063/1.4932594.

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

[28] 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.

[29] Y. Zhang, D. A. Kitchaev, J. Yang, T. Chen, S. T. Dacek, R. A. Sarmiento-Perez, M. A. L. Marques, H. Peng, G. Ceder, J. P. Perdew, and J. Sun. Efficient first-principles prediction of solid stability: Towards chemical accuracy. *npj Comput. Mater.*, 4:9, 2018. doi:10.1038/s41524-018-0065-z.

[30] 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.

[31] K. Lejaeghere, L. Vanduyfhuys, T. Verstraelen, V. V. Speybroeck, and S. Cottenier. Is the error on first-principles volume predictions absolute or relative? *Comput. Mater. Sci.*, 117:390–396, 2016. doi:10.1016/j.commatsci.2016.01.039.

[32] J. Proppe, T. Husch, G. N. Simm, and M. Reiher. Uncertainty quantification for quantum chemical models of complex reaction networks. *Faraday Discuss.*, 195:497–520, 2016. doi:10.1039/c6fd00144k.

[33] 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.