Available online at www.sciencedirect.com

ScienceDirect

Focus Article, *Mendeleev Commun.*, 2018, **28**, 225–235

Mendeleev Communications

Interplay between test sets and statistical procedures in ranking DFT methods: the case of electron density studies

Alexander A. Marjewski,[a,b] Michael G. Medvedev,*[a,c] Igor S. Gerasimov,[a,c,d] Maria V. Panova,[c] John P. Perdew,[e,f] Konstantin A. Lyssenko*[a] and Artem O. Dmitrienko*[a]

[a] *A. N. Nesmeyanov Institute of Organoelement Compounds, Russian Academy of Sciences, 119991 Moscow, Russian Federation. E-mail: kostya@ineos.ac.ru, dmitrienka@gmail.com*
[b] *Higher Chemical College of the Russian Academy of Sciences, D. I. Mendeleev Univeristy of Chemical Technology of Russia, 125047 Moscow, Russian Federation*
[c] *N. D. Zelinsky Institute of Organic Chemistry, Russian Academy of Sciences, 119991 Moscow, Russian Federation. E-mail: medvedev.m.g@gmail.com*
[d] *National University of Science and Technology 'MISIS', 119490 Moscow, Russian Federation*
[e] *Department of Physics, Temple University, Philadelphia, PA, 19122, USA*
[f] *Department of Chemistry, Temple University, Philadelphia, PA, 19122, USA*

DOI: 10.1016/j.mencom.2018.05.001

The task of choosing a reliable density functional (DFT) approximation remains one of the most puzzling ones in quantum chemical modeling and materials simulations. Since DFT functionals are in general not systematically improvable, benchmarking them on specifically designed test sets is the usual way for identifying a method best suited for a particular purpose. To get an answer from a bunch of numbers, statistical analysis should be applied. In this article the possibilities and pitfalls of statistical error analysis are discussed, taking as an example the ranking of approximate functionals by the accuracy of their self-consistent electron densities, which were recently shown to have worsened in the last decade.

Size of a functional is proportional to its accuracy for electron densities

Size of a functional is proportional to predictability of its errors in electron densities

Alexander A. Marjewski, an undergraduate student at Higher Chemical College of the Russian Academy of Sciences (RAS). His research interests comprise symmetry considerations in DFT computations, empirical fitting frameworks, and density cumulant functional theory.

Michael G. Medvedev was graduated from Higher Chemical College RAS (2016), currently he is a graduate student at A. N. Nesmeyanov Institute of Organoelement Compounds RAS. His research interests include quantum chemistry (in particular, density functional theory), reliability of molecular modeling studies, and mechanistic features of chemical reactions.

Igor S. Gerasimov, an undergraduate student at National University of Science and Technology 'MISIS'. His research interests are density functional theory and post-HF methods.

Maria V. Panova was graduated from Higher Chemical College RAS (2016), currently she is a graduate student at N. D. Zelinsky Institute of Organic Chemistry RAS. Her research interests are density functional theory, physical methods in chemistry and mechanistic insights into carbohydrates chemistry.

John P. Perdew received his Ph.D. at Cornell University, and did postdoctoral work at the University of Toronto and Rutgers University. He was a Professor of Physics at Tulane University until 2013, when he became the Laura H. Carnell Professor of Physics and Chemistry at Temple University. His research interests are density functional theory, materials theory, and quantum chemistry.

Konstantin A. Lyssenko received his Ph.D. (2001) and Dr.Sci. (2007) degrees at A. N. Nesmeyanov Institute of Organoelement Compounds RAS. In 2015 he became a Professor of the RAS. His scientific interests include the analysis of electron densities reconstructed from high-resolution X-ray diffraction studies, nature of chemical bonding and supramolecular organization.

Artem O. Dmitrienko was graduated from M. V. Lomonosov Moscow State University (2011) and received his Ph.D. degree (2015) at A. N. Nesmeyanov Institute of Organoelement Compounds RAS. His research interests are statistical methods in chemistry and crystallography, database analysis, accuracy and reliability of geometric and charge density parameters obtained from diffraction experiments and quantum chemistry calculations.

Introduction

Density functional theory (DFT) is a current workhorse of quantum chemical modeling.[1] It is in principle an exact theory, as was proved[2] by Hohenberg and Kohn. According to their theorems, an 'exact functional' exists, which returns exact electronic energy of any system from its exact electron density. The exact electron density, in its turn, can be found by minimizing the electronic energy at fixed electron number and external potential. The exact functional, however, has an unbearable computational cost[3] and cannot be used for practical purposes. This has led to the construction of many hundreds of approximate density functionals intended to be practically useful. The most popular group of them is based on the Kohn–Sham approach,[4] where only a small part (~10%) of the total energy is approximated; this part is called the exchange-correlation energy and is computed by an approximate exchange-correlation functional.

In general, there is no systematic way to determine which of the available Kohn–Sham approximate functionals will do better for a particular task. Thus, a practical way to determine a functional suitable for a given purpose is to benchmark a series of functionals against experimental or high-level theoretical values to determine an 'average' behavior. There are two 'tunable' parameters in such benchmarks: the chemical test set and the metric. In an ideal case, the test set should consist of the whole 'chemical space'. Then a fair metric will be simply the mean error of a method over the chemical space for the property of interest. In reality, however, we have to balance the higher cost of a larger test set against its greater ability to lead to general conclusions. The lack of generalization power of a smaller test set can be compensated by choosing a more sophisticated metric, considering the strengths and weaknesses of the actual test set. Therefore, although seemingly easy, benchmarking requires a deep understanding of the interplay between the statistical procedure and the test set to come to a meaningful conclusion. In this Focus article we discuss this interplay and its effect on the conclusions, taking as an example the recently revived question of the accuracy of density functionals for self-consistent (*i.e.*, converged) electron densities.

A brief historical outline

The outcome of an electron density comparison study always relies on the choice of the way one compares electron densities, and this choice may be crucial to the results one obtains. Incognitant application of a metric may skew the data, providing a misleading output, from which misguided conclusions will be drawn. This kind of mistake is certainly not only a DFT or computational chemistry field problem. Its ubiquitous nature manifests itself throughout every field of science, but in the DFT case it is of absolute significance.

For comparison of electron densities produced by different DFT methods, the oldest approach was visual comparison of absolute density (or its moments) plots. The first DFT electron densities study[5] (Wang and Parr, 1977) presented plots of radial density distribution functions [$4\pi r^2\rho(r)$] of spherically symmetric systems. For more complex systems, absolute density difference contour diagrams[6–8] were employed for visual comparison. This kind of metric, however, does not provide a suitable framework for quantitative examination of differences. While visual examination might seem to be a quick and straightforward way to compare densities in different regions (and it is), it is not clear how to compare performances of many different functionals on different molecules with this approach. It is obvious that such studies would require merging the overall deviations into a single number for comparison to be feasible. Different solutions for this challenge were proposed;[7–18] the most notable are electron density parameters at critical points and on bond paths defined within the Quantum Theory of Atoms in Molecules (QTAIM[19]), QTAIM atomic charges, and the Quantum Molecular Similarity Measure (QMSM[20]), which is based on electron density geometric overlap.

However, most of these solutions did not make their way into modern times, being superseded by approaches which account for the whole electron density and are easily interpretable. Approaches to electron densities comparison can take many mathematical forms, thus creating a variety of metrics that find use in modern density studies. They can be divided into two groups, based on whether the electron density difference function should be computed:

(1) The group of integral metrics. To compute an integral error, one first computes some expectation value on approximate and reference electron densities and then subtracts the latter from the former. There are many expectation values, which can be applied for comparing electron densities in this way: dipole,[21,22] quadrupole and higher moments, 'moments of density'[23,24] $\langle r^n \rangle$, exchange-correlation energy of a trial DFT functional,[25] *etc*. A major pitfall of all integral metrics is the possible internal compensation of errors due to sign changes in the deviation function (*e.g.*, one can compare electron densities by means of $\langle r^0 \rangle$, but as it is equal to the number of electrons in a system, any approximate method would show a zero error with this metric, regardless of its electron density). However, this metric has an advantage of allowing comparison of electron densities between systems with different geometries; this may be useful, for example, for studying self-consistent equilibrium geometries of different functionals.

(2) The group of pointwise metrics. To compute a pointwise error, one first subtracts the reference electron density [or any other related real-space function (descriptor)] from the approximate one and then converts this difference function into a single number with an 'error measure'. Descriptors may vary from local electron density, its gradient and Laplacian to the left Fukui function,[26] electron localization function (ELF[27,28]), density overlap region indicator (DORI[29]), *etc*. In their turn, error measures used in conjunction with pointwise metrics can be either pointwise [*e.g.*, root-mean-square deviation (RMSD)] or integral [*e.g.*, integral of absolute (IAD) or squared (ISD) difference]. Pointwise metrics mitigate the internal compensation error, but limit comparisons to identical geometries and integration grids.

Recently some of us have found,[30] that many of the novel highly-empirical density functionals produce electron density distribution functions that are even worse than those produced by the oldest DFT methods. This was demonstrated on electron densities of fourteen atomic species (Be^0, B^{3+}, B^+, C^{4+}, C^{2+}, N^{5+}, N^{3+}, O^{6+}, O^{4+}, F^{7+}, F^{5+}, Ne^{8+}, Ne^{6+}, Ne^0); these systems are appropriate norms,[31,32] so even semilocal functionals can be highly accurate for them. We call this a test set of 'edge cases', because it is not balanced but the atomic species constituting it possess different difficulties for density functionals, so only a very reliable functional can do well for all of them. In total, 128 DFT methods were tested using aug-cc-pωCV5Z[33,34] basis set and a dense integration grid. [Later it was shown[35] that errors in electron densities do not change even if a huge (999,974) grid is used]. The root-mean-square deviations (RMSD) between radial distribution functions of approximate and reference (CCSD-full[36,37]) electron density descriptors values were calculated as follows:

$$\mathrm{RMSD} = \mathrm{root}(\mathrm{mean}(\mathrm{square}\{g[P_{\mathrm{approx}}(r)] - g[P_{\mathrm{ref}}(r)]\})) =$$

$$= \sqrt{\frac{1}{N}\sum_{i}^{N}\{\pi r_i^2[P_{\mathrm{approx}}(r_i) - P_{\mathrm{ref}}(r_i)]\}^2}\,.$$

Here P_{ref} and P_{approx} are reference and approximate local values of descriptors [local electron density (RHO), its gradient norm (GRD) and Laplacian (LR)], and r_i is the radial point at which

the radial density was calculated. These descriptors are standard ingredients in approximate density functionals; although kinetic energy density is the usual ingredient in meta-GGAs,[38] recently it has been confirmed[39] that it can be safely replaced by a combination of RHO, GRD and LR.

To put errors in different descriptors on the same scale, the RMSD error values were normed to the median error for the given descriptor:

$$\mathrm{MNAE}_{P,a,f} = \frac{\mathrm{RMSD}_{P,a,f}}{\underset{a,f}{\mathrm{median}}\ \mathrm{RMSD}_{P,a,f}}.$$

Here *P* stands for the descriptor, *a* stands for atom and *f* stands for functional. The median error was chosen for normalization because each method makes different errors in different descriptors, so relying on errors of any particular method may make the data unrepresentative (see the discussion below for illustrations). Also, median errors are more stable in the sense that they do not usually change significantly upon addition of several new, even strictly outlying, data points (functionals or systems, in this case). Then, to distinguish methods, which: (1) have acceptable worst-case behavior and (2) work well on average, two additional error measures (not named in the original article[30]) were introduced: maximum error over the atomic species and descriptors (max-max-MNAE or mmMNAE) and maximum error over descriptors after averaging errors over systems (max-average-MNAE or maMNAE):

$$\mathrm{mmMNAE}_f = \max_{P,a}(\mathrm{MNAE}_{P,a,f})$$

$$\mathrm{maMNAE}_f = \max_{P}[\mathrm{mean}_{a}(\mathrm{MNAE}_{P,a,f})].$$

Functionals which are among the best 25% by both mmMNAE and maMNAE were considered as being accurate for electron densities of the studied systems, while those among the worst 25% by both, were considered as inaccurate.[30]

In Figure 1 we plot the mmMNAE values against the year in which a functional was published to provide a better visualization for the conclusions reached in our paper.

Since then, a few studies[21,22,24–26,41–44] have tested and expanded our conclusions; many metrics were applied to measure differences between electron densities, and many statistical procedures were used for normalization and averaging errors over different metrics and chemical systems. Here we will focus on statistical analyses that deal with spherically symmetric electron density or its natural generalizations; a detailed discussion of specific choices on weighting different points of space is beyond the scope of this article.

The first consequent study where a pointwise error measure was applied to the whole system was published by Gould.[26] He criticized the approach of Medvedev *et al.*,[30] pointing out that (1) the applied metric accounts for overall density, which might not be reproduced well by approximations that sacrifice this as unimportant for the sake of reaching good frontier (and so bonding) densities, and (2) the test set is, in his opinion, significantly biased towards functionals designed to incorporate good behavior at the high-density limit. To eliminate the said biases, he suggested comparing left Fukui functions, which are defined as:

$$f^{-}(r) = \rho^{\mathrm{atom}}(r) - \rho^{\mathrm{cation}^{+}}(r).$$

Here $\rho^{\mathrm{atom}}(r)$ and $\rho^{\mathrm{cation}^{+}}(r)$ are the atom and its singly-positively charged cation, respectively. The left Fukui function can be considered as an instance of a descriptor in this article (referred to as 'Left_Fukui'). Application of the left Fukui function, as by Gould, is justified, as it adheres to the spirit of DFT applications by relying on relative properties instead of absolute ones and is approximately equal to the HOMO density in the framework of Kohn–Sham theory. Deviations of absolute electron densities were also included in the study. The Li, F and C atoms were chosen as the test systems because of the interest they pose for chemistry.

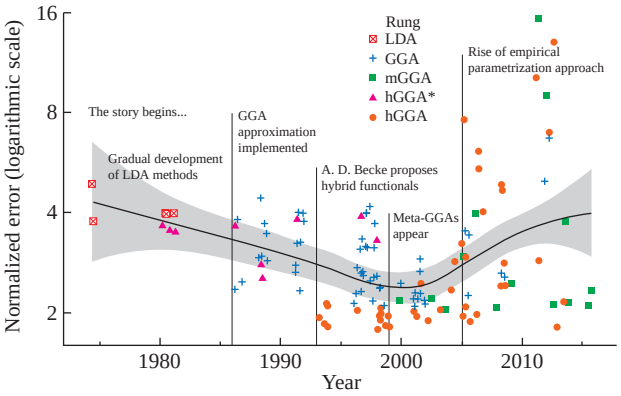

Figure 1 The plot of density-functional errors in electron densities (specifically, mmMNAEs as defined in the text) against the year in which a functional [or its 'final component' (*e.g.*, 'PBE exchange' or 'three-parameter hybridization scheme'), if there is no particular reference] was published. 'LDA' denotes local density approximations (first rung of the Jacob's ladder[40]), 'GGA' denotes generalized gradient approximations (second rung), 'mGGA' denotes meta-GGA functionals (third rung), 'hGGA' denotes hybrid functionals (fourth rung) and 'hGGA*' denotes hGGAs with 100% exact exchange, for which the exchange and correlation components are strongly misbalanced. The black curve shows the rolling average deviation, while the light grey area denoting its 95% confidence interval.

Gould used an integral of squared deviation (ISD) error measure to define errors in electron densities and left Fukui functions:

$$\mathrm{Err}^2_{P,f} \propto 4\pi \int [P_f(r) - P_{\mathrm{ref}}(r)]^2 r^2 \mathrm{d}r.$$

Here *P* stand for descriptor (RHO or Left_Fukui). The most intriguing outcome of the study was the observation that a functional that reproduces left Fukui functions accurately does not necessarily provide good electron densities, and *vice versa*. This finding will be rationalized in the following section.

A major improvement upon the methodology of Medvedev *et al.* was proposed by Mezei *et al.*,[25] who suggested a normed integral of absolute difference (NIAD), which is applicable to molecular systems and is intensive by means of putting the number of electrons in a studied system into the denominator:

$$\mathrm{NIAD}_{P,\mathrm{system},f} = \frac{1}{N} \iiint |P_f(r) - P_{\mathrm{ref}}(r)| \mathrm{d}^3 r.$$

Here *P* stands for descriptor (electron density, its gradient norm or Laplacian) and *f* for functional. The relative average normed error (RANE) was suggested by Mezei *et al.* to put the different errors on the same scale by normalizing them to the error of LDA (SVWN):

$$\mathrm{RANE}_{P,f} = \frac{\sum_{\mathrm{system}} \dfrac{\mathrm{NIAD}_{P,\mathrm{system},f}}{N_{\mathrm{system}}}}{\sum_{\mathrm{system}} \dfrac{\mathrm{NIAD}_{P,\mathrm{system,LDA}}}{N_{\mathrm{system}}}}.$$

The overall method error for three descriptors was defined as their mean value. Concerning atomic densities of the 'edge case' set, this study confirmed general conclusions of Medvedev *et al.*, although displaying some deviations, the most notable of which is an outstanding performance of HF method. As will be shown further, this is the result of applying a statistical procedure constructed for a balanced test set to the 'edge cases' set. Notably, the conclusions of Mezei *et al.* on a balanced molecular test set are in better agreement with the conclusions of Medvedev *et al.*

It is worth noting, that along with RANE, Mezei *et al*. suggested an integral error – density-driven exchange-correlation (DDXC) error. DDXCs and RANEs for different functionals were found to correlate only weakly, although expectedly displaying linear correlation for *ab initio* methods. Many interesting speculations might be made using the data from this paper, but we leave it to be discussed somewhere else.

Interplay between test sets and statistical procedures

In the article of Medvedev *et al*.,[30] there could be about 1.5×10^{226} possible rankings – all permutations of 133 quantum chemical methods; and it is not so hard to invent a comparison method that results in any pre-chosen one. It is clear then that ranking itself – without an understanding of metrics used to calculate it – is completely meaningless. It can be thought of as metrics asking a question to the data and the resulting ranking being an answer. Most of the questions are dumb, as well as the corresponding answers. However, a few can in some cases help to answer The Question 'Which functional should I use?', or at least tell us something useful about functional performance. The Question is disastrously vague, metrics are terribly specific; no surprise that we need several of them to get an insight.

How are useful metrics constructed? At first, one has a set of computed values – in different descriptors, on different systems – associated with every method, and one's goal is to collapse them into a single numeric score per method. For the sake of flexibility, we shall divide this process into four consecutive steps: measuring errors, scaling, normalization, and summarizing over systems and descriptors. Let's discuss these steps in detail.

Measuring errors. The difference between two descriptor functions at each point of space is a good function to plot (especially, for spherical atomic systems). However, for further analysis, one needs to collapse it into one number, which will be an error of the approximation in the used descriptor on the given system. For electron density descriptors this can be (and was) done in different ways in the literature. Regardless of the other notable attempts, in 2017 three distinct pointwise error measures were in favour: RMSD used by Medvedev *et al*.[30] and Wang *et al*.,[43] ISD used by Gould[26] and IAD used by Mezei *et al*.[25] The corresponding formulas are:

$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{i}^{N}\{\pi r_i^{\,2}[P_{\mathrm{approx}}(r_i) - P_{\mathrm{ref}}(r_i)]\}^2},$$

$$\mathrm{ISD} = \iiint [P_{\mathrm{approx}}(r) - P_{\mathrm{ref}}(r)]^2 \mathrm{d}^3 r,$$

$$\mathrm{IAD} = \iiint \left| P_{\mathrm{approx}}(r) - P_{\mathrm{ref}}(r) \right| \mathrm{d}^3 r.$$

Here $P(r)$ denotes descriptor value at the point r. Obviously, ISD and IAD measures are applicable to any molecular system, while RMSD is only applicable to atomic systems because it requires a 'centre' relative to which r is measured. To understand the differences between these error measures, it would be helpful to note that all of them are related to distinct 1D integrals when applied to a spherically-averaged $P(r)$ of an atomic system:

$$\mathrm{RMSD}^2 r_{\max} = \int_0^{r_{\max}} \{\pi r^2 [P_{\mathrm{approx}}(r) - P_{\mathrm{ref}}(r)]\}^2 \mathrm{d}r,$$

$$\mathrm{ISD} = \int_0^{\infty} \pi r^2 [P_{\mathrm{approx}}(r) - P_{\mathrm{ref}}(r)]^2 \mathrm{d}r,$$

$$\mathrm{IAD} = \int_0^{\infty} \pi r^2 \left| P_{\mathrm{approx}}(r) - P_{\mathrm{ref}}(r) \right| \mathrm{d}r.$$

Here $r_{\max}$ is the maximum radius up to which RMSD is calculated.

Figure 2 provides the integrands of these expressions for deviations in four descriptors [RHO, GRD and LR from Medvedev *et al*.[30] and left Fukui function (Left_Fukui) from Gould[26]] of PBE0[45,46] from CCSD-full in the aug-cc-pωCV5Z basis set for beryllium, lithium and neon atoms. All integrands are normalized to their maximum values.

From Figure 2 it is clear that the ISD measure is localized very close to the nucleus and thus seriously overweights the core

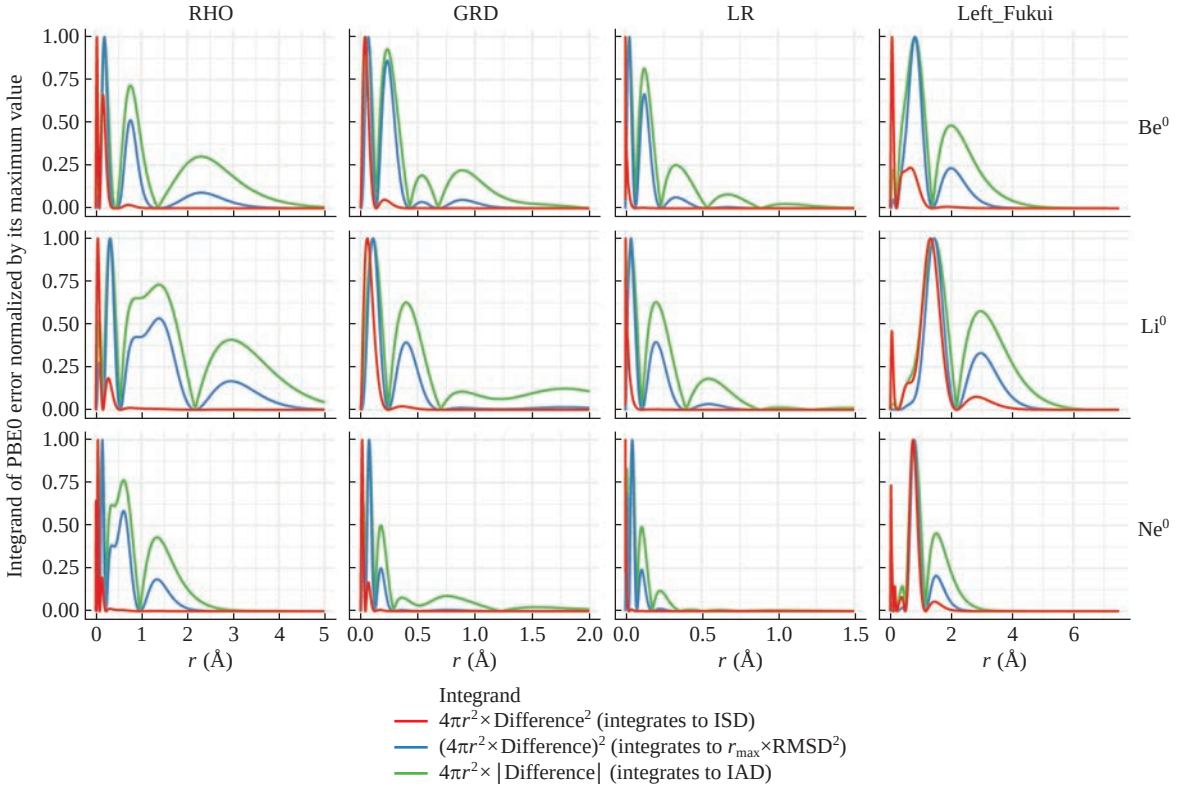

Figure 2 Integrands of RMSD, ISD and IAD for the deviation of PBE0[45,46] descriptors from CCSD-full ones in the aug-cc-pωCV5Z basis set for the beryllium, lithium and neon atoms.

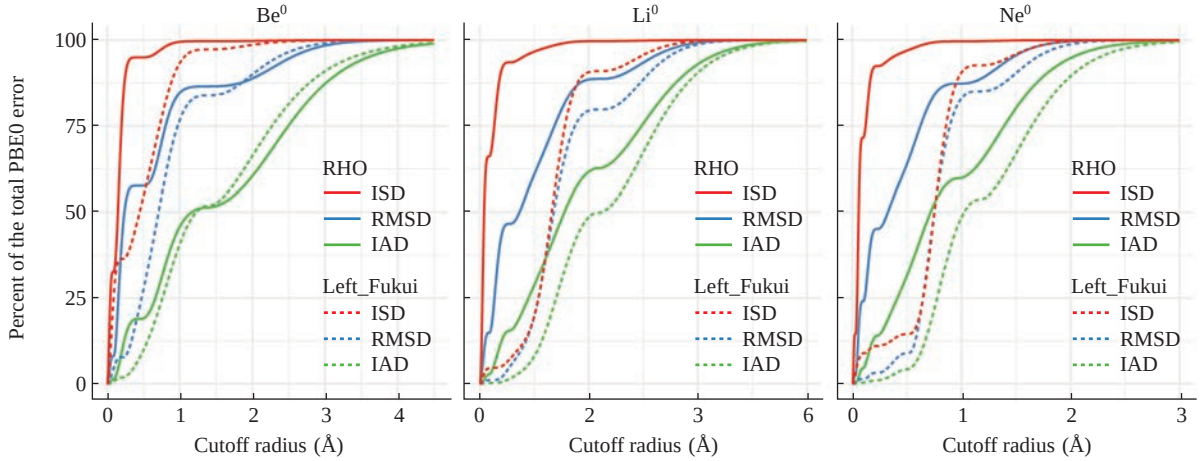

Figure 3 Growth of integrated error (in percents) with cutoff radius for RHO and left_Fukui function descriptors using three error measures (ISD, RMSD and IAD) for the PBE0[45,46] functional on three atomic systems (the Li, Be and Ne atoms).

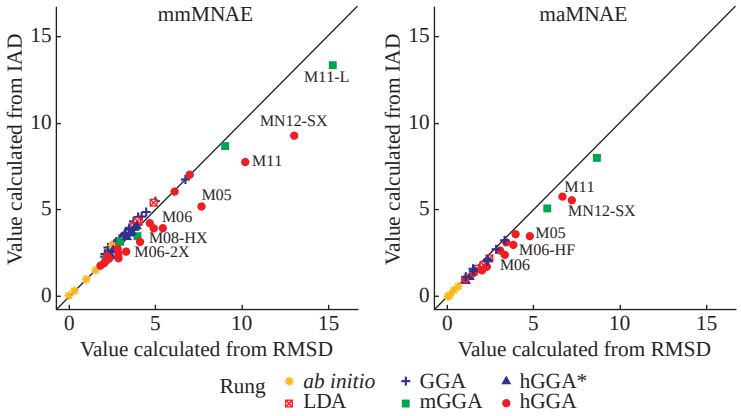

Figure 4 Correlation between mmMNAE and maMNAE values for methods in the Medvedev *et al.*[30] article calculated from RMSD and IAD errors. Diagonal lines on each plot have a slope equal to 1. Functionals for which differences between IAD- and RMSD-based mmMNAEs or maMNAEs are greater than 0.75 are labelled.

region. The ISD measure was used in the study by Gould, who applied it to local electron density and left Fukui function and found no correlation between them, which does not seem strange anymore given the behavior of ISD. To illustrate this, we plot a percent of the final error (reached at 10 Å) gained by a particular error measure for RHO or Left_Fukui descriptor at the distance r from the nucleus for three atoms (Li, Be and Ne) in Figure 3. According to it, ISD(RHO) reaches 95% of its final value at ~0.5 Å for all three atoms; at the same time, at this distance ISD(Left_Fukui) reaches only about 40% on Be and less than 15% on Li and Ne. Applying the ISD to the left Fukui function, thus, seems to be reasonable. With the RMSD error measure, errors in RHO and Left_Fukui grow more simultaneously, and using IAD error measure makes them behave almost synchronously (Figure 3). The left Fukui function applied by Gould is a good descriptor to use in electron densities analysis, but it would weigh the outer electron density better if applied with RMSD or IAD error measures.

The RMSD integrand is about $\pi r^2 \times$ISD one, so it puts visibly more weight to the valence region. The IAD integrand is very similar to the RMSD one but weights the valence region even more (see Figures 2 and 3). As IAD measure is also applicable to molecular systems, it is an advance over the RMSD measure. But how different are RMSD and IAD measures? Could the conclusions of the article of Medvedev *et al*. change if the IAD measure were used? To answer this question, we have calculated mmMNAE and maMNAE (defined above) from IAD and RMSD errors and plotted them against each other (Figure 4).

Figure 4 shows that for most methods mmMNAEs and maMNAEs calculated basing on IAD or RMSD errors are very close, and in many cases identical. The exceptions are, interestingly, some Minnesota functionals, for which mmMNAE and maMNAE values calculated from IAD are usually smaller than those calculated from RMSD. There are two possible explanations for

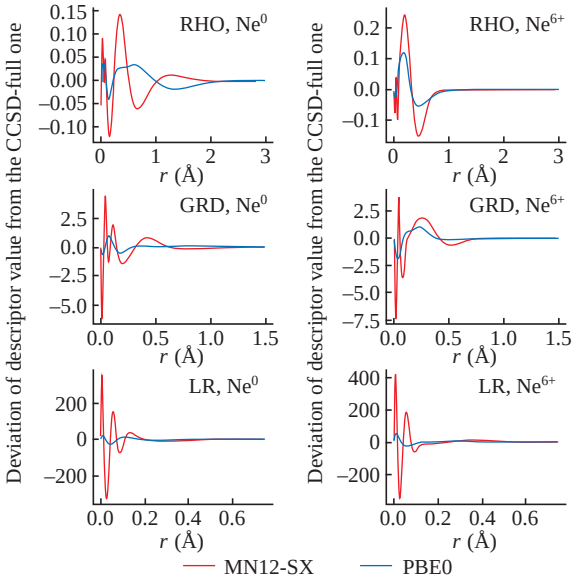

Figure 5 Deviations of electron densities produced by MN12-SX[47] (which has the biggest difference between mmMNAEs calculated from RMSD and IAD) and PBE0[45,46] methods from the CCSD-full ones in the aug-cc-pωCV5Z basis set for Ne^0 and Ne^{6+} systems by means of RHO, GRD and LR descriptors.

this behavior: (1) these Minnesota functionals are more accurate in the bonding region where IAD has greater weight, or (2) these functionals have more oscillating electron differences than on average (*i.e.*, have many narrow but significant deviations), so that squaring in the RMSD formula magnifies their errors. Figure 5 justifies the second hypothesis. Thus, the broad conclusions of the article of Medvedev *et al.*[30] also hold with the IAD error measure, as was also noted by Mezei *et al.*[25]

Scaling step. Next, one may expect that errors in different systems should be reweighted in accordance with their physical properties to make the comparison fairer. From Figure 6 we see, that, indeed, different systems have different error scales – that is, some systems are systematically easier than others. So, easy systems will contribute less to the final score. It may be exactly the behavior we want: if a system is the source of only a small error, then it must contribute less. But we may also decide that a good method should outperform others on easy systems too. Dividing all individual errors by the nuclear charge allows putting more weight on atoms with small nuclear charge, *i.e.*, where electrons are further from the high-density limit. Alternatively, dividing individual errors by the number of electrons in a system allows one to work with per-electron errors, as was done by Mezei *et al*. The effect of scaling on relative IAD and RMSD errors of different atoms is shown in Figure 6.

According to Figure 6, while errors in RHO scale with the electron number, clearly forming a staircase, the errors in GRD and LR heavily depend on the nuclear charge. Moreover, as IAD error measure puts more weight on the outer electron density, which is especially good for 2-electron systems, the gap between two- and four-electron systems is much larger for IAD than for RMSD. So, scaling by electron number nearly does its job (equalizes

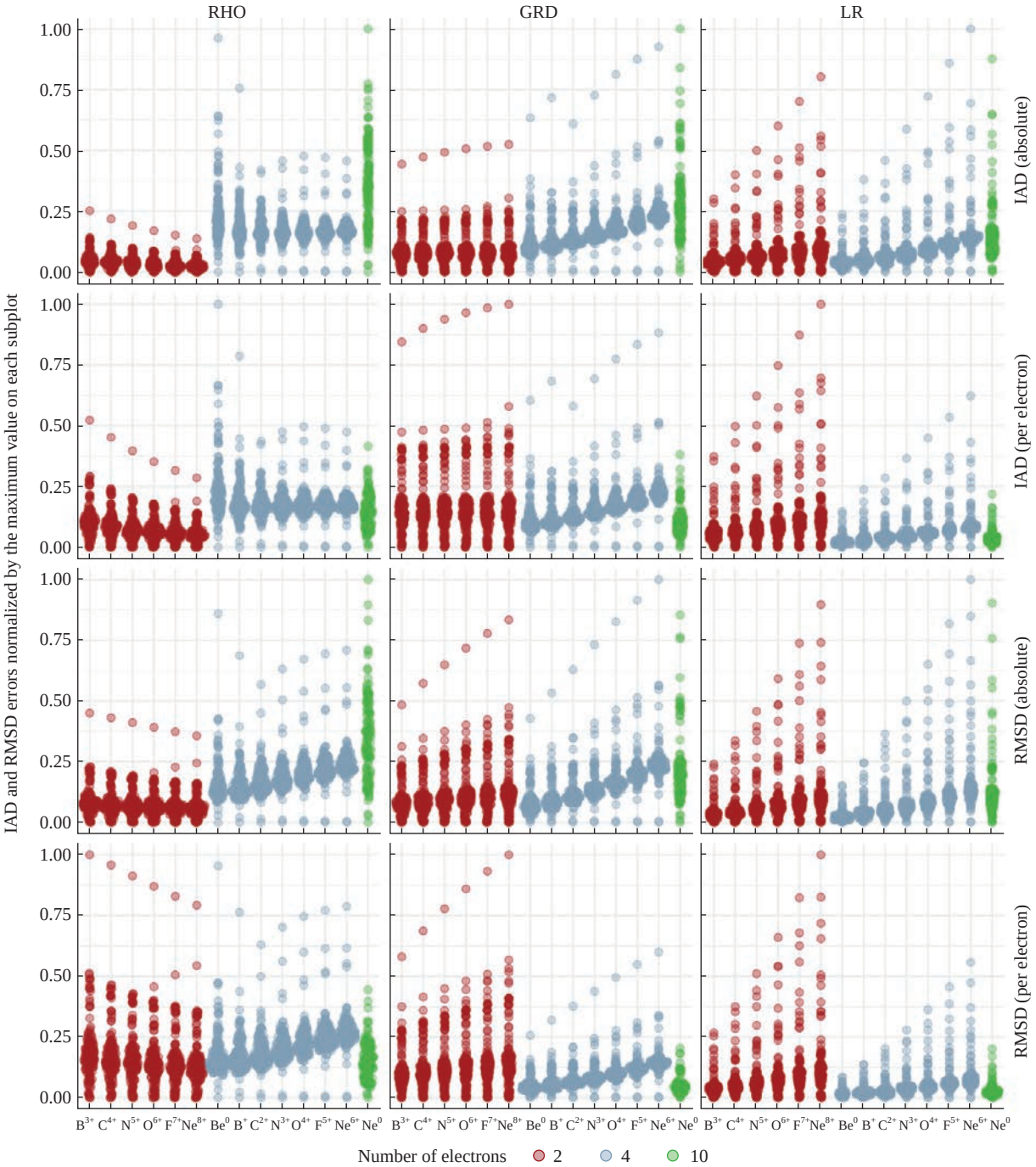

Figure 6 IAD and RMSD errors in different atomic-descriptor combinations with and without scaling by electron number. Points jitter along the *x*-axis is proportional to the density distribution of the data points at a given error value (*i.e.*, the wider jitter corresponds to higher data points accumulation).

median errors for atoms) only in the case of RMSD(RHO). An alternative way to strictly equalize contributions of different atoms to the score is to normalize errors for each combination of system and descriptor separately as discussed in the next section.

Normalization step. In most cases, we are more interested in energy than in electron density. In the ideal case, we could judge errors in electron density and its derivatives according to their effect on the error in energy. As GGA and meta-GGA methods are generally more accurate than LDA, we can conclude that derivatives matter; however, we have no clue to how large their contributions are in comparison to that of the local electron density. A sensible 'zero guess' is that electron density, its gradient and Laplacian contribute equally to the method quality. The normalization step aims to reflect this guess in the metrics.

Unfortunately, RHO, GRD and LR have directly incomparable scales. If we just 'average everything', our metrics will be dominated by the LR error as it is much bigger in absolute value – no matter how well a method reproduces RHO and GRD, it will win if its LR is good. The simplest way to address the problem is to rescale the range of each descriptor to [0,1]. This is achieved by dividing each descriptor by its maximum value. Depending solely on a single maximum value, this normalization approach is extremely vulnerable to outliers. In our case, it results in a severe underestimation of the Laplacian contribution (Figure 7). Mean normalization is less affected by outliers, as it depends on the whole dataset, not on a single value. The median is considered more robust than the mean, as it ignores the outlying values completely. In our case median and mean normalized data are nearly identical; both effectively equalize the contributions of RHO and its derivatives.

Another approach to normalization is to use errors of a particular method instead of those of an 'average' functional. It makes one method special and allows one to examine others in comparison with it. While the interpretation differs significantly, in a practical sense, normalizations in subgroups defined by descriptor based on DFT methods such as LDA,[48] PBE,[49] SCAN,[32,50] PBE0[45,46] and B3LYP[51] result in values very close to those of the mean- or median-based normalizations (Figure 8). Using HF errors as a reference does show some differences – it increases the amount of the LR contribution; on the other hand, using M06-2X[52,53] or M11[54] methods as references magnifies the RHO contribution (Figure 8).

There is also a subtler issue that can be solved with normalization. If we decide that a good method should outperform others on easy systems too, then we can normalize each combination of system and descriptor separately to ensure their equal contributions to the score. While, on a large and balanced (representative) test set, equalizing different systems should make the ranking more universally applicable, on the 'edge cases' test set it has an undesirable effect: since the test set contains 6 (out of 14) two-electron systems, the main effect of equalizing systems is overweighting of two-electron systems.

Summarizing step. Now, when all the errors are reasonably weighted, we need to summarize them into a single score. There are several ways of doing it. The main question here is 'what parameter do we want to be minimized in our future calculations?'

Mean and median are 'central tendencies' aiming to estimate the average of the data. However, there is a fundamental difference. With minimizing score based on the median, we are minimizing the error we get most frequently, no matter how large or small are errors in the extreme cases. The mean-based score does consider extreme cases and minimizes (just as expected) the average error of future calculations. (Strictly speaking, this is true only if the test set is representative for the part of the chemical space where the method is meant to be applied.)

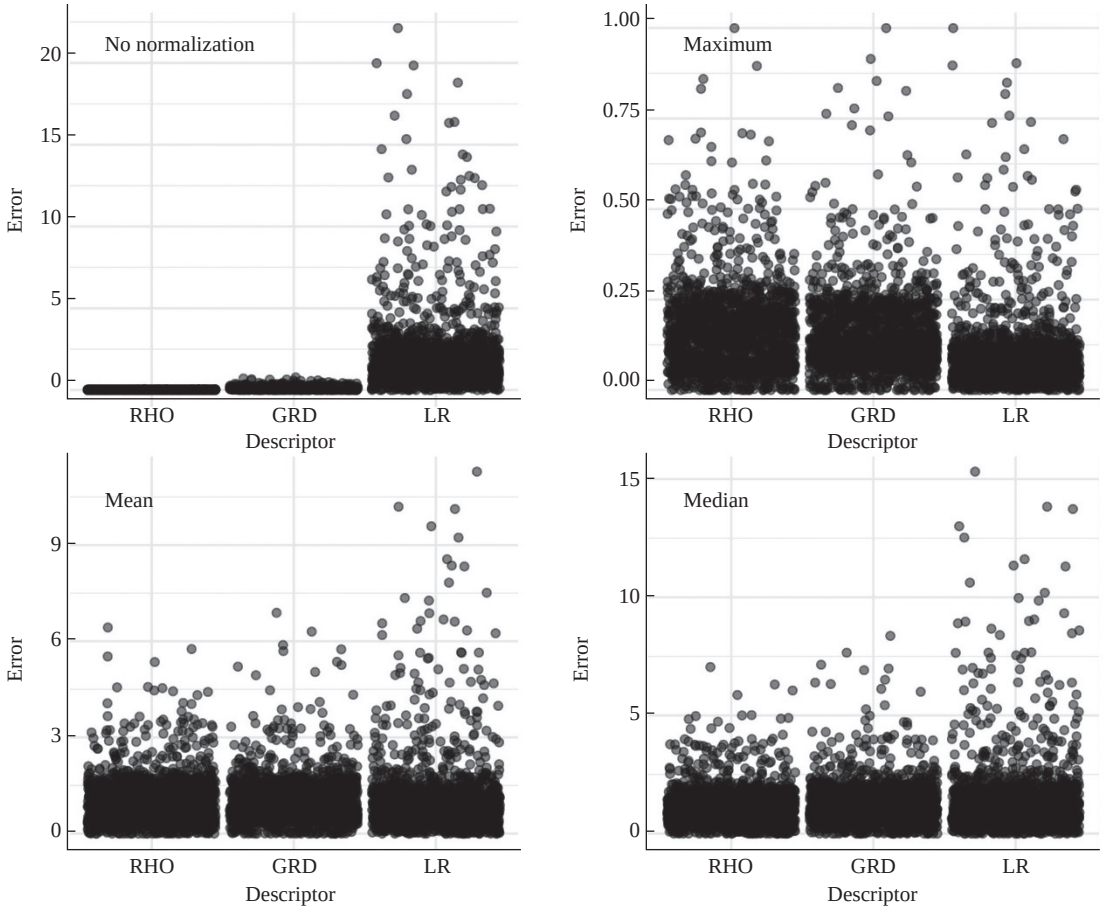

Figure 7 Effect of normalization on descriptor RMSD errors distributions. Normalization is done in subgroups defined by descriptors.

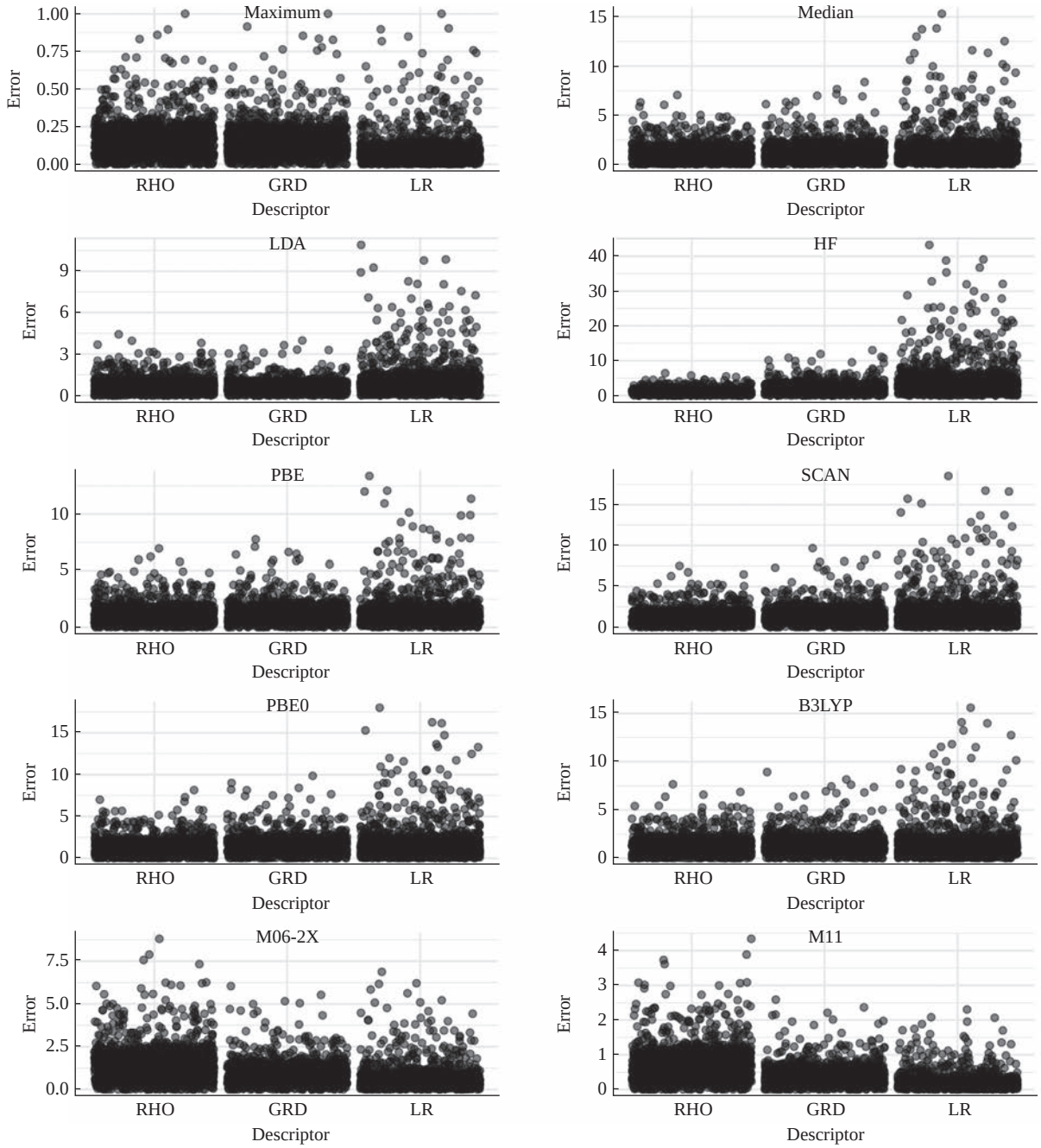

Figure 8 RMSD errors distributions for RHO, GRD and LR after normalization with respect to maximum and median values and to several actual methods. Normalization is done in subgroups defined by descriptor.

While the average behavior is usually of highest interest, in some cases the worst-case behavior is more important. In fact, most computational chemists are not so interested in a long-term average error of all their calculations. They want to be sure they will never get an error large enough to distort the conclusions of the experiment. It is exceptionally difficult to accurately estimate a worst-case error of a quantum chemical method. But using the maximum error as a score aims to get closer to it than average-based techniques.

On the 'edge cases' test set, maximum-based summarization also solves the problem of two-electron-systems overweighting. A mean-based score gives high rankings to HF and hGGA* methods, while a maximum-based score prioritizes more universally applicable methods. An obvious pitfall of the maximum-based method is its instability: adding a single measurement to the test set can change the entire ranking. An alternative approach, that solves the problem without sacrificing robustness, uses a two-step summarization. First, for each descriptor a mean error over atomics is calculated; second, the maximum of three descriptors errors is used as a score. This metric was used in the Medvedev *et al*. article along with the double-maximum one and is denoted as maMNAE here. Note, however, that the maMNAE metric does not address the two-electron-systems overweighting directly, but rather exploits the feature of the test set that HF and related hGGA* methods give relatively inaccurate RHO compared to the other methods.

Another desired parameter for any computational method is predictability of errors. To estimate predictability, one can use a standard deviation as a score. Of course, the resulting ranking alone is useless, as it cannot distinguish between predictably good and predictably bad methods.

Literature metrics. In recent literature, we can find four point-wise density-based metrics used for ranking of DFT methods. The metrics proposed by Brorsen *et al*.[42] use the sophisticated procedure to consider only bonding regions of molecules. Since

the problem this solves does not deal with spherically symmetric electron density or its natural generalizations, a detailed discussion of its specific choices for weighting different points is beyond the scope of this article. The ISD-based metric proposed by Gould is significantly biased to the core region as discussed above, so we will not discuss it further here.

The metrics originally proposed by Medvedev *et al*. use RMSD as an error measure, median-based normalization and maximum-based summarization. Since normalization was performed in subgroups defined by descriptors only, the easier two-electron systems contribute less to the final score than the more difficult four- and ten-electron ones. Maximum-based summarization aims to estimate worst-case behavior and mitigate the test set imbalance. While the metric tries hard to estimate the functional aptitude for everyday use, two things were correctly criticized: (1) RMSD as an error measure lacks clear physical interpretation and – even more importantly – cannot be generalized to non-spherically symmetric systems; (2) maximum-based summarization has its price, being 'noisy' and prone to outliers, so it is not clear if the benefits are worth it.

The slightly different metric applied by Mezei *et al*. uses per-electron IAD as an error measure, LDA-based normalization and mean-based summarization. This metric can be considered an improvement in some parts. IAD is more physically sound than RMSD and its usage enabled the authors to generalize their approach to molecular systems. Mean-based summarization appears to be a sensible compromise for balanced test sets: it is more robust than the maximum-based approach and does not ignore outliers completely as the median-based approach does. However, on the 'edge cases' set of atomic systems, mean-based normalization, together with scaling errors by electron number, exaggerates the contribution of the two-electron systems to the score, so that HF becomes better than all DFT functionals. Although, this result arises for a good reason – HF is, indeed, very accurate for most systems (two-electron cations) in the 'edge cases' test set, this is not the knowledge one intends to gather from the study; one wants to answer The Question (here: 'Which functional should I use to get accurate electron density for an arbitrary system?') – and the extreme-cases behavior (maximum error) of a functional provides more information about its general reliability on an unbalanced test set.

Shiny R web-application. To allow everyone interested to investigate the data from the Medvedev *et al*. article[30] for themselves, we have constructed a website, available at shiny.xrlab.ru/edee. It allows one to apply any of 2880 available (by this article publication date) statistical procedures to analyze the data (or its subsets) from the article and see how it affects the 'historical plot' (Figure 1) and changes the functionals ranking in the tables. The current statistical procedure is outlined at the top of the page, and every bolded entity in it can be modified using selectors below. Under the procedure, there is a red check-line, which indicates if the current statistical procedure is the one proposed in the Medvedev *et al*.[30] article (the default). The statistical procedure proposed by Mezei *et al*.[25] also can be constructed using the web-application by choosing the IAD error measure, dividing individual errors by the number of electrons in each system, normalizing errors by LDA errors within groups defined by descriptor and then averaging errors over both atomics and descriptors by means.

Selectors at the bottom left of the page can be used to modify the statistical procedure as follows:

(1) Any subgroup of atomic species can be selected for analysis with the 'Atomic species to use' selector. Available atomic species are: Be^0, B^{3+}, B^+, C^{4+}, C^{2+}, N^{5+}, N^{3+}, O^{6+}, O^{4+}, F^{7+}, F^{5+}, Ne^{8+}, Ne^{6+} and Ne^0. The default selection is all species.

(2) Any subgroup of descriptors can be selected for analysis with the 'Descriptors to use' selector. Available descriptors are local electron density (RHO), its gradient norm (GRD) and Laplacian (LR). The default selection is all descriptors.

(3) One of the error functions should be selected with the 'Error function' selector. Currently, available error functions are:

– root-mean-square difference between descriptor radial distribution functions (RMSD);

– integral of the absolute difference between descriptor radial distribution functions (IAD). It has a physical meaning of twice the number of 'relocated' electrons and puts more weight on the outer electron density in comparison with the RMSD measure. (See detailed discussion and comparison of these error measures above.)

The default choice is RMSD error function, which was used in Medvedev *et al*. work.

(4) Each error can then be scaled by some number defined by the given atom with the selector 'Divide individual errors by'. Currently available options are:

– skip scaling (do not divide errors at this step);

– divide by the nuclear charge;

– divide by the number of electrons.

The second choice allows one to emphasize atoms with small nuclear charge, *i.e*., where electrons are further from the high-density limit. The last choice allows one to calculate errors per electron. The default choice is to skip scaling.

(5) Calculated errors can be normalized by dividing to some value defined for a subgroup of errors. The 'Normalize all errors within' selector allows one to choose subgroups for normalization. Currently, normalization coefficients can be calculated for subgroups defined as follows:

– by descriptor;

– by a combination of descriptor and electron number (number of electrons in the system);

– by a combination of the descriptor and atomic system.

In the first case, errors in each descriptor for different atoms are divided by the same value, thus preserving distinct complexities of different atoms. On the contrary, in the second and third cases, errors in each descriptor-atom or descriptor-'electron number' pair are divided by independent values calculated within defined subgroups, so distinct complexities of different atoms are not fully preserved. The default choice is to normalize within descriptor-defined subgroups.

(6) 'Normalize all errors by' selector controls the values by which errors in the subgroups specified above are normalized. Current options are to normalize by:

– Median error;

– Mean error;

– Standard deviation of the error;

– Maximum error;

– Values from the Science paper (only available if errors are normalized within subgroups defined by descriptor only);

– LDA[48] error;

– PBE[49] error;

– SCAN[32,50] error;

– HF error;

– B3LYP[51] error;

– PBE0[45,46] error;

– M06-2X[52,53] error;

– M11[54] error.

Thus, errors can be normalized by error values of some 'average functional', *i.e*., one whose errors are equivalent to the median, mean or maximum errors of all studied functionals. Alternatively, errors can be normalized by error values of some real methods: LDA (SVWN), PBE, SCAN, HF, B3LYP, PBE0, M06-2X or M11. If one chooses normalization by errors of a

real method within subgroups defined by descriptor or by a combination of descriptor and electron number, then the errors of the real method need to be averaged somehow for all the systems included in the subgroup. Currently, they are averaged in the same way as atomic errors (chosen by the 'Average atomic errors by' selector).

The default choice is to normalize by the median errors. Together with the default choice for the previous selector, this normalizes all errors for each descriptor by the errors of a 'median functional', preserving distinct complexities of different atoms.

(7) 'Average atomic errors by' and 'Average errors in descriptors by' selectors control errors averaging over all system and descriptors, respectively. Currently available options for both selectors are:

– Maximum error;

– Median error;

– Mean error;

– The standard deviation of the error.

First, errors are averaged over atomic systems, and then system-averaged errors are averaged over descriptors to provide the final 'Electron Density error' ('ED error') for each method, which is used in the plot and tables. The default choice for these selectors is identical – 'Maximum error' – the same as used in the Medvedev *et al*. article.

Under the selectors, there is a field 'To label' which can be used for labelling up to nine density functionals on the plots (tabs 'Historic overview' and 'Individual errors'), and a button 'Restore defaults' which reverts the statistical procedure to the one which was used in the Medvedev *et al*. article.

The bottom right area of the web-page contains four tabs:

(1) 'Historic overview': contains a plot with the historical trend in ED errors calculated using the current procedure. The black curve shows the average deviation, with the light grey area denoting its 95% confidence interval. Each functional is attributed to a randomly selected day within the year when it (or its 'final component', if there is no particular reference for the functional) was published. Hovering a mouse over a point on the plot makes a tooltip with the functional name to appear near the pointer;

(2) 'Individual errors': contains a plot of individual errors of density functionals for selected descriptors and atoms, scaled and normalized according to the chosen procedure (all selectors except 'Average atomic errors by' and 'Average errors in descriptors by' affect this plot). Points jitter along the *x*-axis is proportional to the density distribution of the data points at a given error value (*i.e*., the wider jitter corresponds to higher data points accumulation). On this plot, functionals can be labelled in two ways: (1) using the slider on the tab, which adds names of the functionals with largest errors near corresponding points, and (2) using the 'To label' field on the left – then functionals are labelled with black marking signs on the plot and their names are available in the plot legend;

(3) 'Tables': contains the tables of ED errors calculated using the current procedure and a selector for choosing the number of rows in each table (*i.e*., what number of functionals with lowest and highest errors to show);

(4) 'Changelog': contains the history of the web-application, specifying all notable changes to it and dates they were applied.

Metrics asking 'other' questions. The metrics thoroughly composed to estimate the 'goodness' of methods such as metrics in the Medvedev *et al*.[30] and Mezei *et al*.[25] articles tend to behave similarly on test sets they are designed for. But even among the restricted set of metrics available at the web-application, there are ones that behave differently. These metrics just ask other questions. The ranking they produce must not be interpreted as a ranking of method 'goodness', but can still contain useful information. Here are a few examples.

When choosing a single descriptor and using standard deviation at the first averaging step (*i.e*., over atoms), the interpretation of the ranking is very different from the 'goodness' but still obvious: it estimates predictability of errors for this descriptor. Average predictability for several descriptors (if selected) is calculated according to the 'Average errors in descriptors by' selector. While no DFT functional can be considered 'better' than MP2 on the 'edge cases' set, there are several whose errors are more predictable. Most of them are 'good' as we know from conventional rankings: PBE0, BHHLYP, *etc*. Most of the 'bad' functionals have a large standard deviation of errors as well. An interesting case is the M06-2X functional, which has one of the most 'predictable' errors relative to the other functionals when IAD error measure is used, and one of the most 'unpredictable' with the RMSD error measure (in both cases, mean standard deviation over descriptors was computed); considering descriptors, it shows very 'predictable' errors for RHO and GRD, but is almost 'unpredictable' for LR, independently of error measure.

A whole family of metrics resulting in highly variable rankings can be constructed using system-wise normalization (*i.e*., normalization within subgroups defined by combination of descriptor and atomic system) based on a reference functional and dual-maximum summarization. These metrics give to the reference method a score of 1.0 and a very high rank: PBE, SCAN, PBE0, B3LYP and HF when chosen as reference are ranked #5 – just after post-HF methods. This fact has a precise (but a bit tricky) interpretation: a method is ranked higher than the reference if and only if it outperforms it on all the studied systems and descriptors. So, from these rankings, we know there is no method that outperforms PBE or HF on every descriptor of every system except for *ab initio* ones. However, this is not the case for LDA – when chosen as a reference it is ranked #7; PBEsol and SOGGA functionals do outperform it on all the studied systems for every descriptor. M06-2X is the only functional, among the eight chosen, which beats MP2 method at least once, so it is ranked #4 if selected as a reference; when M11 method is chosen as a reference, it is ranked #43. These few rankings can tell us even more about variability in errors on different systems than standard deviation-based ones. Thus, on a tiny atomic test set, only two DFT methods are uniformly better than LDA, and none of them is uniformly better than PBE or HF. On our 'edge cases' test set, the variability in errors is very large, making the whole ranking so hard and controversial.

Conclusions

Many metrics can be – and were – applied to measure differences between electron densities, and many statistical procedures were used for normalization and averaging errors over different metrics and chemical systems. However, it is not immediately apparent which metric or statistical procedure is better for deriving conclusions concerning accuracies of density functionals for any studied set of systems. In this article, we have demonstrated that every metric and statistical procedure combination, in principle, provides unique information and answers a distinct question. Also, the question a statistical procedure answers depends on the test set it is applied to and should be discussed along with it: While the statistical procedure of Mezei *et al*. is nearly-optimal for a balanced test set, it is completely inappropriate for the 'edge cases' test set, where it puts the HF method ahead of all DFT functionals (although it is perfectly fair for two-electron systems dominating the 'edge cases', this conclusion does not generalize even to four electron systems, let alone molecules). We have presented and described a web-application (available at shiny.xrlab.ru/edee/)

which allows anyone interested to vary the statistical procedure in the analysis of the [*Science*, 2017, **355**, 49] dataset and see how it affects the 'historical plot' and changes the functionals ranking.

Although here we have focused on performance of DFT functionals for electron densities, our discussion is relevant for any benchmarking of quantum chemical methods. Any of the discussed statistical procedures can be applied to an arbitrary (uploaded by the user) dataset of benchmark results using our second web-application, which is available at shiny.xrlab.ru/rank-it/.

M.G.M. is grateful to the Russian Science Foundation (grant no. 17-13-01526) for support. A.A.M., I.S.G., K.A.L. and A.O.D. are grateful to the Russian Foundation for Basic Research (grant no. 17-03-00907) for support. J.P.P. is grateful to the U.S. National Science Foundation for support (grant no. DMR-1607868). The research was carried out using the equipment of the shared research facilities of HPC computing resources at M. V. Lomonosov Moscow State University.[55] The Siberian Branch of the Russian Academy of Sciences Siberian Supercomputer Center is gratefully acknowledged for providing supercomputer facilities.

References

1 R. O. Jones, *Rev. Mod. Phys*., 2015, **87**, 897.
2 P. Hohenberg and W. Kohn, *Phys. Rev*., 1964, **136**, B864.
3 P. Mori-Sánchez and A. J. Cohen, *ArXiv170910284 Cond-Mat Physics-physics Physicsquant-Ph*.
4 W. Kohn and L. J. Sham, *Phys. Rev*., 1965, **140**, A1133.
5 W.-P. Wang and R. G. Parr, *Phys. Rev. A*, 1977, **16**, 891.
6 M. P. C. M. Krijn and D. Feil, *Chem. Phys. Lett*., 1988, **150**, 45.
7 K. B. Wiberg, C. M. Hadad, T. J. LePage, C. M. Breneman and M. J. Frisch, *J. Phys. Chem*., 1992, **96**, 671.
8 J. Wang, L. A. Eriksson, R. J. Boyd, Z. Shi and B. G. Johnson, *J. Phys. Chem*., 1994, **98**, 1844.
9 K. E. Laidig, *Chem. Phys. Lett*., 1994, **225**, 285.
10 J. Wang, Z. Shi, R. J. Boyd and C. A. Gonzalez, *J. Phys. Chem*., 1994, **98**, 6988.
11 R. J. Boyd, J. Wang and L. A. Eriksson, in *Recent Advances in Computational Chemistry*, ed. D. P. Chong, World Scientific Publishing, Singapore, 1995, vol. 1, pp. 369–401.
12 J. Wang, L. A. Eriksson, B. G. Johnson and R. J. Boyd, *J. Phys. Chem*., 1996, **100**, 5274.
13 J. Wang, B. G. Johnson, R. J. Boyd and L. A. Eriksson, *J. Phys. Chem*., 1996, **100**, 6317.
14 M. Solà, J. Mestres, R. Carbó and M. Duran, *J. Chem. Phys*., 1996, **104**, 636.
15 M. Solà, J. Mestres, J. M. Oliva, M. Duran and R. Carbó, *Int. J. Quantum Chem*., 1996, **58**, 361.
16 J. Poater, M. Duran and M. Solà, *J. Comput. Chem*., 2001, **22**, 1666.
17 A. D. Bochevarov and R. A. Friesner, *J. Chem. Phys*., 2008, **128**, 034102.
18 V. Tognetti and L. Joubert, *J. Phys. Chem. A*, 2011, **115**, 5505.
19 R. F. W. Bader, *Atoms in Molecules: A Quantum Theory*, Oxford University Press, Oxford, 1994.
20 R. Carbó, L. Leyda and M. Arnau, *Int. J. Quantum Chem*., 1980, **17**, 1185.
21 P. Verma and D. G. Truhlar, *Phys. Chem. Chem. Phys*., 2017, **19**, 12898.
22 D. Hait and M. Head-Gordon, *J. Chem. Theory Comput*., 2018, **14**, 1969.
23 I. V. Schweigert, V. F. Lotrich and R. J. Bartlett, *J. Chem. Phys*., 2006, **125**, 104108.
24 D. S. Ranasinghe, A. Perera and R. J. Bartlett, *J. Chem. Phys*., 2017, **147**, 204103.
25 P. D. Mezei, G. I. Csonka and M. Kállay, *J. Chem. Theory Comput*., 2017, **13**, 4753.
26 T. Gould, *J. Chem. Theory Comput*., 2017, **13**, 2373.
27 A. D. Becke and K. E. Edgecombe, *J. Chem. Phys*., 1990, **92**, 5397.
28 A. Savin, *J. Mol. Struct. (Theochem.)*, 2005, **727**, 127.
29 L. A. C. Vannay, *PhD Thesis*, École Polytechnique Fédérale de Lausanne, 2018.
30 M. G. Medvedev, I. S. Bushmarinov, J. Sun, J. P. Perdew and K. A. Lyssenko, *Science*, 2017, **355**, 49.
31 J. P. Perdew, A. Ruzsinszky, J. Sun and K. Burke, *J. Chem. Phys*., 2014, **140**, 18A533.
32 J. Sun, A. Ruzsinszky and J. P. Perdew, *Phys. Rev. Lett*., 2015, **115**, 036402.
33 K. A. Peterson and T. H. Dunning, Jr., *J. Chem. Phys*., 2002, **117**, 10548.
34 B. Prascher, D. E. Woon, K. A. Peterson, T. H. Dunning, Jr. and A. K. Wilson, *Theor. Chem. Acc*., 2011, **128**, 69.
35 M. G. Medvedev, I. S. Bushmarinov, J. Sun, J. P. Perdew and K. A. Lyssenko, *Science*, 2017, **356**, 496.
36 R. J. Bartlett and G. D. Purvis, *Int. J. Quantum Chem*., 1978, **14**, 561.
37 G. D. Purvis and R. J. Bartlett, *J. Chem. Phys*., 1982, **76**, 1910.
38 N. Mardirossian and M. Head-Gordon, *Mol. Phys*., 2017, **115**, 2315.
39 D. Mejia-Rodriguez and S. B. Trickey, *Phys. Rev. A*, 2017, **96**, 052512.
40 J. P. Perdew and K. Schmidt, in *Density Functional Theory and Its Applications to Materials*, eds. V. Van Doren, C. Van Alsenoy and P. Geerlings, AIP Publishing, Melville, NY, 2001, vol. 577, pp. 1–20.
41 K. P. Kepp, *Science*, 2017, **356**, 496.
42 K. R. Brorsen, Y. Yang, M. V. Pak and S. Hammes-Schiffer, *J. Phys. Chem. Lett*., 2017, **8**, 2076.
43 Y. Wang, X. Wang, D. G. Truhlar and X. He, *J. Chem. Theory Comput*., 2017, **13**, 6068.
44 N. Q. Su, Z. Zhu and X. Xu, *Proc. Natl. Acad. Sci*., 2018, **115**, 2287.
45 J. P. Perdew, M. Ernzerhof and K. Burke, *J. Chem. Phys*., 1996, **105**, 9982.
46 C. Adamo and V. Barone, *J. Chem. Phys*., 1999, **110**, 6158.
47 R. Peverati and D. G. Truhlar, *Phys. Chem. Chem. Phys*., 2012, **14**, 16187.
48 S. H. Vosko, L. Wilk and M. Nusair, *Can. J. Phys*., 1980, **58**, 1200.
49 J. P. Perdew, K. Burke and M. Ernzerhof, *Phys. Rev. Lett*., 1996, **77**, 3865.
50 J. Sun, R. C. Remsing, Y. Zhang, Z. Sun, A. Ruzsinszky, H. Peng, Z. Yang, A. Paul, U. Waghmare, X. Wu, M. L. Klein and J. P. Perdew, *Nat. Chem*., 2016, **8**, 831.
51 P. J. Stephens, F. J. Devlin, C. F. Chabalowski and M. J. Frisch, *J. Phys. Chem*., 1994, **98**, 11623.
52 Y. Zhao and D. G. Truhlar, *Theor. Chem. Acc*., 2007, **120**, 215.
53 Y. Zhao and D. G. Truhlar, *Acc. Chem. Res*., 2008, **41**, 157.
54 R. Peverati and D. G. Truhlar, *J. Phys. Chem. Lett*., 2011, **2**, 2810.
55 V. Sadovnichy, A. Tikhonravov, V. Voevodin and V. Opanasenko, in *Contemporary High Performance Computing: from Petascale toward Exascale*, CRC Press, Boca Raton, FL, 2013, pp. 283–307.

Received: 26th February 2018; Com. 18/5492