Top Curr Chem (2015) 365: 81–96
DOI: 10.1007/128_2014_600

Published online: 24 February 2015

Judging Density-Functional Approximations: Some Pitfalls of Statistics

Andreas Savin and Erin R. Johnson

Abstract Density-functional theory (DFT) methods have achieved widespread popularity for thermochemical predictions, which has lead to extensive benchmarking of functionals. While the use of statistics to judge the quality of various density-functional approximations is valuable and even seems unavoidable, the present chapter suggests some pitfalls of statistical analyses. Several illustrative examples, focusing on analysis of thermochemistry and intermolecular interactions, are presented to show that conclusions can be heavily influenced by both the data-set size and the choice of the criterion used to assess an approximation's quality. Even with reliable approximations, the risk of publishing inaccurate results naturally increases with the number of calculations reported.

Keywords Density-functional theory · Intermolecular interactions · Statistics · Thermochemistry

Contents

1 Introduction 82
2 Methodology 83
3 Using Statistical Measures to Judge Density Functional Approximations 84
4 Summary 94
References 94

A. Savin
CNRS, UPMC Univ Paris 06, UMR7616, Laboratoire de Chimie Théorique, 75005 Paris, France

E.R. Johnson (✉)
Chemistry and Chemical Biology, School of Natural Sciences, University of California, Merced, 5200 North Lake Road, Merced, CA 95343, USA
e-mail: ejohnson29@ucmerced.edu

1 Introduction

A combination of the high performance of modern computers, fast algorithms, and good accuracy have permitted widespread use of density-functional approximations (DFAs) across chemistry and physics. However, the high number of DFAs often causes users of the theory to pose the question: "Which functional should I use?" Nowadays, large sets of reference data are available to provide valuable help in answering this question. Consequently, there has been a recent surge of benchmarking studies where readers can take their pick of statistical measures to justify use of their chosen functional in a particular application. The rapid growth in the numbers and citations of benchmarking studies is illustrated in Fig. 1.

Habitually, statistical measures such as the mean absolute error are used to indicate the quality of a density-functional approximation. Unfortunately, statistics can also deform reality, and using statistics to judge the quality of an approximation is no exception. A recently published paper [1] sought to answer several questions regarding benchmarking of DFAs:

1. Is the approximation the only source of error and would an exact treatment give the right result?
2. Do the approximations provide the necessary quantitative accuracy?
3. Are we interested in obtaining absolute values or in reproducing trends?

In general, the user is provided with a selection of DFAs and must decide which to choose. However, there is no single objective criterion to determine which DFA is best for a particular problem and a somewhat subjective choice is made regarding, for example, the statistical criteria used to rank the functionals. Which specific functional is then qualified as the best depends on this choice [1].

In the spirit of reference [2], we consider the performance of selected DFAs for thermochemistry and intermolecular interactions. In these cases, accurate data

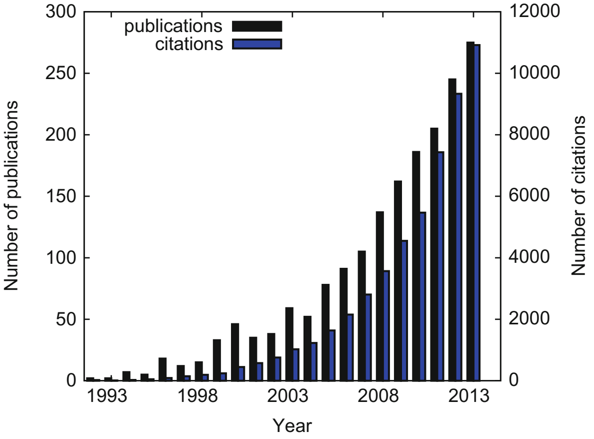

Fig. 1 Growth in the number of published density-functional benchmarking studies and the total citations, as determined from the Thomson Reuters Web of Science Core Database

should be accessible via coupled-cluster theory with large basis sets and extrapolation. We address the following questions:

1. How strong are the effects caused by the finite sample size?
2. Can the same data produce opposite interpretations?
3. What is the probability that all of the results published in an article have sufficient quantitative accuracy?

In the process, we discuss some potential pitfalls of statistics that rank the quality of density-functional approximations.

2 Methodology

In a previous study [3], various dispersion-corrected density functionals were benchmarked against either experimental or high-level ab initio reference data for intermolecular complexes, thermochemistry, and reaction barrier heights. All data sets used in the present work are taken from this prior study. The types and sources of the reference data are summarized in Table 1. These data sets are by no means a comprehensive collection of DFT benchmarks and this chapter focuses on the illustrative examples of thermochemistry and intermolecular interaction data sets only.

The density-functional approximations considered herein are also the same as in our earlier work [3]. The acronyms follow the usual notation related to the names of the authors: BLYP [13, 14], B3LYP [14, 15], BH&HLYP [14, 16], B97-1 [17], CAM-B3LYP [18], LC-ωPBE [19, 20], PBE [21], PBE0 [22], and PW86PBE [21, 23]. To ensure that the conclusions are not adversely impacted by the basis set, aug-cc-pVTZ bases were used as they give results close to the complete basis set limit for conventional density-functional calculations. In all cases, the exchange-hole dipole moment (XDM) dispersion correction was applied [3, 24], but this does *not* appear in the acronyms used below. Note that the two empirical parameters used in the dispersion damping function were fitted separately for each of the functionals, with the aug-cc-pVTZ basis set, which also tends to correct deficiencies of the base DFA.

To assess the quality of the various DFAs for each data set, three statistical error measures are used: the mean absolute errors (MAE):

$$\mathrm{MAE} = \bar{x} = \frac{1}{n}\sum_{i=1,n} x_i, \tag{1}$$

the mean absolute percent errors (MAPE):

Table 1 List of reference data, showing the abbreviation of the data-set name, the source of the data (either calculated or experimental), the number of data points, a brief description of the set, and the relevant literature reference

Name	Type	No.	Description	References
KB49	calc	49	Intermolecular interactions	[3, 4]
S22	calc	22	Intermolecular interactions	[5, 6]
S66	calc	66	Intermolecular interactions	[7, 8]
HSG	calc	21	Intermolecular interactions	[6, 9]
G1	expt	56	Atomization energies	[10]
G2	expt	149	Atomization energies	[11]
G3	expt	222	Atomization energies	[24]

$$\mathrm{MAPE} = \frac{100}{n}\sum_{i=1,n} \frac{x_i}{|x_{\mathrm{ref},i}|} \tag{2}$$

and the root-mean-square errors (RMSE):

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1,n} x_i^2} \tag{3}$$

where n is the number of data points in the set and $x_i = |x_{\mathrm{calc},i} - x_{\mathrm{ref},i}|$ is the absolute error for each point. Finally, the sample variance is

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1,n} (x_i - \overline{x})^2 \tag{4}$$

and the mean absolute error has a variance equal to the sample variance divided by the size of the sample, or σ^2/n. The uncertainty of the MAE is thus $\sigma/\sqrt{n}$.

3 Using Statistical Measures to Judge Density Functional Approximations

When judging the performance of DFAs, the mean error does not tell the full story and further statistical measures are available to describe the distribution of errors for a particular benchmark set. As an example, let us consider the G3/99 data set [12] which gauges the ability of the functionals to describe atomization energies. Histograms of the error distribution for the G3 set are shown in Fig. 2 for selected functionals. From the figure, we see that BLYP and B3LYP have narrow error distributions, while the BHandHLYP distribution is much broader. The maximum errors are 27.3 kcal/mol for BLYP, 34.4 kcal/mol for B3LYP, and 79.7 kcal/mol for

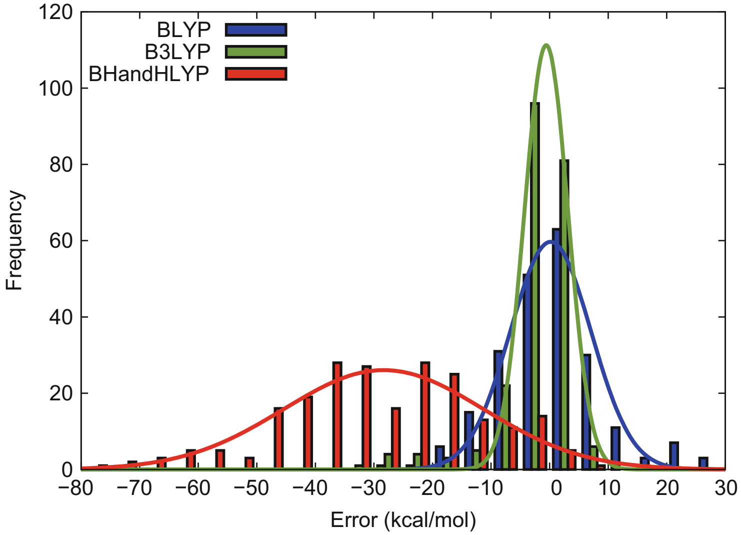

Fig. 2 Histograms of the signed errors for the G3 set with selected functionals. The lines show fits of the data to Gaussian distributions and are included to guide the eye

BHandHYLP. Thus, even for a relatively narrow error distribution, the maximum errors are far from the desired chemical accuracy.

Analogous results can be presented more compactly via "box-and-whisker" plots, where the boxes span the interquartile range of the data (i.e., the range that spans the middle half of the data), the ends of the whiskers show the minimum and maximum errors, and the lines show the median errors. Error distributions are plotted in this fashion for all the DFAs considered in Fig. 3. The figure also shows an analogous plot for the errors expressed on a per-atom basis, which eliminates any bias of the error distribution based on molecular size.

When benchmarking density functionals, mean absolute errors are most often presented to indicate the quality of an approximation. However, the average can be greatly inflated by a few outliers in a data set. The root-mean-square error (RMSE) is even more strongly affected, while the median absolute error reduces this bias. Table 2 collects these statistics for the G3 set. B3LYP gives the lowest MAE and median absolute error, while CAM-B3LYP gives the lowest RMSE and B97-1 the lowest maximum error. As a further alternative, one can also consider the MAE, RMSE, and median or maximum absolute error per atom, to account for differences in molecular size within the benchmark. With the per-atom statistics, B3LYP, CAM-B3LYP, LC-ωPBE, B97-1, and PBE0 all give MAEs of near 1 kcal/mol per atom. B97-1 gives the lowest RMSE, B3LYP gives the lowest median absolute error, and CAM-B3LYP the lowest maximum error. Thus, the choice of which statistical indicator is used to judge the DFAs determines which is ultimately selected as the best functional.

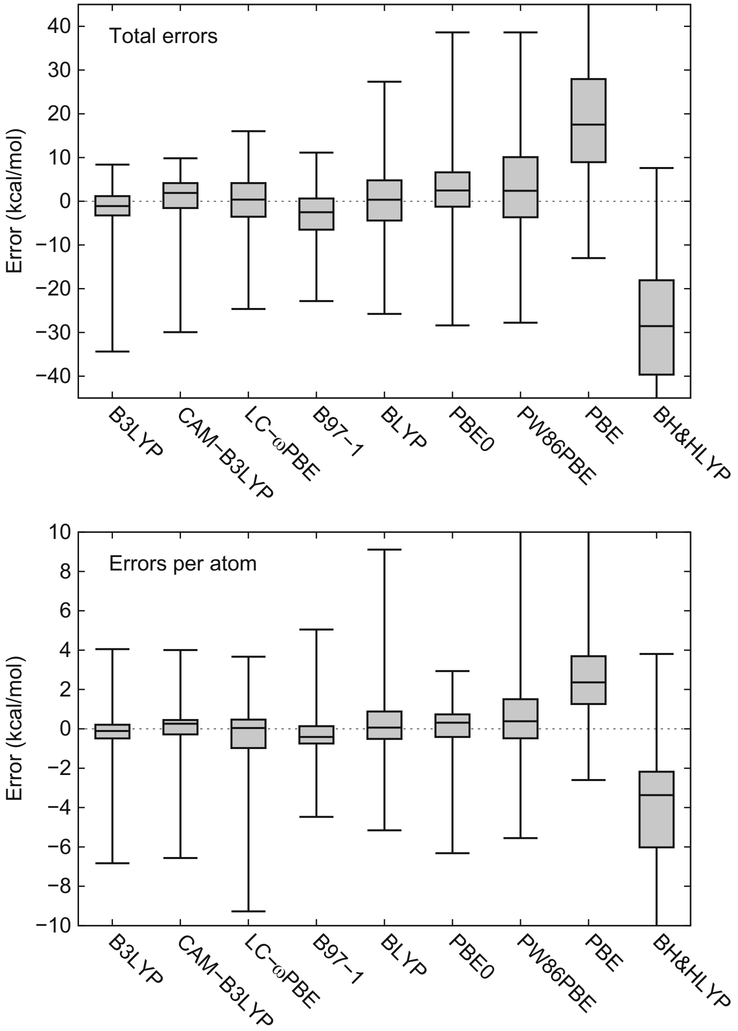

Fig. 3 Box-and-whisker plots of the atomization-energy errors for the G3 set with selected functionals. The *boxes* span the interquartile range, the ends of the *whiskers* show the minimum and maximum errors, and the *lines* show the median errors. The *upper panel* shows the total errors and the *lower panel* shows the errors per atom

With each DFA, the distribution of errors for the G3 set is very broad and the decay of the absolute errors is very slow, so that the mean is not well defined. To demonstrate what is meant by this, consider a linear molecule of $n+1$ atoms, forming n chemical bonds. For each bond energy, we let the error with a particular DFA be x kcal/mol. Then the mean error for chains of size $n = 1, 2, \dots, m$ is

Table 2 MAEs, RMSEs, median absolute errors (Med), and maximum absolute errors (Max) for the G3 set, in kcal/mol. The same quantities, expressed per atom, are also shown

	Total				Per atom			
Functional	MAE	RMSE	Med	Max	MAE	RMSE	Med	Max
B3LYP	4.0	6.7	2.2	34.4	0.8	1.5	0.3	6.8
CAM-B3LYP	4.6	6.5	3.5	29.9	0.9	1.5	0.6	5.0
LC-ωPBE	5.2	7.0	3.9	24.6	1.1	1.7	0.6	9.3
B97-1	5.3	7.1	4.0	22.8	0.8	1.1	0.4	6.6
BLYP	6.5	8.8	4.7	27.3	1.3	2.1	0.6	9.1
PBE0	6.7	9.5	4.2	38.6	1.0	1.5	0.6	6.3
PW86PBE	9.4	12.6	6.7	38.6	1.7	2.6	1.0	11.5
PBE	20.8	25.8	17.5	82.0	3.0	3.8	2.4	13.7
BH&HLYP	29.2	33.5	28.6	79.7	4.5	5.7	3.4	20.4

$$\frac{1}{m}\sum_{n=1}^{m} nx = \frac{m+1}{2}x \tag{5}$$

and this diverges as $m \rightarrow \infty$. Thus, the atomization energy errors increase steadily with molecular size and, as larger systems are added to the benchmark set, the errors increase and the mean is inflated. Conversely, the atomization energy per atom is well defined and approaches x when $m \rightarrow \infty$. The divergence of the MAE with increasing system size can be seen by comparing the error distributions for the G1, G2, and G3 test sets, as shown in Fig. 4. The G1 set was the first set of atomization energy data, compiled for small molecules only. This set was later expanded to the G2, and subsequently the G3, by adding progressively larger organic molecules to the benchmark. From the figure, the errors in total atomization energies increase going from the G1 to the G3 set. On the other hand, the errors per atom remain roughly constant or even decrease slightly going from the G1 to the G3 set. The errors are reduced in some cases since the DFAs tend to perform better for organic molecules, which constitute a larger fraction of the G3 set. Ultimately, the MAE per atom should be a favored statistic over the total MAE when comparing performance of DFAs for atomization energies.

Another consequence of the breadth of the error distribution is that the MAE may have a large associated uncertainty, particularly for small sample sizes. Therefore, the variance of the mean may be larger than the difference in MAEs between two or more functionals, prohibiting use of this metric to make an informed ranking of DFA quality. For the G3 set, MAEs and their uncertainties for each functional are collected in Table 3. The uncertainties are fairly large, ranging from 0.3 to 1.1 kcal/mol, illustrating that the MAEs are not certain beyond the first decimal point. Additionally, we cannot definitively conclude that B3LYP is the optimum functional, despite giving the lowest MAE, because the difference between the MAEs from B3LYP and CAM-B3LYP is smaller than the sum of the uncertainties.

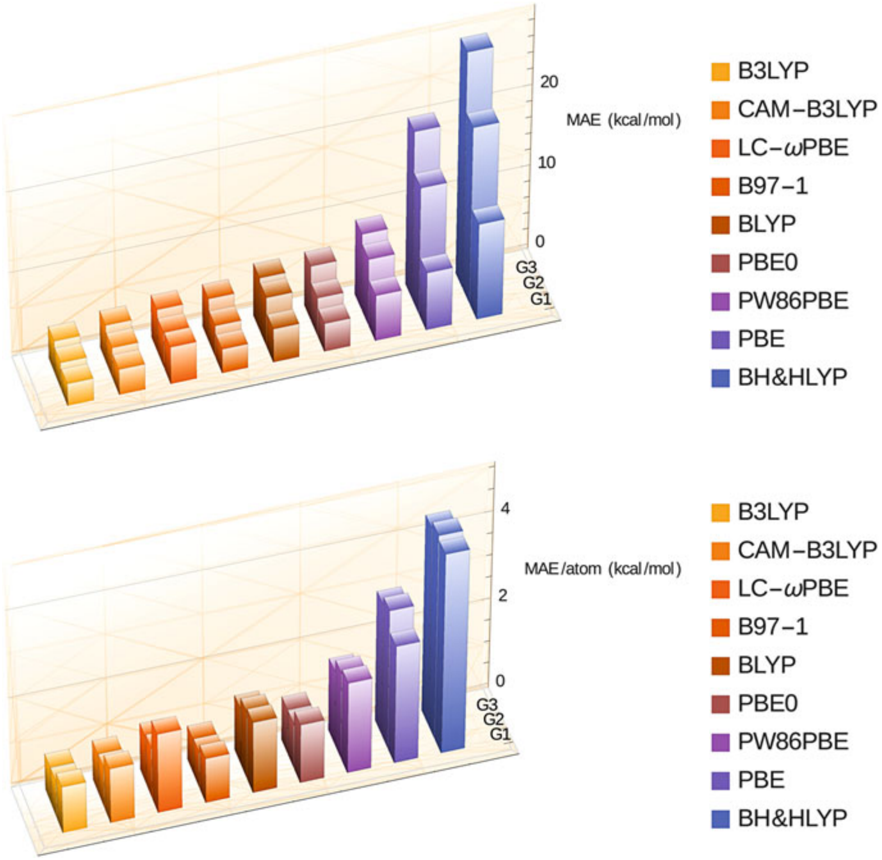

Fig. 4 Histograms of the MAE and MAE per atom with selected functionals for the G1, G2, and G3 sets of atomization energies

Table 3 MAEs and their uncertainties, measured using the square-root of the sample variance, for the full G3 set and three randomly-chosen subsets, in kcal/mol

Functional	G3	Subset 1	Subset 2	Subset 3
B3LYP	4.0 ± 0.4	2.7 ± 0.5	2.4 ± 0.4	3.6 ± 0.8
CAM-B3LYP	4.6 ± 0.3	3.5 ± 0.5	4.0 ± 0.7	4.8 ± 0.8
LC-ωPBE	5.2 ± 0.3	4.6 ± 0.7	5.1 ± 0.7	4.6 ± 0.8
B97-1	5.3 ± 0.3	4.6 ± 0.8	4.4 ± 0.8	5.0 ± 1.0
BLYP	6.5 ± 0.4	5.1 ± 1.1	5.5 ± 1.2	6.3 ± 1.1
PBE0	6.7 ± 0.5	6.6 ± 1.9	4.8 ± 0.8	5.6 ± 1.0
PW86PBE	9.4 ± 0.6	8.7 ± 2.1	7.2 ± 1.4	7.1 ± 1.2
PBE	20.8 ± 1.0	20.6 ± 4.2	20.2 ± 3.8	20.7 ± 2.0
BH&HLYP	29.2 ± 1.1	26.1 ± 4.3	27.4 ± 2.6	30.3 ± 2.2

In general, because of the finite number of systems considered, the mean of a data set is uncertain and has a variance which is inversely dependent on the sample size. To investigate the dependence of the statistics on sample size, we consider the effect of taking three randomly-chosen subsets, each consisting of 22 molecules, from the G3 and evaluating the MAE and its variance for each subset. This procedure demonstrates the danger of using small data sets when benchmarking functionals. For example, using subset 2, the errors for CAM-B3LYP, B97-1, and PBE0 are all equivalent, to within the uncertainty. However, for the full G3 set, CAM-B3LYP is more accurate to within one standard deviation, with an MAE 2 kcal/mol lower than that of PBE0. Additionally, the MAEs obtained with B3LYP for each of the subsets are much lower than the MAE for the full G3 set, showing that it is quite probable that larger errors appear for increasingly larger sets. Indeed, the means for the full set are often worse, because some molecules for which the errors are particularly large are not included in the subsets.

We now turn our focus away from atomization energies and towards intermolecular interactions. Since small sets can give misleading results, are these data sets large enough for us to trust our conclusions? The MAEs for the KB49, S22, S66, and HSG data sets, obtained with the various DFAs, are shown in Table 4, together with their uncertainties, calculated using the square-root of the sample variance. We also consider a superset of 115 intermolecular complexes by combining the KB49 (which already includes the S22 set) with the S66 data set (which is entirely separate from the KB49 set). The MAEs and uncertainties are also shown graphically in Fig. 5. Because of the small size of the benchmarks, particularly for the S22 and HSG sets, definitive statements about the relative quality of the DFAs cannot be made because the uncertainties are so large that the MAEs for many of the functionals are not distinguishable. The statistics for the combined S115 set show, more definitively than for any of the smaller constituent data sets, the separation in performance between the three best-performing functionals (B3LYP, BLYP, and LC-ωPBE) and the rest of the DFAs.

Because of the small values of the interaction energies for dispersion-bound complexes, the mean absolute percent error (MAPE) is often preferred over the MAE. The MAPEs for the intermolecular data sets are shown in Table 5. However, as seen previously for the G3 set, different conclusions regarding which DFA is preferable may be drawn depending on which statistic is used as the selection criterion. The difference between MAE and MAPE is particularly important in sets where the absolute values cover a wide range. For example, considering the HSG set, B3LYP gives the lowest MAE while CAM-B3LYP gives the lowest MAPE. This occurs because, when the binding energies are large in magnitude (typically for H-bonding), CAM-B3LYP generally gives larger errors than B3LYP, while it performs quite well for small binding energies (BEs). The errors for large BEs have less weight in the MAPE than the MAE, so the apparent accuracy of CAM-B3LYP

Table 4 MAEs and their uncertainties for the KB49, S22, S66, and HSG sets, together with the S115 superset (combining KB49 and S66), in kcal/mol

Functional	KB49	S22	S66	HSG	S115
LC-ωPBE	0.27 ± 0.03	0.31 ± 0.06	0.21 ± 0.02	0.23 ± 0.04	0.23 ± 0.02
BLYP	0.29 ± 0.04	0.22 ± 0.04	0.19 ± 0.02	0.20 ± 0.04	0.23 ± 0.02
B3LYP	0.26 ± 0.03	0.31 ± 0.06	0.22 ± 0.02	0.12 ± 0.03	0.24 ± 0.02
PW86PBE	0.39 ± 0.06	0.35 ± 0.09	0.26 ± 0.03	0.17 ± 0.02	0.32 ± 0.03
BH&HLYP	0.35 ± 0.06	0.47 ± 0.11	0.31 ± 0.04	0.18 ± 0.05	0.33 ± 0.04
CAM-B3LYP	0.37 ± 0.06	0.50 ± 0.11	0.35 ± 0.04	0.16 ± 0.05	0.36 ± 0.03
PBE0	0.39 ± 0.06	0.53 ± 0.13	0.36 ± 0.04	0.15 ± 0.03	0.37 ± 0.04
B97-1	0.42 ± 0.09	0.62 ± 0.18	0.38 ± 0.05	0.21 ± 0.05	0.40 ± 0.05
PBE	0.48 ± 0.08	0.57 ± 0.15	0.39 ± 0.04	0.16 ± 0.02	0.43 ± 0.04

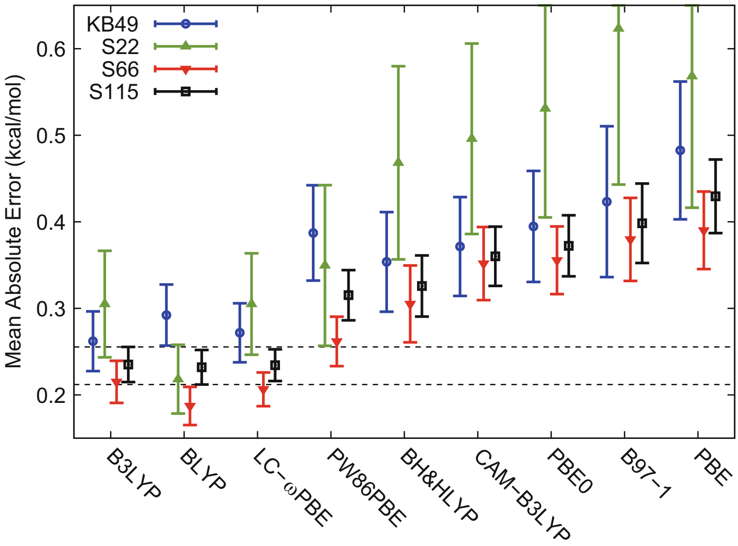

Fig. 5 Mean absolute errors for intermolecular interaction data sets, with the square root of the variance indicated by error bars, for selected DFAs. The numbers in the benchmark names refer the size of each data set

improves. Thus, the choice of statistical indicator determines which would be selected as the best functional and this is an example where the same data can produce opposite interpretations. It is also interesting that B3LYP, BLYP, and LC-ωPBE all give equivalent MAEs for the S115 set, but B3LYP gives a significantly lower MAPE.

As we have demonstrated, using a sufficiently large data set to ensure that the differences between functionals are less than the sum of the uncertainties is of key importance. However, other hidden problems may be encountered when enlarging the test set. For example, a method can perform well for one part of the set, but not

Table 5 MAPEs for the KB49, S22, S66, and HSG sets, together with the S115 superset

Functional	KB49	S22	S66	HSG	S115
B3LYP	6.3	5.0	3.9	10.0	4.9
LC-ωPBE	7.6	5.0	4.3	23.7	5.7
BLYP	9.4	4.8	3.9	9.9	6.2
BH&HLYP	8.1	6.4	5.2	9.5	6.4
CAM-B3LYP	8.2	7.2	6.1	8.4	7.0
PW86PBE	11.3	5.9	6.0	10.5	8.3
PBE0	9.8	8.3	7.4	13.5	8.4
B97-1	11.9	11.7	8.8	14.9	10.1
PBE	13.8	10.5	8.5	10.7	10.8

Table 6 MAPEs for selected methods, for the combined S115 set, divided into two subsets of 35 H-bonded (HB) complexes and 80 other weakly-interacting (WI) complexes

Functional	HB	WI
B97-1	3.0	13.3
B3LYP	3.4	5.6
BLYP	3.4	7.5
LC-ωPBE	3.5	6.7
PW86PBE	3.7	10.3
PBE0	3.9	10.4
PBE	4.2	13.7
BH&HLYP	5.6	6.8
CAM-B3LYP	6.5	7.2

for another. This is the case for intermolecular complexes, where the functionals behave differently depending on whether a dimer is bound by dispersion interactions or hydrogen-bonding. To illustrate this, consider the combined KB49 and S66 sets (called S115 above), which can then be divided into two new subsets: a subset of 35 H-bonded (HB) complexes and a subset of 80 weakly-interacting (WI) complexes. The MAPEs for each subset are given in Table 6 for the selected DFAs. For the WI subset, the lowest MAPE is obtained with B3LYP, and it also gives the second-lowest MAPE for the HB subset, so it performs best for the combined set. However, many of the other functionals do not provide such balanced performance for both interaction types. For the hydrogen-bonding complexes, the best result is obtained with B97-1, although it performs much worse for dispersion-bound complexes. Conversely, CAM-B3LYP and BH&HLYP give the largest MAPEs for H-bonding, but perform much better than B97-1 for dispersion-bound complexes.

When combining data sets, we often wish to determine which functional gives the best balance of errors. As an example of a potential pitfall in such an assessment, let us attempt to judge whether B97-1 or BH&HLYP is most accurate for the union of the HB and WI sets, using ratios of the MAPEs.

On one hand, the defender of B97-1 makes the following argument. True, the ratio of MAPEs for the WI subset ($6.8/13.3 \approx 0.51$) is <1 and favors BH&HLYP. However, for the HB subset, the ratio is >1 ($5.6/3.0 \approx 1.87$) and BH&HLYP is worse for this set. Thus, on average, we get $(1.87 + 0.51)/2 \approx 1.19$ and conclude that the errors of BH&HLYP are larger than those of B97-1 so the latter should be preferred.

On the other hand, the proponent of BH&HLYP makes the following, analogous argument. True, the ratio of MAPEs for the HB subset $(3.0/5.6 \approx 0.54)$ is <1 and favors B97-1. However, for the WI subset, the ratio is >1 $(13.3/6.8 \approx 1.96)$ and B97-1 is worse for this set. Thus, on average, we get $(1.96 + 0.54)/2 \approx 1.24$ and conclude that the errors of B97-1 are larger than those of BH&HLYP so the latter should be preferred.

This illustrates how the same data can support two different conclusions. To understand how it is possible to reach two contradictory conclusions in this fashion, consider the more general case, where the two functionals to be compared are labeled methods 1 and 2 and the data sets are A and B. We denote the average error (MAPE in this case) for functional 1 on set A as $\overline{A_1}$ and follow an analogous notation for the other functional and data set. To compare the errors of the two functionals we average the ratios for both data sets: $\overline{A_1}/\overline{A_2}$ and $\overline{B_1}/\overline{B_2}$. If the result is <1, then functional 1 is judged to perform better. However, this analysis can also be performed using the *inverse* ratios: $\overline{A_2}/\overline{A_1}$ and $\overline{B_2}/\overline{B_1}$. The result depends on which ratio was used because of the different order of operations:

$$\frac{1}{2}\left(\frac{\overline{A}_1}{\overline{A}_2} + \frac{\overline{B}_1}{\overline{B}_2}\right) \neq \frac{1}{2}\left(\frac{\overline{A}_2}{\overline{A}_1} + \frac{\overline{B}_2}{\overline{B}_1}\right) \tag{6}$$

Thus, we see that the way the mean has been produced (arithmetic vs harmonic) can sometimes lead to different conclusions. In particular, using the harmonic mean can favor a method that gives near zero error for one data set.

Finally, we consider the probability that all the results published in an article are significant. As we have seen, based on statistics for the G3 set, density-functional approximations work very well for general thermochemistry. However, there is always the possibility of obtaining an error larger than one wants to accept for useful chemical predictions and this risk increases as more results are generated.

For example, assume that the distribution of the errors for the G3 is representative for the chemical systems under study and that the required accuracy is 0.5 kcal/mol per atom. Then the number of cases where this accuracy is reached in the G3 data set, divided by the size of the set, gives the probability that a single additional calculation yields an accurate result. The highest probability to obtain the desired accuracy is provided by B3LYP, but it is only 0.64. Thus, if ten new B3LYP calculations are performed, for systems having the same error distribution as the G3 data set, the probability that all results possess the needed accuracy is extremely small $(0.64^{10} = 0.01)$.

One can also impose less strict requirements to judge a result as accurate. Figure 6 shows how the probability of obtaining one, two, five, or ten accurate results for the G3 set evolves when gradually increasing the acceptance bar. For

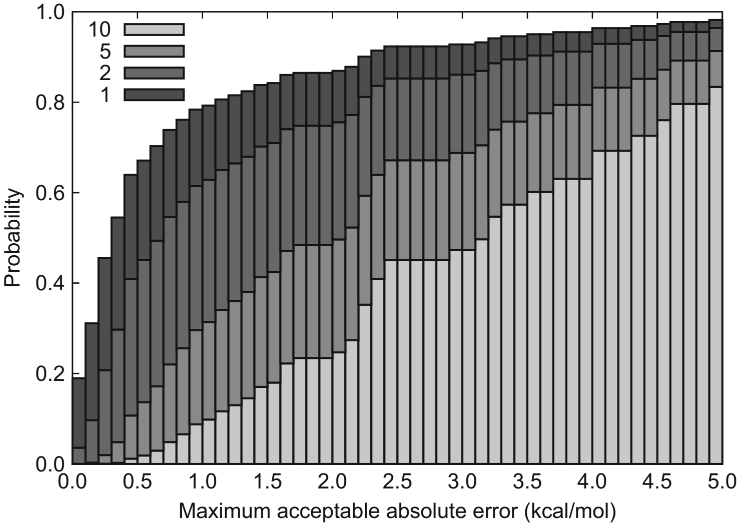

Fig. 6 The probability to obtain B3LYP atomization energies with absolute errors per atom less than a chosen maximum acceptable value, for a set of *n* systems and assuming the same error distribution as the G3 data set

Table 7 Probabilities of a single calculated result having an error less than a particular accuracy threshold (in kcal/mol) for either the G3 or KB49 data sets, with selected functionals. The values for the G3 set refer to errors per atom

	G3		KB49
Functional	<0.5	<1	<0.5
B3LYP	0.64	0.78	0.86
CAM-B3LYP	0.57	0.77	0.73
B97-1	0.43	0.80	0.78
LC-ωPBE	0.45	0.68	0.84
BLYP	0.41	0.68	0.80
PBE0	0.38	0.65	0.73
PW86PBE	0.29	0.48	0.67
PBE	0.05	0.15	0.63
BH&HLYP	0.03	0.08	0.71

example, if one wants to accept errors within ±2 kcal/mol per atom, the probability of reaching the desired accuracy for ten calculations is now 0.23.

Similar effects are observed for other data sets and functionals, as shown in Table 7. For example, if we repeat this analysis for the KB49 set, the probability of obtaining a required accuracy of 0.5 kcal/mol from a single B3LYP calculation is very high at 0.86. However, the probability of having ten sufficiently accurate results decreases to 0.22, and to 0.05 for 20. Even if the probability of obtaining a reliable result is high, the probability that all future calculations are accurate becomes low as the number of published results increases.

4 Summary

There is no doubt that density-functional approximations have enhanced the field of computational chemistry. This would not have been the case without the ability of DFAs to produce interesting and reliable data. As the use of statistical measures to assess the quality of DFAs is valuable and necessary, the number of benchmarking studies has been growing rapidly. It should be pointed out, however, that such statistics-based judgments are subject to several potential pitfalls. For atomization energies, commonly used in parameterization of new functionals, the mean absolute error is not well defined in the limit of large molecular size and errors per atom are a preferable statistic. Large data sets are critical for ranking of functionals to minimize the variance, although they can include a variety of effects and can thus blur one's judgment of the functionals. It is even possible to reach opposite conclusions using the same data, for example depending on the choice of statistical measure. Finally, publishing more data naturally augments the risk of including some data with unsatisfactory accuracy. Statistical data used to judge the quality of density-functional approximations must be carefully analyzed and understood in advance of drawing conclusions.

Acknowledgements The authors would like to thank Prof. J. Percus and Prof. O. Percus (both NYU, USA), Dr. Alberto Otero-da-la-Roza (NINT, Canada), and Stephen Dale (UC Merced) for stimulating discussions.

References

1. Civalleri B, Presti D, Dovesi R, Savin A (2012) On choosing the best density functional approximation, specialist periodical reports. Chem Model Appl Theory 9:168–185
2. Hao P, Sun J, Xiao B, Ruzsinszky A, Csonka G, Tao J, Glindmeyer S, Perdew JP (2013) J Chem Theory Comput 9:355
3. Otero-de-la Roza A, Johnson ER (2013) Non-covalent interactions and thermochemistry using XDM-corrected hybrid and range-separated hybrid density functionals. J Chem Phys 138:204109
4. Kannemann FO, Becke AD (2010) van der Waals interactions in density-functional theory: intermolecular complexes. J Chem Theory Comput 6:1081–1088
5. Jurečka P, Šponer J, Černý J, Hobza P (2006) Benchmark database of accurate (MP2 and CCSD (T) complete basis set limit) interaction energies of small model complexes, DNA base pairs, and amino acid pairs. Phys Chem Chem Phys 8:1985–1993
6. Marshall MS, Burns LA, Sherrill CD (2011) Basis set convergence of the coupled-cluster correction, $\Delta_{MP2}^{CCSD(T)}$: best practices for benchmarking noncovalent interactions and the attendant revision of the S22, NBC10, HBC6, and HSG databases. J Chem Phys 135:194102
7. Rezac J, Riley KE, Hobza P (2011) S66: a well-balanced database of benchmark interaction energies relevant to biomolecular structures. J Chem Theory Comput 7:2427–2438
8. DiLabio GA, Johnson ER, Otero-de-la Roza A (2013) An evaluation of the performance of conventional and dispersion-corrected density-functional theory methods for hydrogen bonding interaction energies. Phys Chem Chem Phys 15:12821

9. Faver JC, Benson ML, He X, Roberts BP, Wang B, Marshall MS, Kennedy MR, Sherrill CD, Merz KM (2011) Formal estimation of errors in computed absolute interaction energies of protein-ligand complexes. J Chem Theory Comput 7:790–797
10. Curtiss LA, Jones C, Trucks GW, Raghavachari K, Pople JA (1990) Gaussian-1 theory of molecular energies for second-row compounds. J Chem Phys 93:2537–2545
11. Curtiss LA, Raghavachari K, Redfern PC, Pople JA (1997) Assessment of Gaussian-2 and density functional theories for the computation of enthalpies of formation. J Chem Phys 106:1063–1079
12. Curtiss LA, Raghavachari K, Redfern PC, Pople JA (2000) Assessment of Gaussian-3 and density functional theories for a larger experimental test set. J Chem Phys 112:7374–7383
13. Becke AD (1988) Density-functional exchange-energy approximation with correct asymptotic behavior. Phys Rev A 38:3098–3100
14. Lee C, Yang W, Parr RG (1988) Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. Phys Rev B 37:785–789
15. Becke AD (1993) Density-functional thermochemistry. III. The role of exact exchange. J Chem Phys 98:5648–5652
16. Becke A (1993) A new mixing of Hartree–Fock and local density-functional theories. J Chem Phys 98:1372
17. Hamprecht F, Cohen A, Tozer D, Handy N (1998) Development and assessment of new exchange-correlation functionals. J Chem Phys 109:6264
18. Yanai T, Tew DP, Handy NC (2004) A new hybrid exchange-correlation functional using the Coulomb-attenuating method (CAM-B3LYP). Chem Phys Lett 393:51–57
19. Vydrov OA, Scuseria GE (2006) Assessment of a long-range corrected hybrid functional. J Chem Phys 125:234109
20. Vydrov OA, Heyd J, Krukau AV, Scuseria GE (2006) Importance of shortrange versus long-range Hartree-Fock exchange for the performance of hybrid density functionals. J Chem Phys 125:074106
21. Perdew J, Burke K, Ernzerhof M (1996) Generalized gradient approximation made simple. Phys Rev Lett 77:3865–3868
22. Adamo C, Barone V (1999) Toward reliable density functional methods without adjustable parameters: the PBE0 model. J Chem Phys 110:6158–6170
23. Perdew JP, Wang Y (1986) Accurate and simple density functional for the electronic exchange energy: generalized gradient approximation. Phys Rev B 33:8800
24. Becke AD, Johnson ER (2007) Exchange-hole dipole moment and the dispersion interaction revisited. J Chem Phys 127:154108