Prediction Uncertainty of Density Functional Approximations for Properties of Crystals with Cubic Symmetry

Pascal Pernot, Bartolomeo Civalleri, Davide Presti, Andreas Savin

Prediction uncertainty of density functional approximations for properties of crystals with cubic symmetry

Prediction uncertainty of density functional approximations for properties of crystals with cubic symmetry

Pascal Pernot,[*,†,‡] Bartolomeo Civalleri,[¶] Davide Presti,[§] and Andreas Savin[‖,⊥]

CNRS, UMR8000, Laboratoire de Chimie Physique, F-91405 Orsay, France, Univ. Paris-Sud, UMR000, Laboratoire de Chimie Physique, F-91405 Orsay, France, Department of Chemistry and NIS Center, University of Torino, Via P. Giuria 7, I-10125 Torino, Italy, Department of Chemical and Geological Sciences, University of Modena and Reggio-Emilia, Via Campi 183, I-41125 Modena, Italy, CNRS, UMR7616, Laboratoire de Chimie Théorique, F-75005 Paris, France, and UPMC Univ Paris 06, UMR7616, Laboratoire de Chimie Théorique, F-75005 Paris, France

E-mail: pascal.pernot@u-psud.fr

Abstract

The performance of a method is generally measured by an assessment of the errors between the method's results and a set of reference data. The *prediction uncertainty* is a measure of the confidence that can be attached to a method's prediction. Its estimation is based on the *random* part of the errors not explained by reference data uncertainty, which implies an evaluation of the *systematic* component(s) of the errors. As the predictions of most density functional approximations (DFA) present systematic errors, the standard performance statistics such as the mean of the absolute errors (MAE or MUE), cannot be directly used to infer prediction uncertainty. We investigate here an *a posteriori* calibration method to estimate the prediction uncertainty of DFAs for properties of solids. A linear model is shown to be adequate to address the systematic trend in the errors. The applicability of this approach to modest-size reference sets (28 systems) is evaluated for the prediction of band gaps, bulk moduli and lattice constants with a wide panel of DFAs.

Keywords: virtual measurement; calibration statistics; band gap; lattice constant; bulk modulus.

[*]To whom correspondence should be addressed

[†]CNRS, UMR8000, Laboratoire de Chimie Physique, F-91405 Orsay, France

[‡]Univ. Paris-Sud, UMR000, Laboratoire de Chimie Physique, F-91405 Orsay, France

[¶]Department of Chemistry and NIS Center, University of Torino, Via P. Giuria 7, I-10125 Torino, Italy

[§]Department of Chemical and Geological Sciences, University of Modena and Reggio-Emilia, Via Campi 183, I-41125 Modena, Italy

[‖]CNRS, UMR7616, Laboratoire de Chimie Théorique, F-75005 Paris, France

[⊥]UPMC Univ Paris 06, UMR7616, Laboratoire de Chimie Théorique, F-75005 Paris, France

1 Introduction

The success of density functional theory, of modern algorithms and computers has produced not only a large amount of numerical results, but of also a large number of Density Functional approximations (DFA). To choose amongst those, benchmark data sets are increasingly used. Although this should be seen as a quantification of experience, one should be also warned that using statistical tools to quan-

tify DFAs performance has its pitfalls, and care is needed[1].

If ranking is a concern for DFA designers to assess the overall performance of new developments, it is less practically useful to end users, who need to select a method with criteria such as code availability, computing performance, and, most important, *prediction uncertainty*. The later provides a confidence measure on the results of a DFA for a given property. If, in addition to performance statistics, users are informed of the prediction uncertainty of DFAs, they might have a better rationale to select a method satisfying their specific requirements.

The definition of prediction uncertainty for computational chemistry methods has been formalized by Irikura *et al.*[2] in the *Virtual Measurement* (VM) framework. The interest of VM is to define a statistical approach in agreement with international standards for the evaluation of measurement uncertainty, as recommended by the *Guide to the Expression of Uncertainty in Measurement* (GUM)[3]. The VM approach has been adopted by the National Institute of Standards and Technology (NIST), notably for its Computational Chemistry Comparison and Benchmark Database (CCCBDB)[2]. The VM framework has been reported in the computational chemistry literature mostly to estimate prediction uncertainty for scaled harmonic and anharmonic vibrational frequencies and zero-point energies[2,4–10]. Recently, Ruscic[11] strongly recommended its use to improve the uncertainty evaluation of predicted thermochemical quantities. The interest of this approach has also been demonstrated for molecular simulation[12–14].

In the GUM approach to uncertainty estimation, "it is assumed that the result of a measurement has been corrected for all recognized significant systematic effects and that every effort has been made to identify such effects"[3]. This is a key point which is challenging for computational chemistry, where most error sources are known to be systematic, due to the various approximations in the chemistry models. The correction of systematic errors can only be achieved by comparison with reference data. The assessment of a prediction uncertainty requires therefore either an *internal* calibration (adjustment of parameters) of a method against a reference data set, or an *a posteriori* calibration of the results of this method. We address the latter approach in this article.

The *internal* calibration of semi-empirical DFAs followed by propagation of the uncertainty on calibrated parameters to predictions has been sparsely reported[15–17], and recently applied to computational catalysis[18,19]. Few studies along similar lines have also been reported for molecular simulation force fields[20–22].

In the *a posteriori* approach, calibration is used to remove the predictable part of the errors (*systematic errors*). Prediction uncertainty for a method is then derived from the remaining, unpredictable part of the errors (*random errors*)[3]. In the aforementioned vibrational frequency applications, correction of systematic errors is done through the scaling of the calculated data, and the root-mean-square of the errors (RMSE) of *scaled* vibrational frequencies has been shown to provide, under mild conditions, a good approximation of prediction uncertainty[8,9]. This scaling approach has been used recently by Lejaeghere *et al.*[23] to estimate the prediction errors of solid-state DFAs for elemental crystals. As will be shown below, two points need to be addressed to complement these scaling studies for other systems and properties: (i) the calibration model cannot always be reduced to a simple scaling, and (ii) the reference data uncertainties are not always small enough to be neglected in the statistical analysis.

We present therefore a detailed derivation of prediction uncertainty of computational methods by *a posteriori* calibration in a more general framework than for a single scaling factor, moreover taking into account the uncertainty on the reference data. The method is applied to the calculation of lattice constants, bulk moduli and band gaps for a set of 28 crystals (semiconductors and insulators) with cubic symmetry, by 18 different DFAs (local, semi-local, and hybrid).

The paper consists of four main sections. In the first part, we present the difficulty of deriving prediction uncertainty from common per-

formance statistics provided in the benchmark literature and the need to design a specific approach. In the second part, we illustrate a typical distribution of errors observed in the application cases and derive an adequate stochastic calibration/prediction model. In the third section, we apply the calibration/prediction model to the reference data and study its validity. Discussion of the advantages and limitations of the VM approach in the context of this study is the object of the fourth section.

2 Method performance evaluation

Performance evaluation of computational chemistry methods relies on two ingredients: a benchmark data set used as a reference to assess calculation accuracy, and performance statistics on the differences between calculations and reference data (see *e.g.* Peverati and Truhlar[24] for a recent review). Both ingredients play a crucial role in performance assessment.

2.1 Definitions

We thereafter call *error* the difference between the value of a property, $c_{m,s}$, calculated for a system s by a method (*e.g.* DFA) m, and the corresponding reference value, o_s (observed or calculated):

$$e_{m,s} = c_{m,s} - o_s. \tag{1}$$

For performance assessment of a method, one uses statistics summarizing the error sets containing the error values of all systems for a given method, $E_m = \{e_{m,s};\, s = 1, N_s\}$, where N_s is the number of systems in the reference set.

In the following, we consider deterministic methods and assume that all sources of code uncertainty are controlled at a negligible level (numerical errors, convergence thresholds effects, etc.[2]).

In this case, the errors can be attributed (i) to reference data uncertainty, u_s, and, if this source alone cannot explain the amplitude of the errors, (ii) to *method inadequacy* errors, characterizing the inability of a method to predict the reference data within their error bars. The uncertainty of the reference data is therefore a key information to properly assess method inadequacy errors.

2.2 Performance estimators: MAD vs. MAD

Several performance statistics are commonly used in the benchmark literature to rank methods. We review these estimators in order to appreciate their usability, or lack thereof, in the estimation of prediction uncertainty.

First, there is some confusion in the computational chemistry literature about the nomenclature of the performance statistics. In particular, the use of some acronyms conflicts with the standard use in the statistical literature. The main example is the *mean absolute deviation* (MAD), which is commonly used in the community to refer to the *mean of the absolute errors* (MAE)

$$MAE = \frac{1}{N_s} \sum_{s=1}^{N_s} |e_{m,s}|\,, \tag{2}$$

whereas for statisticians[25] MAD is a measure of dispersion around a reference point, either the *mean absolute deviation* (from the arithmetic mean $\overline{E_m}$),

$$M[ean]AD = \frac{1}{N_s} \sum_{s=1}^{N_s} \left|e_{m,s} - \overline{E_m}\right|, \tag{3}$$

or the *median absolute deviation* (from the median $\mathrm{med}(E_m)$)

$$M[edian]AD = \mathrm{med}\left|E_m - \mathrm{med}(E_m)\right|. \tag{4}$$

Synonyms of MAE in the computational chemistry literature are the *mean unsigned error/deviation* (MUE/D) and the *average absolute error/deviation* (AAE/D). The occasional occurrence in this corpus of a meaningless definition of MAD as *mean average deviation* is even more confusing.[26]

The arithmetic mean is often referred to as

mean signed error (MSE)

$$MSE = \overline{E_m} = \frac{1}{N_s}\sum_{s=1}^{N_s} e_{m,s}. \qquad (5)$$

Uncertainty is defined in the International Vocabulary of Metrology[27] as a "non-negative parameter characterizing the dispersion of the quantity values being attributed to a measurand". With regard to this definition, it is important to acknowledge that MAE and $MeanAD$ are different statistics:

- $MeanAD$ is a measure of dispersion (for a normal distribution of standard deviation σ, one has $MeanAD = \sqrt{2/\pi}\sigma$);
- for $\overline{E_m} = 0$, MAE and $MeanAD$ are identical, but when $\overline{E_m} \neq 0$, MAE is a non-invertible mixture of dispersion and location statistics. In the extreme case where all errors are positive, MAE is equal to MSE, a measure of location.

In a recent paper intended on clarifying the difference between MAE and prediction uncertainty, Ruscic[11] addresses MAE (called MAD in the paper, but unambiguously synonymized with MUE) as a dispersion measure, which it is not for non-zero-centered error samples, the standard case in computational chemistry. The MAE can be used, amongst many other criteria, to rank methods, but *should not be used to assess the uncertainty associated with a given method*.

The same remarks apply to the *root-mean-square error* (RMSE)

$$RMSE = \sqrt{\frac{1}{N_S}\sum_{s=1}^{N_s} e_{m,s}^2}, \qquad (6)$$

which is commonly used alongside the MAE in the benchmark literature. The corresponding measure of dispersion is the *root-mean-square deviation* (RMSD)

$$RMSD = \sqrt{\frac{1}{N_S}\sum_{s=1}^{N_s} (e_{m,s} - \overline{E_m})^2}. \qquad (7)$$

The equality

$$RMSE^2 = RMSD^2 + MSE^2 \qquad (8)$$

lets clearly appear the $RMSE$ as a mixture of location and dispersion measures. The interest of $RMSE$ in the context of performance measures is also disputed, because of the better robustness of MAE to outliers, but the debate is ongoing[28,29].

2.3 From performance estimators to prediction uncertainty

If an error set is affected by a *constant* (*i.e.* system-independent) systematic contribution, then the MSE estimates the mean value of the systematic error, and the $RMSD$ provides the standard deviation of the remaining (random) part of the errors. In the case of a negligible contribution of the reference data uncertainty, the $RMSD$ and the uncertainty on the MSE could then be combined to estimate a prediction uncertainty.

As will be illustrated in the next section, DFAs do not generally produce only constant systematic errors[23]. Additional corrections are necessary to access the random contribution of the errors. Moreover, dispersion statistics ($MeanAD$ and $RMSD$) are not always provided in the benchmark literature, preventing the estimation of prediction uncertainty from existing benchmarks. An exception is for scaled harmonic frequencies, where the $RMSD$ of scaled frequencies is generally available[9,30].

An additional issue with the MAE and $RMSE$ estimators, is that they do aggregate reference data uncertainty with the errors due to the method. If the reference data uncertainty is not negligible before method inadequacy, these estimators, as they do not average out the random reference data errors, underestimate method performance. The applicability of MAE and $RMSE$ requires therefore the use of high accuracy reference data[11,31], which might be a severe restriction for some properties.

We show in the following how to circumvent these difficulties in a practical way and estimate prediction uncertainty by statistical modeling

of the errors.

2.4 Some issues in the use of benchmark data sets

Besides the requirement for high quality data in the reference sets, one can be confronted with more challenging issues: the experimental data do not necessarily reflect the best, exact reference to be used. There are several reasons for this:

1. The calculated quantities do not necessarily correspond to the experimental data. This can be the case for the fundamental band gaps, see, *e.g.*, Civalleri *et al.*[1].

2. The theoretical method is not necessarily supposed to provide the quantity analyzed (Kohn-Sham orbital energies, even exact, are not supposed to provide fundamental band gaps[32–34]).

3. The experimental data are subject to factors that are not properly taken into account (*e.g.*, temperature, in particular for bulk modulus).

4. The inclusion of the systems into the benchmark data set is conditioned to data availability, which introduces a bias in the representativity of the data set.

3 Prediction uncertainty estimation

In order to estimate a prediction uncertainty, a four-steps procedure is used:

1. build and validate a statistical model of the errors from the benchmark set (*calibration model*),

2. evaluate the uncertainties of the parameters involved in this model,

3. propagate the uncertainties of the parameters in the calibration model to the *prediction model*, and

4. validate the prediction model.

Validation in steps 1 and 4 is necessary to ensure that calculated values and reference data agree within the error bars defined by the calibration or prediction model. One generally faces the case where reference data uncertainty alone cannot explain the errors amplitude, and corrections to the calibration model have to be done, either by updating its deterministic part (representing the systematic errors), or its stochastic part (representing the random errors).

3.1 Distribution of errors

Designing a statistical model requires to examine the data and their distribution. We illustrate the process on the case of the B3LYP DFA for lattice constants (LC), extracted from the full application set described in Section 4.1. This example is well representative of the other cases considered in the present article.

3.1.1 Systematic and random errors

Fig. 1 (a) displays a scatter plot of the reference data *vs.* the calculated data. One observes that the points are grouped along a line which is not the identity line. This is an evidence of the presence of systematic errors and of a trend in the systematic errors, which in this case increase as a function of the calculated property value.

Plotting the errors (set $E_{LC,B3LYP}$) against the calculated values (Fig. 1 (b)) reveals more clearly the trend: the errors increase more or less linearly with the value of the lattice constant and have a non-null mean value. The orange line represents the least squares linear fit of the errors. The MSE and $MSE \pm RMSD$ values of these data are represented by horizontal full and dashed lines, respectively.

The trend line represents a systematic effect (a deterministic contribution) in the errors, which has to be corrected before we can estimate the random contribution on which the uncertainty estimation is based[3]. The corrected errors, obtained by subtraction of the least-squares regression line are shown in Fig. 1 (b) (triangles) and present a zero-centered, more or less symmetrical distribution. One can see that

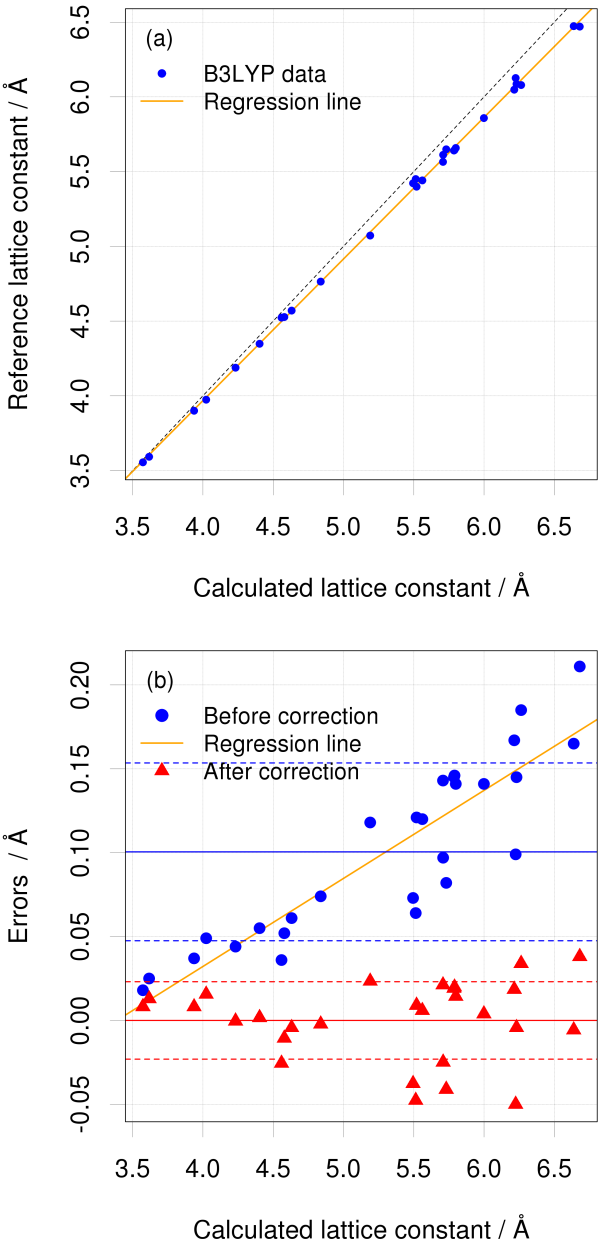


Figure 1: Structure of the errors for lattice constants (LC) calculated by the B3LYP method: (a) Reference *vs.* calculated data (bullets), the least-squares regression line through the points is the red/solid line, while the black/dashed line represents the identity line; (b) Errors *vs.* calculated values, before and after linear correction. A linear trend (orange line) can be assigned to a systematic (predictable) component of the initial errors. The horizontal lines in (b) represent the MSE (full line) and $MSE \pm RMSD$ (dashed lines) of the corresponding data sets.

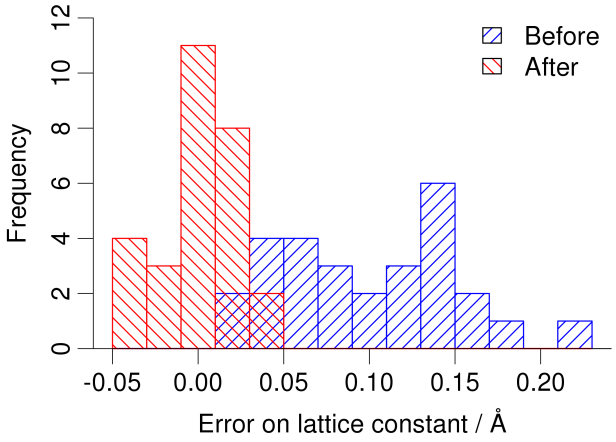


Figure 2: Distribution of errors obtained using the B3LYP functional to predict lattice constants, before and after linear correction of systematic errors. The histograms are produced by distributing the data into bins of 0.02Å.

they do not present any obvious trend. Moreover, their $RMSD$ is smaller than for the uncorrected errors.

Histograms of the errors before and after linear correction can be compared in Fig. 2. It is interesting to contrast the width of the corrected distribution (about 0.02Å) with the typical measurement errors on lattice constants, which are considered to be an order of magnitude smaller (about 0.001Å, see Section 4.1). One has thus to face the fact that, even after a linear correction, the B3LYP DFA cannot predict the reference data within their uncertainty range.

3.1.2 From deterministic calculations to random errors

Considering the very small uncertainty on lattice constants, the errors in the $E_{LC,B3LYP}$ set can be mainly attributed to the method's inability to reproduce reference data, and decomposed in Fig. 1 into predictable/systematic and unpredictable/random contributions. In the following, we will refer to the random part of method inadequacy as *method inadequacy error*, the systematic part being addressed through corrections.

The method inadequacy error has a random-like trace as a function of lattice constant value

(Fig. 1 (b)), despite the fact that the model chemistry (*i.e.* method and basis set) calculations are deterministic[23]. This is not truly a random process, in the sense that repeated calculations with a model chemistry for the same systems would provide the same values, but its variation with the lattice constant value is practically unpredictable without doing the calculation. Moreover, for a given basis set, this random contribution is irreducible without changing the DFA or splitting the reference data set (if it appears heterogeneous).

The method inadequacy error represents therefore our *lack of knowledge* on the prediction of properties of new systems, due to the use of an approximate method, and to some extent, of a limited reference data set.

This pattern can be exploited to define and estimate prediction uncertainty by modeling method inadequacy error by a random variable, as detailed in the next section.

3.2 Calibration/prediction statistical modeling

In this section we present the implementation of the VM framework by an *a posteriori* calibration model, enabling (1) to correct for the systematic errors of a method, (2) to evaluate the method inadequacy uncertainty, and (3) to estimate the prediction uncertainty of the calibrated method.

3.2.1 Calibration

Let us start with the simplest statistical model linking the calculated values ($c_{m,s}$) and the uncertain reference values ($o_s \pm u_s$)

$$o_s = c_{m,s} + \epsilon_s \, (s = 1,\, N_s), \tag{9}$$

where the ϵ_s are independent random variable of mean 0 and *known*, finite, standard deviation u_s. This model is a generalization of Eq. 1: it uses random variables ϵ_s to describe stochastic processes from which one assumes that the actual errors $e_{m,s}$ are realizations.

In most cases of interest in the present study and many others, this model is *invalid*, in the sense that the values calculated by a given DFA are not compatible with the reference values within their uncertainty range.[35]

In order to get a valid calibration model, one has to account for the structure of the errors set. A *systematic trend* observed in the error sets of the benchmark data can be corrected by a transformation of the calculated values $c_{m,s}$, providing a new (calibration) model

$$o_s = f_m(c_{m,s}; \boldsymbol{\vartheta}_m) + \epsilon_s, \tag{10}$$

where $\boldsymbol{\vartheta}_m$ represents the set of parameters defining f_m. The functional form and parameters values of f_m are method-dependent. It is important for the prediction ability of the model to choose a functional form which does not overfit the data. One can always find a high-degree polynomial fitting exactly all points in the errors set. However this kind of correction has no generalizability, *i.e.* it performs poorly at the prediction stage. Low-order polynomials or functions with few parameters (compared to N_s) should be preferred.

After optimization of the parameters $\boldsymbol{\vartheta}_m$ ($\hat{\boldsymbol{\vartheta}}_m$ represents the set of optimal parameters), the validity of the model depends on the comparison between the residual errors

$$r_{m,s} = o_s - f_m(c_m; \hat{\boldsymbol{\vartheta}}_m) \tag{11}$$

and the reference data uncertainties u_s. In the least-squares optimization framework, one compares the chi-square value

$$\chi^2 = \sum_{s=1,N_s} \left(\frac{r_{m,s}}{u_s} \right)^2 \tag{12}$$

to the number of degrees of freedom $n_{df} = N_s - N_\vartheta$, where N_ϑ is the number of free parameters in f_m[36,37].

If $\chi^2 \simeq n_{df}$ the corrected model can be considered as valid, and the prediction uncertainty will be limited to the parametric uncertainty of the correction function, as defined below (Eq. 16).

In most practical cases however, the reference data uncertainties are small compared to the residual errors, which invalidates this calibra-

tion model ($\chi^2 \gg n_{df}$). If the residual errors still present discernible trends, the correction function f_m has to be updated.

If the residual errors $r_{m,s}$ present a random-like pattern, for which no further deterministic correction appears suitable, one introduces a new stochastic term δ_m, to describe the dispersion of the errors in excess of the reference data uncertainty, that we attribute to method inadequacy

$$o_s = f_m(c_{m,s}; \hat{\boldsymbol{\vartheta}}_m) + \epsilon_s + \delta_m, \qquad (13)$$

where δ_m is a random, unpredictable, variable of mean 0 (systematic errors are corrected by f_m) and *unknown*, finite, standard deviation d_m. The value of d_m is estimated in order to ensure the statistical validity of Eq. 13.

Practically, d_m^2 can be chosen as the difference between the variance of the residual errors $r_{m,s}$ and the mean variance of the reference data (see Appendix A). With this choice, the corrected calculated values and the reference data are compatible within the combination of their respective error bars.

3.2.2 Prediction

For the estimation of a new value of a property knowing a calculated value c^* (*i.e.* for a system not in the benchmark set), the prediction model and prediction variance are[3]

$$\begin{aligned} p_m(c^*) &= f_m(c^*; \hat{\boldsymbol{\vartheta}}_m) + \hat{\delta}_m \qquad (14) \\ u^2_{p_m}(c^*) &= u^2_{f_m}(c^*; \hat{\boldsymbol{\vartheta}}_m) + d_m^2 \qquad (15) \end{aligned}$$

where $\hat{\delta}_m \equiv 0$ has been left in the prediction equation as a reminder of the occurrence of d_m^2 in the prediction variance.

The term $u_{f_m}(c^*; \hat{\boldsymbol{\vartheta}}_m)$ represents the parametric uncertainty on the value of the function f_m at c^*. This contribution results from the uncertainty in the optimal parameters set due to the stochastic terms in Eq. 13. For functional forms of f_m linear in $\boldsymbol{\vartheta}_m$ or showing weak non-linearity on the variation domain of $\boldsymbol{\vartheta}_m$, it can be estimated by combination of variances[3,38]

$$u^2_{f_m}(c^*; \hat{\boldsymbol{\vartheta}}_m) = J^T \Sigma^2_{\vartheta_m} J, \qquad (16)$$

where J is a vector of sensitivity coefficients evaluated at $\boldsymbol{\vartheta}_m = \hat{\boldsymbol{\vartheta}}_m$,

$$J_i = \left. \frac{\partial f_m(c^*; \boldsymbol{\vartheta}_m)}{\partial \vartheta_{m,i}} \right|_{\hat{\boldsymbol{\vartheta}}_m}, \qquad (17)$$

and $\Sigma^2_{\vartheta_m}$ is the variance-covariance matrix of the parameters. For highly non-linear functions, Monte Carlo uncertainty propagation can be used[39].

3.2.3 The linear case

A linear transformation function, $f_m(x; a_m, b_m) = a_m + b_m(x)$, will be used in this study, leading to the calibration model

$$o_s = a_m + b_m c_{m,s} + \epsilon_s + \delta_m. \qquad (18)$$

Weighted least-squares regression can be used to estimate the optimal values of all parameters, $\hat{a}_m$, $\hat{b}_m$, $\hat{d}_m$, and the uncertainties and covariance of the line parameters $u(a_m)$, $u(b_m)$ and $u(a_m, b_m)$. The details are provided in Appendix A.

The prediction model and prediction variance are

$$\begin{aligned} p_m(c^*) &= \hat{a}_m + \hat{b}_m c^* \qquad (19) \\ u^2_{p_m}(c^*) &= u^2_{f_m}(c^*; \hat{a}_m, \hat{b}_m) + \hat{d}_m^2 \qquad (20) \\ u^2_{f_m}(c^*; \hat{a}_m, \hat{b}_m) &= u^2(a_m) + c^{*2} u^2(b_m) \qquad (21) \\ &\quad + 2c^* u(a_m, b_m). \qquad (22) \end{aligned}$$

The prediction uncertainty u_{p_m} depends on the calculated value c^*. However, if the benchmark set is large enough and if c^* lies within the range covered by the benchmark set (no extrapolation), the uncertainty on the calibration model can become negligible before d_m [9], and Eq. 20 reduces to

$$u_{p_m}(c^*) \simeq d_m. \qquad (23)$$

This convenient approximation will be tested in the next section.

We insist on the fact that the prediction uncertainty u_{p_m} has two contributions: the method inadequacy error d_m and the correction model uncertainty u_{f_m}. An example of the relative contributions of these quantities is shown in Fig. 3, where the major contribution of d_m can

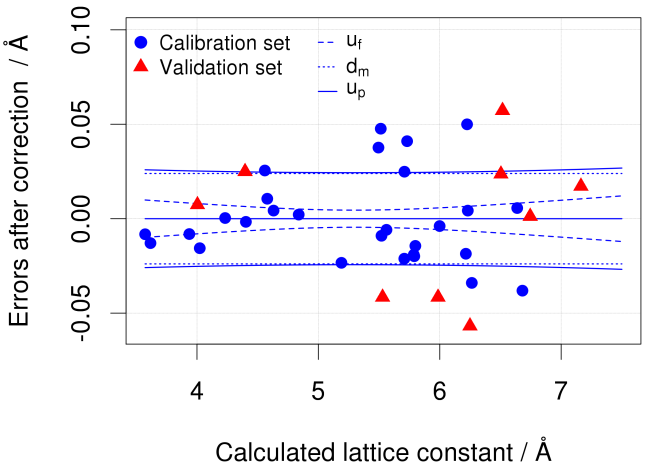


Figure 3: Prediction uncertainty using the B3LYP functional for lattice parameters: the dashed lines represent the contribution of the calibration model uncertainty, $\pm u_f$; the dotted lines represents the method inadequacy error contribution, $\pm d_m$; the full lines is the total prediction uncertainty, $\pm u_p$; the blue bullets are the residual errors for the reference set used for calibration and the red triangles are the residual errors for the data in the validation set.

be appreciated. In terms of variance, the contribution of $u_{f_m}^2$ to $u_{p_m}^2$ is about 20 % at the extremities of the plotted lattice constant range, which corresponds to a value of u_{p_m} larger than d_m by about 10 %. As shown by Pernot and Cailliez[8,9], ignoring method inadequacy errors leads to unreliable prediction uncertainty estimations.

4 Application

4.1 Benchmark and validation data

We analyze the lattice constant, bulk modulus and band gap for a set of 28 crystals with cubic symmetry (semiconductors and insulators) and compare 18 different density functional approximations (local, semi-local, and hybrid functionals). All calculations have been carried out with the CRYSTAL14 code.[40,41] All-electron and effective-core potentials calculations have been done by using atom-centerd Gaussian-type basis sets. The latter have been taken from Ref.[42], except for alkali halides and $SrTiO_3$ for which a triple-zeta quality basis set has been employed. The full set of data is reported in the Supplementary Material.

4.1.1 Choice of reference data

Reference data were collected for the following crystals (*Strukturbericht* designation[66] in parentheses): 22 semiconductors, also present in the SC40 data set[42], namely : C(A4), Si(A4), Ge(A4), SiC(B3), BN(B3), BP(B3), BAs(B3), AlP(B3), AlAs(B3), AlSb(B3), GaN(B3), GaN(B4), GaP(B3), GaAs(B3), GaSb((B3), InP(B3), InAs(B3), InSb(B3), ZnS(B3), ZnSe(B3), ZnTe(B3), CdTe(B3), MgS(B1); 4 alkali halides: LiF(B1), LiCl(B1), NaF(B1) and NaCl(B1); and two oxides: MgO(B1), $SrTiO_3$(E2_1).

The reference dataset includes: (1) experimental lattice constant values corrected for the zero-point anharmonic expansion, as reported in Ref.[67]; (2) experimental bulk modulus values, taken from Refs.[57,68–70]; and (3) low temperature (below 77 K) experimental (fundamental) band gap values[42,69,71,72].

For bulk modulus, we referred to low temperature data[57,68,69], if available, and, when possible, the zero-point anharmonic expansion correction has been included from Ref[57]. The band gaps considered cover two orders of magnitude, between ≈0.2 and ≈12 eV.

Validation data. A set of 9 system has been set aside for validation purpose. These are systems for which we did not find bulk modulus reference data: AlN(B3), CdS(B3), CdSe(B3), MgSe(B1), MgTe(B1), BaS(B1), BaSe(B1), BaTe(B1), and LiH(B1).

Reference data uncertainties. Concerning the error bars for lattice constants, the uncertainty from X-ray diffraction experiments depends on the sample (*i.e.* powder or single crystals) and on the instrument/detector. It is claimed that the uncertainty can reach 0.0001Å or even smaller[73,74]. However, due to the procedure adopted to obtain the reference ZPAE-corrected data, that mixes experimental lattice

Table 1: List of the DFT methods assessed in the present work. Parameters are also reported for global (GH) and range-separated hybrid (RSH) exchange functionals.

Method	Name	Exchange	c_{SR}	c_{MR}	c_{LR}	ω_{SR}	ω_{LR}	Correlation	Ref.
HF	HF	HF	-	-	-	-	-	-	
LDA	SVWN	S	-	-	-	-	-	VWN	[43,44]
GGA	PBE	PBE	-	-	-	-	-	PBE	[45]
	PBEsol	PBEsol	-	-	-	-	-	PBEsol	[46]
mGGA	M06-L	M06-L	-	-	-	-	-	M06-L	[47]
GH-GGA	B3LYP	B88	0.20	0.20	0.20	0.00	0.00	LYP	[43,44,48–50]
	B97	B97	0.21	0.21	0.21	0.00	0.00	B97	[51,52]
	PBE0	PBE	0.25	0.25	0.25	0.00	0.00	PBE	[45,53,54]
	PBEsol0	PBEsol	0.25	0.25	0.25	0.00	0.00	PBEsol	[46]
GH-mGGA	M06	M06	0.27	0.27	0.27	0.00	0.00	M06	[55]
SC-RSH	HSE06	PBE	1.00	0.00	0.00	0.11	0.11	PBE	[45,56]
	HSEsol	PBEsol	1.00	0.00	0.00	0.11	0.11	PBEsol	[46,57]
MC-RSH	HISS	PBE	0.00	0.60	0.00	0.84	0.20	PBE	[45,58,59]
LC-RSH	LC-ωPBE	PBE	0.00	0.00	1.00	0.40	0.40	PBE	[45,60]
	LC-ωPBEsol	PBEsol	0.00	0.00	1.00	0.60	0.60	PBEsol	[46,60]
	RSHXLDA	S	0.00	0.00	1.00	0.40	0.40	VWN	[44,61–64]
	ωB97	B97	0.00	0.00	1.00	0.40	0.40	B97	[52,65]
	ωB97-X	B97	0.157706	0.00	1.00	0.30	0.30	B97	[52,65]

constants and computed ZPAE corrections, we assume that an uncertainty of 0.001Å is more representative.

For band gaps, most of the reference data correspond to low temperature (LT) values, but some of them have been measured at room temperature (RT). When comparing LT and RT data, as reported by Lucero *et al.*[75], the former are systematically larger than the latter by 0.10 eV on average (23 systems), with a maximum difference of 0.30 eV. However, from Ref.[71] and reference therein, the error bar for the band gaps ranges from 0.001 eV, or less, up to 0.01 eV. This depends on the experimental technique adopted to measure it (e.g. diffuse reflectance, photoluminescence spectroscopy,...) but is more or less independent from the temperature. Therefore, we consider 0.01 eV as the uncertainty for reference experimental band gaps.

The experimental uncertainties for bulk modulus range from a few tenths of GPa up to 4-5 GPa. Again, it depends on the measurement approach: either from equation of state data by means of x-ray diffraction measurements, usually for a given hydrostatic path, or through the knowledge of elastic constants. For the latter, various techniques can be employed (e.g. Brillouin scattering, ultrasonic resonance...) and measurements can be carried out at different temperatures thus allowing extrapolation at the static limit. Here, we refer to an average estimated experimental uncertainty of 2 GPa.

One should consider these global estimations of reference data uncertainties as optimistic. They often result from the simple transcription of experimental repeatability statistics[3,27], without considering additional uncertainties resulting from sample preparation, materials impurities, uncertainty in various corrections, apparatus calibration, etc. A more pessimistic scenario will be explored in Section 4.5.

Choice of density functional approximations. The DF approximations used in the present work can be classified into the following groups:

- local and semi-local density functionals (*i.e.* LDA, GGA and mGGA),

- linear global hybrids (GH, where the density functional exchange is mixed up linearly with the Hartree-Fock exchange), and
- range separated hybrids (RSH).

In the latter class of functionals, the amount of HF exchange included depends on the distance between electrons. They are obtained from the separation of the Coulomb operator in different ranges (three ranges in the current implementation) by means of the *error function* as:

$$\begin{aligned}\frac{1}{r_{12}} &= \underbrace{\frac{\mathrm{erfc}(\omega_{SR} r_{12})}{r_{12}}}_{SR} \\ &+ \underbrace{\frac{1-\mathrm{erfc}(\omega_{SR} r_{12})-\mathrm{erf}(\omega_{LR} r_{12})}{r_{12}}}_{MR} \\ &+ \underbrace{\frac{\mathrm{erf}(\omega_{LR} r_{12})}{r_{12}}}_{LR} \end{aligned} \quad (24)$$

where ω is the length scale of separation. Range separated hybrids can be subdivided in: long-range corrected (LC-RSH), middle-range hybrids (MC-RSH) and short-range corrected (SC-RSH) functionals, also known as screened Coulomb. In these approximations, the long-, middle- and short-range part of the exchange, respectively, is described by Hartree-Fock.

The general form of a range-separated hybrid is:

$$\begin{aligned}E_{xc}^{RSH} &= E_{xc}^{DFA} + c_{SR}(E_{x,SR}^{HF} - E_{x,SR}^{DFA}) \\ &+ c_{MR}(E_{x,MR}^{HF} - E_{x,MR}^{DFA}) \\ &+ c_{LR}(E_{x,LR}^{HF} - E_{x,LR}^{DFA}) \end{aligned} \quad (25)$$

According to the values of c_{SR}, c_{MR}, c_{LR}, ω_{SR} and ω_{LR}, short-, middle- and long-range corrected RSH functionals can be defined. When $\omega = 0$ and $c_{SR} = c_{MR} = c_{LR}$, range separated hybrids reduce to linear global hybrids.

Mixtures of RSH and linear hybrid functional have also been considered, as for the ωB97-X. The list of DFT methods considered in the present work is summarized in Table 1, including HF.

Table 2: Sample statistics for the methods on the benchmark set for band gap: mean absolute error (MAE), root-mean-square error (RMSE), mean signed error (MSE) and root-mean-square deviation (RMSD). The minimal absolute values in each column are in bold type.

	Band Gap (eV)			
	MAE	RMSE	MSE	RMSD
HF	6.1	6.2	6.1	1.2
LDA	1.4	1.7	-1.4	1.0
PBE	1.4	1.8	-1.4	1.1
PBEsol	1.4	1.8	-1.4	1.1
B97	0.46	0.67	**-0.06**	0.66
B3LYP	0.59	0.75	-0.20	0.73
PBE0	0.55	0.63	0.28	0.56
PBEsol0	0.53	0.60	0.28	0.53
HSE06	0.45	0.72	-0.30	0.65
HSEsol	**0.42**	0.68	-0.32	0.60
HISS	0.48	**0.53**	0.32	**0.41**
RSHXLDA	4.5	4.5	4.5	0.53
ωB97	4.3	4.4	4.3	0.52
ωB97X	3.9	4.0	3.9	0.54
LC-ωPBE	4.4	4.4	4.4	**0.41**
LC-ωPBEsol	5.4	5.4	5.4	0.72
M06-L	0.89	1.2	-0.88	0.86
M06	0.52	0.65	0.19	0.62

Note that in contrast to the generally used B3LYP, we used the variant implemented in the CRYSTAL code, where the local functional is fitted to the accurate correlation energy of the uniform electron gas, i.e. VWN5[44], and not to the VWN3 random phase approximation.

The statistical calculation presented in this paper have been done in the R environment[76], either with core functions, additional packages as mentioned in the text, or specifically developed routines.

Table 3: Same as Table 2 for bulk modulus.

	Bulk Modulus (GPa)			
	MAE	RMSE	MSE	RMSD
HF	10.0	14.0	6.2	13.0
LDA	7.1	9.1	3.9	8.3
PBE	12.0	15.0	-12.0	9.9
PBEsol	6.2	9.2	-4.2	8.2
B97	8.0	11.0	-6.9	7.9
B3LYP	9.8	12.0	-9.3	8.3
PBE0	**5.1**	**7.4**	0.52	**7.4**
PBEsol0	7.5	10.0	5.8	8.3
HSE06	5.2	**7.4**	**-0.34**	**7.4**
HSEsol	6.8	9.6	4.8	8.4
HISS	9.1	13.0	7.6	11.0
RSHXLDA	12.0	14.0	11.0	9.5
ωB97	11.0	12.0	7.5	9.5
ωB97X	8.3	9.5	4.6	8.4
LC-ωPBE	14.0	17.0	13.0	11.0
LC-ωPBEsol	24.0	28.0	23.0	17.0
M06-L	7.2	10.0	-4.4	9.1
M06	7.1	9.5	-0.95	9.5

Table 4: Same as Table 2 for lattice constants.

	Lattice Constant (Å)			
	MAE	RMSE	MSE	RMSD
HF	0.100	0.130	0.100	0.078
LDA	0.035	0.044	-0.035	0.026
PBE	0.089	0.096	0.089	0.038
PBEsol	0.024	0.029	0.024	0.016
B97	0.088	0.097	0.088	0.041
B3LYP	0.100	0.110	0.100	0.052
PBE0	0.040	0.047	0.039	0.027
PBEsol0	0.013	0.019	-0.006	0.018
HSE06	0.044	0.051	0.043	0.028
HSEsol	**0.012**	**0.015**	**-0.001**	**0.015**
HISS	0.020	0.026	0.011	0.024
RSHXLDA	0.032	0.040	-0.013	0.038
ωB97	0.029	0.037	0.026	0.027
ωB97X	0.041	0.051	0.041	0.032
LC-ωPBE	0.035	0.041	-0.019	0.037
LC-ωPBEsol	0.062	0.072	-0.062	0.037
M06-L	0.067	0.092	0.066	0.064
M06	0.058	0.074	0.057	0.047

4.2 Benchmark statistics

The estimations of MAE, RMSE, MSE and RMSD for all methods and properties are reported in Tables 2-4. The comparison of MSE and RMSE values tells us that most methods present significant systematic errors ($|MSE| \simeq RMSE$). Note that because of the presence of trends in the systematic errors, a small absolute value of the MSE is not an indicator of the absence of systematic errors.

Results in Tables 2-4 agree with previous benchmarks studies[24,57,67,68,75,77]. Hybrid methods are by far superior to semilocal functionals in the prediction of the band gap of solids as expected because of the inclusion of some HF exchange within the generalized Kohn-Sham formalism. In this respect, results for global, short and middle range-separated hybrids are not far from each other. Surprisingly, long-range corrected hybrids tend to systematically overestimate band gaps although HF exchange is included at long-range to recover the correct decay of the exchange potential. For lattice parameters and bulk moduli, GGA and related hybrid functionals for solids (e.g. PBEsol family) give improved results with respect to common functionals devised for molecules (e.g. PBE family). Interestingly, we confirm results by Lucero *et al.*[75] for the HISS functional which gives overall good results when compared to other hybrids. Highly parametrized mGGA functionals such as M06-L and M06 do not significantly improve results with respect to other examined functionals. As expected, inclusion of HF exchange in the M06 hybrid functional leads to a better prediction of band gaps than the semilocal counterpart.

Overall, computed MAE and MSE for LDA,

PBE, PBEsol, M06-L, HSE06, HSEsol and HISS on similar set of solids agree with the ones reported in other works[24,57,67,68,75].

DFAs ranking. One could attempt a ranking based on MSE and RMSD statistics. At this level, the best performing methods are characterized by two criteria: (1) the errors are nearly centered on zero; and (2) they have a small dispersion. With such criteria, there is no important distinction for band gaps between HSEsol and other hybrids, like PBE0, PBEsol0, and even HSE06, ωB97, M06, and B3LYP. The situation is less clear for bulk moduli, with a slight advantage to PBE0 and HSE06, and, for lattice parameters, PBEsol0 and HSEsol are the best contenders.

Whatever the performance statistics, none of the methods seems optimal for all the properties.

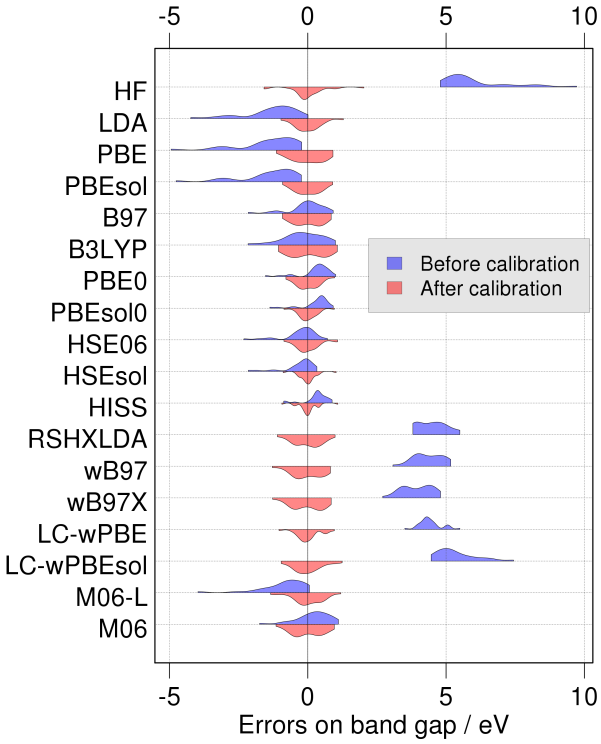


Figure 4: Probability density of the $E_{p,m}$ error sets for band gaps: for each method, the upper density represents the errors for the raw calculation data and the lower density the residual errors for the calibrated data.

4.3 Statistical modeling

Figures 4-6 show the probability densities of the E_m error sets for the three properties, before and after linear calibration (Eq. 18).

The errors distributions for the raw data (before calibration) confirm or reveal a few features relevant for the following developments:

- most methods provide biased estimates for some or all properties;
- the shape of the distributions varies considerably between methods and properties, some distributions are strongly asymmetric while others are bimodal; and
- some points seem to lie far of the main batch (outliers) and many distributions present a long tail.

4.3.1 Calibration

In order to determine the polynomial degree of the trend in systematic errors, we used Bayesian Model Selection (BMS)[78] for all error sets. BMS calculates the *posterior probability distribution* over a set of models, combining a parsimony criterion (Occam's razor) with a goodness-of-fit criterion. It avoids to overfit the data with overly complex models. We used the algorithm described in Mana *et al.*[79], over polynomial degrees from 1 to 3. BMS shows that the linear model is the most probable, except for RSHXLDA, ωB97 and ωB97-X in the prediction of band gaps, where second order polynomials have slightly higher posterior probability. As the linear model is not fully rejected for these cases, we considered a linear correction for all cases.

As can be seen on Figs. 4-6, linear correction, besides eliminating prediction bias, contributes often to produce more symmetrical distributions (*e.g.,* HF for band gaps), albeit without always resulting in normal distributions (*e.g.,* ωB97 for band gaps). In many cases, the dispersion of the errors is notably reduced (*e.g.,* B3LYP for lattice constants), along with the distribution tails (*e.g.,* PBE for band gaps). For some methods, one observes a mere shift of the distribution due to bias correction, as

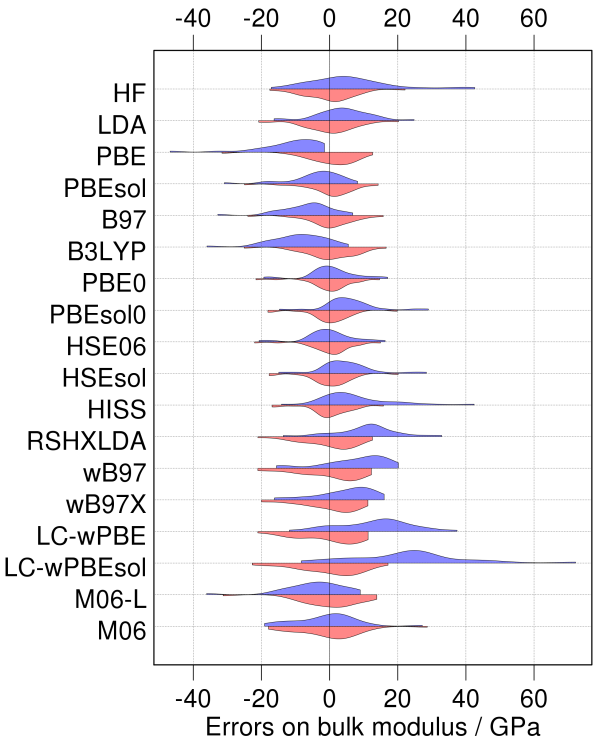


Figure 5: Same as Fig 4 for bulk modulus.

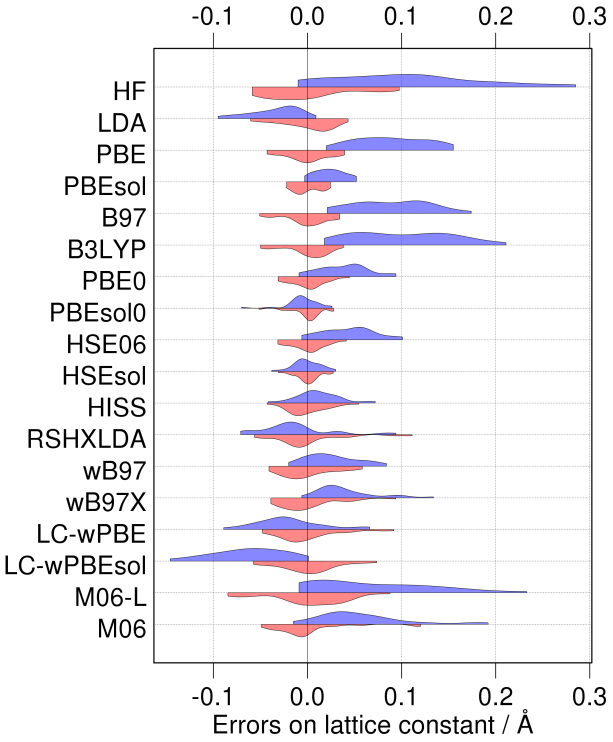


Figure 6: Same as Fig 4 for lattice constant.

for ωB97 for band gaps; this corresponds to methods for which the points were originally distributed along a line nearly parallel to the identity line.

Despite the linear correction, there might remain a variation of the dispersion of the errors as a function of the property value. In Figure 1(b), the corrected errors display no significant trend, but get larger in absolute value for systems with increasing lattice parameters. Capturing this behavior in the method inadequacy model might contribute to improve the quality of prediction uncertainty. This can however be considered as a second order effect and its correction has not been attempted in the present study, where, considering the small size of our calibration samples, we aimed at testing the simplest correction setup.

4.3.2 Internal validation of the calibration model

In order to validate the linear correction, we calculated the q^2 statistics through the Leave-One-Out cross-validation method[80]

$$q^2 = 1 - \frac{\sum_{s=1}^{N_s} \left(o_s - f_m(c_{m,s}; \hat{\vartheta}_m^{(-s)}) \right)^2}{\sum_{s=1}^{N_s} (o_s - \overline{o})^2}, \quad (26)$$

where one performs N_s linear regressions for sets of data without one of the points (regression parameters noted $\hat{\vartheta}_m^{(-s)}$), and compares the the sum-of-squares of the prediction errors of these regressions for the left-out data (numerator) to the sum-of-squares of the deviations of the sample points from their mean (denominator).

The q^2 statistics goes from 0 to 1, the larger the better. The q^2 statistics for all methods/properties pairs are reported in Fig. 7. Values of q^2 above 0.95 can be considered as excellent. There is therefore no evidence at this stage against the choice of a linear correction function.

4.4 Prediction uncertainty analysis

We calculated the contribution of the calibration model uncertainty u_f to the total variance

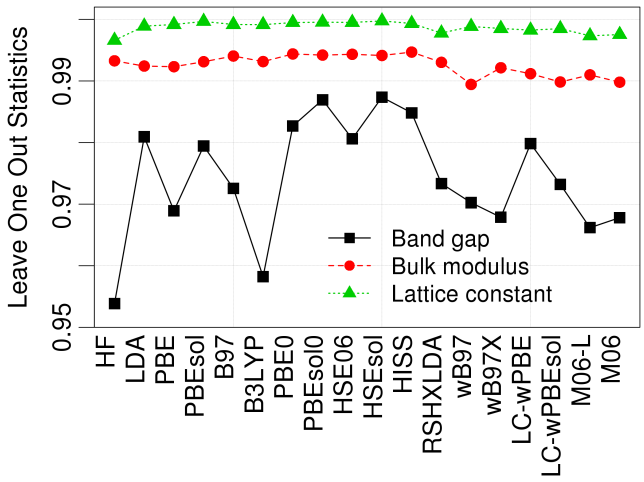


Figure 7: Validation of linear calibration by Leave One Out (q^2) statistics.

in Eq. 20. This uncertainty is typically larger at the extremes of the calibration range, and minimal around the mean value of the calibration data (Fig. 3). Its relative contribution for any calculated value x of a property is (ignoring the method index)

$$r(x) = \frac{u_f^2(x)}{u_p^2(x)}. \tag{27}$$

The maximal and minimal values of $r(x)$ over the calibration range have been found to be very weakly dependent of the DFA for a given property, but strongly dependent on the property. The DFA-averaged maximal value of $r(x)$ is about 25 % for band gaps, 30 % for bulk moduli and 15 % for lattice constants, and its minimal value is about 5 % for all properties.

The approximation of the prediction uncertainty by method inadequacy uncertainty d_m alone (Eq. 23) is therefore too optimistic, notably for band gaps and bulk moduli, in the sense that it underestimates uncertainty, notably at the extremities of the calibration range. In terms of uncertainty, using d_m represents a maximal underestimation of prediction uncertainty of about 8 % for lattice constants, 15 % for band gaps and 20 % for bulk moduli. These values are to be compared with the the relative uncertainty on the standard deviation of a normal-distributed sample of size N_s, $\Delta u/u \simeq 1/\sqrt{2(N_s-1)}$, which for $N_s = 28$ is about 15 %. Therefore, except maybe for lattice constants, one cannot consider these differences as negligible. One has also to keep in mind that they are obtained for values of the reference data uncertainty which are plausibly underestimated. Increasing the reference data uncertainties can only reduce the the method discrepancy uncertainty d_m (Eq. 33), and increase the relative contribution of the correction model uncertainty, u_f, to the prediction uncertainty, u_p.

The non-negligible contribution of the correction model to the total uncertainty is mainly due to the small size of the benchmark sets (28 values). For instance,the calibration uncertainty contribution has been shown by Pernot and Cailliez[8] to be negligible for calibration sets with 500 and 2500 harmonic vibrational frequencies, and around 12 % for sets of 39 zero point vibrational energies.

A thorough way to improve the constant uncertainty approximation (Eq. 23) is therefore to complement the benchmark set by new reference data, which is not always possible at short term. Another solution is to provide users with all the parameters required to use Eq. 20 ($u(a)$, $u(b)$ and $u(a,b)$; see Supporting Material). Although accurate, this solution does not enable a quick assessment of a set of DFAs. In order to provide a reliable uncertainty estimate while maintaining simplicity, we use the mean value of the prediction uncertainty

$$\overline{u_p} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} u_p^2(x_i)} \tag{28}$$

calculated on a regular grid of values of the property x covering its calibration range.

We reported in Tables 5-7 the linear correction factors to apply to the different DFAs of this study and their approximate prediction uncertainties d_m (Eq. 23) and $\overline{u_p}$ (Eq. 28; $n = 1000$). Uncertainties are conventionally reported with two significant digits[3], but considering the small sample size and the approximations involved, one should not attach too much credit to the last digit.

Comparison with Tables 2-4 shows that cor-

rection of the trend in systematic errors enables a strong reduction of the higher values of the standard deviation of residual errors: the methods have prediction uncertainties between 0.4 and 0.8 eV for band gaps ($RMSD$ between 0.4 and 1.2 eV), 7.2 and 10 GPa for bulk moduli ($RMSD$ between 7.4 and 17 GPa), 0.013 and 0.049 Å for lattice constants ($RMSD$ between 0.015 and 0.078 Å).

DFAs ranking. The best preforming DFAs do not see their dispersion notably improved, but Tables 5-7 offer a new landscape for method selection: it can be used to select methods according to uncertainty requirements. For instance, assuming that one wants to be able to estimate the lattice constant for a new compound with an uncertainty smaller than 0.02 Å, the last column of Table 7 tells us that one can choose amongst five (corrected) methods: PBEsol, PBE0, PBEsol0, HSE06, and HSEsol. If our requirement is an uncertainty below 0.05 Å, basically all methods should be able to comply.

Prediction uncertainty estimation. Tables 5-7 can also be used to estimate the uncertainty of a calculated value and to elaborate an uncertainty budget, for instance in the comparison of the properties of various compounds for screening studies. As an example of uncertainty evaluation, let us assume that we calculated a band gap of 10 eV for a new compound using the PBE method. The prediction of the 'true' value for this property is calculated from Table 5 as 0.502 + 1.385*10 = 14.35 eV (Eq. 20), with an uncertainty of 0.60 eV (the value of the uncertainty using the exact expression Eq. 20 is only slightly larger: 0.67 eV; the full sets of coefficients for Eq. 20 are provided as Supplementary Material).

Table 5: Linear correction factors a, b and approximate prediction uncertainties d and $\overline{u_p}$ for all methods on band gaps.

	Band Gap (eV)			
	a	$b^\dagger$	d	$\overline{u_p}$
HF	-3.773	0.770	0.70	0.74
LDA	0.500	1.347	0.47	0.49
PBE	0.502	1.385	0.57	0.60
PBEsol	0.503	1.385	0.46	0.48
B97	-0.476	1.142	0.53	0.55
B3LYP	-0.196	1.108	0.67	0.70
PBE0	-0.819	1.132	0.40	0.42
PBEsol0	-0.835	1.134	0.35	0.37
HSE06	-0.293	1.167	0.44	0.46
HSEsol	-0.268	1.167	0.34	0.36
HISS	-0.594	1.065	0.37	0.39
RSHXLDA	-4.052	0.948	0.52	0.54
ωB97	-4.328	0.998	0.54	0.57
ωB97-X	-3.909	0.996	0.56	0.59
LC-ωPBE	-4.312	0.987	0.43	0.45
LC-ωPBEsol	-4.175	0.872	0.52	0.55
M06-L	0.166	1.240	0.56	0.59
M06	-0.568	1.093	0.56	0.59

†the slope parameter b is dimensionless

4.4.1 External validation of prediction model

In order to validate the prediction uncertainty model derived in the previous section, we use a validation set of 9 systems not included in the calibration set. For this systems, we have data only for band gaps and lattice constants. No external validation is done on bulk modulus.

The principle of the validation is, for each DFA and property to:

1. correct the calculated values by the appropriate linear factors (Tables 5-7 and Eq. 19);
2. calculate the residual errors with the validation reference values (Eq. 11); and
3. calculate the number of errors falling within a prediction confidence interval extended to account for reference data uncertainty (Eq. 41). Here, we count the errors within a 2-σ confidence interval

Table 6: Same as Table 5 for bulk modulus.

	Bulk Modulus (GPa)			
	a	$b^\dagger$	d	$\overline{u_p}$
HF	5.300	0.907	7.8	8.4
LDA	-2.880	0.992	8.3	8.9
PBE	5.745	1.057	8.4	9.0
PBEsol	1.259	1.026	7.8	8.4
B97	4.004	1.027	7.5	8.1
B3LYP	6.673	1.024	8.1	8.6
PBE0	2.248	0.977	7.0	7.5
PBEsol0	-0.830	0.960	7.1	7.6
HSE06	2.902	0.978	7.0	7.6
HSEsol	0.188	0.959	7.1	7.7
HISS	2.368	0.920	6.7	7.2
RSHXLDA	-4.091	0.948	7.7	8.3
ωB97	-10.670	1.025	9.3	10.0
ωB97-X	-5.779	1.010	8.4	9.0
LC-ωPBE	-5.248	0.940	8.8	9.5
LC-ωPBEsol	-5.818	0.878	9.4	10.0
M06-L	3.656	1.007	9.2	9.9
M06	0.291	1.006	9.6	10.0

†the slope parameter b is dimensionless

Table 7: Same as Table 5 for lattice constants.

	Lattice Constant (Å)			
	a	$b^\dagger$	d	$\overline{u_p}$
HF	0.266	0.930	0.047	0.049
LDA	0.031	1.001	0.027	0.028
PBE	0.092	0.966	0.024	0.024
PBEsol	0.017	0.992	0.015	0.015
B97	0.115	0.962	0.024	0.024
B3LYP	0.178	0.948	0.024	0.025
PBE0	0.082	0.977	0.019	0.019
PBEsol0	0.054	0.991	0.017	0.018
HSE06	0.091	0.974	0.018	0.019
HSEsol	0.046	0.991	0.013	0.013
HISS	0.063	0.986	0.021	0.022
RSHXLDA	-0.033	1.009	0.039	0.040
ωB97	-0.045	1.004	0.027	0.028
ωB97-X	0.014	0.990	0.031	0.033
LC-ωPBE	-0.088	1.020	0.034	0.035
LC-ωPBEsol	-0.069	1.025	0.032	0.033
M06-L	0.225	0.945	0.042	0.043
M06	0.102	0.970	0.041	0.043

†the slope parameter b is dimensionless

(about 95 % coverage for a normal distribution).

Two difficulties are to be expected: (1) with small validation samples, the statistics might be far from their asymptotic values (*i.e.* one should not expect that 95 % of the points fall within a 95 % confidence interval), and (2) the definition of confidence intervals requires to know the shape of the errors distribution.

Considering the latter point, the part of the prediction uncertainty due to the correction function (u_f) can be assumed to have a normal distribution: the optimal regression coefficients being the combination of many uncertain contributions (o_s) with finite variance (Eqn. 29-30), the Central Limit Theorem ensures that they are normally distributed. This is not the case for the method inadequacy term d_m, representing essentially the residual errors distributions, which are often non normal (Figs. 4-6). It is therefore to be expected that confidence intervals built on a normality hypothesis will not be fully consistent with the validation data.

The results of the validation test are reported in Table 8. For band gaps, 9 points fall within the predicted interval in 12 cases, 8 points in 4 cases, and 7 in 2 cases. We can consider that the predicted uncertainties are satisfying, even if they are probably slightly overestimated. This is on the safe side: users have a very small risk to overestimate the accuracy of their calculations.

For lattice constants, there are 7 cases where only 6 points or less are correctly predicted. For two DFAs (RSHXLDA and ωB97), the problem persist if one uses a 3-σ confidence interval.

Table 8: Number of points of the validation set within the predicted 2-σ error range after linear correction. The validation set contains 9 points.

	Band Gap	Lattice Constant
HF	9	7
LDA	8	8
PBE	9	7
PBEsol	8	6
B97	9	6
B3LYP	9	7
PBE0	9	7
PBEsol0	8	8
HSE06	9	7
HSEsol	7	7
HISS	7	7
RSHXLDA	9	5
ωB97	9	4
ωB97-X	9	5
LC-ωPBE	9	6
LC-ωPBEsol	9	5
M06-L	8	9
M06	9	8

Referring to the results on Bayesian Model Selection (Section 4.3.1), this could suggest that the linear correction model is insufficient. However, investigation of the errors distributions for these cases shows that there is a small overlap between the calibration and validation error sets. The linear correction does not contribute to shift those points towards the center of the distribution (Fig. 8). For these DFAs, it seems that the calibration set is not fully representative of the species in the validation set.

Globally, we checked that, despite the caveats of small sample size and non-normal distribution, the prediction models provides reasonable confidence intervals, except for a few DFAs (RSHXLDA and ωB97),for which the lattice constants calibration set is poorly representative of the data in the validation set. Ever after linear correction, theses DFAs should not be recommended to predict lattice constants.

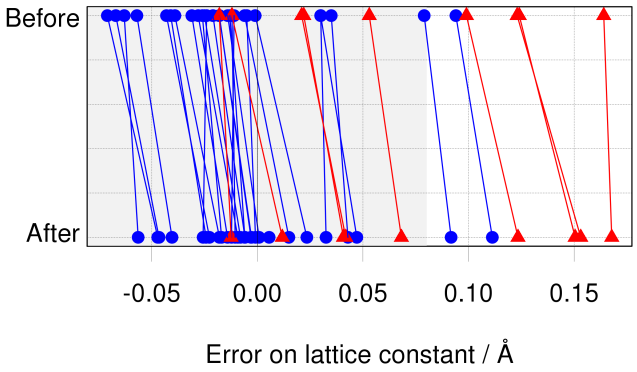


Figure 8: Errors redistribution by linear correction, for calibration (blue dots) and validation (red triangles) sets of lattice constants estimated by the RSHXLDA method. The gray area represents the 2-σ interval used for validation.

4.5 Sensitivity to reference data uncertainty

Up to this point, all evaluations have been done with reference data uncertainty values u_s which are plausibly underestimated (Section 4.1.1). In order to assess the impact of u_s on the prediction uncertainty u_p, we reevaluate $\overline{u_p}$ (Eq. 28) with values of u_s more akin to account for various perturbations such as temperature effects, corrections uncertainty...

We consider indeed a worst case scenario, *i.e.* the largest values of u_s that do not compromise the least-squares regression validity. As overestimated values of u_s would produce unlikely small values of χ^2 (Eq. 12), we request χ^2 to be above the 5 % quantile of the standard chi-squared distribution with $N_s - 2$ degrees of freedom ($\chi^2_{min} \simeq 15.4$). The corresponding values are $u_s = 0.3\,\text{eV}$ for band gaps, 7 GPa for bulk moduli, and 0.015 Å for lattice constants.[81]

The new values of $\overline{u_p}$ are shown in Table 9 alongside those issued from Tables 5-7. For all properties, the effect is more visible for methods which had a prediction uncertainty close or below the worst case value of u_s. One reached a point where some methods (most of them for bulk modulus, indicating that the worst case value of $u_s = 7\,\text{GPa}$ might be too large) have their prediction uncertainty smaller than refer-

ence data uncertainty. In this scenario, such methods, after *a posteriori* correction, could be selected to replace advantageously costly and/or difficult measurements, with the same level of confidence.

Globally, the prediction uncertainty presents a low sensitivity to the reference data uncertainty: for band gaps and lattice constants, an increase of more than one order of magnitude in u_s results at most in a reduction by a factor 2 (band gaps) or 4 (lattice constants) of $\overline{u_p}$. These factors are however still much larger than the 15 % of relative uncertainty on a standard deviation one can expect for samples of this size (see Section 4.4).

This analysis shows that for a reliable estimation of method prediction uncertainty one needs an adequate evaluation of reference data uncertainty, which is probably the most sensitive issue in the implementation of the VM framework.

4.6 Looking back at the reference data

Figures 9-11 show the distributions of errors *per system*, $E_{p,s} = \{E_{p,m,s}; m \in DFAs\}$, before and after the linear correction applied in the previous section.

At the system level, the reduction of dispersion is remarquable, confirming that calibrated DFAs produce more consistent results. After correction, some systems present a significant bias, *i.e.* their error distribution does not overlap the zero axis. Outstanding examples are the bulk moduli of GaN, MgS, and $SrTiO_3$, or the band gap for LiF, NaF and NaCl. These systems are contributing to the "outliers" in the errors distributions per method in Figs. 4-6, and the fact that all methods are unable to predict these systems properties deserves further attention.

For the bulk modulus, the reason is probably the experimental temperature. In fact, data for GaN, MgS, and $SrTiO_3$ correspond to room temperature values. For GaN, there is also a problem with the zinc-blende phase (B3), because there are some discrepancies among available experimental data. Moreover, the bulk

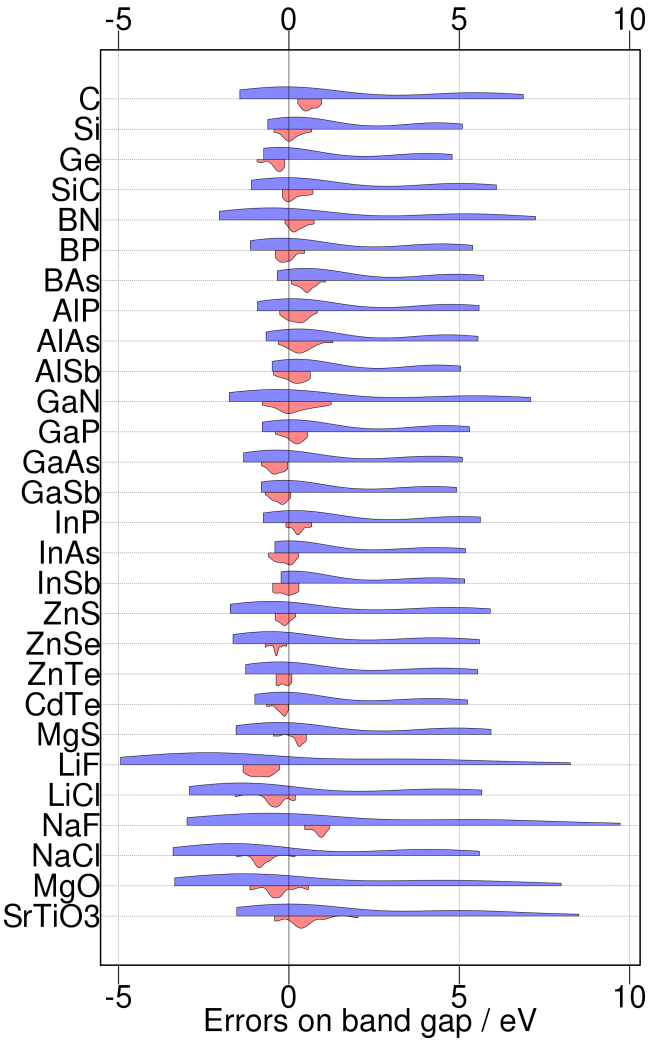


Figure 9: Effect of linear calibration on errors distribution per system for band gap. Above/blue: before; below/red: after.

modulus has been derived from experimental data for the wurzite phase (B4). These data should either be corrected for temperature effects, and their uncertainty increased accordingly, or rejected from the calibration set[23].

Concerning the band gaps, alkali halides are the systems with the largest values in the dataset (from 9 eV to 14 eV) and all tested methods systematically fail in predicting the band gap of wide band gap insulators. Here again, one has to consider if these data should be rejected.

One might be tempted to treat this problem by making subsets of the reference data (such as done for vibrational frequencies[30], or intermolecular potentials[12]), using a different calibration models for each subset. This is not a

Table 9: Effect of reference data uncertainty (u_s) on prediction uncertainty $\overline{u_p}$. The values in boldface are the one where $\overline{u_p} \leq u_s$.

	Band Gap (eV)		**Bulk Modulus** (GPa)		**Lattice Constant** (Å)	
u_s	0.01	0.30	2	7	0.001	0.015
HF	0.74	0.67	8.4	**5.0**	0.049	0.047
LDA	0.49	0.39	8.9	**5.9**	0.028	0.024
PBE	0.60	0.52	9.0	**6.0**	0.024	0.019
PBEsol	0.48	0.38	8.4	**5.1**	0.015	**0.004**
B97	0.55	0.46	8.1	**4.5**	0.024	0.019
B3LYP	0.70	0.63	8.6	**5.5**	0.025	0.020
PBE0	0.42	**0.30**	7.5	**3.4**	0.019	**0.012**
PBEsol0	0.37	**0.21**	7.6	**3.6**	0.018	**0.009**
HSE06	0.46	0.35	7.6	**3.5**	0.019	**0.011**
HSEsol	0.36	**0.19**	7.7	**3.7**	0.013	**0.004**
HISS	0.39	**0.25**	7.2	**2.7**	0.022	0.016
RSHXLDA	0.54	0.45	8.3	**4.9**	0.040	0.037
ωB97	0.57	0.48	10.0	7.5	0.028	0.024
ωB97X	0.59	0.51	9.0	**6.0**	0.033	0.029
LC-ωPBE	0.45	0.33	9.5	**6.7**	0.035	0.032
LC-ωPBEsol	0.55	0.46	10.0	7.6	0.033	0.029
M06-L	0.59	0.50	9.9	7.2	0.043	0.041
M06	0.59	0.51	10.0	7.9	0.043	0.040

viable solution in the present case for two reasons: (i) the resulting subsets would become too small to enable significant statistical analysis; and (ii) one would then calibrate the calculation methods to correct for different reference biases and ruin the prediction ability of the calibrated methods.

5 Discussion

We have derived a statistical model in the VM framework to estimate the uncertainty on a property value predicted by a DFA. Calculating an uncertainty required us to correct the DFA results for systematic errors. We have seen that for the solids properties studied here, a linear correction was generally sufficient. The residual errors of a corrected DFA are the basis for estimating its prediction uncertainty, which also includes a part due to the linear correction model. The calibration/prediction procedure uses standard statistical tools (WLS regression, uncertainty propagation by combination of variances) and is simple to implement.

The essential contribution of the present derivation is to introduce a *method inadequacy* error term in the calibration model to acknowledge the fact that a corrected DFA is typically unable to reproduce reference data within their uncertainty range. This additional error term, modeled by a stochastic variable, ensures the statistical consistency of the calibration and prediction stages. We has shown that it is generally the major contribution to prediction uncertainty.

We want to address here a few points regarding the assumptions and limits of this approach.

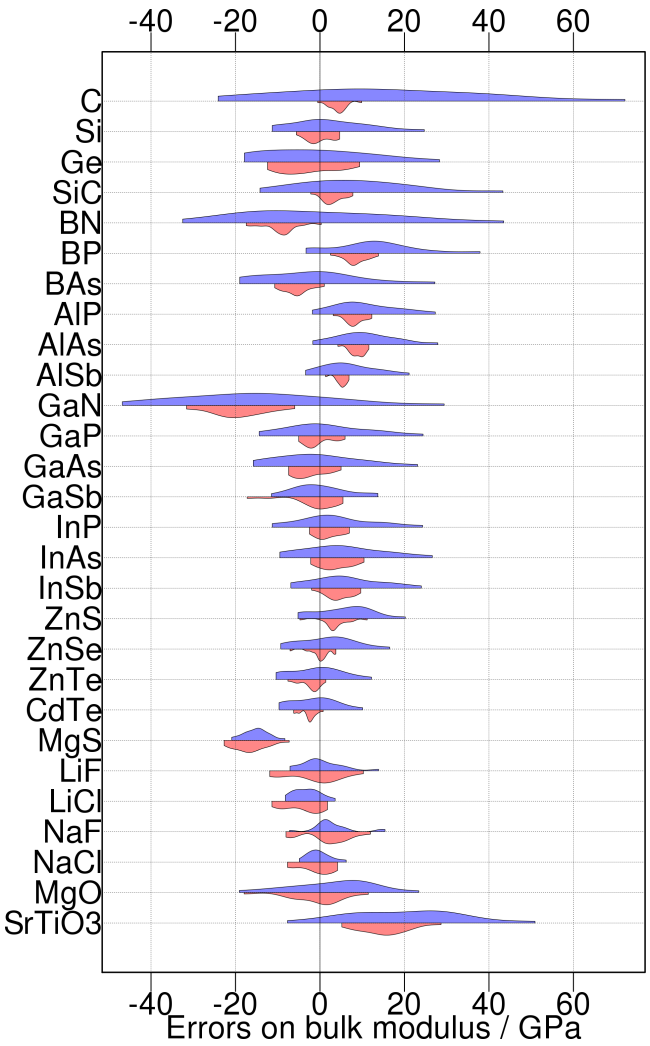


Figure 10: Effect of linear calibration on errors distribution per system for bulk modulus. Above/blue: before; below/red: after.

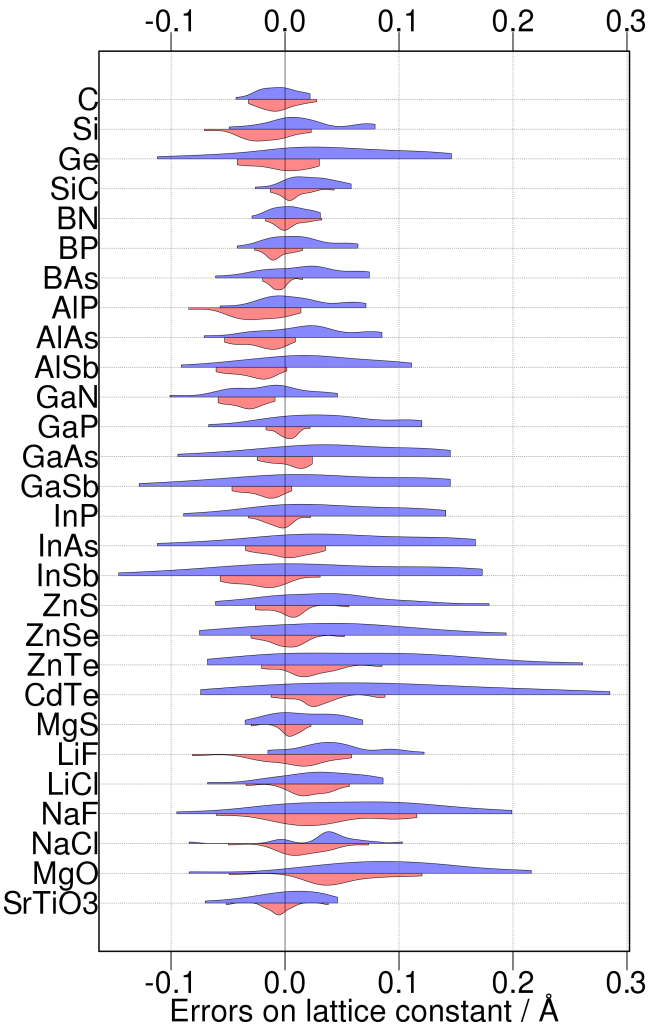


Figure 11: Effect of linear calibration on errors distribution per system for lattice constant. Above/blue: before; below/red: after.

5.1 Weigted least-squares regression

The WLS regression formulae rely on few assumptions: the errors have to arise from distributions of mean zero and finite variance and they have to be uncorrelated. There is therefore no constraint on the specific shape of the errors distributions. Nevertheless, two sensitive points of the method should be considered:

- *Dependence on the reference data uncertainty.* If the reference data uncertainty is not negligible before method inadequacy errors, it might play a significant role through the weights in the WLS procedure. The present study was based on the assumption of uniform uncertainty for each property. We have seen in this case that the prediction uncertainty has a non-negligible dependence on the reference data uncertainty value. A more detailed budget of reference data uncertainties has to be established, notably by prioritizing the outlier systems identified in Sec. 4.6.

- *Sensitivity to outliers.* Least-squares procedures are well-known to be sensitive to outliers, *i.e.* points with much larger *weighted* residual errors than the other points in the set. Outliers can be dealt with at different levels: they can be rejected from the reference set, maybe on the basis of an heterogeneity in experimental methods or physico-chemical

properties, or they can be given less importance by using robust regression methods. A preliminary study using a rank-based robust method (package Rfit[82]) revealed only non-significative differences with the least-squares results, but this has to be further explored.

5.2 Calibration model

The calibration model is based on two choices:

- *The correction function* f_m. It is striking that the observed trend in the errors of most DFAs is linear. In the present study, this might result from the small sample size, *i.e.* a lack of information for selecting more complex trends. Note however that the linear trend is also observed for very large sets of harmonic vibrational frequencies, albeit following the discrimination of low frequency from high frequency modes[30], and for large sets of elemental solids[23]. If necessary, more complex correction functions, such higher order polynomials could be considered while preserving the WLS regression method.

- *The method inadequacy stochastic model* δ_m. We have chosen to describe method inadequacy errors by a random variable with a uniform standard deviation across the calibration range. However, we noted from Fig. 1(b) that, for some DFAs, the dispersion of residual errors seems to increases with the calculated value of the property. We have considered this as a secondary order effect, but a property-value-dependent model could be directly inserted in the WLS procedure and its parameters optimized iteratively. One would then have to deal explicitly with property-value-dependent prediction errors, more complex to communicate to the end users. Here again, larger samples would be necessary to assess the necessity of this refinement.

5.3 From standard uncertainty to enlarged uncertainty

It is recommended by Ruscic[11], following the *de facto* standard in thermochemical data tabulations, that computed data should be provided with an enlarged uncertainty, u_{95}, enabling to define a 95 % confidence interval for the true value. A major difficulty we evidenced in the present study is that converting a standard uncertainty, such as u_p, to a confidence interval requires the knowledge of the errors distribution. For instance, for a normal distribution, one would have $u_{95} \simeq 2 * u_p$, whereas for a uniform distribution $u_{95} \simeq 1.65 * u_p$. Considering the varied and non-standard shapes of the errors distributions observed for the calibrated DFAs (Figs. 4-6), the estimation of u_{95} cannot be done as simply as for standard distributions. A numerical estimation of a confidence interval based on the 2.5 % and 97.5 % quantiles could be done, but one is facing again the problem of small sample size, even more sharply for the estimation of extreme quantiles than for the standard deviation (for a 28 points sample, there is in average less than one point in each of the 2.5 % external intervals).

We would like also to stress out that there is no reason why we should expect errors produced by (calibrated) DFAs to follow normal distributions. Model discrepancy generates property-dependent systematic errors with no predictable distribution. Normality is only to be expected as a limit case for methods with null inadequacy errors, when the errors are dominated by assumed normal reference data uncertainties. Moreover, the small sample size and the selection process of reference data might also play a role in the observed deviance from normality.

6 Conclusions

Benchmarks have their limitations, but are a condensate of numerical experience, and are thus useful. They can provide information, but it should be treated with care.

We tested in this study the applicability of

the Virtual Measurement framework to Density Functional Approximations for the estimation of various properties of solids. In this approach, each computational result has to be qualified by an uncertainty or confidence interval. Informed of the prediction uncertainty of various methods, users should be able to choose a suitable method, in terms of accuracy, availability and costs, which is not necessarily the "best" method highlighted by standard performance statistics. Users would also be able to assess the contribution of each calculation in an uncertainty budget, for instance in DFT-based multiscale simulations.

We have shown that the measures of performance commonly used in computational chemistry benchmarks do not provide directly the prediction uncertainty for a method, mostly because they do not disambiguate the predictable/systematic components of the errors from their unpredictable/random component. In fact, statistical analysis of the benchmark error sets reveals notable systematic components, presenting a regular trend as function of the property value, which needs to be corrected in order to get reliable uncertainty estimates. In the present study, a linear correction of the calculated values was found sufficient to reach this goal.

Pernot and Cailliez[9] have shown that for large benchmark sets the prediction uncertainty can be safely derived from the standard deviation of the errors of the scaled properties. We treated here reference sets with about 30 values, and we observed that this approximation underestimates prediction uncertainty. A corrected value of prediction uncertainty has been proposed and validated on an external set of reference data.

As usual performance statistics do not account for reference data uncertainty, their use requires high quality (meaning *negligible uncertainty*) reference data. In contrast, the statistical models of the virtual measurement approach can deal with reference data presenting uncertainties of the same order as model prediction uncertainty. They offer also a practical correction method for those cases where the calculated property does not exactly correspond to the experimental one, such as for band gap (Section 2.4). The *a posteriori* calibration models are therefore of very general applicability, *at the additional cost of a reliable estimation of reference data uncertainty*, which is not a minor issue.

Another difficulty identified in this study for the successful application of the VM approach is the estimation of reliable confidence intervals. We have shown that the estimation of a prediction uncertainty is rather straightforward, but the estimation of an enlarged uncertainty to define a 95 % confidence interval is made difficult by the small sample sizes and the arbitrary shape of the errors distributions.

A drawback of the *a posteriori* calibration approach is the lack of generalizability: it is not possible to estimate the prediction uncertainty of a DFA for a property against which the DFA has not been calibrated. On the other hand, there is no evidence in the conventional approach that a DFA with good performance statistics will perform as well for untested properties, hence the need of exhaustive benchmark tests[24].

On the positive side, we have shown that methods rejected on their MAE or RMSE performance because of large systematic errors can, after calibration, become competitive with the "best" benchmark performers in terms of prediction uncertainty. This considerably widens the choice of methods for the end users. Low-cost calibrated methods with well characterized prediction uncertainty could be promising for high-throughput studies.

It is too early to suggest that the correction parameters provided here for the band gaps, bulk moduli and lattice constants of crystals with cubic symmetry should be used confidently. The database still needs to grow and to be groomed, including a better assessment of reference data uncertainties. Nevertheless, the methodology to estimate the calibration parameters and prediction uncertainties can easily be applied to any other benchmark set, and we consider that it would be a very valuable complement to the usual performance statistics.

A Estimation of the calibration parameters

We provide below the expressions of the optimal parameters, their uncertainty and covariance for the problem of weighted least squares linear regression with a stochastic method inadequacy contribution (Eq. 18). The derivation of the basic formulae can be found in data analysis textbooks[36], although most of them do not provide the covariance formula, which is essential for uncertainty propagation[83,84]. The main difference of our derivation resides in the interpretation and management of the variance contributions, and the need for an iterative procedure (Iteratively Reweighted Least Squares).

A.1 Calibration

We first treat the general case of linear regression of reference data with known uncertainty, $(o_s \pm u_s,\ s = 1, N_s)$, and then we consider the particular case of negligible reference data uncertainty. We do not address the case of correlated reference data, because such correlation information is practically never available in reference data sets. We remove the method index m in the equations for concision.

Let us assume, in a first stage, that reference data uncertainty is the sole source of dispersion of the points around the regression line. The optimal parameters for Weighted Least Squares (WLS) linear regression have closed-form expressions

$$\hat{b} = \frac{\sum w \sum wco - \sum wc \sum wo}{\Delta} \quad (29)$$

$$\hat{a} = \frac{\sum wo}{\sum w} - \hat{b}\frac{\sum wc}{\sum w} \quad (30)$$

$$\Delta = \sum w \sum wc^2 - (\sum wc)^2, \quad (31)$$

where all sums run over $s = 1, N_s$ (*i.e.* $\sum wx \equiv \sum_{s=1,N_s} w_s x_s$), and the weights are defined as $w_s = 1/u_s^2$.

If this regression model is valid, one should have

$$\chi^2 = \sum w(o - \hat{a} - \hat{b}c)^2 \simeq N_s - 2. \quad (32)$$

If this is the case, the residuals variance is fully explained by the reference data uncertainty, and there is no need to consider method inadequacy: the calibrated method is able to predict reference data within their error bars.

For many approximated methods, this scenario is unlikely and would occur for reference sets of very uncertain data, improper to evaluate model prediction uncertainty. Similarly, if $\chi^2 \ll N_s - 2$, the reference data uncertainty have probably been overestimated and are also improper for our purpose.

The case which interests us here is when $\chi^2 \gg N_s - 2$, *i.e.* when the residuals variance is significantly larger than what is expected from reference data uncertainty. In the hypothesis where the weighted residuals are randomly distributed, one can estimate the variance due to method inadequacy as the difference between the residuals variance and the mean reference variance

$$d^2 = \frac{1}{N_s - 2}\sum_s (o_s - \hat{a} - \hat{b}c_s)^2 - \frac{1}{N_s}\sum_s u_s^2, \quad (33)$$

which cannot be negative if $\chi^2 \gg N_s - 2$. Knowing d, one is now able to specify the full calibration model (Eq. 18). We solve it by redefining the weights as

$$w_s = 1/(u_s^2 + d^2) \quad (34)$$

and inserting them in the formulae giving Δ, $\hat{a}$ and $\hat{b}$ (Eqn. 29-31).

For uniform reference data uncertainty ($u_s = const.$), this reweighting will not change the values of $\hat{a}$ and $\hat{b}$, and one can proceed directly to the evaluation of the variance-covariance of the parameters with the updated value for Δ. Otherwise, a few iterations of the reweighting procedure (Eqn. 29-31, 33, 34) will be necessary to reach convergence.

The chi-square test (Eq. 32) is verified by construction, and we can derive the parameters uncertainty and covariance by the standard WLS

formulae

$$u^2(b) = \frac{\sum w}{\Delta} \quad (35)$$

$$u^2(a) = \frac{\sum wc^2}{\Delta} \quad (36)$$

$$u^2(a,b) = -\frac{\sum wc}{\sum w} u^2(b) \quad (37)$$

To summarize, in the general case where we have nonuniform reference data uncertainties, one must apply an Iteratively Reweighted Least Squares procedure to determine (1) the excess variance d^2 attributed to method inadequacy errors; and (2) the optimal parameters of the calibration function, their uncertainty and covariance. If the reference data uncertainty are uniform, only one step of the reweighting procedure is necessary.

If the reference data uncertainty is negligible before the fit residuals, one recovers the ordinary least squares method[36], but where the full residuals variance is explained by method inadequacy, *i.e.*

$$d^2 = \frac{1}{N_s - 2} \sum_{\mathrm{s}} (o_s - \hat{a} - \hat{b}c_s)^2. \quad (38)$$

A.2 Prediction

For the estimation of a new value of a property knowing a calculated value c^* (*i.e.* for a system not in the benchmark set), the prediction model and prediction variance are

$$p(c^*) = \hat{a} + \hat{b}c^* + \hat{\delta} \quad (39)$$

$$u_p^2(c^*) = \hat{d}^2 + u^2(a) + c^{*2}u^2(b) + 2c^* u^2(a,b), \quad (40)$$

where $\hat{\delta} \equiv 0$ has been left in the prediction equation as a reminder of the occurrence of d^2 in the prediction variance. The expression of u_p^2 is obtained by combination of variances applied to p[3,38]. The uncertainty on d^2 has been shown to be of secondary importance[8], and has not be considered here. u_p^2 accounts for uncertainties linked to model calibration and method inadequacy errors.

Note that for the comparison of a model prediction with reference data (as in cross- or external validation procedures), or the prediction of an experimental result, this variance has to be further combined with the corresponding reference/experimental data uncertainty

$$u^2 = u_p^2 + u_s^2. \quad (41)$$

Acknowledgments

The authors thank Roberto Dovesi for useful discussions.

Supporting Information

The data sets used in this article and the full sets of parameters for evaluating the prediction uncertainties using Eq. 20. This material is available free of charge via the Internet at `http://pubs.acs.org`.

References

(1) Civalleri, B.; Presti, D.; Dovesi, R.; Savin, A. *Chemical Modelling: Applications and Theory Volume 9*; Royal Soc. Chem., 2012; Vol. 9; pp 168–185.

(2) Irikura, K. K.; Johnson, R. D.; N., K. R. Uncertainty Associated with Virtual Measurements from Computational Quantum Chemistry Models. *Metrologia* **2004**, *41*, 369–375.

(3) BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML, *Evaluation Of Measurement Data - Guide to the Expression of Uncertainty in Measurement (GUM)*; 2008.

(4) Irikura, K. K.; Johnson, R. D.; N., K. R. Uncertainties in Scaling Factors for ab Initio Vibrational Frequencies. *J. Phys. Chem. A* **2005**, *109*, 8430–8437.

(5) Irikura, K. K.; Johnson, R. D.; N., K. R.; Kessel, R. Uncertainties in Scaling Factors for ab Initio Vibrational Zero-Point Energies. *J. Chem. Phys.* **2009**, *130*, 114102.

(6) Johnson, R. D.; Irikura, K. K.; Kacker, R. N.; Kessel, R. Scaling Factors and Uncertainties for ab Initio Anharmonic Vibrational Frequencies. *J. Chem. Theory Comput.* **2010**, *6*, 2822–2828.

(7) Teixeira, F.; Melo, A.; Cordeiro, M. N. D. S. Calibration Sets and the Accuracy of Vibrational Scaling Factors: A Case Study with the X3LYP Hybrid Functional. *J. Chem. Phys.* **2010**, *133*, 114109.

(8) Pernot, P.; Cailliez, F. Semi-Empirical Correction of ab Initio Harmonic Properties by Scaling Factors: a Validated Uncertainty Model for Calibration and Prediction. *ArXiv e-prints* **2010**, *1010.5669*.

(9) Pernot, P.; Cailliez, F. Comment on 'Uncertainties in scaling factors for ab initio vibrational zero-point energies' [J. Chem. Phys. 130, 114102 (2009)] and 'Calibration Sets and the Accuracy of Vibrational Scaling Factors: A case study with the X3LYP Hybrid Functional' [J. Chem. Phys. 133, 114109 (2010)]. *J. Chem. Phys.* **2011**, *134*, 167101.

(10) Jacobsen, R. L.; Johnson, R. D.; Irikura, K. K.; Kacker, R. N. Anharmonic Vibrational Frequency Calculations Are Not Worthwhile for Small Basis Sets. *J. Chem. Theory Comput.* **2013**, *9*, 951–954.

(11) Ruscic, B. Uncertainty Quantification in Thermochemistry and Benchmarking Electronic Structure Computations and Active Thermochemical Tables. *Int. J. Quantum Chem.* **2014**, *114*, 1097–1101.

(12) Faver, J. C.; Benson, M. L.; He, X.; Roberts, B. P.; Wang, B.; Marshall, M. S.; Kennedy, M. R.; Sherrill, C. D.; Merz Jr, K. M. Formal Estimation of Errors in Computed Absolute Interaction Energies of Protein-ligand Complexes. *J. Chem. Theory Comput.* **2011**, *7*, 790–797.

(13) Faver, J. C.; Yang, W.; Merz Jr, K. M. The Effects of Computational Modeling Errors on the Estimation of Statistical Mechanical Variables. *J. Chem. Theory Comput.* **2012**, *8*, 3769–3776.

(14) Ucisik, M. N.; Zheng, Z.; Faver, J. C.; Merz, K. M. Bringing Clarity to the Prediction of Protein-Ligand Binding Free Energies via "Blurring". *J. Chem. Theory Comput.* **2014**, *10*, 1314–1325.

(15) Frederiksen, S. L.; Jacobsen, K. W.; Brown, K. S.; Sethna, J. P. Bayesian Ensemble Approach to Error Estimation of Interatomic Potentials. *Phys. Rev. Lett.* **2004**, *93*, 165501.

(16) Mortensen, J. J.; Kaasberg, K.; Frederiksen, S. L.; Norksov, J. K.; Sethna, J. P.; Jacobsen, K. W. Bayesian Error Estimation in Density Functional Theory. *Phys. Rev. Lett.* **2005**, *95*, 216401.

(17) Petzold, V.; Bligaard, T.; Jacobsen, K. W. Construction of New Electronic Density Functionals with Error Estimation Through Fitting. *Topics in Catalysis* **2012**, *55*, 402–417.

(18) Wellendorff, J.; Lundgaard, K. T.; Møgelhøj, A.; Petzold, V.; Landis, D. D.; Nørskov, J. K.; Bligaard, T.; Jacobsen, K. W. Density Functionals for Surface Science: Exchange-Correlation Model Development with Bayesian Error Estimation. *Phys. Rev. B* **2012**, *85*, 235149.

(19) Medford, A. J.; Wellendorff, J.; Vojvodic, A.; Studt, F.; Abild-Pedersen, F.; Jacobsen, K. W.; Bligaard, T.; Norskov, J. K. Assessing the Reliability of Calculated Catalytic Ammonia Synthesis Rates. *Science* **2014**, *345*, 197–200.

(20) Cailliez, F.; Pernot, P. Statistical Approaches to Forcefield Calibration and Prediction Uncertainty in Molecular Simulation. *J. Chem. Phys.* **2011**, *134*, 054124.

(21) Rizzi, F.; Najm, H.; Debusschere, B.; Sargsyan, K.; Salloum, M.; Adalsteinsson, H.; Knio, O. Uncertainty Quantification in MD Simulations. Part II: Bayesian Inference of Force-Field Parameters. *Multiscale Mod. Sim.* **2012**, *10*, 1460–1492.

(22) Chernatynskiy, A.; Phillpot, S. R.; LeSar, R. Uncertainty Quantification in Multiscale Simulation of Materials: A Prospective. *Ann. Rev. Mat. Res.* **2013**, *43*, 157–182.

(23) Lejaeghere, K.; Speybroeck, V. V.; Van Oost, G.; Cottenier, S. Error Estimates for Solid-State Density-Functional Theory Predictions: An Overview by Means of the Ground-State Elemental Crystals. *Critical Reviews in Solid State and Materials Sciences* **2014**, *39*, 1–24.

(24) Peverati, R.; Truhlar, D. G. Quest for a Universal Density Functional: the Accuracy of Density Functionals Across a Broad Spectrum of Databases in Chemistry and Physics. *Phil. Trans. R. Soc. A* **2014**, *372*, 20120476.

(25) Pham-Gia, T.; Hung, T. L. The Mean and Median Absolute Deviations. *Math. Comput. Mod.* **2001**, *34*, 921–936.

(26) In the core functions of major mathematical/statistical softwares (Matlab, R...) the `mad()` function refers to the $MeanAD$ and/or $MedianAD$ definitions, not to MAE. In Mathematica, the ambiguity is somewhat relieved (at the expense of the 'absolute' term): `MeanDeviation[]` calculates the $MeanAD$, and `MedianDeviation[]` the $MedianAD$. In Excel, the `AVEDEV()` function calculates the *average absolute deviation* (AAD) according to the $MeanAD$ definition.

(27) BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML, *International Vocabulary of Metrology : Basic and General Concepts and Associated Terms*; 2012.

(28) Willmott, C. J.; Matsuura, K. Advantages of the Mean Absolute Error (MAE) Over the Root Mean Square Error (RMSE) in Assessing Average Model Performance. *Climate Res.* **2005**, *30*, 79–82.

(29) Chai, T.; Draxler, R. R. Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)? - Arguments Against Avoiding RMSE in the Literature. *Geosci. Model Dev.* **2014**, *7*, 1247–1250.

(30) Scott, A. P.; Radom, L. Harmonic Vibrational Frequencies: An Evaluation of Hartree-Fock, Moller-Plesset and Quadratic Configuration Interaction and Density Functional Theory and and Semiempirical Scale Factors. *J. Phys. Chem.* **1996**, *100*, 16502–16513.

(31) Karton, A.; Daon, S.; Martin, J. M. L. W4-11: A High-Confidence Benchmark Dataset for Computational Thermochemistry Derived From First-Principles {W4} Data. *Chem. Phys. Lett.* **2011**, *510*, 165–178.

(32) Mott, N. F. Metal-Insulator Transition. *Rev. Mod. Phys.* **1968**, *40*, 677–683.

(33) Perdew, J. P.; Levy, M. Physical Content of the Exact Kohn-Sham Orbital Energies: Band Gaps and Derivative Discontinuities. *Phys. Rev. Lett.* **1983**, *51*, 1884–1887.

(34) Sham, L. J.; Schluter, M. Density-Functional Theory of the Energy Gap. *Phys. Rev. Lett.* **1983**, *51*, 1888–1891.

(35) If this model is valid, there is no model inadequacy and no prediction uncertainty: one cannot reject the hypothesis that the calculation method provides the true value of the property. Note that this statement might have to be revised when more accurate reference data become available.

(36) Bevington, P. R.; Robinson, D. K. *Data Reduction and Error Analysis for the Physical Sciences*; McGraw-Hill: New York, 1992.

(37) The sum of the squares of N_s independent *normal* random variables of mean 0 and variance 1 has a chi-square distribution with $n_{df} = N_s$ degrees of freedom; the mean value of this distribution is equal to N_s. If, instead of normal distributions, we consider *non-normal* distributions of mean 0 and variance 1, there is no known distribution for the sum of squares, but the value of the mean is still N_s. When one treats the sum of squared residuals, the constraints imposed by the model reduce the number of degrees of freedom to $n_{df} = N_s - N_\vartheta$.

(38) Tellinghuisen, J. Statistical Error Propagation. *J. Phys. Chem. A* **2001**, *105*, 3917–3921.

(39) BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML, *Evaluation Of Measurement Data - Supplement 1 To The "Guide to the Expression of Uncertainty In Measurement" - Propagation of Distributions Using a Monte Carlo Method*; 2008.

(40) Dovesi, R.; Saunders, V. R.; Roetti, C.; Orlando, R.; Zicovich-Wilson, C. M.; Pascale, F.; Civalleri, B.; Doll, K.; Harrison, N. M.; Bush, I. J. et al. `CRYSTAL14` User's Manual. 2014; http://www.crystal.unito.it.

(41) Dovesi, R.; Orlando, R.; Erba, A.; Zicovich-Wilson, C. M.; Civalleri, B.; Casassa, S.; Maschio, L.; Ferrabone, M.; Pierre, M. D. L.; D'Arco, P. et al. CRYSTAL14: A Program for the Ab initio Investigation of Crystalline Solids. *Int. J. Quantum Chem.* **2014**, *114*, 1287–1317.

(42) Heyd, J.; Peralta, J. E.; Scuseria, G. E.; Martin, R. L. Energy Band Gaps and Lattice Parameters Evaluated with the Heyd-Scuseria-Ernzerhof Screened Hybrid Functional. *J. Chem. Phys.* **2005**, *123*, 174101.

(43) Slater, J. C. A Simplification of the Hartree-Fock Method. *Phys. Rev.* **1951**, *81*, 385–390.

(44) Vosko, S. H.; Wilk, L.; Nusair, M. Accurate Spin-Dependent Electron Liquid Correlation Energies for Local Spin Density Calculations: A Critical Analysis. *Can. J. Phys.* **1980**, *58*, 1200–1211.

(45) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.

(46) Perdew, J. P.; Ruzsinszky, A.; Csonka, G. I.; Vydrov, O. A.; G. E. Scuseria, L. A. C.; Zhou, X.; Burke, K. Restoring the Density-Gradient Expansion for Exchange in Solids and Surfaces. *Phys. Rev. Lett.* **2008**, *100*, 136406.

(47) Zhao, Y.; Truhlar, D. G. A new local density functional for main-group thermochemistry, transition metal bonding, thermochemical kinetics, and noncovalent interactions. *J. Chem. Phys.* **2006**, *125*, 194101.

(48) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *J. Phys. Chem.* **1994**, *98*, 11623–11627.

(49) Becke, A. D. Density-Functional Exchange-Energy Approximation with Correct Asymptotic Behavior. *Phys. Rev. A* **1988**, *38*, 3098.

(50) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron Density. *Phys. Rev. B* **1988**, *37*, 785–789.

(51) Becke, A. D. Density-Functional Thermochemistry. V. Systematic Optimization of Exchange-Correlation Functionals. *J. Chem. Phys.* **1997**, *107*, 8544–8560.

(52) Hamprecht, F. A.; Cohen, A. J.; Tozer, D. J.; Handy, N. C. Development and Assessment of New Exchange-

Correlation Functionals. *J. Chem. Phys.* **1998**, *109*, 6264–6271.

(53) Perdew, J. P.; Ernzerhof, M.; Burke, K. Rationale for Mixing Exact Exchange with Density Functional Approximations. *J. Chem. Phys.* **1996**, *105*, 9982–9985.

(54) Adamo, C.; Barone, V. Toward Reliable Density Functional Methods Without Adjustable Parameters: The PBE0 Model. *J. Chem. Phys.* **1999**, *110*, 6158–6170.

(55) Zhao, Y.; Truhlar, D. G. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theor. Chem. Acc.* **2008**, *120*, 215–241.

(56) Krukau, A. V.; Vydrov, O. A.; Izmaylov, A. F.; Scuseria, G. E. Influence of the Exchange Screening Parameter on the Performance of Screened Hybrid Functionals. *J. Chem. Phys.* **2006**, *125*, 224106.

(57) Schimka, L.; Harl, J.; Kresse, G. Improved Hybrid Functional for Solids: The HSEsol Functional. *J. Chem. Phys.* **2011**, *134*, 024116.

(58) Henderson, T. M.; Izmaylov, A. F.; Scuseria, G. E.; Savin, A. The Importance of Middle-Range Hartree-Fock-Type Exchange for Hybrid Density Functionals. *J. Chem. Phys.* **2007**, *127*, 221103.

(59) Henderson, T. M.; Izmaylov, A. F.; Scuseria, G. E.; Savin, A. Assessment of a Middle-Range Hybrid Functional. *J. Chem. Theory Comput.* **2008**, *4*, 1254–1262.

(60) Vydrov, O. A.; Scuseria, G. E. Assessment of a Long-Range Corrected Hybrid Functional. *J. Chem. Phys.* **2006**, *125*, 234109.

(61) Savin, A.; Toulouse, J.; Flad, H. J. Short-range Exchange-Correlation Energy of a Uniform Electron Gas with Modified Electron-Electron Interaction. *Int. J. Quantum Chem.* **2004**, *100*, 1047–1056.

(62) Gerber, I. C.; Angyan, J. G. London Dispersion Forces by Range-Separated Hybrid Density Functional with Second Order Perturbational Corrections: The Case of Rare Gas Complexes. *J. Chem. Phys.* **2007**, *126*, 044103.

(63) Savin, A. In *Recent Development and Applications of Density Functional Theory*; Seminario, J., Ed.; Elsevier: Amsterdam, 1996; pp 327–357.

(64) Gill, P. M. W.; Adamson, R. D.; Pople, J. A. Coulomb-Attenuated Exchange Energy Density Functionals. *Mol. Phys.* **1996**, *88*, 1005–1009.

(65) Chai, J.-D.; Head-Gordon, M. Systematic Optimization of Long-Range Corrected Hybrid Density Functionals. *J. Chem. Phys.* **2008**, *128*, 084106.

(66) *Strukturbericht* designation is taken from the Crystal Lattice Structures web page: http://cst-www.nrl.navy.mil/lattice/struk/index.html.

(67) Hao, P.; Fang, Y.; Sun, J.; Csonka, G. I.; Philipsen, P. H. T.; Perdew, J. P. Lattice Constants from Semilocal Density Functionals with Zero-Point Phonon Correction. *Phys. Rev. B* **2012**, *85*, 014111.

(68) Haas, P.; Tran, F.; Blaha, P.; Schwarz, K. Construction of an Optimal GGA Functional for Molecules and Solids. *Phys. Rev. B* **2011**, *83*, 205117.

(69) Madelung, O. *Semiconductors: Data Handbook*; Springer-Verlag Berlin Heidelberg and New York, 3rd ed. and 2004.

(70) El-Mellouhi, F.; Brothers, E. N.; Lucero, M. J.; Scuseria, G. E. Modeling of the Cubic and Antiferrodistortive Phases of SrTiO3 with Screened Hybrid

Density Functional Theory. *Phys. Rev. B* **2011**, *84*, 155122.

(71) Pässler, R. Parameter Sets Due to Fittings of the Temperature Dependencies of Fundamental Bandgaps in Semiconductors. *Phys. Status Solidi B* **1999**, *216*, 975–1007.

(72) Roessler, D. M.; Walker, W. C. Electronic Spectra of Crystalline NaCl and KCl. *Phys. Rev.* **1968**, *166*, 599–606.

(73) Kaduk, J. A. Chemical Accuracy and Precision in Structural Refinements from Powder Diffraction Data. *Adv. X-Ray Anal.* **1996**, *40*, 352.

(74) Herbstein, F. H. How Precise Are Measurements of Unit-Cell Dimensions from Single Crystals? *Acta Cryst. B* **2000**, *56*, 547–557.

(75) Lucero, M. J.; Henderson, T. M.; Scuseria, G. Improved Semiconductor Lattice Parameters and Band Gaps from a Middle-Range Screened Hybrid Exchange Functional. *J. Phys.: Condens. Matter* **2012**, *24*, 145504.

(76) R Core Team, R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing: Vienna, Austria, 2013; http://www.R-project.org/.

(77) Csonka, G. I.; Perdew, J. P.; Ruzsinszky, A.; Philipsen, P. H. T.; Lebègue, S.; Paier, J.; Vydrov, O. A.; Ángyán, J. G. Assessing the performance of recent density functionals for bulk solids. *Phys. Rev. B* **2009**, *79*, 155107.

(78) Sivia, D. S. *Data Analysis: A Bayesian Tutorial*, 2nd ed.; Oxford Univ. Press: New York, 2006.

(79) Mana, G.; Giuliano Albo, P. A.; Lago, S. Bayesian Estimate of the Degree of a Polynomial Given a Noisy Data Sample. *Measurement* **2014**, *55*, 564–570.

(80) Hawkins, D. M.; Basak, S. C.; Mills, D. Assessing Model Fit by Cross-Validation. *J. Chem. Inf. Comp. Sci.* **2003**, *43*, 579–586.

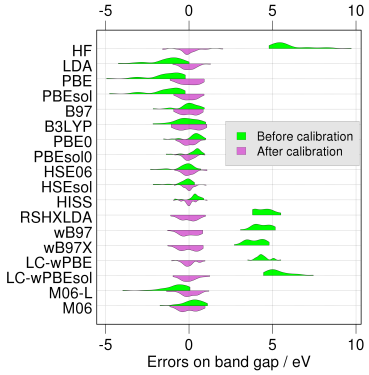
(81) Note that these values cannot be larger than the smallest *RMSD* for each property (Tables 2-4). Otherwise, one would have to conclude that some DFAs are overfitting the reference data, which is rather unlikely, except if these DFAs had been internally calibrated on these same data, with some level of overfitting.

(82) Kloke, J.; McKean, J. Rfit: Rank Estimation for Linear Models. 2014; R package version 0.18 - http://CRAN.R-project.org/package=Rfit.

(83) Heberger, K.; Kemeny, S.; Vidoczy, T. On The Errors Of Arrhenius Parameters And Estimated Rate-Constant Values. *Int. J. Chem. Kinet.* **1987**, *19*, 171–181.

(84) Hébrard, E.; Dobrijevic, M.; Pernot, P.; Carrasco, N.; Bergeat, A.; Hickson, K. M.; Canosa, A.; Le Picard, S. D.; Sims, I. R. How Measurements of Rate Coefficients at Low Temperature Increase the Predictivity of Photochemical Models of Titan's Atmosphere. *J. Phys. Chem. A* **2009**, *113*, 11227–11237.

TOC Image


-5
0
5
10
HF
LDA
PBE
PBEsol
B97
B3LYP
PBE0
PBEsol0
HSE06
HSEsol
HISS
RSHXLDA
wB97
wB97X
LC-wPBE
LC-wPBEsol
M06-L
M06
Before calibration
After calibration
Errors on band gap / eV