Uncertainty associated with virtual measurements from computational quantum chemistry models

Karl K Irikura, Russell D Johnson III and Raghu N Kacker

National Institute of Standards and Technology, Gaithersburg, MD 20899, USA
E-mail: karl.irikura@nist.gov, russell.johnson@nist.gov and raghu.kacker@nist.gov

Received 19 April 2004
Published 21 September 2004
Online at stacks.iop.org/Met/41/369
doi:10.1088/0026-1394/41/6/003

Abstract
A value for the measurand determined from a computational model is frequently referred to as a virtual measurement to distinguish it from a physical measurement, which is determined from a laboratory experiment. Any measurement, physical or virtual, is incomplete without a quantitative statement of its associated uncertainty. The science and technology of making physical measurements and quantifying their uncertainties has evolved over many decades. In contrast, the science and technology of making virtual measurements is evolving. We propose an approach for quantifying the uncertainty associated with a virtual measurement of a molecular property determined from a computational quantum chemistry model. The proposed approach is based on the Guide to the Expression of Uncertainty in Measurement, published by the International Organization for Standardization, and it uses the Computational Chemistry Comparison and Benchmark Database maintained by the National Institute of Standards and Technology.

1. Introduction
By virtual measurement we mean the output of a computational model as an alternative to a physical measurement, which is determined from a laboratory experiment. As computational models improve, virtual measurements are being increasingly treated on the same footing as physical measurements. Interest in virtual measurements is growing for reasons of economics and safety. Comparatively rapid computational approaches are gaining importance as the demand for property data increasingly exceeds the capacity for making physical measurements. Virtual measurements are becoming critical in research and development for chemical processes, new materials, and drug discovery.

Any measurement, whether physical or virtual, is incomplete without a quantitative and valid expression of its associated uncertainty. We address the problem of quantifying the uncertainty associated with a virtual measurement for a molecular property determined from a computational quantum chemistry model. For the present discussion, a virtual measurement is a scalar quantity with an uncertainty that arises primarily, but not necessarily exclusively, from its bias (systematic error) with respect to the value of the molecular property.

The International Organization for Standardization (ISO) Guide to the Expression of Uncertainty in Measurement [1] was developed primarily for physical measurements. However, the Guide is especially useful for virtual measurements because the Guide has established an approach for quantifying the uncertainty arising from bias, which is the primary source of uncertainty associated with a virtual measurement. The approach is as follows. Apply a correction for bias, thus obtaining a corrected virtual measurement. The bias is unknown, and so a correction for bias carries uncertainty. Quantify the standard uncertainty associated with the correction and include it in the combined standard uncertainty.

1 Many chemists and some metrologists prefer the terms calculated and experimental values rather than virtual and physical measurements.

2 There was no generally accepted approach to account for the uncertainty arising from bias before publication of the Guide.
uncertainty associated with the corrected measurement [1, 2]. The Guide treats all components of uncertainty exactly the same way, whether arising from random effects or arising from corrections for biases. Also, the Guide makes no distinction between the components of uncertainty evaluated by statistical methods (Type A) and those evaluated by other means (Type B).

To specify a correction for bias in a virtual measurement from a computational quantum chemistry model and to specify its associated standard uncertainty, we propose to use the Computational Chemistry Comparison and Benchmark Database (CCCBDB) maintained by the National Institute of Standards and Technology (NIST). The CCCBDB [3] is a large, web-accessible database of virtual measurements from many computational quantum chemistry models. Various properties of hundreds of molecules are included, along with the corresponding high-quality physical measurements and their associated uncertainties, where available. For the molecular properties addressed in this paper, the difference between a virtual measurement and the corresponding high-quality physical measurement characterized in the CCCBDB is generally an order of magnitude larger than the uncertainty associated with the high-quality physical measurement. Therefore, a signed difference between a virtual measurement and the corresponding high-quality physical measurement is a useful estimate for the bias in the virtual measurement. In summary, the CCCBDB provides estimated biases in virtual measurements for a number of molecular properties.

We refer to the molecule of interest, for which we require a virtual measurement, as the target molecule to distinguish it from the molecules already characterized in the CCCBDB. Suppose that a class of molecules can be identified in the CCCBDB for which the estimated biases are believed to be similar in sign and magnitude to the bias for the target molecule. Then the arithmetic mean of these estimated biases may be used to specify the correction for bias in the target molecule, and their standard deviation may be used to specify the uncertainty associated with the correction. The correction for bias and its associated uncertainty so determined may then be used to determine a corrected virtual measurement and its associated uncertainty. This approach is feasible when a suitable class of molecules can be identified in the CCCBDB.

In section 2, we describe the proposed approach for quantifying the uncertainty arising from the bias in a virtual measurement determined from a computational quantum chemistry model. In section 3, we give a brief description of the CCCBDB. In section 4, we describe a simple procedure for specifying a correction for bias in a virtual measurement and its associated uncertainty using the CCCBDB. In section 5, we illustrate this procedure. A summary appears in section 6.

2. Uncertainty from bias in computational quantum chemistry

The virtual measurements addressed here are determined from quantum chemistry. For a number of reasons, quantum chemistry is an area where progress in quantifying uncertainties will have immediate and significant impact. (i) It has achieved remarkable success in the past decade, often replacing certain types of laboratory measurements for isolated, gas-phase molecules. (ii) Commercial software packages are proliferating. (iii) Its use is expanding so rapidly that many users of commercial software lack a thorough understanding of the methods and the limitations thereof. (iv) The leaders in the field have not seriously attempted to quantify the uncertainties in quantum chemistry virtual measurements. (v) Practical considerations, such as finite resources, force one to use more strongly approximate theories and/or more severely truncated basis sets than one might prefer, leading to substantial uncertainties.

A formal theory in quantum chemistry is an analytical theory based upon an approximate Hamiltonian, which may be simple or highly complex [4]. The Hamiltonian specifies the physics that is included in the computational model. For a given property and molecule there are many theories that may be selected. Some theories can be ordered according to theoretical rigour, while others cannot be so ordered; the most careful work usually employs theories that can be ordered. To obtain an actual result, the equations of the theory must be solved numerically. This requires a basis set, which is a set of functions that are used in linear combinations to express the molecular orbitals in functional form. Products of the molecular orbitals, in turn, are used in linear combinations to express the electronic wavefunction for the molecule in functional form. Some implementations require a choice of grid size instead of or in addition to the basis set. Some basis sets can be ordered according to completeness and some cannot be so ordered; the most careful work usually employs basis sets that can be ordered. The numerical results tend to converge as the basis set is enlarged [5]. The rate of convergence changes with the property and the molecule under study in a way that is not understood quantitatively.

The computational cost of quantum chemistry calculations increases rapidly as the basis set is enlarged. For example, energy calculations for the molecule H2O using the sophisticated CCSD(T) theory [6, 7] and the series of basis sets aug-cc-pVnZ (n = 2–6) [5], which are the most popular in careful, quantitative work, have computational times of about 0.2 × 10^6 s on a desktop personal computer. Computational difficulty also increases rapidly as the complexity of the formal theory increases. Again for H2O, on the same computer, the computational time using the n = 3 basis set approximately doubles with each step in the sequence of theories HF, MP2, MP3, MP4, CCSD(T). Since the cost of calculations increases so quickly as the basis set or theory is improved, in practice one is usually forced to accept strong approximations in both, leading to significant bias in the output of a computational quantum chemistry model.

2.1. Bias in computational quantum chemistry

Suppose the measurand is a particular property of a specific molecule and its value is Y. The value Y is a statistical parameter. A computational quantum chemistry model is defined by a combination of a formal theory and a basis set [8]. Suppose x(t,b) is a virtual measurement for Y based on a computational quantum chemistry model, where t is the ordinal number for the formal theory in some hierarchy and b is the ordinal number for the basis set in some hierarchy. For example, if n_e is the number of electrons in the molecule under
Suppose \( X_{t,b} \) is the expected value of the sampling probability distribution for \( x_{t,b} \). The difference \( X_{t,b} - X_{t,b} \) is the random error in \( x_{t,b} \). The ratio \( x_{t,b} / X_{t,b} \) is termed the fractional random error. The random error arises from a variety of small contributions, such as the non-zero convergence thresholds that create some dependence upon the choice of initial geometry and wavefunction. Such a random error is generally negligible. If not negligible, its associated uncertainty must be quantified and included as a component of the uncertainty associated with \( x_{t,b} \). The difference \( X_{t,b} - Y \) is the additive bias in \( x_{t,b} \). The ratio \( x_{t,b} / Y \) is termed the fractional (or multiplicative) bias in \( x_{t,b} \). In this paper, we deal with the additive bias only, and so we will drop the adjective ‘additive’. The bias \( X_{t,b} - Y \), denoted by \( B_{t,b} \), is a statistical parameter. The bias is unknown because \( Y \) is unknown.

The bias \( B_{t,b} \) has two components: the bias \( B_t \), arising from the choice of an approximate formal theory, and the bias \( B_b \), arising from the choice of an incomplete basis set. Convergent behaviour is assumed, i.e. that \( \lim_{t \to \infty} B_t = 0 \) as \( t \to \infty \) (at least for some theoretical hierarchies) and that \( \lim_{b \to \infty} B_b = 0 \) as \( b \to \infty \). There have been few investigations directed towards evaluating the biases \( B_t \) and \( B_b \), attributable to theory and the basis set, respectively. To reveal \( B_t \), results are needed in the limit of a complete basis set, at which point \( B_b = 0 \). This is usually done using semi-empirical extrapolation methods [5] but remains too expensive computationally to be of widespread practical use. To reveal \( B_b \), results are needed in the limit of a rigorous theory, at which point \( B_t = 0 \). So-called ‘full configuration-interaction’ (FCI) calculations are even more expensive, and see occasional application only for benchmarking approximate theories [9]. There have been few studies of the correlations between \( B_t \) and \( B_b \) [10]. It is often assumed that \( B_t \) and \( B_b \) are independent and additive. This ‘additivity approximation’ is believed to be most valid for ‘high-level’ models, i.e. models that combine a refined formal theory and a large basis set. Its assumed validity underlies the popular semi-empirical procedure known as ‘G3’ [11] and the related ‘focal point’ [12] and ‘W3’ [13] approaches. Of the few studies of uncertainties in quantum chemistry models, nearly all consider only the aggregate bias, \( B_{t,b} \). This approach stems from the difficulty in obtaining \( B_t \) and \( B_b \) independently, as explained above. Furthermore, the most popular models are relatively crude, for which \( B_t \) and \( B_b \) are interrelated anyway. The sign and magnitude of the aggregate bias, \( B_{t,b} \), depends on the choices made for the formal theory and the basis set. In section 2.2, we describe an approach based on the Guide [1] to quantify the uncertainty arising from the bias \( B_{t,b} \) in \( x_{t,b} \).

2.2. Uncertainty from bias

The Guide is based on the concept of a measurement equation. In its simplest form, this is a mathematical function, \( Y = f(Q_1, \ldots, Q_N) \), that represents the process used for estimating the value, \( Y \), of the measurand and its associated standard uncertainty\(^3\) from various input quantities \( Q_1, \ldots, Q_N \) [2]. Each input and output quantity of a measurement equation is regarded as a variable with a state-of-knowledge probability distribution having a finite expected value and a finite standard deviation. The input variables \( Q_1, \ldots, Q_N \) may themselves be viewed as measurands and functions of additional input variables [2]. Thus the measurement equation may actually be a hierarchical system of equations.

The Guide recommends that \( x_{t,b} \), be corrected to counter its bias \( B_{t,b} \), thus providing a corrected virtual measurement \( y \) for \( Y \). From this viewpoint, we refer to \( x_{t,b} \) as an uncorrected virtual measurement for \( Y \). A measurement equation is required to incorporate a correction for bias. The measurement equation that corresponds to the bias \( B_{t,b} = X_{t,b} - Y \) is

\[
y = X_{t,b} + C_{t,b},
\]

where \( C_{t,b} \) is a variable representing the state-of-knowledge about the expression \( Y - X_{t,b} \) for the negative of bias. In the measurement equation (1), the input quantity \( X_{t,b} \) is regarded as a variable with a state-of-knowledge probability distribution about the expected value \( X_{t,b} \), and the output quantity \( Y \) is regarded as a variable with a state-of-knowledge distribution about the value \( Y \) of the measurand\(^4\). The expected value, \( E(X_{t,b}) \), of a state-of-knowledge distribution for \( X_{t,b} \) is identified with the uncorrected virtual measurement \( x_{t,b} \). The standard deviation \( S(X_{t,b}) \) of a state-of-knowledge distribution for \( X_{t,b} \) is referred to as the standard uncertainty associated with \( x_{t,b} \) and is denoted by \( u(x_{t,b}) \). We will discuss evaluation of \( u(x_{t,b}) \) in section 2.3. The expected value \( E(C_{t,b}) \) and standard deviation \( S(C_{t,b}) \) of a state-of-knowledge distribution for \( C_{t,b} \) are denoted by \( c_{t,b} \) and \( u(c_{t,b}) \), respectively. We will discuss in section 4 how the CCCDBB [3] may be used to specify the correction \( c_{t,b} \) and its associated uncertainty \( u(c_{t,b}) \).

A corrected virtual measurement \( y \) for \( Y \) is determined by substituting the expected value \( x_{t,b} \) for the variable \( X_{t,b} \) and the expected value \( c_{t,b} \) for the variable \( C_{t,b} \) in the measurement equation (1). Thus

\[
y = x_{t,b} + c_{t,b}.
\]

That is, \( c_{t,b} \) is the correction applied to the uncorrected virtual measurement \( x_{t,b} \) to counter its possible bias. Following the Guide, the combined standard uncertainty, \( u(y) \), associated with the corrected virtual measurement \( y \) is determined by propagating the standard uncertainties \( S(X_{t,b}) = u(x_{t,b}) \), \( S(C_{t,b}) = u(c_{t,b}) \), and the covariance \( C(X_{t,b}, C_{t,b}) \). A distribution for \( C_{t,b} \) is specified independent of the state-of-knowledge distribution for \( X_{t,b} \) after \( x_{t,b} \) and \( u(x_{t,b}) \) have been evaluated. So the state-of-knowledge distributions for \( X_{t,b} \) and \( C_{t,b} \) are independent. Consequently, the covariance \( C(X_{t,b}, C_{t,b}) \) is zero. Therefore, the expression for propagating uncertainties based on the measurement equation (1) is

\[
u(y) = \left[ u^2(x_{t,b}) + u^2(c_{t,b}) \right]^{1/2}.
\]

The corrected virtual measurement \( y \) and uncertainty \( u(y) \) so determined are interpreted as the expected value and standard deviation of a state-of-knowledge distribution for \( Y \).

---

\(^3\) Standard uncertainty is the standard deviation of a state-of-knowledge distribution for \( Y \).

\(^4\) As in the Guide [1], we use the same symbols for both the statistical parameters and the variables with state-of-knowledge probability distributions about the parameters.
The input quantities for determining \( x(t,b) \) from a computational quantum chemistry model are the fundamental physical constants and a few (or zero) empirically derived parameters. A computational model for \( x(t,b) \) may be expressed in the form of a measurement equation as follows:

\[
X(t,b) = g(W_1, \ldots, W_n | t, b) + E(t,b),
\]

where \( W_1, \ldots, W_n \) are variables with state-of-knowledge distributions for the fundamental physical constants and any empirically derived parameters, and \( E(t,b) \) is a variable with a state-of-knowledge distribution for the random error. The expected values of state-of-knowledge distributions for \( W_1, \ldots, W_n \) are identified with the input values of the fundamental physical constants and any empirically derived parameters denoted by \( w_1, \ldots, w_n \). The standard deviations of state-of-knowledge distributions for \( W_1, \ldots, W_n \) are identified with the input values of the fundamental physical constants and any empirically derived parameters denoted by \( w_1, \ldots, w_n \). The expected value of a state-of-knowledge distribution for \( E(t,b) \) is \( e(t,b) = 0 \). The standard deviation of a state-of-knowledge distribution for \( E(t,b) \) is referred to as the standard uncertainty associated with \( e(t,b) = 0 \) and denoted by \( u(e(t,b)) \). The uncorrected virtual measurement \( x(t,b) \) is

\[
x(t,b) = g(w_1, \ldots, w_n | t, b) + 0 = g(w_1, \ldots, w_n | t, b).
\]

The standard uncertainty, \( u(x(t,b)) \), associated with \( x(t,b) \) is determined from a linear approximation, \( X(t,b) \approx X(t,b,\text{linea}) = x(t,b) + \sum d_i(W_i - w_i) + (E(t,b) - e(t,b)) \), of the measurement equation (4), where \( d_i \) is the partial derivative of \( X(t,b) \) with respect to \( W_i \) evaluated at \( w_i \) for \( i = 1, \ldots, n \) and the partial derivative of \( X(t,b) \) with respect to \( E(t,b) \) evaluated at \( e(t,b) = 0 \) is one. The variable \( E(t,b) \) is uncorrelated with the variables \( W_1, \ldots, W_n \). Thus the standard deviation \( \sigma(X(t,b)) \), denoted by \( u(x(t,b)) \), is

\[
u(x(t,b)) = \left[ u^2 + \sum_{i<j} d_i d_j u(w_i) u(w_j) \times r(w_i, w_j) + u^2(e(t,b)) \right]^{1/2},
\]

where \( r(w_i, w_j) \) is the correlation coefficient between \( W_i \) and \( W_j \) for \( i, j = 1, \ldots, N \), \( i \neq j \). The uncorrected virtual measurement \( x(t,b) \) and uncertainty \( u(x(t,b)) \) determined from equations (5) and (6), respectively, are components of the corrected virtual measurement \( x(t,b) \) for \( X \) and uncertainty \( u(x) \) defined by equations (2) and (3), respectively.

3. Computational chemistry comparison and benchmark database

The CCCBDB is a web-accessible database of differences between virtual measurements and the corresponding high-quality physical measurements. The initial focus was on gas-phase thermochemistry. Values derived from physical measurements, including uncertainties where available, have been collected for the enthalpies of formation, entropies, heat capacities, geometries, and vibrational frequencies of 640 molecules. The uncertainties associated with the physical measurements are generally an order of magnitude smaller than the differences between virtual measurements and the corresponding high-quality physical measurements. Thus the high-quality physical measurements are appropriate for benchmarking the virtual measurements.

The quantum chemistry calculations in the CCCBDB have been performed using a variety of computational models. As of August 2004, results from more than 85,000 quantum chemistry calculations are available in the database. The CCCBDB includes web pages for examining the differences between virtual and physical measurements. Thus, the CCCBDB provides the estimated biases in virtual measurements from many computational models. Because some interesting properties are neither calculated nor measured directly (such as enthalpy of formation), tools are provided for designing customized chemical reactions. The corresponding
4. Specification of a correction for bias and its associated uncertainty

The correction $c_{(t,b)}$ and its associated standard uncertainty $u(c_{(t,b)})$ are determined from a state-of-knowledge distribution for $C_{(t,b)}$ that is specified from all available information including relevant data and scientific judgment. We propose for $C_{(t,b)}$ a mixture probability distribution whose expected value and standard deviation are determined using the CCCBDB. Then $c_{(t,b)}$ and $u(c_{(t,b)})$ are identified with the expected value $E(C_{(t,b)})$ and standard deviation $S(C_{(t,b)})$ of the distribution for $C_{(t,b)}$, respectively. The correction $c_{(t,b)}$ and uncertainty $u(c_{(t,b)})$ are then used to determine the corrected virtual measurement $y$ and uncertainty $u(y)$ from equations (2) and (3), respectively.

Suppose, for the same theory and basis set as for the target molecule, a class of molecules in the CCCBDB can be identified with the following three characteristics.

(i) The bias $B_{(t,b)}$ for the target molecule is believed to be of the same sign and of similar magnitude as the estimated biases for the molecules in the class. The negatives of these estimated biases are estimated corrections for the molecules in the class. Suppose the number of molecules in this class is $m$ and the estimated corrections are $c_1, \ldots, c_m$.

(ii) The estimated corrections $c_1, \ldots, c_m$ appear to have an approximately normal distribution and do not have an excessively large spread. An approximately normal distribution is desired because we treat $c_1, \ldots, c_m$ as a set of randomly distributed values about their arithmetic mean, $\mu = (1/m) \sum c_i$. A distribution that has an excessively large spread is often a mixture of narrower distributions that may be separated. The approximate normality and spread can be assessed from a histogram of $c_1, \ldots, c_m$, the standard deviation $\sigma = [(\sum (c_i - \mu)^2/m]^{1/2}$, and the coefficient of skewness $\eta_1 = [(\sum (c_i - \mu)^3/m]/\sigma^3$ [16], which is zero for a normal distribution.

(iii) The number, $n$, of molecules in the class is sufficiently large.

Suppose the values of the molecular property for the identified class of molecules in the CCCBDB are $y_1, \ldots, y_m$, the corresponding uncorrected virtual measurements are $x_1, \ldots, x_m$ with standard uncertainties $u(x_1), \ldots, u(x_m)$, and the high-quality physical measurements are $z_1, \ldots, z_m$ with standard uncertainties $u(z_1), \ldots, u(z_m)$, respectively. The uncertainties $u(x_1), \ldots, u(x_m)$, like $u(x_{(t,b)})$, do not include the components of uncertainty arising from the biases in $x_1, \ldots, x_m$, respectively. Suppose the expected values of the sampling distributions for $x_1, \ldots, x_m$ are $X_1, \ldots, X_m$, respectively. Then the bias in $x_i$ is $X_i - Y_i$ for $i = 1, 2, \ldots, m$. The virtual measurement $x_i$ is an estimate for $X_i$ and the high-quality physical measurement $z_i$ is an estimate for $Y_i$, and so $x_i - z_i$ is an estimate for the bias $X_i - Y_i$ in $x_i$ and the estimated correction for bias is $c_i = z_i - x_i$.

In accordance with the Guide, the virtual measurement $x_i$ and the uncertainty $u(x_i)$ are regarded as the expected value and the standard deviation of a state-of-knowledge distribution for $X_i$. The physical measurement $z_i$ and the uncertainty $u(z_i)$ are regarded as the expected value and the standard deviation of a state-of-knowledge distribution for $Y_i$. Let $C_i = Y_i - X_i$ be a variable representing the correction for bias in $x_i$. Then the expected value of a state-of-knowledge distribution for $C_i$ is $c_i = z_i - x_i$. Since the state-of-knowledge distributions for $Y_i$ and $X_i$ are determined independently, the covariance between $Y_i$ and $X_i$ is zero. Thus the standard deviation of a state-of-knowledge distribution for $C_i$ is $S(C_i) = [u^2(z_i) + u^2(x_i)]^{1/2}$. We will use the symbol $u(c_i)$ for $S(C_i)$. Thus

$u(c_i) = [u^2(z_i) + u^2(x_i)]^{1/2}$.

The uncertainty $u(x_i)$ is generally negligible relative to the uncertainty $u(z_i)$ (see section 2.3). Thus, to a reasonable approximation $u(c_i) \approx u(z_i)$, for $i = 1, 2, \ldots, m$.

According to the belief that the bias $B_{(t,b)}$ for the target molecule is similar to the estimated biases for the class of molecules identified in the CCCBDB, each of the $m$ state-of-knowledge distributions for $C_1, \ldots, C_m$ may be attributed to $C_{(t,b)}$. Suppose the probability density function (PDF) for $C_i$ is $p(y)$. We propose that the PDF $p(\cdot)$ attributed to $C_{(t,b)}$ be defined as a linear combination $p(y) = \sum_i p_i(y)$ of the PDFs $p_i(\cdot)$, where $a_1 / \sum_i a_i$ and $a_1, \ldots, a_m$ are non-negative ‘weights’ attributed to $p_1(\cdot), \ldots, p_m(\cdot)$, respectively. A combined probability distribution with PDF $p(\cdot) = \sum_i a_i p_i(\cdot)$ is referred to as a mixture probability distribution. The expected value and standard deviation of the PDF $p(\cdot)$ are $\sum_i a_i c_i$ and $[\sum_i a_i c_i^2 + \sum_i a_i (c_i - \mu)^2]/m^{1/2}$, respectively. A combined probability distribution with equal weights, the correction $c_{(t,b)}$ and uncertainty $u(c_{(t,b)})$ may be specified as

$$c_{(t,b)} = \mu = \frac{1}{m} \sum_i c_i \quad \text{(7)}$$

and

$$u(c_{(t,b)}) = \left[ \frac{1}{m} \sum_i c_i^2 + \frac{1}{m} \sum_i (c_i - \mu)^2 \right]^{1/2}$$

$$= \left[ \frac{1}{m} \sum_i c_i^2 + \sigma^2 \right]^{1/2}, \quad \text{(8)}$$

respectively. In equation (8), $u(c_i)$ is approximated by $u(z_i)$, the uncertainty associated with the high-quality physical measurement, for $i = 1, 2, \ldots, m$.

5. Illustration of the procedure

We consider a computationally inexpensive quantum chemistry model\(^6\) for calculating enthalpy changes for atomization reactions (e.g., $\text{H}_2\text{O} \rightarrow \text{2H} + \text{O}$). The virtual measurement $x_{(t,b)}$ is the atomization enthalpy for a target molecule, at the

---

\(^6\) The density-functional method denoted mPW1PW91 [18] combined with the basis set denoted 6-31G(d).
Suppose the target molecule is the sulfur-containing organic molecule ethyl thioformate (C\textsubscript{2}H\textsubscript{5}SCHO). Figure 1 shows a histogram of estimated corrections for the biases in the atomization enthalpies for a class of 65 sulfur-containing organic molecules for which data are available in the CCCBDB. Figure 2 separates the histogram of figure 1 into two histograms, corresponding to two smaller classes, one for the molecules containing S–O bonds and the other for the molecules lacking S–O bonds. The entries in columns 2, 3, 4, and 5 of table 1 are the number, \(m\), of molecules in the three classes, the arithmetic mean, \(\mu\), the standard deviation, \(\sigma\), and the coefficient of skewness, \(\eta_1\), for the three distributions of estimated corrections. Figure 2 and table 1 illustrate the benefit of recognizing better ways of classifying molecules. When we distinguish between the molecules based upon whether they contain S–O bonds, the two resulting distributions for estimated corrections are more symmetric than is their combined distribution. It is clear from table 1 that the finer classification leads to a marked reduction in the coefficient of skewness and the standard deviation. The entries in columns 3 and 6 of table 1 are the corrections \(c(t,b)\) based on equation (7) and the uncertainties \(u(c(t,b))\) based on equation (8) for the three classes of molecules.

The target molecule, ethyl thioformate (C\textsubscript{2}H\textsubscript{5}SCHO), is a member of the class of sulfur-containing organic molecules without an S–O bond. Therefore the summary statistics for the class of molecules without S–O bonds apply. The atomization enthalpy for ethyl thioformate using the same model as above is \(x(t,b) = 4093.8\text{ kJ mol}^{-1}\). Therefore the appropriate correction \(c(t,b)\) and uncertainty \(u(c(t,b))\) are \(c(t,b) = \mu = 21.8\text{ kJ mol}^{-1}\) and \(u(c(t,b)) = 19.2\text{ kJ mol}^{-1}\), respectively. Thus the corrected virtual measurement for ethyl thioformate is \(y = x(t,b) + c(t,b) = 4115.6\text{ kJ mol}^{-1}\) and the standard uncertainty associated with the correction is \(u(y) \approx u(c(t,b)) = 19.2\text{ kJ mol}^{-1}\). In summary, after applying the conventional coverage factor \(k = 2\), the result of measurement determined from the computational model is \((4115.6 \pm 38.4)\text{ kJ mol}^{-1}\). This is in agreement with the corresponding physical measurement \((4129.2 \pm 5)\text{ kJ mol}^{-1}\) [19]. We note that many quantum chemistry models are superior to this one, although they generally carry a higher computational cost.

As illustrated by table 1, the central problem is discovering useful classification schemes for molecules. Identifying a useful class of molecules requires some understanding of the relationship between the property being modelled and the limitations of quantum chemistry models. Fortunately, a few simple considerations are appropriate for many properties. For example, molecules can be divided into large classes based upon the chemical elements of which they are composed. An example is shown in figure 1. A finer distinction, which is often beneficial, is to distinguish molecules based upon the types of chemical bonds that they contain. This is illustrated in figure 2, where the initial class has been divided into two smaller classes. Further distinctions might be made based upon how many bonds of a given type are in the molecule, upon the existence of low-lying excited states, etc. Smaller classes are expected to be more reliable since they may resemble the target
molecule more closely. However, classes must be sufficiently large for the arithmetic mean and standard deviation to be useful. Furthermore, the distribution of estimated biases for the class of molecules should be approximately normal. Classification schemes need not be unique and classes need not be disjoint. Different classification schemes may yield different uncertainties. Although the present discussion deals with discrete classification, classification schemes may also be continuous. For example, the correction for bias may be a function of an electron density or the length of a bond [20].

6. Summary

The uncertainty associated with a virtual measurement from a computational quantum chemistry model arises primarily from its bias, which results from the choice of the theory and the basis sets used for computation. According to the Guide, a correction for bias must be applied to the virtual measurement, thus obtaining a corrected virtual measurement. The uncertainty associated with the correction must be quantified and then included in the combined standard uncertainty associated with the corrected virtual measurement. We propose the following procedure for determining a corrected virtual measurement and its associated uncertainty:

Step 1. Determine the virtual measurement $x(t,b)$ and its associated uncertainty $u(x(t,b))$ for the target molecule. The uncertainty $u(x(t,b))$ includes the components of uncertainty associated with the fundamental physical constants, any empirically derived parameters, and a variety of small contributions, such as the non-zero convergence thresholds that create some dependence upon the choice of initial geometry and wavefunction.

Step 2. Identify a suitable class of molecules in the CCCBDB database that are believed to have biases similar to that of the target molecule. The database provides the estimated corrections (negative of estimated biases) in the virtual measurements of the same property for the selected class of molecules. Suppose $c_1, \ldots, c_m$ are the estimated corrections with standard uncertainties $u(c_1), \ldots, u(c_m)$, respectively, for the class of molecules. The uncertainties $u(z_1), \ldots, u(z_m)$ are approximated by the uncertainties $u(c_1), \ldots, u(c_m)$ associated with the high-quality physical measurements used to benchmark the virtual measurements characterized in the database. Compute the mean $\mu = (1/m) \sum c_i$ and standard deviation $\sigma = \left[ \left( \sum (c_i - c_A)^2 / m \right)^{1/2} \right]$ of the estimated corrections. Then the correction to be applied to the virtual measurement $x(t,b)$ is $c(t,b) = \mu$ and the standard uncertainty associated with the correction is $u(c(t,b)) = \left[ (1/m) \sum u(c_i)^2 + \sigma^2 \right]^{1/2}$.

Step 3. Determine the corrected virtual measurement $y = x(t,b) + c(t,b)$ and its associated standard uncertainty $u(y) = \left[ u^2(x(t,b)) + u^2(c(t,b)) \right]^{1/2}$. Frequently, the uncertainty $u(x(t,b))$ is negligible relative to $u(c(t,b))$. In that case $u(y) \approx u(c(t,b))$. The corrected virtual measurement $y$ and uncertainty $u(y)$ are regarded as the expected value and standard deviation of a state-of-knowledge distribution for $Y$, the value of the molecular property for the target molecule.

Acknowledgments

We thank both anonymous referees for especially helpful comments.

References