Interval Prediction of Molecular Properties in Parametrized Quantum Chemistry

David E. Edwards,[1,2] Dmitry Yu. Zubarev,[3] Andrew Packard,[1] William A. Lester, Jr.,[4,5] and Michael Frenklach[1,2,*]
[1]Department of Mechanical Engineering, University of California at Berkeley, Berkeley, California 94720-1740, USA
[2]Environmental Energy Technologies Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA
[3]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02138, USA
[4]Kenneth S. Pitzer Center for Theoretical Chemistry, Department of Chemistry, University of California at Berkeley, Berkeley, California 94720-1460, USA
[5]Chemical Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA
(Received 10 April 2014; published 26 June 2014)

The accurate evaluation of molecular properties lies at the core of predictive physical models. Most reliable quantum-chemical calculations are limited to smaller molecular systems while purely empirical approaches are limited in accuracy and reliability. A promising approach is to employ a quantum-mechanical formalism with simplifications and to compensate for the latter with parametrization. We propose a strategy of directly predicting the *uncertainty interval* for a property of interest, based on training-data uncertainties, which sidesteps the need for an optimum set of parameters.

DOI: 10.1103/PhysRevLett.112.253003 PACS numbers: 31.15.A-, 31.15.E-, 32.10.Hq

The accurate evaluation of molecular properties lies at the core of predictive physicochemical models. Examples can be found in current research in atmospheric, materials, energy, and combustion processes. Quantum-chemistry tools provide a solid base for molecular computations. However, the most reliable quantum-chemical calculations are limited to smaller molecular systems while practical interests reside in increasingly larger molecules.

In a purely empirical approach, a model is trained on—i.e., parameters fit to—a limited set of chemical compounds. The ability of such empirical models to make reliable predictions far beyond the training set is obviously indeterminate. To remove some of the empiricism, a general strategy has been pursued of employing quantum mechanics with simplifications that enable faster calculations. Most notable in this quest are the so-called semi-empirical methods [1,2] where numerically costly parts of the Hamiltonian are parametrized with values determined by fits to experimental or high-level quantum-chemical data. Over the years, several such optimized parameter sets have been developed that showed improved performance [1–5]. Another example is density functional theory (DFT). Development of DFT functionals that tend toward the formally exact nature of the underlying theory is an uphill battle [6–13], and the lack of capacity for systematic improvement is a recognized deficiency [6]. Fitting procedures are employed both at the level of exchange-correlation functional design and construction of hybrid functionals. The landscape of available functionals is diverse and continues to grow.

All parametrized quantum-chemical methods share one problem: uncertainty in their predictions is understood in some "averaged" way, as an intrinsic value associated with the specific method, e.g., a mean average deviation over a reference data set. It is assumed that any follow-up predictive calculation inherits the same effective uncertainty. However, estimation of uncertainty in the predicted value from the average deviation over the training set is not a reliable strategy. Additionally, discrimination among alternative models on the basis of effective uncertainty is inefficient and leads to an abundance of parametrized quantum-chemical methods similar to the currently frustrating proliferation of combustion chemistry models [14].

Here we propose a different strategy: predicting directly the uncertainty interval for a given property of interest based on training-data uncertainties, thereby sidestepping the need for an optimum set of parameters. The proposed approach is based on the uncertainty quantification framework developed in a series of studies [15–20], called the bound-to-bound data collaboration (B2BDC). It is an optimization-based framework for combining models and training data from multiple sources to ascertain the collective information content. The details of the B2BDC methodology are described in the literature cited. We describe those features pertinent to the present application, using the specific system employed, thus making the presentation more tractable. We begin with the system description.

In this proof-of-principle study we consider vertical ionization potentials (VIPs) of water clusters. Emergence of these quantities in the context of atmospheric chemistry, charging processes, and solvation phenomena confirms their importance [21]. We pose a specific objective of predicting the VIP of a water hexamer from VIP data of the dimer to the pentamer. We choose to work with the double-hybrid form of DFT that incorporates Hartree-Fock exchange and correlation by second-order perturbation theory [22,23],

$$E_{XC} = (1 - C_{HF})E_{X\text{-}GGA} + C_{HF}E_{X\text{-}HF} + (1 - C_{MP2})E_{C\text{-}GGA} + C_{MP2}E(2) \quad (1)$$

where E_{XC} is the exchange-correlation energy, $E_{X\text{-}GGA}$ and $E_{C\text{-}GGA}$ are the exchange and correlation energies, respectively, in the generalized gradient approximation (GGA) of DFT, $E_{X\text{-}HF}$ is the Hartree-Fock exchange energy, $E(2)$ is the correlation energy from second-order Moller-Plesset perturbation theory, C_{HF} is the amount of Hartree-Fock exchange, and C_{MP2} is the amount of MP2 correlation. The calculations were performed using the GAMESS *ab initio* package [5], employing a Pople 6-311G basis set with $2d$- and $1f$-type polarization functions and a diffuse sp shell on O atoms and $1p$ polarization functions added to H atoms [24]. We do not expect the methodology presented here to be very sensitive to the choice of a basis set. As an indicator, the calculations were performed with basis sets differing only by the inclusion of p functions on H. The results showed that while some details changed, the sought-after predicted intervals remained essentially the same. A comparison is provided in the Supplemental Material [25].

The VIP is computed as the difference in total energy of the ionized and neutral forms of the water cluster at the geometry of the neutral form; the total energy in this case is the sum of the nuclear energy, which remains unchanged, and the electronic energy given by Eq. (1). The parameters C_{HF} and C_{MP2} in Eq. (1) can range from 0 to 1, and the uncertainty is presumed to be encapsulated in these two parameters. In doing so, analysis is limited to two dimensions, facilitating visualization.

Following the usual optimization approach, one would determine the optimal C_{HF} and C_{MP2} values by fits to the VIPs of a water dimer, trimer, tetramer, and pentamer and then from these best-fit coefficients would proceed to predict the VIP of the water hexamer. Following the B2BDC framework, we ask not for the single set of the best-fit parameter values, but for all possible parameter combinations that predict the VIPs of dimer to pentamer within their respective uncertainty bounds. This set of all such parameter combinations is referred to as the feasible set, F. If F is empty, i.e., there is no single combination (C_{HF}, C_{MP2}) that satisfies inequalities $\mathrm{VIP}_{\mathrm{lowerbound}} \leq \mathrm{VIP} \leq \mathrm{VIP}_{\mathrm{upperbound}}$ for all the training data (namely, dimer, trimer, tetramer, and pentamer), then our data-model system is said to be *inconsistent*. There could be several reasons for this: the training data could be incorrect, the uncertainty bounds could be overoptimistic, or the model represented by Eq. (1) is inadequate to capture the underlying physics. Identification of the inconsistency and the factors causing or affecting it is revealing in its own right and constitutes one of the distinguishing features of the B2BDC approach [15]. Here, as we will see shortly, our data-model system is consistent and, hence, we can proceed with the next step of the analysis, prediction of the unknown values.

Every point of a nonempty F is a combination of parameters C_{HF} and C_{MP2} that maps to a value for the property we seek to predict, i.e., the VIP of the hexamer, which is consistent with the training VIPs. Collectively, all points in F result in an interval of predicted hexamer VIP values. The latter encompasses all the uncertainties of the input data, and is obtained without identifying the best-fit (C_{HF}, C_{MP2}) point. Such a direct transfer of uncertainties is more accurate than the two-step process of first fitting the parameters [20]. A further benefit of the strategy outlined of prediction on the feasible set is the relative ease of adding new training data without the need for determination of new best-fit C_{HF} and C_{MP2} parameters. We next demonstrate these features with the water-cluster example.

First, we select the training data. The accurate measurement of VIPs is difficult [26], so the training-data uncertainty ranges were taken from previously reported high-level quantum chemical calculations including CCSD (T), CASPT2//CASSCF, and CASPT2 [26] and fixed node diffusion Monte Carlo (FNDMC) calculations. The geometries used for the VIP double-hybrid single point energy calculations were taken from the MP2/aug-cc-pVDZ calculations of Segarra-Marti *et al.* [26] (abbreviated as SMR throughout the text).

In the next step, we construct a response surface [27] relating the two double-hybrid parameters C_{HF} and C_{MP2} to the VIP values calculated through the double-hybrid DFT for each molecular system. Representing the relationships between the parameters and properties via simple functions facilitates rapid evaluation during the analysis. The response surfaces for the present study were second-order polynomials. They were created by Latin hypercube sampling [27,28] of the $C_{HF} - C_{MP2}$ space. The initial sampling was performed over the complete variation ranges of C_{HF} and C_{MP2}, namely [0, 1], and using smaller ranges as information on F boundaries became available, see below. The calculated VIPs at the sampled points and the fitted response surface for the dimer are shown in Fig. 1(a). The response-surface fits, fitting errors, points sampled, and VIPs calculated at those points for all cases of the present study, are included in the Supplemental Material [25].

We are now ready to begin the analysis. The intersection of the dimer VIP bounds with the response surface identifies the feasible set for the dimer VIP [Fig. 1(b)]. Any combination of the C_{HF} and C_{MP2} parameter values from the feasible set will yield values consistent with the given dimer VIP range. The feasible sets obtained for all molecular structures of the training set are shown in Fig. 1(c). The dimer through pentamer VIP bounds, taken from SMR, are shown in Table I. As can be seen in Fig. 1(c), the individual feasible sets have a common, overlapping area, outlined in red. This is the feasible set of our combined data-model system. Its existence, as discussed above, indicated that our data-model system is

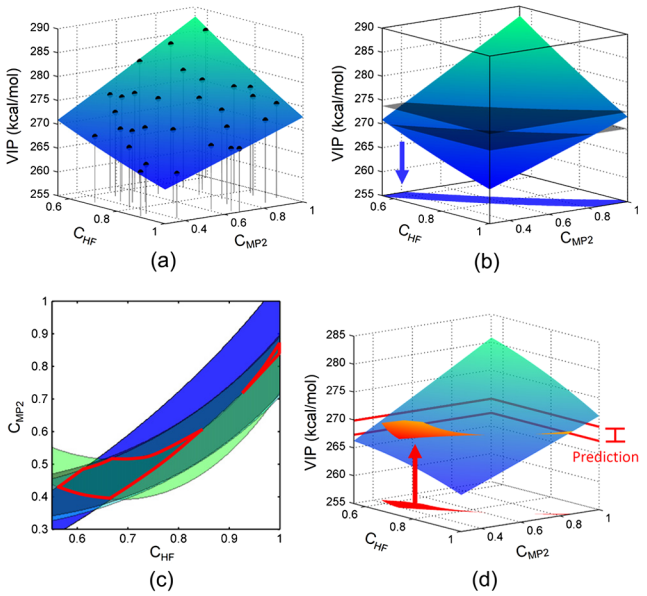

FIG. 1 (color online). (a) Response-surface construction: the stem plot points are from double-hybrid calculations for the dimer VIP at the sampled points; the surface (blue) is the fitted response surface. (b) Feasible set construction: the two horizontal planes intersecting the response surface are the dimer VIP upper and lower bounds; the projection of the response surface patch between these values onto the C_{HF}-C_{MP2} plane is the dimer feasible set. (c) Intersection of the feasible sets for the dimer (blue), trimer (light green), tetramer, and pentamer (overlapped by dark green) yields an overlapping feasible set (outlined in red). (d) The overlapping feasible set (red area in the C_{HF}-C_{MP2} plane) is projected onto the response surface of hexamer-book geometry (blue). The lowest and highest points of the response-surface patch (red) constructed over the overlapping feasible set determine the predicted range of hexamer-book VIP.

consistent and we can proceed to a prediction of a modeled unknown.

For the unknown quantity we chose the VIP of water hexamers, shown in Fig. 1(d) for the hexamer-book geometry [26]. In a manner similar to the water dimer above, we developed a response surface for the hexamer-book VIP, colored in blue in Fig. 1(d). The projection of the established data-model feasible set onto this response surface selects a patch (in red) of the hexamer-book response surface. Each point of this patch, through its relation to F, identifies the hexamer-book VIP values that are consistent with the given VIP uncertainty ranges of dimer, trimer, tetramer, and pentamer. The lowest and highest points of the patch specify the predicted range for the hexamer-book VIP, thus answering the posed question.

TABLE I. VIP bounds for the dimer through pentamer water clusters [26].

Water cluster	VIP bounds (kcal/mol)
Dimer	270.3–273.7
Trimer	282.7–285.0
Tetramer	281.1–283.9
Pentamer	279.0–281.8

It is pertinent to mention again that limiting the analysis to only two parameters, C_{HF} and C_{MP2}, makes possible graphical visualization of response surfaces, direct identification and display of feasible sets, and hence a more descriptive illustration of the main ideas of the approach. The B2BDC framework, which is the source of the ideas presented here, was developed to handle high-dimensional systems and response surfaces of various mathematical forms [17]; for instance, one of the examples was natural-gas combustion with a training set of 77 experimental observations and 102 active model parameters. The computational methodology of B2BDC is built on constrained optimization methods that generate uncertainty bounds from training data for the quantity of interest [18].

To obtain response surfaces of higher fidelity, the process described can be repeated iteratively by shrinking the sampling domain. Initially, the sampling was done over the entire domain of the parameter space. This may lead to differences in response-surface fits, but does provide an approximate mapping of the overlap region, and it is the overlap region that is the ultimate outcome from the analysis. In the next round of response-surface building, sampling is performed over a smaller space of parameters limited to the area surrounding the overlapping feasible set identified in the first round—results presented in Fig. 1 were calculated at this level. If the quality of this next-round response surfaces is still inadequate, improvements can be made by dividing the domain into parts and creating response surfaces for each, i.e., a piecewise sampling strategy [17,29]. The final results of the present work were obtained using this strategy.

The iterative domain restriction and piecewise building of response surfaces were necessitated by encountered difficulties. For small water cluster ionization, at C_{HF} values below ~0.6, the double-hybrid DFT calculations for cations converged to solutions where the unpaired electron was highly delocalized over O atoms, leading to ~20 kcal/mol shift in VIP and, in turn, to poorly fit response surfaces and to data-model inconsistency. This behavior was resolved by using an initial guess with high spin localization and slight perturbations to the cluster geometries, thereby increasing the range of spin-localized cationic solutions from ~0.6–1 to ~0.5–1 for the C_{HF} parameter. As shown in Fig. 1(c), the overlap region does not extend below $C_{HF} = 0.55$, and so leaving out the lower domain of C_{HF} did not affect the final results. The effect of the geometry perturbation on the neutral energy calculations was negligible, not exceeding 0.07 kcal/mol for all cases.

The average and largest deviations of the response surfaces shown in Fig. 1 were 0.14 and 1.87 kcal/mol, respectively. With the domain divided into two, at $C_{HF} = 0.65$, the respective deviations decreased to 0.03

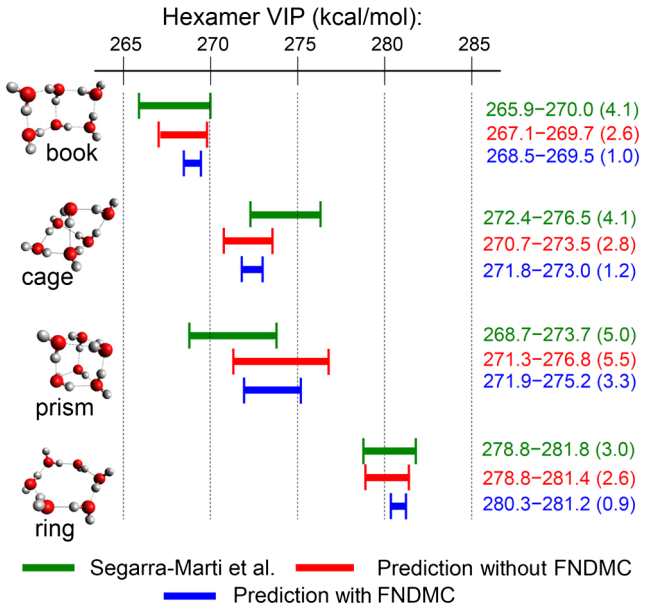

FIG. 2 (color online). VIP intervals of the four hexamer conformers. First, green: intervals reported by Segarra-Marti *et al.* [26]; second, red: predictions of this work without the inclusion of the FNDMC results; third, blue: predictions of this work including the FNDMC calculations. The numerical values are listed on the right, and the values in parentheses are the predicted intervals.

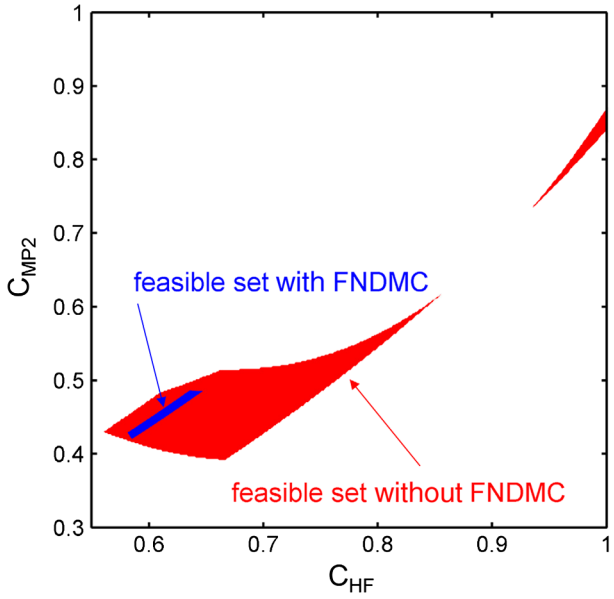

FIG. 3 (color online). Feasible sets obtained before (red, larger area) and after (blue, small area inside red) the inclusion of the dimer FNDMC calculations.

and 0.36 kcal/mol. As discussed above, reducing the domain size is one of the possible strategies: namely, start with a large domain and a coarse fit of response surfaces, narrow to a smaller domain with a better fit, and finally, if needed, split the domain. In this way, high-fidelity response surfaces can be constructed.

Following the above methodology, predictions were computed for four geometries of the water hexamer and are shown in Fig. 2. The top intervals (colored green) for each hexamer geometry represent the VIP results reported by SMR. The next intervals (colored red) are the predictions based on the feasible set calculated with the piecewise domain. The average and maximum response-surface fitting errors for these four hexamer geometries were 0.04 and 0.39 kcal/mol, respectively (the details are included in the Supplemental Material [25]).

Comparison of the predicted with SMR data indicates general consistency among the two sets, thus corroborating the proposed methodology. In all four cases, the two sets of intervals have overlapping regions, which indicates that if the hexamer data were included in the data-model system (to predict, say, a heptamer), the system would be consistent. Further comparison shows that the predicted intervals for the prism and ring geometries are essentially of the same length as the SMR data and, more significantly, the predicted intervals for the book and cage geometries are half the SMR values. The latter indicates that the proposed methodology may offer a possible strategy for improving predictive accuracy.

To test this supposition further, we recomputed the VIP of the water dimer but now with a higher-level quantum chemistry, namely, FNDMC calculations (the details are reported in Supplemental Material [25]). The FNDMC calculations narrowed the SMR dimer VIP values from 270.3–273.7 to 272.8–273.1 kcal/mol. The narrower uncertainty interval for the dimer VIP resulted in the much smaller overlapping feasible set, colored blue in Fig. 3. The smaller feasible set, in turn, yielded increased response-surface accuracy, reducing the average and maximum fit errors to 0.014 and 0.15 kcal/mol, respectively, for the dimer through pentamer, and to 0.026 and 0.25 kcal/mol for the hexamers (the details are included in the Supplemental Material [25]).

The VIP intervals predicted with the new feasible set are shown in Fig. 2 as the third interval for each isomer, colored blue. Inspection of these results indicates that with the smaller feasible set the predictions for the hexamer VIPs are significantly smaller, by about a factor of 2 compared to the SMR data.

We conclude with the following remarks. Parametrized methods are likely to play an increasingly central role in theoretical and computational chemistry. Moving toward reliable model predictions requires rigorous quantification of uncertainties, from placement of realistic bounds on the training data to their effect on model outcomes. The approach presented here offers such capability. Also, the methodology demonstrated suggests a path for model validation via analysis of data-model consistency and data integration via merging individual feasible sets.

The present approach replaces the ill-conditioned and humanly tedious process of model re-optimization [29] with

every new piece of evidence by a methodologically simpler process of rendering feasible sets. The response surfaces underlying the methodology are reusable and readily adaptable to parallel computing. Altogether, these features are amenable to crowdsourcing and collaborative science.

D. E. E., W. A. L., and M. F. were supported by the Director, Office of Energy Research, Office of Basic Energy Sciences, Chemical Sciences, Geosciences and Biosciences Division of the U.S. Department of Energy, under Contract No. DE AC03-76F00098. M. F. was supported by the Air Force Office of Scientific Research, Grant No. FA9550-12-1-0165. This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE AC02 05CH11231.

*frenklach@berkeley.edu

[1] M. J. S. Dewar, E. G. Zoebisch, E. F. Healy, and J. J. P. Stewart, J. Am. Chem. Soc. **107**, 3902 (1985).
[2] J. J. P. Stewart, J. Mol. Model. **19**, 1 (2013).
[3] A. S. Hukkerikar, R. J. Meier, G. Sin, and R. Gani, Fluid Phase Equilib. **348**, 23 (2013).
[4] G. B. Rocha, R. O. Freire, A. M. Simas, and J. J. P. Stewart, J. Comput. Chem. **27**, 1101 (2006).
[5] K. Sastry, D. D. Johnson, A. L. Thompson, D. E. Goldberg, T. J. Martinez, J. Leiding, and J. Owens, Materials and Manufacturing Processes **22**, 553 (2007).
[6] K. Burke, J. Chem. Phys. **136**, 150901 (2012).
[7] L. Goerigk and S. Grimme, Phys. Chem. Chem. Phys. **13**, 6670 (2011).
[8] S. Grindy, B. Meredig, S. Kirklin, J. E. Saal, and C. Wolverton, Phys. Rev. B **87**, 075150 (2013).
[9] M.-C. Kim, E. Sim, and K. Burke, Phys. Rev. Lett. **111**, 073003 (2013).
[10] R. Peverati and D. G. Truhlar, Phil. Trans. R. Soc. A **372**, 20120476 (2014).
[11] J. C. Snyder, M. Rupp, K. Hansen, K.-R. Müller, and K. Burke, Phys. Rev. Lett. **108**, 253002 (2012).
[12] S. F. Sousa, P. A. Fernandes, and M. J. Ramos, J. Phys. Chem. A **111**, 10439 (2007).
[13] M. Steinmetz and S. Grimme, Chemistry Open **2**, 115 (2013).
[14] M. Frenklach, Proc. Combust. Inst. **31**, 125 (2007).
[15] R. Feeley, P. Seiler, A. Packard, and M. Frenklach, J. Phys. Chem. A **108**, 9573 (2004).
[16] M. Frenklach, A. Packard, P. Seiler, and R. Feeley, International Journal of Chemical Kinetics **36**, 57 (2004).
[17] R. Feeley, M. Frenklach, M. Onsum, T. Russi, A. Arkin, and A. Packard, J. Phys. Chem. A **110**, 6803 (2006).
[18] P. Seiler, M. Frenklach, A. Packard, and R. Feeley, Optim. Eng. **7**, 459 (2006).
[19] T. Russi, A. Packard, R. Feeley, and M. Frenklach, J. Phys. Chem. A **112**, 2579 (2008).
[20] T. Russi, A. Packard, and M. Frenklach, Chem. Phys. Lett. **499**, 1 (2010).
[21] R. N. Barnett and U. Landman, J. Phys. Chem. A **101**, 164 (1997).
[22] S. Grimme, J. Chem. Phys. **124**, 034108 (2006).
[23] Y. Zhao, B. J. Lynch, and D. G. Truhlar, J. Phys. Chem. A **108**, 4786 (2004).
[24] R. Krishnan, J. S. Binkley, R. Seeger, and J. A. Pople, J. Chem. Phys. **72**, 650 (1980).
[25] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.112.253003, which includes Refs. [30–43], for the details on response surfaces, the fixed node diffusion Monte Carlo (FNDMC) method, and the comparison of the results with the p function removed from the basis set.
[26] J. Segarra-Marti, M. Merchan, and D. Roca-Sanjuan, J. Chem. Phys. **136**, 244306 (2012).
[27] A. Forrester, A. Sobester, and A. Keane, *Engineering Design via Surrogate Modelling. A Practical Guide* (Wiley, Chippenham, England, 2008).
[28] R. H. Myers, D. C. Montgomery, and C. M. Anderson-Cook, *Response Surface Methodology: Process and Product Optimization Using Designed Experiments* (Wiley, Hoboken, NJ, 2009).
[29] M. Frenklach, A. Packard, and R. Feeley, in *Comprehensive Chemical Kinetics*, edited by R. W. Carr (Elsevier., New York, 2007), p. 243.
[30] J. B. Anderson, Int. Rev. Phys. Chem. **14**, 85 (1995).
[31] A. Aspuru-Guzik, A. C. Kollias, R. Salomon-Ferrer, and W. A. Lester, Jr., in *Handbook of Theoretical and Computational Nanotechnology*, edited by M. Rieth and W. Schommers (American Scientific Publishers, Los Angeles, 2005).
[32] B. M. Austin, D. Yu. Zubarev, and W. A. Lester, Jr., Chem. Rev. **112**, 263 (2012).
[33] D. M. Ceperley and L. Mitas, in *New Methods in Computational Quantum Mechanics*, edited by I. Prigogine and S. A. Rice, Advances in Chemical Physics Vol. XCIII (Wiley, New York, 1996).
[34] C. Filippi and C. J. Umrigar, in *Recent Advances in Quantum Monte Carlo Methods, Part II*, edited by W. A. Lester, Jr., S. M. Rothstein, and S. Tanaka (World Scientific, New Jersey, 2002).
[35] B. L. Hammond, W. A. Lester, Jr., and P. J. Reynolds, *Monte Carlo Methods in Ab Initio Quantum Chemistry* (World Scientific, Singapore, 1994).
[36] C. Filippi and C. J. Umrigar, J. Chem. Phys. **105**, 213 (1996).
[37] M. Burkatzki, C. Filippi, and M. Dolg, J. Chem. Phys. **126**, 234105 (2007).
[38] S. F. Boys and N. C. Handy, Proc. R. Soc. A **310**, 63 (1969).
[39] K. E. Schmidt and J. W. Moskowitz, J. Chem. Phys. **93**, 4172 (1990).
[40] J. Toulouse and C. J. Umrigar, J. Chem. Phys. **126**, 084102 (2007).
[41] C. J. Umrigar, J. Toulouse, C. Filippi, S. Sorella, and R. G. Hennig, Phys. Rev. Lett. **98**, 110201 (2007).
[42] C. J. Umrigar, M. P. Nightingale, and K. J. Runge, J. Chem. Phys. **99**, 2865 (1993).
[43] P. J. Reynolds, D. M. Ceperley, B. J. Alder, and W. A. Lester, Jr., J. Chem. Phys. **77**, 5593 (1982).