Title

Atomic Radius and Charge Parameter Uncertainty in Biomolecular Solvation Energy Calculations

Permalink

https://escholarship.org/uc/item/9qx90765

Journal

Journal of Chemical Theory and Computation, 14(2)

ISSN

1549-9618

Authors

Yang, Xiu
Lei, Huan
Gao, Peiyuan
et al.

Publication Date

2018-02-13

DOI

10.1021/acs.jctc.7b00905

Atomic radius and charge parameter uncertainty in biomolecular solvation energy calculations

Xiu Yang,[†,‖] Huan Lei,[†,‖] Peiyuan Gao,[†,‖] Dennis G. Thomas,[‡] David Mobley,[¶] and Nathan A. Baker[*,†,§]

†*Advanced Computing, Mathematics, and Data Division, Pacific Northwest National Laboratory, Richland, WA 99352, USA*

‡*Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA 99352, USA*

¶*Department of Pharmaceutical Sciences, University of California Irvine, Irvine, CA 92697, USA*

§*Division of Applied Mathematics, Brown University, Providence, RI 02912, USA*

‖*These authors contributed equally.*

E-mail: nathan.baker@pnnl.gov

Phone: +1-509-375-3997

Abstract

Atomic radii and charges are two major parameters used in implicit solvent electrostatics and energy calculations. The optimization problem for charges and radii is under-determined, leading to uncertainty in the values of these parameters and in the results of solvation energy calculations using these parameters. This paper presents a method for quantifying this uncertainty in solvation energies using surrogate models based on generalized polynomial chaos (gPC) expansions. There are relatively few atom types used to specify radii parameters in implicit solvation calculations; therefore, surrogate models for these low-dimensional spaces could be constructed using least-squares fitting. However, there are many more types of atomic charges; therefore, construction of surrogate models for the charge parameter space required compressed sensing combined with an iterative rotation method to enhance problem sparsity. We present results for the uncertainty in small molecule solvation energies based on these approaches. Additionally, we explore the correlation between uncertainties due to radii and charges which motivates the need for future work in uncertainty quantification methods for high-dimensional parameter spaces. The method presented in this paper is a promising approach for efficiently quantifying uncertainty in a wide range of force field parameterization problems, including those beyond continuum solvation calculations.

1 Introduction

Implicit solvent models and their applications have been the subject of numerous previous reviews.[1–3] Such solvation models require the coordinates of the solute atoms as well as atomic charge distributions and a representation of the solute-solvent interface. Charges and interfaces are generally modeled through parameterized empirical representations; however, these parameterizations are often under-determined, leading to uncertainty in the resulting parameter sets.[4–6] The Poisson equation is a popular model for implicit solvent electrostatics and serves as a good example for exploring the influence of this uncertainty on properties such as molecular solvation energy.[1–3] This is a partial differential equation for the electrostatic potential $\varphi : \Omega \mapsto \mathbb{R}$

$$-\nabla \cdot \epsilon(\boldsymbol{x}) \nabla \varphi(\boldsymbol{x}) = \rho(\boldsymbol{x}) \text{ for } \boldsymbol{x} \in \Omega \tag{1}$$

$$\varphi(\boldsymbol{s}) = \varphi_D(\boldsymbol{s}) \text{ for } \boldsymbol{s} \in \partial\Omega, \tag{2}$$

where $\Omega \subset \mathbb{R}^3$ is the problem domain, $\partial\Omega$ is the domain boundary, $\epsilon : \Omega \mapsto [1, \infty)$ is a dielectric coefficient, $\rho : \Omega \mapsto \mathbb{R}$ is the charge distribution, and φ_D is a reference potential function (e.g., Coulomb's law) used for the Dirichlet boundary condition. The dielectric coefficient ϵ is usually defined implicitly[7–10] with respect to the solute atomic radii $\{\sigma_i\}$ and solvent properties such that the coefficient reaches two limiting constant values: ϵ_u inside the solute and ϵ_v away from the solute in bulk solvent. The solvation energy is calculated by

$$\Delta G = \int_\Omega \rho(\boldsymbol{x}) \left(\varphi(\boldsymbol{x}) - \varphi_0(\boldsymbol{x})\right) d\boldsymbol{x}, \tag{3}$$

where φ is the Poisson equation solution for the system with a bulk value of ϵ corresponding to the solvent of interest and φ_0 is the solution for the system with a bulk value of ϵ corresponding to a vacuum. For atomic monopoles, the solute charge distribution has the (numerically unfortunate) form $\rho(\boldsymbol{x}) = \sum_i^{N_A} q_i \delta(\boldsymbol{x} - \boldsymbol{x}_i)$ for N_A solute atoms with positions $\{\boldsymbol{x}_i\}$ and charges q_i. The δ terms are formally defined as Dirac delta functionals but usually approximated by functions with finite support (e.g., when projected onto a grid or finite element basis). The delta functional approximation leads to a simplified form for the solvation energy in Eq. 3,

$$\Delta G = \sum_i^{N_A} q_i \left(\varphi(\boldsymbol{x}_i) - \varphi_0(\boldsymbol{x}_i)\right). \tag{4}$$

Atomic charge models are designed to approximate the "true" vacuum electrostatic potential due to quantum mechanical electron and nuclei charge distributions. While quantum mechanical charge distributions can be incorporated directly in implicit solvent models,[11,12] atomic point charge distributions are generally used.[2] These point charges can include inducible and fixed multipoles[13,14] but monopoles are the most common form. For the purposes of assigning charges, atoms are grouped into sets based on molecular connectivity and environment.[15] The charge values for atoms in these sets are usually determined by numerical fitting to quantum mechanical vacuum electrostatic potentials. Such charge optimization is ill-posed and fitting requires careful choice of the objective function and regularization constraints.[16–19] While sophisticated fitting procedures have been developed, significant information reduction occurs in the transformation of the continuous quantum mechanical electron density into a discrete set of atomic point charges.

Solute-solvent interface models are much more empirical than the charge distribution models; the definition of a solvent "interface" is imprecise at length scales comparable to the size of water molecules. Therefore, such models are generally developed to represent a reasonable description of the solute geometry while also optimizing agreement with experimental quantities such as solvation energy. A large number of solute-solvent interface models exist, including van der Waals,[10] solvent-accessible,[20] solvent-excluded (or Connolly),[21] Gaussian-based,[22] spline-based,[23] and differential geometry surfaces.[9,24–28] All of these interface models represent atoms as spheres and require information about the radii of these spheres. These radii are generally assigned to sets of atoms based on their "type" as determined by the local molecular connectivity. Unlike atomic charges, there are relatively few sets of atom types used to assign radii.[15,29] These radii parameters are determined by optimization of properties such as solvation energy against experimental data.[15,29] Additionally, many of these models also require information about solvent characteristics, generally in the form of a solvent radius, characteristic solvent length scales, or bulk solvent pressure/surface tension properties.

In the present work, we quantify the uncertainty in solvation energy calculated by the Poisson equation and induced by the uncertainty of the input radii and charge parameters. In particular, we construct two surrogate (or statistical regression) models of the solvation energy in terms of the radii and the atomic charges, respectively. These surrogate models enable us to estimate the solvation energy with different input parameters quickly and to evaluate the statistical information of the target properties (e.g., probability density function) efficiently. We model the input parameters as independent (i.i.d.) Gaussian random variables with different means and standard deviations. To construct the surrogate of the Poisson model, we use a generalized polynomial chaos (gPC)[30,31] expansion to represent the dependence of the solvation energy on uncertain parameters such as the atomic charge and radii. The efficacy of the gPC method for elliptic problems such as the Poisson equation has been extensively studied with robust results for its efficiency and accuracy.[32,33] This approach is straightforward to apply to the relatively low-dimensional parameter sets. However, the main challenge of applying this method to implicit solvent calculation parameter uncertainty is the high-dimensionality of parameter sets (especially the atomic charges): the surrogate models require more basis functions and, therefore, more expansion coefficients need to be identified. To address this challenge, we adopt a compressive sensing method combined with the rotation-based sparsity-enhancing method first proposed by Lei et al.[34] and extended by Yang et al.,[35] which enable us to construct the surrogate with relatively few sample outputs of the numerical Poisson solver.

2 Methods

We demonstrated the framework using a test set of 17 compounds from the SAMPL computational challenge for solvation energy prediction[15] (see Table 1).

Table 1: List of 17 compounds from the SAMPL computational challenge for solvation energy prediction.

ind.	compound	ind.	compound
1	glycerol triacetate	10	$1,4$-dioxane
2	benzyl bromide	11	diethyl propanedioate
3	benzyl chloride	12	dimethoxymethane
4	m-bis(trifluoromethyl)benzene	13	ethylene glycol diacetate
5	N,N-dimethyl-p-methoxybenzamide	14	$1,2$-diethoxyethane
6	$N,N-4$-trimethylbenzamide	15	diethyl sulfide
7	bis-2-chloroethyl ether	16	phenyl formate
8	$1,1$-diacetoxyethane	17	imidazole
9	1,1-diethoxyethane		

2.1 Uncertain parameters

Many parameterization approaches for atomic charge use ESP (electrostatic potential)[36] or related methods (e.g., RESP[18]). These methods optimize atomic charges by least-squares fitting of the charges' Coulombic potential to the electrostatic potential obtained from quantum mechanical calculations. This under-determined optimization is performed subject to various constraints, including the requirement that the atomic charges sum to the integer formal charge of the molecule. More specifically, the calculated ESP $\hat{V}_i$ at the i-th grid point is the electrostatic potential given by Coulomb's law summed over the charge q_j at the centers of the j-th atoms. Least-squares fitting is performed by minimizing $\sum_i (V_i - \hat{V}_i)^2$ with constraints, where V_i is the electrostatic potential computed by *ab initio* calculations. Least-squares fitting implies a Gaussian noise model wherein the atomic charges q_j can be modeled as Gaussian random variables.

In the present work, we modeled the uncertainty in atomic charges by considering atomic charges obtained by 11 different approaches: AM1BCC,[37] CHELP,[38] CHELPG,[39] CM2,[40] ESPMK,[36] Gasteiger,[41] PCMESP,[42] QEQ,[43] RESP,[18] MMFF94,[44] Mulliken.[45] The Hartree-Fock method and the 6-31G*basis set were used to optimize molecular geometries. The methods we selected here are popular ones from different approaches that have been proposed for the derivation of atomic charges. For instance the General Amber force field use atomic charges which are fitted by RESP, the GLYCAM force field employs the CHELPG method, and CHARMM force field developers used both approaches. We note that some of these resutls are known to be sensitive to details in the molecular modeling, which increases the uncertainties in the input (atomic charge). We demonstrate that even with this less favorite scenario, our method is still able to investigate the uncertainty in the solvation energy with relatively few Monte Carlo samples.

We *assumed* that the variation of atomic charges across different methods can be modeled by a Gaussian random field with covariance kernel

$$\mathrm{Cov}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \eta_i \eta_j \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^p}{\theta}\right), \tag{5}$$

where η_i is the standard deviation of the i-th atomic charge, $\boldsymbol{x}_i$ is the position of the i-th atom, and $0 < p < 2$. We used atomic charges from 11 different methods to estimate η_i and then used the maximum likelihood estimate (MLE) method to estimate θ and p. Since the sum of N_A charges in a molecule is constrained (to its formal molecular charge $Q \in \mathbb{Z}$), we modeled the Gaussian random field with $N_A - 1$ atoms by removing the last hydrogen in the PDB file. Additionally, we use symmetry in the molecular structure to reduce the number of independent atomic charge types before applying the MLE to identify the random field. For example, in a benzene, there is only one type of carbon and one type of hydrogen due to the symmetry of this molecule. Therefore, we considered the charges of its atoms as a Gaussian random field with only two entries instead of 12 ones (the total number of atoms in benzene).

After obtaining the covariance matrix by integrating across methods, we represented the atomic charge as

$$\boldsymbol{q} = \langle \boldsymbol{q} \rangle + \boldsymbol{L}_c \boldsymbol{\gamma}, \tag{6}$$

where $\boldsymbol{q} = (q_1, q_2, \cdots, q_{N_A - 1})$ are the atomic charges, $\langle \boldsymbol{q} \rangle$ is the mean of $\boldsymbol{q}$ estimated from the 11 different charge values, $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \cdots, \gamma_{N_A-1})$ are i.i.d. zero-mean unit-variance Gaussian random variables, and $\boldsymbol{L}_c$ is a lower triangular matrix from the Cholesky decomposition of the covariance matrix (Eq. 5). We note that for the atoms in the test set used in the present work, the covariance matrices of these random field are almost diagonal: the off-diagonal entries are smaller than 10^{-12}. This suggests the correlation between atomic charges is effective removed during their symmetry-based grouping. The atomic charge for the remaining atom is obtained by summation of the other random charge variables based on the constraint $q_i = Q - \sum_{j \neq i}^{N_A} q_j$.

Similarly, we used multiple force fields (ZAP-9,[15] OPLSAA,[46] Bondi[47] and PARSE[29]) to model uncertainty in the radii parameters in the same manner. Although radii are non-negative, we did not explicitly impose constraints on the radii. After obtaining the covariance matrix, we represented the radii as

$$\boldsymbol{\sigma} = \langle \boldsymbol{\sigma} \rangle + \boldsymbol{L}_r \boldsymbol{\zeta}, \tag{7}$$

where $\boldsymbol{\sigma} = (\sigma_1, \cdots, \sigma_{N_A})$, σ_i is the radius of atom (type) i, $\boldsymbol{\zeta} = (\zeta_1, \cdots, \zeta_{N_A})$ are independent zero-mean unit-variance Gaussian random variable and $\boldsymbol{L}_r$ is a lower triangular matrix from the Cholesky decomposition of the covariance matrix. We note that the standard deviations here are smaller than 10% of the mean values which implies very low probabilities for unphysical negative radii values. Therefore, by employing truncated Gaussian random variables within 4 standard deviations (capturing more than 99.99% of the probability), we guaranteed that the radii are always positive and that the distributions of the truncated Gaussian variables were almost identical to the original Gaussian variates.

Although we use γ and ζ to denote the random variables used for modeling the uncertainties in q_j and σ_j, in what follows, we still use $\boldsymbol{\xi} = (\xi_1, \xi_2, \cdots)$ to denote general uncertain inputs when introducing the algorithm and reporting results.

2.2 Solvation energy surrogate models

We used generalized polynomial chaos (gPC) expansions as surrogate models for the solvation energy. The goal of surrogate construction is to estimate the variations in quantities of interest, such as solvation energy, much more efficiently than solving the original problem, such as solving the Poisson equation. The details for these expansions are provided in Supporting Information.

2.3 Poisson equation solver

We used the Adaptive Poisson-Boltzmann Solver (APBS)[48] to solve the Poisson equation for solvation energies.

3 Results and discussion

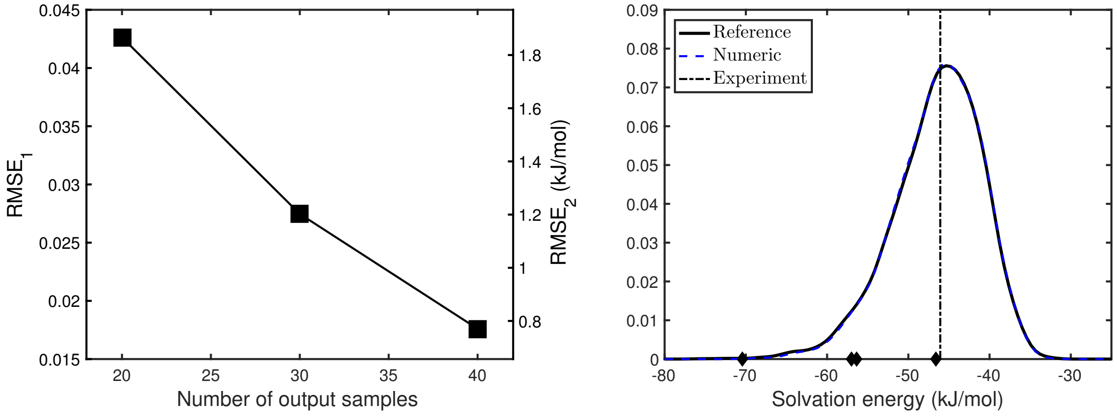
For each test case, we used Monte Carlo simulations to generate $10,000$ samples of the input parameters $\boldsymbol{\xi}^q$ and then solved PB equation using APBS to obtain output samples of the solvation energy $E^q = E(\boldsymbol{\xi}^q)$. We used these outputs as "ground truth" reference solutions to examine the performance of the surrogate models. More precisely, given a surrogate model $\tilde{E}$, we use two different root-mean-squared error (RMSE) measures to examine its accuracy:

$$RMSE_1 = \sqrt{\frac{\sum_{q=1}^{10000}\left(\tilde{E}(\boldsymbol{\xi}^q) - E^q\right)^2}{\sum_{q=1}^{10000}(E^q)^2}}, \quad RMSE_2 = \sqrt{\frac{\sum_{q=1}^{10000}\left(\tilde{E}(\boldsymbol{\xi}^q) - E^q\right)^2}{10000}}. \qquad (8)$$

We also use box-whisker plots to demonstrate the statistics. The line in the middle is the median of 16 molecules, the tops and bottoms of the boxes are 25th and 75th percentiles, and the whisker plots cover more than 99% probability.

3.1 Influence of radii uncertainties on solvation energies

We investigated the effect of the uncertainties in the radii with fixed atomic charges obtained from AM1-BCC.[37] As an example, there are eight different sets of radii for N,N-dimethyl-p-methoxybenzamide across the ZAP-9, Bondi, OPLSAA, and PARSE parameter sets, as shown in the support material. We modeled the solvation as a function of eight i.i.d. Gaussian random variables. We constructed gPC surrogate models with multi-variate normalized Hermite polynomials up to third order. The surrogate model consisted of $C_{8+4}^4 = 495$ basis functions. Figure 1 (a) presents the RMSE obtained by our method with respect to different numbers of samples E^q. Figure 1 (b) compares the solvation energy probability distribution function (PDF) obtained by our method and the reference solutions. The numerical results are obtained by constructing the surrogate model with the 36 output samples first, then sampling the surrogate model $10,000$ times with random samples to estimate the PDF. The reference solution is computed from the $10,000$ outputs of E^q.



(a) (b)

Figure 1: Performance of the surrogate model for radii uncertainties for N,N-dimethyl-p-methoxybenzamide. (a): RMSE with different numbers of samples M. (b): comparison of the solvation energy PDFs estimated by the numerical surrogate method ("Numeric") based on 40 output samples of APBS; dash line ("Experiment") is the experimental result; diamonds are the results by using radii from ZAP-9, Bondi, OPLSAA and PARSE, respectively. The diamond closest to the experiment was obtained from ZAP-9.

We performed the same analysis for all the molecules in the test set and present the results in Figure 2. For most molecules, we can build an accurate surrogate model (RMSE< 0.05) for the solvation energy with only a few samples (less than 40) of the input parameters. However, m-bis-trifluoromethylbenzene (TFMB) required significantly more samples. In particular, the RMSE for the TFMB solvation energy surrogate model was close to 0.15 with 40 samples and required 100 samples to reduce the RMSE to less than 5%. This variability arises from the radius of fluorine: in the ZAP force field it is 2.4 Å; however, it is only ~ 1.4 Å for the other force fields. Hence, the standard deviation of this radius is around 25% of the mean and fluorine requires more terms in the surrogate model for an accurate description and therefore more samples to parameterize those terms. The influences of the uncertainties in the input radii on the solvation energy for each molecule are demonstrated in box-whisker plots in Figure 3. The experiment results are presented for comparison. We note that some experiments results are "outliers" of the box-whisker plots, this is because that the atomic charges are computed from AM1BCC for the purpose of fixing the atomic

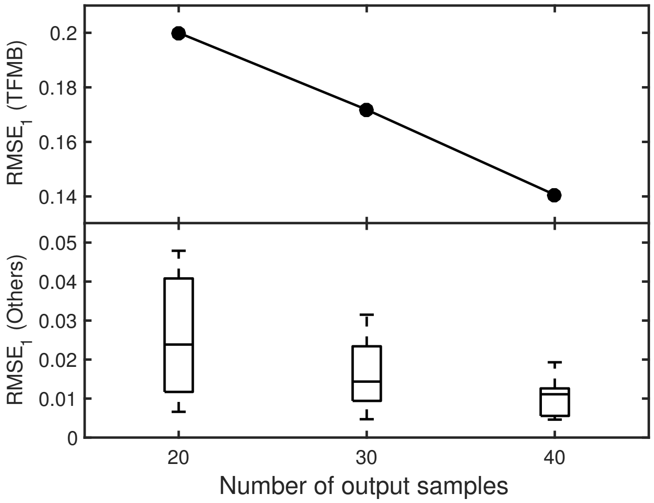


Figure 2: Performance of surrogate models with respect to number of samples. Circles are the $RMSE_1$ of m-bis-trifluoromethylbenzene (TFMB), box-whisker plots are the $RMSE_1$ of the remaining 16 molecules.

charges and it does not guarantee that the computed solvation energy is sufficiently close to the experiment results. For example, for the m-bis(trifluoromethyl)benzene AM1BCC charges yield negative solvation energy while the experiment result is positive.

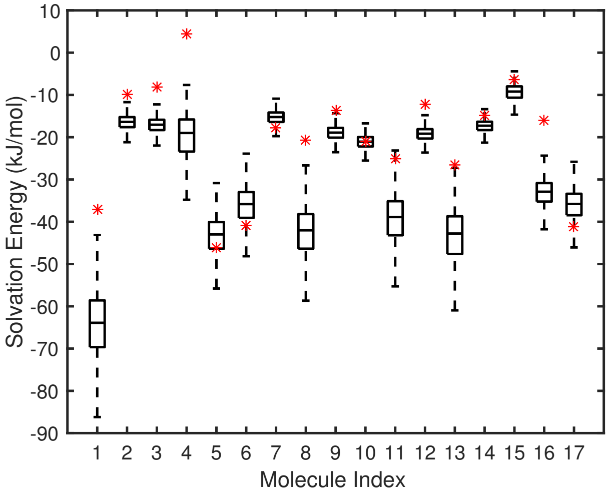


Figure 3: Influence of radii uncertainties on molecular solvation energies for the 17-molecule test set. The red stars are the experiment results.

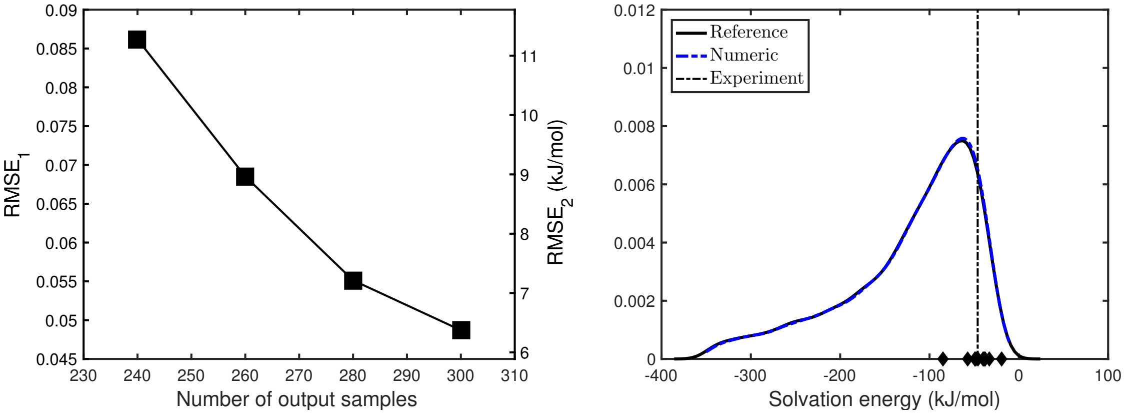
3.2 Influence of atomic charge uncertainties on solvation energies

We also examined the influence of charge perturbation for solvation energy calculations with fixed radii (ZAP-9). As an example, there are 14 different types of atoms in N,N-dimethyl-p-methoxybenzamide as shown in Supporting Material. We note that we model the surrogate with 13 inputs due to the constraint on the summation of the charges. The mean and standard deviation are computed from the results of 11 different charge fitting approaches. We used no more than 3000 multi-variate normalized Hermite polynomials (up to fourth order) in the gPC surrogate model for E_g for all the molecules. We use N,N-dimethyl-p-methoxybenzamide as an example. Figure 4 (a) presents the RMSE obtained by our method with respect to different numbers of samples E^q. It illustrates that 300 output samples are needed to reduce the RMSE to less than 5%. Figure 4 (b) compares the PDF obtained by our method and the reference solution. The numerical results are obtained by constructing the surrogate model with the 300 output samples first, then sampling the surrogate model $10,000$ times with random samples to estimate the PDF. The reference solution is computed from the $10,000$ outputs of E^q.

The influences of the uncertainties in the input atomic charges on the solvation energy for each molecule are demontrated in Figure 5. For most molecules, the experiment results lie in the whisker plots and some of them are in the box. We also present the number of output samples needed to construct a surrogate with RMSE less than 5% with respect to the number of atom types in Figure 6.

3.3 Combined influence of radius and atomic charge uncertainties

Comparing the PDFs in Figures 1 (b) and 4 (b), we notice that the uncertainty in the solvation energy induced by the atomic charges is stronger than that induced by the radii. The atomic charges vary significantly across different methods while the variation in the radii is much smaller. To understand the combined influence of charges and radii on solvation energies, we modeled the correlated uncertainties for these two types of parameters can be



(a)

(b)

Figure 4: Performance of surrogate models for charge uncertainties for N,N-dimethyl-p-methoxybenzamide. (a): RMSE for surrogate model with different number of output samples. (b): comparison of the PDFs estimated by the numerical surrogate method ("Numeric") based on 300 output samples of APBS; dash line ("Experiment") is the the result by the experiment; diamonds are results by using atomic charges from AM1BCC, CHELP, CHELPg, CM2, ESPMK, Gasteiger, PCMESP, QEQ, RESP, MMFF94, Mulliken.

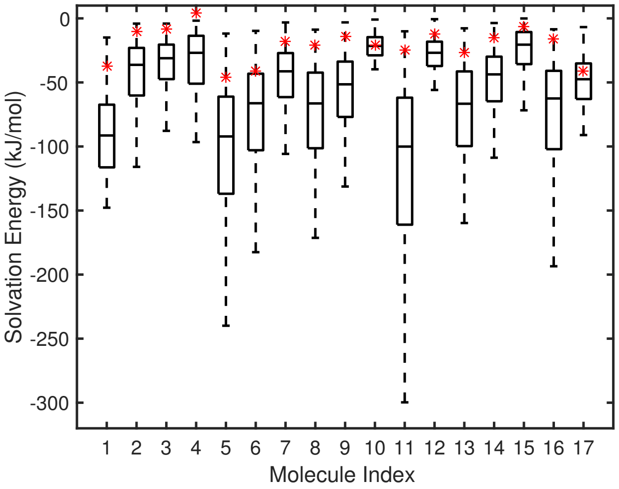


Figure 5: Results of atmoic charge uncertainties. Box-whisker plots demonstraing the uncertainties in the numerical results of the solvation energy for 17 compounds. The red stars are the experiment results.

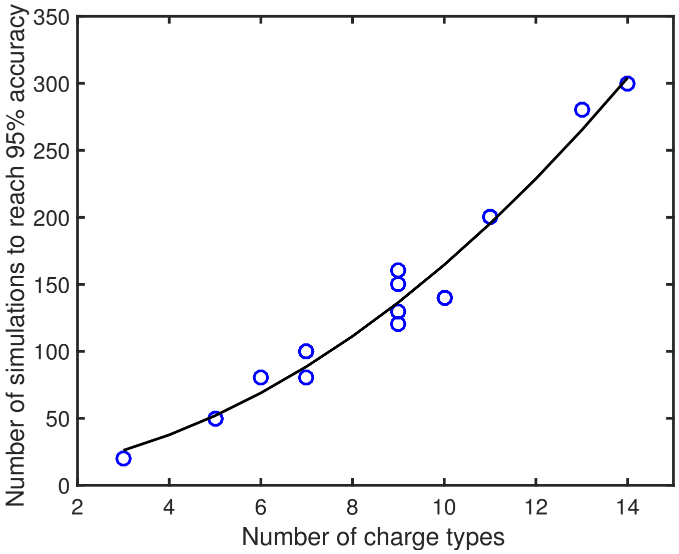
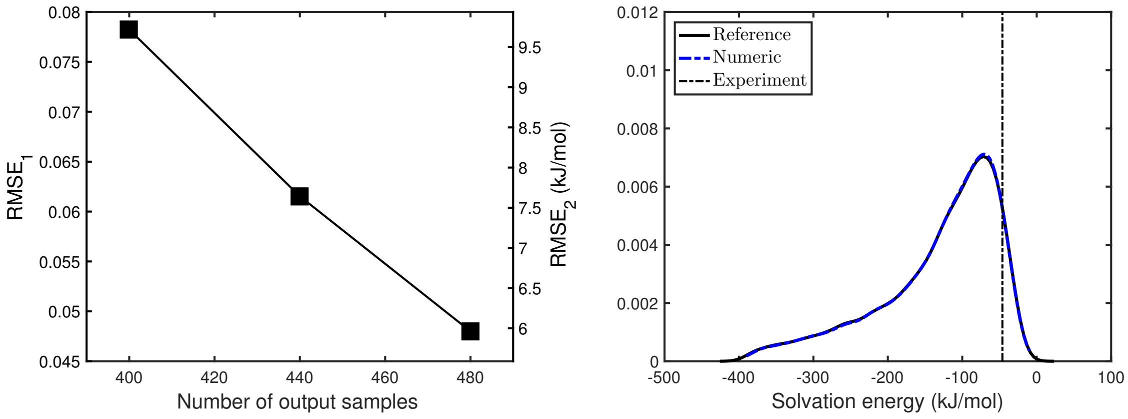


Figure 6: "∘" : number of output samples need to construct a surrogate model with RMSE less than 5% with respect to the number of atom types; "-" is the best-fit curve $1.4x^2 + 1.9x + 7.9$.

modeled with i.i.d. Gaussian random variables. We use N, N-dimethyl-p-methoxybenzamide as an example. 480 output samples are needed to reduce the RMSE to less than 5%. Figure 7 (a) presents the RMSE obtained by our method with respect to different numbers of samples E^q. Figure 7 (b) compares the PDF obtained by our method and the reference solution. The numerical results are obtained by constructing the surrogate model from the 480 output samples and then sampling the surrogate model $10,000$ times with random samples to estimate the PDF. The reference solution is computed from the $10,000$ outputs of E^q. Not surprisingly, the number of output samples needed to construct an accurate surrogate increases as we take into account both uncertainties in the charges and radii. The shape of the solvation energy changes PDF also slightly as the radii variation of the radii across different methods are much smaller than charge variations.

The influences of the uncertainties in the input atomic charges on the solvation energy for each molecule are demontrated in Figure 8. This figure is similar to Figure 5 since the uncertainties in the atomic charges dominate the results. Figure 9 shows the number of output samples needed to construct a surrogate with less than 5% RMSE for all 17 molecules



(a) (b)

Figure 7: Results of radii and charges uncertainties for N, N-dimethyl-p-methoxybenzamide. (a): RMSE with different number of output samples M. (b): comparison of the PDFs estimated by the numerical method ("Numeric") based on 480 output samples of APBS; dashed line ("Experiment") is the experimental result.

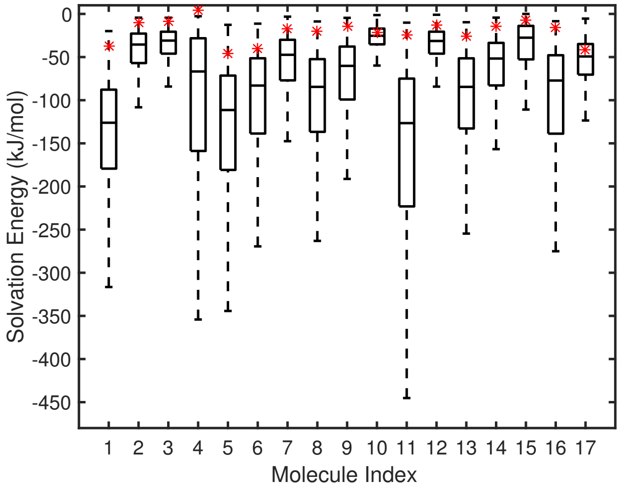


Figure 8: Results of radius and atomic charge uncertainties. Box-whisker plots demonstraing the uncertainties in the numerical results of the solvation energy for 17 compounds. Red stars are the experiment results.

in the test set. This figure illustrates the approximately quadratic scaling with the respect

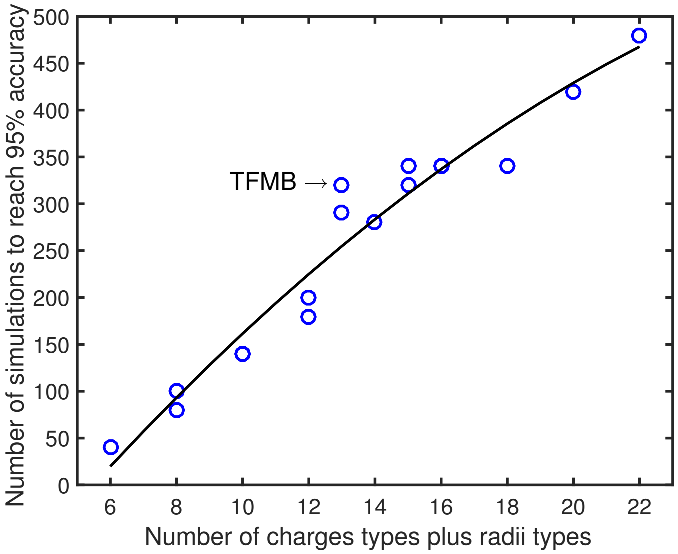


Figure 9: "○" : number of output samples need to construct a surrogate model with RMSE less than 5% with respect to the number of atom charge types plus radius types; "-" fitting curve $-0.6x^2 + 45x - 188$.

to the number of atom types in the molecule.

4 Conclusions

We used a newly developed extension of compressive sensing method to construct surrogate models of solvation energy based on gPC expansions. These surrogate models allow us to efficiently and accurately estimate the variation in solvation energy due to uncertainty in parameter input. Our results demonstrate that for the data sets used in the present work, the variation of radii across different approaches are small. On the other hand, the variations of the atomic charges obtained by different methods are much larger. Therefore, the number of output samples needed for accurate UQ analysis requires are much larger, growing quadratically with respect to the number of atom types. In addition, modeling the radii uncertainty and the atomic charges uncertainty are different in that the former is transferable while the latter is not. More specifically, the radii of atoms are identified disregard of which molecule or residual they belong to in practice, hence, the uncertainty of the radius of a specific atom can be applied to different molecules. However, the atomic charges are computed by fitting the *ab initio* results, which does not guarantee that the same atom in different molecules has the same uncertainty in the charges. This framework can be applied to estimate the statistics (e.g., mean, variance), PDF, confidence interval, Chernoff-like bounds,[49] etc. of solvation computing and other chemical computing when the inputs are uncertain. The current study focused on uncertainty in solute charges and radii; however, this framework could also be applied to other solvation model characteristics such as dielectric coefficient, solvent radius, and biomolecular surface definition.

In the future, we anticipate that this approach could be used for a much wider range of force field parameterization activities, including both coarse-grained and atomistic representations of biomolecules. Uncertainty quantification methods have begun to be used in force field parameterization of simple alkane systems;[50] this paper demonstrates the ability to extend the methods to higher-dimensional systems with more diversity of atom types. Application of these methods offer the benefit of efficiently characterizing parameter space and understanding the impact of parameter variation on quantities of interest. Additionally, the iterative method we used in the present work is very suitable for this type of problem, as the accuracy of the surrogate models are improved significant after iterations. Especially, the error of the surrogate models for the atomic charge induced uncertainties are reduced by 40% $\sim$ 50% compared with the standard compressive sensing method. Also, there is significant room for development in the numerical methods. For example, the sparsity-enhancing approaches can be combined with other techniques including improved sampling strategies,[51,52] adaptive basis selection,[53,54] and advanced optimization methods.[55,56] These approaches improve the accuracy of the compressive sensing method from different aspects. As such, they will help to reduce the number of expensive simulations or quantum mechanics calculations needed for constructing accurate surrogates.

Acknowledgments

This work was supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific-Computing Research as part of the Collaboratory on Mathematics for Mesoscopic Modeling of Materials (CM4) and by NIH grant GM069702. Pacific Northwest National Laboratory is operated by Battelle for the DOE under Contract DE-AC05-76RL01830.

References

(1) Lamm, G. *Reviews in Computational Chemistry*; John Wiley & Sons, Inc., 2003; pp 147–365.

(2) Ren, P. Y.; Chun, J. H.; Thomas, D. G.; Schnieders, M. J.; Marucho, M.; Zhang, J. J.; Baker, N. A. *Quarterly Reviews of Biophysics* **2012**, *45*, 427–491.

(3) Grochowski, P.; Trylska, J. *Biopolymers* **2008**, *89*, 93–113.

(4) Ponder, J. W.; Case, D. A. *Advances in Protein Chemistry* **2003**, *66*, 27–85.

(5) Gosink, L. J.; Overall, C. C.; Reehl, S. M.; Whitney, P. D.; Mobley, D. L.; Baker, N. A. *The Journal of Physical Chemistry B* **2016**, doi:10.1021/acs.jpcb.6b09198.

(6) Swanson, J. M. J.; Adcock, S. A.; McCammon, J. A. *Journal of Chemical Theory and Computation* **2005**, *1*, 484–493.

(7) Swanson, J. M. J.; Mongan, J.; McCammon, J. A. *The Journal of Physical Chemistry B* **2005**, *109*, 14769–14772.

(8) Swanson, J. M. J.; Wagoner, J. A.; Baker, N. A.; McCammon, J. A. *Journal of Chemical Theory and Computation* **2007**, *3*, 170–183.

(9) Bates, P. W.; Wei, G. W.; Zhao, S. *Journal of Computational Chemistry* **2008**, *29*, 380–391.

(10) Dong, F.; Zhou, H.-X. *Proteins* **2006**, *65*, 87–102.

(11) Eckert, F.; Diedenhofen, M.; Klamt, A. *Molecular Physics* **2010**, *108*, 229–241.

(12) Tomasi, J.; Mennucci, B.; Cammi, R. *Chemical Reviews* **2005**, *105*, 2999–3094.

(13) Schnieders, M. J.; Ponder, J. W. *Journal of Chemical Theory and Computation* **2007**, *3*, 2083–2097.

(14) Schnieders, M. J.; Baker, N. A.; Ren, P.; Ponder, J. W. *J Chem Phys* **2007**, *126*, 124114.

(15) Nicholls, A.; Mobley, D. L.; Guthrie, J. P.; Chodera, J. D.; Bayly, C. I.; Cooper, M. D.; Pande, V. S. *Journal of Medicinal Chemistry* **2008**, *51*, 769–779.

(16) Besler, B. H.; Merz, K. M.; Kollman, P. A. *Journal of Computational Chemistry* **1990**, *11*, 431–439.

(17) Haschka, T.; Hénon, E.; Jaillet, C.; Martiny, L.; Etchebest, C.; Dauchez, M. *Computational and Theoretical Chemistry* **2015**, *1074*, 50–57.

(18) Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A. *The Journal of Physical Chemistry* **1993**, *97*, 10269–10280.

(19) Bader, R. F. W. *Chemical Reviews* **1991**, *91*, 893–928.

(20) Lee, B.; Richards, F. M. *Journal of Molecular Biology* **1971**, *55*, 379–400.

(21) Connolly, M. L. *Journal of Applied Crystallography* **1983**, *16*, 548–558.

(22) Grant, J. A.; Pickup, B. T.; Nicholls, A. *Journal of Computational Chemistry* **2001**, *22*, 608–640.

(23) Im, W.; Beglov, D.; Roux, B. *Computer Physics Communications* **1998**, *111*, 59–75.

(24) Bates, P. W.; Chen, Z.; Sun, Y.; Wei, G.-W.; Zhao, S. *Journal of Mathematical Biology* **2009**, *59*, 193–231.

(25) Chen, Z.; Baker, N. A.; Wei, G. W. *J Comput Phys* **2010**, *229*, 8231–8258.

(26) Cheng, L.-T.; Dzubiella, J.; McCammon, J. A.; Li, B. *The Journal of Chemical Physics* **2007**, *127*, 084503–084503.

(27) Dzubiella, J.; Swanson, J. M. J.; McCammon, J. A. **2006**, *96*.

(28) Dzubiella, J.; Swanson, J. M.; McCammon, J. A. *The Journal of Chemical Physics* **2006**, *124*, 84905–84905.

(29) Sitkoff, D.; Sharp, K. A.; Honig, B. *The Journal of Physical Chemistry* **1994**, *98*, 1978–1988.

(30) Ghanem, R. G.; Spanos, P. D. *Stochastic finite elements: a spectral approach*; Springer-Verlag: New York, 1991.

(31) Xiu, D.; Karniadakis, G. E. *SIAM J. Sci. Comput.* **2002**, *24*, 619–644.

(32) Todor, R. A.; Schwab, C. *IMA J. Numer. Anal.* **2007**, *27*, 232–261.

(33) Babuška, I.; Nobile, F.; Tempone, R. *SIAM Rev.* **2010**, *52*, 317–355.

(34) Lei, H.; Yang, X.; Zheng, B.; Lin, G.; Baker, N. A. *SIAM Multiscale Model. Simul.* **2015**, *13*, 1327–1353.

(35) Yang, X.; Lei, H.; Baker, N. A.; Lin, G. *Journal of Computational Physics* **2016**, *307*, 94–109.

(36) Singh, U. C.; Kollman, P. A. *Journal of Computational Chemistry* **1984**, *5*, 129–145.

(37) Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. *Journal of Computational Chemistry* **2000**, *21*, 132–146.

(38) Chirlian, L. E.; Francl, M. M. *Journal of Computational Chemistry* **1987**, *8*, 894–905.

(39) Breneman, C. M.; Wiberg, K. B. *Journal of Computational Chemistry* **1990**, *11*, 361–373.

(40) Li, J.; Zhu, T.; Cramer, C. J.; Truhlar, D. G. *The Journal of Physical Chemistry A* **1998**, *102*, 1820–1831.

(41) Gasteiger, J.; Marsili, M. *Tetrahedron* **1980**, *36*, 3219–3228.

(42) Cammi, R.; Tomasi, J. *Journal of Computational Chemistry* **1995**, *16*, 1449–1458.

(43) Rappe, A. K.; Goddard III, W. A. *The Journal of Physical Chemistry* **1991**, *95*, 3358–3363.

(44) Halgren, T. A. *Journal of Computational Chemistry* **1996**, *17*, 490–519.

(45) Mulliken, R. S. *The Journal of Chemical Physics* **1955**, *23*, 1833–1840.

(46) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.

(47) Bondi, A. *The Journal of Physical Chemistry* **1964**, *68*, 441–451.

(48) Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; McCammon, J. A. *Proceedings of the National Academy of Sciences* **2001**, *98*, 10037–10041.

(49) Rasheed, M.; Clement, N.; Bhowmick, A.; Bajaj, C. Statistical Framework for Uncertainty Quantification in Computational Molecular Modeling. Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. 2016; pp 146–155.

(50) Messerly, R. A.; Knotts, T. A.; Wilding, W. V. *Journal of Chemical Physics* **2017**, *146*, 194110.

(51) Rauhut, H.; Ward, R. *J. Approx. Theory* **2012**, *164*, 517–533.

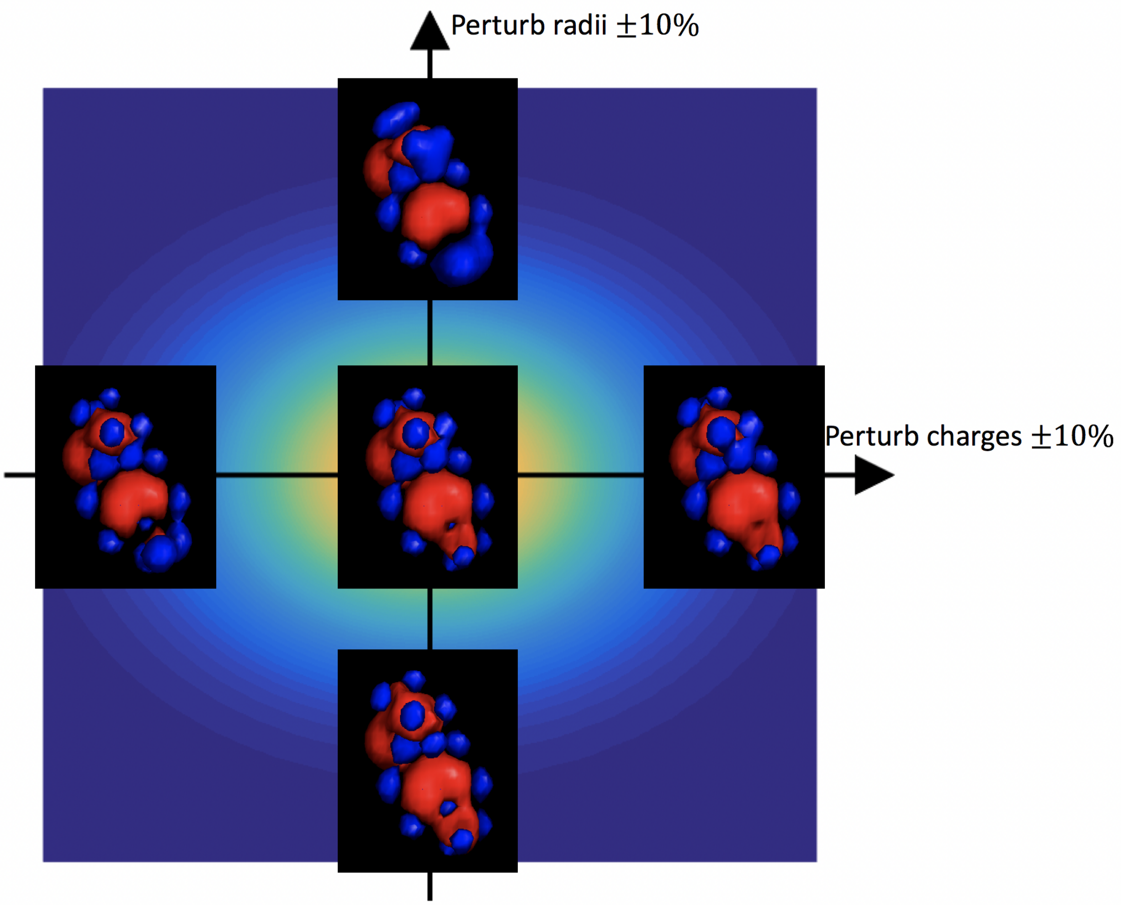
(52) Peng, J.; Hampton, J.; Doostan, A. *J. Comput. Phys.* **2014**, *267*, 92 – 111.

(53) Yang, X.; Choi, M.; Lin, G.; Karniadakis, G. E. *J. Comput. Phys.* **2012**, *231*, 1587–1614.

(54) Jakeman, J. D.; Eldred, M. S.; Sargsyan, K. *J. Comput. Phys.* **2015**, *289*, 18–34.

(55) Candès, E. J.; Wakin, M. B.; Boyd, S. P. *J. Fourier Anal. Appl.* **2008**, *14*, 877–905.

(56) Yang, X.; Karniadakis, G. E. *J. Comput. Phys.* **2013**, *248*, 87–108.



Supporting Information Available

The following files are available free of charge. An appendix PDF provides detailed information on the mathematical numerical methods used in this study.

Graphical TOC Entry

See below.

Supporting Information: Atomic radius and charge parameter uncertainty in biomolecular solvation energy calculations

Xiu Yang,[†,‖] Huan Lei,[†,‖] Peiyuan Gao,[†,‖] Dennis G. Thomas,[‡] David Mobley,[¶] and Nathan A. Baker[*,†,§]

†*Advanced Computing, Mathematics, and Data Division, Pacific Northwest National Laboratory, Richland, WA 99352, USA*

‡*Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA 99352, USA*

¶*Department of Pharmaceutical Sciences, University of California Irvine, Irvine, CA 92697, USA*

§*Division of Applied Mathematics, Brown University, Providence, RI 02912, USA*

‖*These authors contributed equally.*

E-mail: nathan.baker@pnnl.gov

Phone: +1-509-375-3997

A Pairwise correlation between different fitting methods

The pairwise correlation of the atomic charges obtained by different fitting methods was modeled by the covariance matrix $\mathbf{Cov}^{\{m_a,m_b\}}$, described in the main text, where m_a and m_b denote two different fitting methods. We used MLE to identify the parameters p and θ

based on charges from two different methods. The values of p and θ are presented in Tables S1 and S2, respectively.

Table S1: Estimate of p from MLE. (1-AM1BCC, 2-CHELP, 3-CHELPG, 4-CM2, 5-ESPMK, 6-ANTECHAMBER, 7-PCMESP, 8-QEQ, 9-RESP, 10-MMFF94, 11-Mulliken

	1	2	3	4	5	6	7	8	9	10	11
1	-	1.398	1.429	1.431	1.430	1.430	1.460	1.362	1.435	1.419	1.430
2		-	1.398	1.398	1.449	1.402	1.450	1.400	1.397	1.398	1.398
3			-	1.429	1.351	1.371	1.450	1.430	1.343	1.444	1.314
4				-	1.463	1.495	1.449	1.445	1.421	1.371	1.439
5					-	1.390	1.474	1.430	1.397	1.441	1.424
6						-	1.995	1.430	1.393	1.400	1.396
7							-	1.431	1.426	1.444	1.427
8								-	1.447	1.476	1.430
9									-	1.393	1.397
10										-	1.395
11											-

Table S2: Estimate of θ from MLE. (1-am1bcc, 2-chelp, 3-chelpg, 4-cm2, 5-espmk, 6-antechamber, 7-pcmesp, 8-qeq, 9-resp, 10-mmff94, 11-mulliken)

	1	2	3	4	5	6	7	8	9	10	11
1	-	0.048	0.043	0.044	0.043	0.043	0.057	0.060	0.058	0.055	0.043
2		-	0.048	0.048	0.052	0.059	0.052	0.055	0.063	0.048	0.048
3			-	0.043	0.066	0.060	0.052	0.043	0.059	0.056	0.067
4				-	0.064	0.061	0.059	0.058	0.056	0.065	0.066
5					-	0.058	0.063	0.043	0.064	0.058	0.060
6						-	0.283	0.043	0.065	0.065	0.065
7							-	0.042	0.060	0.056	0.064
8								-	0.059	0.059	0.043
9									-	0.061	0.064
10										-	0.060
11											-

The entries of correlation matrix $\mathbf{C}^{m_a,m_b}$ are computed as $C_{ij}^{m_a,m_b} = \mathrm{Cov}_{ij}^{m_a,m_b}/\eta_i\eta_j$, where η_i is the standard deviation of the i-th (type) atomic charge. For each matrix $\mathbf{C}^{\{m_a,m_b\}}$, we define the correlation magnitude $M^{\{m_a,m_b\}}$ by

$$M^{\{m_a,m_b\}} = \|\mathbf{C}^{\{m_a,m_b\}}\|_F/N_A,$$

where $\| \cdot \|_F$ represents the Frobenius norm and N_A is the total number of atom types. The correlation magnitude between different fitting methods is presented in Figure S1.

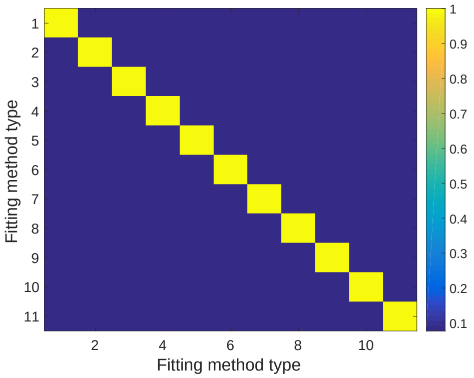


Figure S1: Correlation magnitude of the atom charge distribution for N,N-dimenthyl-p-methoxybenzamide between different fitting methods: am1bcc, chelp, chelpg, cm2, espmk, antechamber, pcmesp, qeq, resp, mmff94, mulliken.

B Setup of the input uncertainties

Example uncertainties for N,N-dimenthyl-p-methoxybenzamide charges and radii are given in Tables S4 and S3, respectively.

Table S3: Uncertainties in the radii of N,N-dimenthyl-p-methoxybenzamide.

atoms (index)	mean	std.
C ({1,2,3,4,5,6})	1.7613	0.0807
C ({8})	1.7863	0.0996
C ({9,12,13})	1.7550	0.0802
O ({7})	1.4725	0.0585
O ({10})	1.5400	0.1549
N ({11})	1.5188	0.0944
H ({14,15,16,17})	1.1275	0.0984
H ({18,19,20,21,22,23,24,25,26})	1.1375	0.1109

Table S4: Uncertainties in the charges of N,N-dimenthyl-p-methoxybenzamide.

atoms (index)	mean	std.
C ({1,3})	-0.2654	0.1328
C ({2})	0.3606	0.2128
C ({4,6})	-0.0307	0.0937
C ({5})	-0.1578	0.1639
O ({7})	-0.4294	0.1239
C ({8})	0.6500	0.2122
C ({9})	0.1026	0.2666
O ({10})	-0.5643	0.1241
N ({11})	-0.4023	0.1705
C ({12,13})	-0.0580	0.2083
H ({14,15})	0.1585	0.0432
H ({16,17})	0.1327	0.0460
H ({18,19,20})	0.0442	0.0805
H ({21,22,23,24,25,26})	-	-

C Generalized polynomial chaos expansions

Let $E(\boldsymbol{\xi})$ denote a quantity of interest, such as the solvation energy, which depends on uncertain variables $\boldsymbol{\xi}$. The gPC expansion for E can be written as

$$E(\boldsymbol{\xi}) = E_g(\boldsymbol{\xi}) + \varepsilon(\boldsymbol{\xi}) = \sum_{n=1}^{N} c_n \psi_n(\boldsymbol{\xi}) + \varepsilon(\boldsymbol{\xi}), \tag{1}$$

where ε is the model error, N is a positive integer, c_n are expansion coefficients, and ψ_n are multivariate polynomials which are orthonormal with respect to the distribution of $\boldsymbol{\xi}$:

$$\int_{\mathbb{R}^d} \psi_i(\boldsymbol{x})\psi_j(\boldsymbol{x})\rho_{\boldsymbol{\xi}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \delta_{ij}, \tag{2}$$

where $\rho_{\boldsymbol{\xi}} : \Gamma \mapsto [0, \infty)$ is the probability distribution function (PDF) of $\boldsymbol{\xi}$ over domain $\Gamma \subseteq \mathbb{R}^d$ and δ_{ij} is the Kronecker delta. In this work, we study systems relying on d-dimensional Gaussian random vector $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. Therefore, the gPC basis functions are constructed by tensor products of univariate orthonormal Hermite polynomials. For a multi-index $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_d), \alpha_i \in \mathbb{N} \cup \{0\}$, we set

$$\psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi}) = \psi_{\alpha_1}(\xi_1)\psi_{\alpha_2}(\xi_2)\cdots\psi_{\alpha_d}(\xi_d). \tag{3}$$

For two different multi-indices $\boldsymbol{\alpha}_i = ((\alpha_i)_1, (\alpha_i)_2, \cdots, (\alpha_i)_d)$ and $\boldsymbol{\alpha}_j = ((\alpha_j)_1, (\alpha_j)_2, \cdots, (\alpha_j)_d)$, we have the property

$$\int_{\mathbb{R}^d} \psi_{\boldsymbol{\alpha}_i}(\boldsymbol{x})\psi_{\boldsymbol{\alpha}_j}(\boldsymbol{x})\rho_{\boldsymbol{\xi}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \delta_{\boldsymbol{\alpha}_i\boldsymbol{\alpha}_j} = \delta_{(\alpha_i)_1(\alpha_j)_1}\delta_{(\alpha_i)_2(\alpha_j)_2}\cdots\delta_{(\alpha_i)_d(\alpha_j)_d}, \tag{4}$$

where

$$\rho_{\boldsymbol{\xi}}(\boldsymbol{x}) = \left(\frac{1}{\sqrt{2\pi}}\right)^d \exp\left(-\frac{x_1^2 + x_2^2 + \cdots + x_d^2}{2}\right). \tag{5}$$

For simplicity, we denote $\psi_{\boldsymbol{\alpha}_i}$ as ψ_i and set $\psi_i = 1$. The mean and the variance can be estimated very easily from this expansion: $\mathbb{E}\{E\} \approx \mathbb{E}\{E_g\} = c_1, \mathrm{Var}\{E\} \approx \mathrm{Var}\{E_g\} = \sum_{n=2}^{N} c_n^2$.

After the basis functions are selected, the surrogate model is constructed by computing the coefficients c_n. The first step is to generate input parameter samples $\boldsymbol{\xi}^q, q = 1, 2, \cdots, M$; e.g., by using Monte Carlo sampling, and to obtain corresponding output samples E^q by running the Poisson solver with the input parameters $\boldsymbol{\xi}^q$. With this data, the simplest approach to compute c_n is linear regression. However, such approaches only work when $M >$

N; i.e., if the compound consists of many atoms and the number of uncertain parameters are large, we need to include many ψ_n in the surrogate model to obtain an accurate surrogate model for the PB solver. In general, we can only obtain relatively small number of output samples due to limited computational resources such that $M < N$ or $M \ll N$. Therefore, we generally need to solve the under-determined linear system:

$$\boldsymbol{\Psi}\boldsymbol{c} = \boldsymbol{E} + \boldsymbol{\varepsilon}, \tag{6}$$

where $\boldsymbol{E} = (E^1, E^2, \cdots, E^M)^T$ is the vector of output samples, $\boldsymbol{\Psi}$ is an $M \times N$ matrix with $\Psi_{ij} = \psi_j(\boldsymbol{\xi}^i)$, and $\boldsymbol{\varepsilon} = (\varepsilon^1, \varepsilon^2, \cdots, \varepsilon^M)^T$ is a vector of error samples with $\varepsilon^i = \varepsilon(\boldsymbol{\xi}^i)$. The compressive sensing method is effective at solving this type of under-determined problem when $\boldsymbol{c}$ is *sparse* and $\boldsymbol{\Psi}$ satisfies some condition.[? ? ?] Here "sparse" means that many c_n are close to 0, and only a small number of c_n are of large magnitude. There are several approaches to enhance sparsity;[? ? ? ?] the compressive sensing method typically approximates $\boldsymbol{c}$ by solving the following ℓ_1 minimization problem:

$$(P_{1,\tau}): \quad \arg\min_{\boldsymbol{c}} \|\boldsymbol{c}\|_1, \quad \text{subject to} \quad \|\boldsymbol{\Psi}\boldsymbol{c} - \boldsymbol{E}\|_2 \le \tau, \tag{7}$$

where $\tau = \|\boldsymbol{\varepsilon}\|_2$ is the magnitude of the truncation error estimated by the cross-validation.[?] Standard convex optimization methods are applicable[?] to this type of minimization problem. In this work, we used a specifically designed MATLAB package `spgl1`[? ?] to solve $(P_{1,\tau})$. We implement a newly developed sparsity-enhancing compressive sensing technique[?] to further improve the accuracy of the solution to $(P_{1,\tau})$.

Given a fixed M, the accuracy of the compressive sensing method relies on the structure of $\boldsymbol{\Psi}$ and the sparsity of $\boldsymbol{c}$. The former can be improved by special sampling strategies,[? ?] and the latter can be improved by sparsity-enhancing techniques.[? ?] The sparsity-enhancing techniques find another set of random variables $\boldsymbol{\eta} = (\eta_1, \eta_2, \cdots, \eta_d)^T$ such that the vector $\tilde{\boldsymbol{c}}$, which are the gPC coefficients of u with respect to $\boldsymbol{\eta}$, is sparser. In order words, our goal is to seek $\boldsymbol{\eta}(\boldsymbol{\xi})$ with

$$E_g(\boldsymbol{\xi}) = \sum_{n=1}^{N} c_n \psi_n(\boldsymbol{\xi}) = \sum_{n=1}^{N} \tilde{c}_n \psi_n(\boldsymbol{\eta}) = E_g(\boldsymbol{\eta}),$$

such that $\tilde{\boldsymbol{c}}$ is sparser than $\boldsymbol{c}$. We use a rotation matrix $\boldsymbol{A}$ to identify $\boldsymbol{\eta}$: $\boldsymbol{\eta} = \boldsymbol{A}\boldsymbol{\xi}$. Since $\boldsymbol{\xi}$ are i.i.d. Gaussian, and $\boldsymbol{A}$ is orthonormal (i.e., $\boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{I}$), we have $\boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. Therefore, $E_g(\boldsymbol{\eta})$ is a gPC expansion with respect to $\boldsymbol{\eta}$ because ψ_n are orthonormal Hermite polynomials and $\rho_{\boldsymbol{\eta}}(\boldsymbol{x}) = \rho_{\boldsymbol{\xi}}(\boldsymbol{x})$, hence the orthonormal condition Eq. (1) holds. In order to obtain the rotation matrix $\boldsymbol{A}$, we first define the "gradient matrix"? ? G:

$$\boldsymbol{G} = \mathbb{E}\left\{\nabla E(\boldsymbol{\xi}) \cdot \nabla E(\boldsymbol{\xi})^T\right\} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T, \quad \boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{I}, \tag{8}$$

where $\boldsymbol{G}$ is symmetric, $\nabla E(\boldsymbol{\xi}) = (\partial E/\partial \xi_1, \partial E/\partial \xi_2, \cdots, \partial E/\partial \xi_d)^T$ is a column vector, $\boldsymbol{U} = (\boldsymbol{U}_1, \boldsymbol{U}_2, \cdots, \boldsymbol{U}_d)$ is an orthonormal matrix consisting of eigenvectors $\boldsymbol{U}_i$, and $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_d)$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ is a diagonal matrix with elements representing decreasing variation of the system along the respective eigenvectors. We choose $\boldsymbol{A} = \boldsymbol{U}^T$, which is a orthonormal matrix. Consequently, when the differences between λ_i are large, g helps to concentrate the dependence of u primarily on the first few new random variables η_i due to the larger variation of u along the directions of the corresponding eigenvectors. Therefore, we obtain a sparser $\tilde{\boldsymbol{c}}$ than $\boldsymbol{c}$. Since E is unknown, we use E_g computed from standard ℓ_1 minimization method to approximate $\boldsymbol{G}$:

$$\boldsymbol{G} \approx \mathbb{E}\left\{\nabla\left(\sum_{n=1}^{N} c_n \psi_n(\boldsymbol{\xi})\right) \cdot \nabla\left(\sum_{n'=1}^{N} c_{n'} \psi_{n'}(\boldsymbol{\xi})\right)^T\right\}. \tag{9}$$

The entries of $\boldsymbol{G}$ can be approximated as:

$$G_{ij} \approx \mathbb{E}\left\{\frac{\partial}{\partial \xi_i}\left(\sum_{n=1}^{N} c_n \psi_n(\boldsymbol{\xi})\right) \cdot \frac{\partial}{\partial \xi_j}\left(\sum_{n'=1}^{N} c_{n'} \psi_{n'}(\boldsymbol{\xi})\right)\right\} = \boldsymbol{c}^T \boldsymbol{K}_{ij} \boldsymbol{c}, \tag{10}$$

where $\boldsymbol{K}_{ij}$ is a "stiffness" matrix with entries

$$(K_{ij})_{kl} = \mathbb{E}\left\{\frac{\partial\psi_k(\boldsymbol{\xi})}{\partial\xi_i}\cdot\frac{\partial\psi_l(\boldsymbol{\xi})}{\partial\xi_j}\right\}. \tag{11}$$

We note that $\boldsymbol{K}_{ij}$ can be pre-computed since $\{\psi_i\}$ are normalized Hermite polynomials. Here we provide the formula of $(K_{ij})_{kl}$ and more details can be found in.[?]

$$\begin{aligned}(K_{ij})_{kl} &= \mathbb{E}\left\{\frac{\partial\psi_{\boldsymbol{\alpha}_k}(\boldsymbol{\xi})}{\partial\xi_i}\cdot\frac{\partial\psi_{\boldsymbol{\alpha}_l}(\boldsymbol{\xi})}{\partial\xi_j}\right\}\\ &= \mathbb{E}\left\{\left(\psi_{(\alpha_k)_i}(\xi_i)'\prod_{\substack{m=1\\ m\neq i}}^{d}\psi_{(\alpha_k)_m}(\xi_m)\right)\cdot\left(\psi_{(\alpha_l)_j}(\xi_j)'\prod_{\substack{m=1\\ m\neq j}}^{d}\psi_{(\alpha_l)_m}(\xi_m)\right)\right\}\\ &= \sqrt{(\alpha_k)_i(\alpha_l)_j}\,\delta_{(\alpha_k)_i-1(\alpha_l)_i}\delta_{(\alpha_k)_j(\alpha_l)_j-1}\cdot\prod_{\substack{m=1\\ m\neq i,m\neq j}}\delta_{(\alpha_k)_m(\alpha_l)_m}.\end{aligned} \tag{12}$$

After identifying $\boldsymbol{A}$ and consequently $\boldsymbol{\eta}$ we solve $P(1,\gamma)$ to obtain $\tilde{\boldsymbol{c}}$. This procedure can be performed iteratively; i.e., every time we find a $\tilde{\boldsymbol{c}}$, we identify a new $\boldsymbol{A}$, then consequently define new $\boldsymbol{\eta}$. Hence, we can solve $(P_{1,\gamma})$ to find a new, possibly sparser $\tilde{\boldsymbol{c}}$. The entire iterative procedure is summarized in Algorithm 1.

In this algorithm, steps 1-4 construct the surrogate model by directly applying the compressive sensing method to the Monte Carlo sampling output. Steps 5-8 use sparsity-enhancing techniques to improve the accuracy of the compressive sensing method iteratively. In step 8, we use $\tau^{(l+1)}$ because the estimate of the truncation error by the cross-validation can be different in each iteration. In step 9, a termination criterion is required such as the empirical test $|\sum_{ij}|U_{ij}|-d|<\kappa$ where we chose $\kappa = 0.1d$.[?] This criterion implies that if the $(l+1)$-th rotation matrix is close to identity matrix or permutation matrix, we should stop the iteration. It is usually sufficient to use one or two iterations since further rotations do not improve the accuracy significantly.[?] [?]

Algorithm 1 Compressive sensing method with iterative rotations

```
1: Generate samples of independent Gaussian random variables $\boldsymbol{\xi}^q, q = 1, 2, \cdots, M$.
2: Generate samples of solvation energy $E^q = E(\boldsymbol{\xi}^q)$ by solving the PB equation with input $\boldsymbol{\xi}^q$.
3: Select multi-variate normalized Hermite polynomials to be used in the gPC surrogate model (e.g., Hermite polynomials up to a certain order) and construct the measurement matrix $\boldsymbol{\Psi}$ by setting $\Psi_{ij} = \psi_j(\boldsymbol{\xi}^i)$.
4: Solve the optimization problem $(P_{1,\tau})$ to obtain the gPC surrogate model of solvation energy $E_g(\boldsymbol{\xi}) = \sum_{n=1}^{N} \psi_n(\boldsymbol{\xi})$.
5: Set counter $l = 0$, $\boldsymbol{\eta}^{(0)} = \boldsymbol{\xi}$, $\tilde{\boldsymbol{c}}^{(0)} = \boldsymbol{c}$.
6: Construct $\boldsymbol{G}^{l+1}$ with $\boldsymbol{c}^{(l)}$ according to Eq. (10). Then decompose $\boldsymbol{G}^{(l+1)}$ as
       $\boldsymbol{G}^{(l+1)} = \boldsymbol{U}^{(l+1)}\boldsymbol{\Lambda}^{(l+1)}(\boldsymbol{U}^{(l+1)})^T, \quad \boldsymbol{U}^{(l+1)}(\boldsymbol{U}^{(l+1)})^T = \boldsymbol{I}.$
7: Define $\boldsymbol{\eta}^{(l+1)} = (\boldsymbol{U}^{(l+1)})^T\boldsymbol{\eta}^{(l)}$, and compute samples $(\boldsymbol{\eta}^{(l+1)})^q = (\boldsymbol{U}^{(l+1)})^T(\boldsymbol{\eta}^{(l)})^q, q = 1, 2, \cdots, M$. Also, construct the new measurement matrix $\boldsymbol{\Psi}^{(l+1)}$ with $\Psi_{ij}^{(l+1)} = \psi_j((\boldsymbol{\eta}^{(l+1)})^i)$.
8: Solve the optimization problem $(P_{1,\tau^{(l+1)}})$:
       $\arg\min_{\boldsymbol{c}} \|\boldsymbol{c}\|_h, \quad \text{subject to} \|\boldsymbol{\Psi}^{(l+1)}\boldsymbol{c} - \boldsymbol{E}\|_2 \le \tau^{(l+1)},$
   and set $\tilde{\boldsymbol{c}}^{(l+1)} = \boldsymbol{c}$.
9: Set $l = l + 1$. If the termination criterion is satisfied, set $\boldsymbol{A} = \left(\boldsymbol{U}^{\{1\}}\boldsymbol{U}^{\{2\}} \cdots \boldsymbol{U}^{\{l\}}\right)^T$ and stop. Otherwise, go to Step 5.
```