Multiscale Model. Simul.
Vol. 13, No. 4, pp. 1327–1353

CONSTRUCTING SURROGATE MODELS OF COMPLEX SYSTEMS WITH ENHANCED SPARSITY: QUANTIFYING THE INFLUENCE OF CONFORMATIONAL UNCERTAINTY IN BIOMOLECULAR SOLVATION[*]

H. LEI[†], X. YANG[†], B. ZHENG[†], G. LIN[‡], AND N. A. BAKER[†]

Abstract. Biomolecules exhibit conformational fluctuations near equilibrium states, inducing uncertainty in various biological properties in a *dynamic* way. We have developed a general method to quantify the uncertainty of target properties induced by conformational fluctuations. Using a generalized polynomial chaos (gPC) expansion, we construct a surrogate model of the target property with respect to varying conformational states. To alleviate the high dimensionality of the corresponding stochastic space, we propose a method to increase the sparsity of the gPC expansion by defining a set of conformational "active space" random variables. With the increased sparsity, we employ the compressive sensing method to accurately construct the surrogate model. We demonstrate the performance of the surrogate model by evaluating fluctuation-induced uncertainty in solvent-accessible surface area for the bovine trypsin inhibitor protein system and show that the new approach offers more accurate statistical information than standard Monte Carlo approaches. Furthermore, the constructed surrogate model also enables us to *directly* evaluate the target property under various conformational states, yielding a more accurate response surface than standard sparse grid collocation methods. In particular, the new method provides higher accuracy in high-dimensional systems, such as biomolecules, where sparse grid performance is limited by the accuracy of the computed quantity of interest. Our new framework is generalizable and can be used to investigate the uncertainty of a wide variety of target properties in biomolecular systems.

Key words. uncertainty quantification, biomolecular conformation fluctuation, polynomial chaos, compressive sensing method, model reduction

AMS subject classifications. 92C05, 74F05, 82D99, 82D60

DOI. 10.1137/140981587

1. Introduction.

Biomolecular structures are inherently uncertain due to thermal fluctuations and experimental limits in structural characterization. At equilibrium, a biomolecule samples an ensemble of states governed by an energy landscape. For a biomolecule with well-defined native structure at an energetic global minimum, these states are generally located in the neighborhood of the native structure. While the native equilibrium structure of a biomolecule provides essential insight, it is also important to understand conformational fluctuations of biomolecular systems and their impact on molecular properties. In particular, it is of great interest to accurately

[*] Received by the editors August 8, 2014; accepted for publication (in revised form) August 31, 2015; published electronically November 5, 2015. 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). Pacific Northwest National Laboratory is operated by Battelle for the U.S. Department of Energy under contract DE-AC0576RL01830. The U.S. Government retains, and the publisher by accepting this article for publication, acknowledges that the U.S. Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for U.S. Government purposes.

http://www.siam.org/journals/mms/13-4/98158.html

[†] Pacific Northwest National Laboratory, Richland, WA 99352 (huan.lei@pnnl.gov, xiu.yang@pnnl.gov, bin.zheng@pnnl.gov, Nathan.Baker@pnnl.gov). The first author's research was supported by a travel grant from the IMA for a workshop on Uncertainty Quantification in Materials Modeling and a travel grant from the DOE for the Conference on Data Analysis 2014.

[‡] Department of Mathematics, Purdue University, West Lafayette, IN 47906 (guanglin@purdue.edu).

quantify the uncertainty in these properties caused by stochastic conformational fluctuations.

Molecular dynamics (MD) simulations offer a powerful tool for examining the influence of conformational uncertainty on biomolecular properties [17, 2]. Over the past few decades, this approach has made great progress in the development of accurate empirical force fields as well as efficient simulation algorithms [38]. However, despite these advances, MD is still a very computationally expensive simulation approach, particularly for large biomolecular complexes. Moreover, the finite durations of MD simulations are plagued with uncertainty in calculated properties due to nonergodic sampling. Many coarse-grained (CG) models and methods have been developed to facilitate molecular simulation at larger length scales and longer time scales. One popular approach is the elastic network model (ENM), which involves a harmonic approximation of molecular energy landscape. It has been observed that the low-frequency normal modes of a biomolecular system can be reproduced using a single-parameter Hookean potential between neighboring residues [49, 25, 47]. In particular, by only modeling interactions between the neighboring α-carbon (C_α), ENMs are able to predict structural fluctuations (e.g., Debye–Waller or B-factors) with surprising accuracy [25, 4].

The simplified potentials used by CG models such as the ENM allow us to examine structural fluctuations in a semianalytical manner. However, there does not exist an *analytical formula* that directly leads from the structural fluctuations to target biomolecular properties computed from the structure. Instead, given a specific biomolecular conformation (e.g., one snapshot of biomolecule structure under fluctuation), we still need further numerical computation to obtain the target properties. This leads to an important practical question: how do we utilize the stochastic information obtained from these models to efficiently quantify the uncertainty of the target property induced by the biomolecular conformational (structural) fluctuation? In many applications, a single native conformation of a molecule is used when computing properties such as molecular volume and area [27, 41, 12], electrostatic and solvation properties [40, 44], titration states [3], and other quantities. However, these quantities are all sensitive to the structure of the molecule and therefore subject to uncertainty induced by conformational fluctuations. Many studies neglect this uncertainty; those which attempt to assess it are forced to resort to time-consuming Monte Carlo sampling over the numerous biomolecular conformation states.

In the present work, we address this issue by providing a general framework to quantify conformation-induced uncertainty on various biomolecular properties. In particular, we construct a surrogate model of a target quantity in terms of the molecular conformational states. The constructed surrogate model enables us to efficiently evaluate the statistical information of the target property, e.g., probability density function. To the best of our knowledge, this is the first demonstration of how a target property response surface—including property uncertainty—can be directly evaluated from the biomolecular conformational distribution.

To construct the surrogate model, we adopt the generalized polynomial chaos (gPC) [22, 53] and formulate the target property as an expansion of a set of gPC basis functions determined by the specific conformation states, where the gPC coefficients are determined by the values of the target properties on a number of sampling conformation states. Within this framework, numerical quantification of the conformation-induced uncertainty is formulated as the following problem: how can we accurately and efficiently construct the gPC based surrogate model of the target property using limited sampling points within the high-dimensional conformational space? Several probabilistic collocation methods (PCMs) such as ANOVA [31, 18, 58, 55] and sparse grid methods [52, 20, 19, 36, 30] have been proposed to accurately construct gPC expansions by selecting specific collocation points for sampling. However, there are two fundamental barriers when directly applying these approaches to high-dimensional biomolecular systems with hundreds to thousands of degrees of freedom in CG representations. The first barrier is the required number of sampling points, which can be too large for any gPC approach beyond a linear approximation. Moreover, empirical evidence indicates that sparse grid methods are often limited to dimensions less than ~ 40 (e.g., see [37]). The second barrier is the presence of limited accuracy in the calculation of target properties—even in the absence of structural uncertainty. For example, many calculations related to biomolecular solvation properties are subject to errors in the discretization and numerical solution of the associated partial differential equations [5, 26]. The error between the true values and the computed values of these target properties can lead to erroneous results due to inhomogeneous weight distribution over the sampling points, as illustrated in this paper. To circumvent these difficulties, we adopt an alternative noncollocation method based on compressive sensing [11, 15, 7] which reduces the influence of the limited accuracy of the target property while taking advantage of the sparsity of the gPC expansion. The compressive sensing method was initially proposed for signal processing and later applied to wide range of applications, including uncertainty quantification frameworks [29, 16, 54, 56].

2. Stochastic model.

In this section, we briefly introduce a semianalytical stochastic model based on the ENM presented in [49, 25, 47]. The resulting harmonic system yields a Gaussian probability distribution for conformational states [4] that is straightforward to use in stochastic models for uncertainty quantification. In addition to the dimensionality reduction provided by the CG ENM, we note that further dimensionality reduction can be obtained for biomolecular target properties that have local dependence on structure, i.e., where the values associated with a particular property depend only on a subset of atoms in the molecule.

2.1. Full stochastic model of conformational fluctuation.

We construct the stochastic conformation space of the biomolecular system based on the CG anisotropic network model (ANM) [4], a variant of the ENM where each amino acid residue is modeled as a single CG particle connected to neighboring residues by anisotropic harmonic potentials. The ANM can be viewed as a simplified CG model of normal mode analysis [23, 6, 28, 32], where the model potential does not rely on the complex atomic-detail force field. Considering a biomolecule of N residues: we denote the $3N$-dimensional equilibrium position vector by $\overline{\mathbf{R}}^T = \left[\overline{\mathbf{r}}_1^T \; \overline{\mathbf{r}}_2^T \cdots \; \overline{\mathbf{r}}_N^T\right]$, where $\overline{\mathbf{r}}_i$ is a three-dimensional vector representing the equilibrium position of residue i. Similarly, we denote the $3N$-dimensional instantaneous position vector $\mathbf{R}^T = \left[\mathbf{r}_1^T \; \mathbf{r}_2^T \cdots \; \mathbf{r}_N^T\right]$, where $\mathbf{r}_i$ represents the instantaneous position vector of residue i. The fluctuation vector can then be defined by $\Delta\mathbf{R} = \mathbf{R} - \overline{\mathbf{R}}$. The harmonic approximation for the potential energy V with respect to the instantaneous position $\mathbf{R}$ is given by

$$V(\mathbf{R}) = \frac{\gamma}{2} \sum_{i<j} (r_{ij} - \overline{r}_{ij})^2 h(r_c - \overline{r}_{ij}), \tag{2.1}$$

where $\overline{r}_{ij}$ and r_{ij} represent the equilibrium and instantaneous distances between residue i and j, γ is a model parameter representing the elastic coefficient of the harmonic potential, r_c is the cut-off distance of the harmonic potential, and h is the Heaviside function.

Given the potential defined by (2.1), the $3N \times 3N$ Hessian matrix has the form

$$\mathbf{H} = \begin{pmatrix} \mathbf{H}_{11} & \mathbf{H}_{12}\cdots & \mathbf{H}_{1N} \\ \mathbf{H}_{21} & \mathbf{H}_{22}\cdots & \mathbf{H}_{2N} \\ \vdots & & \\ \mathbf{H}_{N1} & \mathbf{H}_{N2}\cdots & \mathbf{H}_{NN} \end{pmatrix}$$

with the element $\mathbf{H}_{ij}$ defined by

$$\mathbf{H}_{ij} = \begin{pmatrix} \partial^2 V/\partial X_i \partial X_j & \partial^2 V/\partial X_i \partial Y_j & \partial^2 V/\partial X_i \partial Z_j \\ \partial^2 V/\partial Y_i \partial X_j & \partial^2 V/\partial Y_i \partial Y_j & \partial^2 V/\partial Y_i \partial Z_j \\ \partial^2 V/\partial Z_i \partial X_j & \partial^2 V/\partial Z_i \partial Y_j & \partial^2 V/\partial Z_i \partial Z_j \end{pmatrix},$$

where X_i, Y_i, and Z_i represent the Cartesian coordinates of residues i. We note that the rank of $\mathbf{H}$ is $3N-6$ since V is translationally and rotationally invariant. This harmonic form for the potential leads to Gaussian statistics for the conformational probability distribution (e.g., individual residue position distribution). The correlation between individual residue fluctuation can be determined by the pseudoinverse of the Hessian matrix $\mathbf{H}$ as [4]

$$\mathbf{C} = \mathbb{E}\left[\Delta\mathbf{R}\Delta\mathbf{R}^T\right] = \frac{k_B T}{\gamma}\mathbf{H}^{-1}, \tag{2.2}$$

where $\mathbb{E}\left[\cdot\right]$ denotes the expectation, k_B is the Boltzmann constant, and T is the temperature.

We perform an eigendecomposition of $\mathbf{H}$,

$$\mathbf{H} = \mathbf{W}\mathbf{\Lambda}\mathbf{W}^T, \quad \mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_{3N-6}), \tag{2.3}$$

where λ_i is the ith nonzero eigenvalue of $\mathbf{H}$. $\mathbf{W}$ is a $3N \times (3N-6)$ matrix defined by

$$\mathbf{W} = \left[\mathbf{w}_1 \mathbf{w}_2 \cdots \mathbf{w}_{3N-6}\right], \tag{2.4}$$

where $\mathbf{w}_i$ is the corresponding ith eigenvector of $\mathbf{H}$. Then, the correlation matrix can be written as

$$\mathbf{C} = \frac{k_B T}{\gamma}\mathbf{W}\mathbf{\Lambda}^{-1}\mathbf{W}^T = \mathbf{U}\mathbf{U}^T, \tag{2.5}$$

where $\mathbf{U} = \left(\frac{k_B T}{\gamma}\right)^{\frac{1}{2}}\mathbf{W}\mathbf{\Lambda}^{-\frac{1}{2}}$. The stochastic conformation space can be given by

$$\mathbf{R}(\boldsymbol{\xi}) = \overline{\mathbf{R}} + \Delta\mathbf{R}(\boldsymbol{\xi}), \tag{2.6a}$$

$$\Delta\mathbf{R}(\boldsymbol{\xi}) = \mathbf{U}\boldsymbol{\xi}, \tag{2.6b}$$

where $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_{3N-6})$ is an independent and identically distributed (i.i.d.) Gaussian random vector. Given a value of $\boldsymbol{\xi}$, the corresponding CG conformation is fully determined by (2.6), allowing us to calculate target properties, denoted by $X(\boldsymbol{\xi})$.

2.2. Reduced-dimensionality stochastic model for conformational fluctuations. The dimension of the stochastic conformation space constructed by (2.6) is $3N-6$, which can still be high when N is large. However, we note that some target properties of interest, particularly those related to a specific residue p, may have only

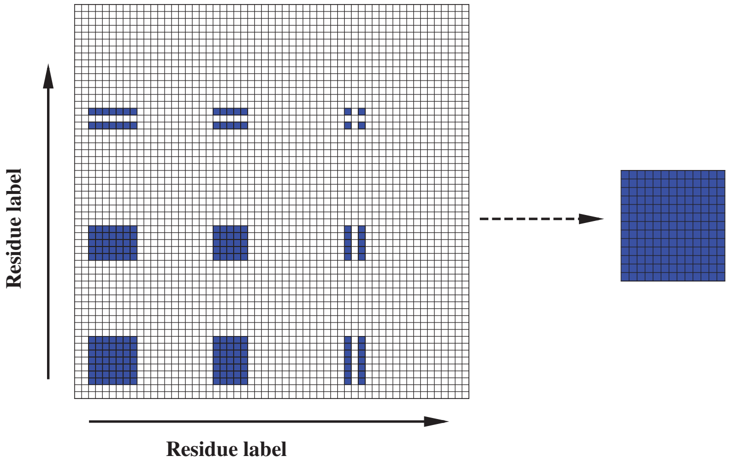


Fig. 1. *Sketch of a typical reduced property correlation matrix. The square on the left-hand side represents the full correlation matrix. Each block represents nine elements (in the x, y, and z directions) of a residue in the correlation matrix* $\mathbf{C}$*. The blue (solid) blocks represent the matrix elements associated with some* local *target property* X*. The square on the right-hand side represents the reduced correlation matrix* $\mathbf{C}^{\{p\}}$ *with lower dimensionality.*

local dependence on the neighboring residues' conformation rather than depending on all degrees of freedom in the biomolecule. For example, the solvent-accessible surface area (SASA) of a specific residue p depends only on its own positions as well as its neighboring residues within a certain cutoff distance $r_c^{\{p\}}$. Under such circumstances, the full position fluctuation correlation matrix $\mathbf{C}$ can be replaced by a $3N \times 3N$ matrix $\mathbf{C}'$ of larger sparsity where the (i,j) element (a 3×3 matrix) is given by

$$\mathbf{C}'_{ij} = \mathbf{C}_{ij} h(r_c^{\{p\}} - r_{ip}) h(r_c^{\{p\}} - r_{jp}), \tag{2.7}$$

$r_{ip} = |\mathbf{r}_p - \mathbf{r}_i|$, $r_{jp} = |\mathbf{r}_p - \mathbf{r}_j|$, and $r_c^{\{p\}}$ is a cutoff distance of residue p such that $X^{\{p\}}$ is independent of the residue i if $r_{ip} > r_c^{\{p\}}$.

Figure 1 illustrates this dimensionality reduction procedure for local properties as discussed above. Similar to (2.6), we can construct the reduced stochastic conformation space by

$$\mathbf{C}^{\{p\}} = \mathbf{U}^{\{p\}} {\mathbf{U}^{\{p\}}}^T, \tag{2.8a}$$

$$\mathbf{R}^{\{p\}}(\boldsymbol{\xi}^{\{p\}}) = \overline{\mathbf{R}^{\{p\}}} + \mathbf{U}^{\{\mathbf{p}\}} \boldsymbol{\xi}^{\{p\}}, \tag{2.8b}$$

where $\boldsymbol{\xi}^{\{p\}}$ is a d-dimensional i.i.d. normal random vector.

In summary, the value of a target property X is determined by the specific conformation state, corresponding to a point $\boldsymbol{\xi}$ (or $\boldsymbol{\xi}^{\{p\}}$) in the full (or reduced) random space. Our goal is to systematically quantify the uncertainty in X with respect to the conformational fluctuations through gPC expansion, as introduced in the next section. The rest of the paper focuses on local properties, so we will omit the superscript $\{p\}$ in the following text and use $X(\boldsymbol{\xi})$ and $\boldsymbol{\xi}$ to represent the target property and the d-dimensional random vector, respectively.

3. Numerical methods. In this section, we first review the gPC expansion with a brief discussion on possible difficulties with probabilistic collocation methods. Next, we introduce a noncollocation method to construct the gPC expansion based on compressive sensing. We note that the sparsity of gPC coefficients will affect the performance of the compressive sensing method (see [10]). Hence, we propose a method to elevate the sparsity of the gPC expansion by defining a new set of random variables according to the direction of variability in the target properties.

3.1. gPC expansion and collocation method. We use the gPC expansion to construct the surrogate model of the target property X with respect to the model parameter $\boldsymbol{\xi}$ (e.g., the molecular conformation) by

$$X(\boldsymbol{\xi}) = \sum_{|\boldsymbol{\alpha}|=0}^{\infty} c_{\boldsymbol{\alpha}} \psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi}), \tag{3.1a}$$

$$\psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi}) = \psi_{\alpha_1}(\xi_1)\psi_{\alpha_2}(\xi_2)\cdots\psi_{\alpha_d}(\xi_d), \quad \alpha_i \in \mathbb{N} \cup \{0\}, \tag{3.1b}$$

where d is the number of random variables, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_d)$ is a multi-index, and $c_{\boldsymbol{\alpha}}$ is the gPC coefficient to be determined. $\psi_{\alpha_i}(\xi_i)$ are univariate normalized Hermite polynomials, which satisfy the orthonormality condition:

$$\int_{-\infty}^{\infty} \psi_k(\xi_i)\psi_l(\xi_i)\rho(\xi_i)\mathrm{d}\xi_i = \delta_{kl}, \quad k, l \in \mathbb{N} \cup \{0\}, \tag{3.2}$$

where δ_{kl} is the Kronecker's delta and $\rho(\xi_i) = \frac{1}{\sqrt{2\pi}} \exp(-\xi_i^2/2)$ is the normal distribution function. The ENM described in section 2.1 relies on d i.i.d. standard normal random variables, and hence the gPC basis functions are constructed as the tensor products of univariate normalized Hermite polynomials, as shown in (3.1b).

We truncate the expression (3.1) up to polynomial order P, and hence X is approximated as

$$X(\boldsymbol{\xi}) \approx \widetilde{X}(\boldsymbol{\xi}) = \sum_{|\boldsymbol{\alpha}|=0}^{P} c_{\boldsymbol{\alpha}} \psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi}), \tag{3.3}$$

using a total number of n gPC terms with $n = (P + d)!/(P!d!)$.

Ideally, we would construct the truncated gPC expansion of $X(\boldsymbol{\xi})$ by computing $c_{\boldsymbol{\alpha}}$ using the orthonormality of $\psi_{\boldsymbol{\alpha}}$, e.g.,

$$c_{\alpha} = \int X(\boldsymbol{\xi})\psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi})\rho(\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi}, \tag{3.4}$$

where $\rho(\boldsymbol{\xi})$ is the probability density function (PDF) of $\boldsymbol{\xi}$.

The integration can be accomplished by utilizing probabilistic collocation approaches such as tensor product [42, 43] or sparse grid [52, 20, 19] methods. Specifically, by evaluating X on specific collocation points $\boldsymbol{\xi}^1, \boldsymbol{\xi}^2, \ldots, \boldsymbol{\xi}^S$, we have

$$c_{\boldsymbol{\alpha}} = \int X(\boldsymbol{\xi})\psi_{\alpha}(\boldsymbol{\xi})\rho(\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi} \approx \sum_{i=1}^{S} X(\boldsymbol{\xi}^i)\psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi}^i)w^i, \tag{3.5}$$

where w^i is the corresponding weight associated with collocation point $\boldsymbol{\xi}^i$.

However, for the high-dimensional biomolecular systems considered in the present work, the required number of collocation points S can be computationally intractable.

For example, a small biomolecular system with a 27-dimensional reduced conformation random space would require $S = 7.6 \times 10^{12}$ tensor product collocation points to construct a quadratic-order gPC expansion. The standard sparse grid method based on Gaussian quadrature and Smolyak construction reduces this to 1513 sampling points; however, the required number of sampling points is fixed for each order of the gPC approximation (e.g., the required number of sampling points for a third-order approximation is 27829), which makes it difficult to incorporate adaptive sampling strategies.

Also we note that X is generally accompanied by numerical error, e.g.,

$$X(\boldsymbol{\xi}) = \bar{X}(\boldsymbol{\xi}) + \phi, \tag{3.6}$$

where $\bar{X}(\boldsymbol{\xi})$ is the true value of the target property and ϕ represents the numerical error. In this work, we require ϕ to satisfy

$$|\phi| \ll |X|, \tag{3.7}$$

so we can systematically study the accuracy of the constructed surrogate model using different numbers of sampling data. In general, the condition $|\phi| \ll |X|$ is not essential to the application of gPC expansion to construct the surrogate model. ϕ provides a lower bound of the numerical error of the surrogate model; e.g., we should not expect the error of the surrogate model to be less than the error accompanied by the sampling data.

Moreover, as will be shown in section 4, the aliasing error and the numerical error ϕ associated with X may lead to poor approximation of $c_{\boldsymbol{\alpha}}$ by using the probabilistic collocation method even if (3.7) is satisfied. To overcome the above difficulties, we compute the gPC expansion by applying compressive sensing as described in section 3.2.

3.2. Compressive sensing method.

To construct the gPC expansion in (3.3), we compute $X(\boldsymbol{\xi})$ on M sampling points $(\boldsymbol{\xi}^1, \boldsymbol{\xi}^2, \ldots, \boldsymbol{\xi}^M)$, which are generated according to the distribution of random variables $\boldsymbol{\xi}$. In the present work, $\boldsymbol{\xi}$ are d-dimensional i.i.d. standard normal random variables. We discretize (3.3) as a linear system

$$\begin{pmatrix} \psi_{\boldsymbol{\alpha}_1}(\boldsymbol{\xi}^1) & \psi_{\boldsymbol{\alpha}_2}(\boldsymbol{\xi}^1) & \cdots & \psi_{\boldsymbol{\alpha}_n}(\boldsymbol{\xi}^1) \\ \psi_{\boldsymbol{\alpha}_1}(\boldsymbol{\xi}^2) & \psi_{\boldsymbol{\alpha}_2}(\boldsymbol{\xi}^2) & \cdots & \psi_{\boldsymbol{\alpha}_n}(\boldsymbol{\xi}^2) \\ \vdots & \vdots & & \vdots \\ \psi_{\boldsymbol{\alpha}_1}(\boldsymbol{\xi}^M) & \psi_{\boldsymbol{\alpha}_2}(\boldsymbol{\xi}^M) & \cdots & \psi_{\boldsymbol{\alpha}_n}(\boldsymbol{\xi}^M) \end{pmatrix} \begin{pmatrix} c_{\boldsymbol{\alpha}_1} \\ c_{\boldsymbol{\alpha}_2} \\ \vdots \\ c_{\boldsymbol{\alpha}_n} \end{pmatrix} = \begin{pmatrix} X(\boldsymbol{\xi}^1) \\ X(\boldsymbol{\xi}^2) \\ \vdots \\ X(\boldsymbol{\xi}^M) \end{pmatrix} + \begin{pmatrix} \varepsilon^1 \\ \varepsilon^2 \\ \vdots \\ \varepsilon^M \end{pmatrix}$$

or equivalently,

$$\boldsymbol{\Psi c} = \boldsymbol{X} + \boldsymbol{\varepsilon}, \tag{3.8}$$

where $\boldsymbol{\Psi}$ is the "measurement matrix" with entries $\boldsymbol{\Psi}_{i,j} = \psi_{\boldsymbol{\alpha}_j}(\boldsymbol{\xi}^i)$, $\boldsymbol{c} = (c_{\boldsymbol{\alpha}_1}, c_{\boldsymbol{\alpha}_2}, \ldots, c_{\boldsymbol{\alpha}_n})^T$ is the vector of the gPC coefficients, $\boldsymbol{X} = (X(\boldsymbol{\xi}^1), X(\boldsymbol{\xi}^2), \ldots, X(\boldsymbol{\xi}^M))^T$ is the vector consisting of the outputs, and $\boldsymbol{\varepsilon} = (\varepsilon^1, \varepsilon^2, \ldots, \varepsilon^M)^T$ is related to the truncation error.

Notice that $\boldsymbol{\Psi}$ is an $M \times n$ matrix, and we are interested in the case when $M < n$ or even $M \ll n$. It is presented in [10] that if $\boldsymbol{\Psi}$ satisfies the restricted isometry property (RIP), we can estimate $\boldsymbol{c}$ by solving the following optimization problem:

$$(P_{1,\epsilon}): \qquad \arg\min_{\boldsymbol{c}^*} \|\boldsymbol{c}^*\|_1 \quad \text{subject to} \quad \|\boldsymbol{\Psi c}^* - \boldsymbol{X}\|_2 \le \epsilon, \tag{3.9}$$

where $\epsilon = \|\boldsymbol{\varepsilon}\|_2$. The upper bound of the error $\|\boldsymbol{c} - \boldsymbol{c}^*\|_2$ is decided by ϵ and the sparsity of $\boldsymbol{c}$:

$$\|\boldsymbol{c} - \boldsymbol{c}^*\|_2 \leq C_1\epsilon + C_2\frac{\|\boldsymbol{c} - \boldsymbol{c}_s\|_1}{\sqrt{s}}, \tag{3.10}$$

where C_1, C_2 are constants, s is a positive integer, and $\boldsymbol{c}_s$ is $\boldsymbol{c}$ with all but the s-largest entries set to zero. For $\boldsymbol{c}$ in the present work, "sparse" means small $\|\boldsymbol{c} - \boldsymbol{c}_s\|_1$ with s being smaller (or much smaller) than the length of $\boldsymbol{c}$. The $(P_{1,\epsilon})$ optimization problem can be solved using classical convex optimization solvers (e.g., `CVX` [14]), sparse recovery software packages (e.g., `SPGL1` package [50], ℓ_1-`MAGIC` [1]), or the split Bregman method [24, 57, 9, 8]. In this paper, we use `SPGL1`.

To solve (3.9), we need the value of ϵ, which is generally not known a priori. In this work, we estimate ϵ using a cross-validation method [16, 56]. We first divide M sampling data into two parts, denoted by M_r and M_v. Second, $\boldsymbol{c}$ is computed with M_r sample points with a chosen series of tolerance error ϵ_r. Next, an optimized estimate $\hat{\epsilon}_r$ is determined such that $\|\boldsymbol{\Psi}_v\boldsymbol{c} - \boldsymbol{X}_v\|_2$ is minimized, where $\boldsymbol{\Psi}_v$ and $\boldsymbol{X}_v$ represent the submatrix of $\boldsymbol{\Psi}$ and the subvector of $\boldsymbol{X}$ corresponding to the validation portion of the sampling data. Finally, we repeat the above process for different replicas of the sample points and determine the optimal ϵ as $\epsilon = \sqrt{M/M_r}\hat{\epsilon}_r$. In this work, we set $M_r = 2M/3$ and performed the cross-validation for three replications. We note that verifying the RIP for a given matrix is an NP-hard problem. The aforementioned cross-validation procedure also serves as the verification for applying the ℓ_1 minimization method to approximate $\boldsymbol{c}$ as it estimates the error of ϵ.

3.3. Sparsity recovery via a "renormalized active" random space.

The performance of the compressive sensing method introduced above is closely related to the ratio between the numbers of sampling points M and basis functions n, as well as the sparsity of the linear system in (3.8). In general, accuracy improves with either larger M/n ratios or sparser target vectors $\boldsymbol{c}$. One way to increase M/n (for a given M) is to reduce the dimension of stochastic space; hence n is reduced. Unfortunately, for biomolecular systems, the dimension of the stochastic conformation space is determined by the structure of the molecule and is not always amenable to direct reduction. Constantine, Dow, and Wang [13] have developed an alternative approach that can be used to increase sparsity by analysis of variability in the target properties. For the target $X(\boldsymbol{\xi})$ with respect to PDF $\rho(\boldsymbol{\xi})$, we define gradient matrix $\mathbf{G}$ by [13]

$$\mathbf{G} = \mathbb{E}\left[\nabla X(\boldsymbol{\xi})\nabla X(\boldsymbol{\xi})^T\right], \tag{3.11}$$

where $\nabla X(\boldsymbol{\xi})$ is the gradient vector defined by $\nabla X(\boldsymbol{\xi}) = \left(\frac{\partial X}{\partial \xi_1}, \frac{\partial X}{\partial \xi_2}, \ldots, \frac{\partial X}{\partial \xi_d}\right)^T$. We conduct the eigendecomposition

$$\mathbf{G} = \mathbf{Q}\mathbf{K}\mathbf{Q}^T, \qquad \mathbf{Q} = \left[\mathbf{q}_1\ \mathbf{q}_2 \cdots\ \mathbf{q}_d\right], \tag{3.12a}$$

$$\mathbf{K} = \mathrm{diag}(k_1, \ldots, k_d), \quad k_1 \geq \cdots \geq k_d \geq 0, \tag{3.12b}$$

where $\mathbf{q}_i$ is the ith eigenvector of $\mathbf{G}$. Therefore, the target property X exhibits the largest variability along the direction $\mathbf{q}_1$ while it exhibits the smallest variability along the direction $\mathbf{q}_d$. This motivates the definition of a new random vector

$$\boldsymbol{\chi} = \mathbf{Q}^T\boldsymbol{\xi}, \tag{3.13}$$

where $\mathbf{Q}$ is unitary and $\boldsymbol{\chi} = (\chi_1, \chi_2, \ldots, \chi_d)^T$ are i.i.d. Gaussian variables since $\boldsymbol{\xi}$ are i.i.d. Gaussian (also similar to [48]). Dependence of the target property X on χ_i decreases from χ_1 to χ_d. Therefore, if we represent X by a gPC expansion with respect to $\boldsymbol{\chi}$, X may depend primarily on the first few random variables. Then the gPC coefficients associated with other variables exhibit a much smaller value (or even close to 0), yielding sparser $\boldsymbol{c}$ for the linear system defined in (3.8). Hence, if we recover the gPC coefficients with respect to $\boldsymbol{\chi}$ in (3.8) by the compressive sensing method, we expect a more accurate result than directly recovering gPC coefficients with respect to $\boldsymbol{\xi}$.

Unfortunately, the gradient vector $\nabla X(\boldsymbol{\xi})$ is generally not known a priori. Direct evaluation of $\mathbb{E}\left[\nabla X(\boldsymbol{\xi}) \nabla X(\boldsymbol{\xi})^T\right]$ is very computationally expensive: the cost of evaluation of $\nabla X(\boldsymbol{\xi})$ is proportional to the dimension of the $\boldsymbol{\xi}$. Therefore, we evaluate $\nabla X(\boldsymbol{\xi})$ by approximating it via the gPC expansion recovered from $\boldsymbol{\xi}$, e.g.,

$$\mathbf{G} \approx \mathbb{E}\left[\nabla X^{\mathrm{gPC}}(\boldsymbol{\xi}) \nabla X^{\mathrm{gPC}}(\boldsymbol{\xi})^T\right], \tag{3.14a}$$

$$X^{\mathrm{gPC}}(\boldsymbol{\xi}) = \sum_{|\boldsymbol{\alpha}|=0}^{P} c_{\boldsymbol{\alpha}}^{\{\boldsymbol{\xi}\}} \psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi}), \tag{3.14b}$$

where the superscript $\{\boldsymbol{\xi}\}$ represents gPC coefficients directly recovered from $\boldsymbol{\xi}$. Evaluation of $\mathbb{E}\left[\nabla X^{\mathrm{gPC}}(\boldsymbol{\xi}) \nabla X^{\mathrm{gPC}}(\boldsymbol{\xi})\right]$ is straightforward with respect to PDF $\rho(\boldsymbol{\xi})$, which can be used to define $\mathbf{Q}$ and, therefore, the new random basis $\boldsymbol{\chi} = \mathbf{Q}^T \boldsymbol{\xi}$. Finally, new basis functions associated with new random variables $\boldsymbol{\chi}$ can be used to reconstruct the gPC expansion of X with respect to the $\boldsymbol{\chi}$ which, in general, yields greater sparsity.

Remark 3.1. We do not reduce the dimension of the conformational space in the above procedure. Instead, we define a new basis spanning the random space based on the variability direction of the target property. This set of basis functions is *not* universal; it depends on the specific target property X.

Remark 3.2. The gradient matrix $\mathbf{G}$ is approximated by (3.14). Therefore, eigenvectors $[\mathbf{q}_1\ \mathbf{q}_2 \cdots\ \mathbf{q}_d]$ may not correspond exactly to the steepest decay directions of variability for the target property X. Nevertheless, we adopt (3.14) to construct a "rotated" space that provides larger (if not optimal) sparsity.

Remark 3.3. Notice that $\boldsymbol{\xi}$ are i.i.d. Gaussian random variables and so are $\boldsymbol{\chi}$ since the matrix $\mathbf{Q}$ is unitary, and the new basis functions associated with $\boldsymbol{\chi}$ are still tensor products of Hermite polynomials, i.e., of the same form as in (3.1).

We summarize the entire procedure presented above (sections 2 and 3) in Algorithm 1. In the next section, we apply this framework to quantify uncertainty in biomolecular solvent accessible surface area properties in the presence of conformational fluctuations.

ALGORITHM 1 (procedure to construct the gPC response surface of a given target quantity X with respect to a stochastic biomolecular conformation space).

Step 1. *For a biomolecular system, we model the potential energy using the harmonic elastic network approach so that the conformation fluctuation is Gaussian-distributed. We construct the full stochastic conformation space given in* (2.6). *For "local" target properties, we further reduce the dimension of the stochastic conformation space as in* (2.8). *We conduct eigenvalue decomposition of the correlation matrix and represent the fluctuation by a d-dimensional i.i.d. standard normal random vector denoted by* $\boldsymbol{\xi}$.

Step 2. *Generate M sampling points $\boldsymbol{\xi}^1, \boldsymbol{\xi}^2, \ldots, \boldsymbol{\xi}^M$ based on the distribution of $\boldsymbol{\xi}$. Numerically compute X on $\boldsymbol{\xi}^1, \boldsymbol{\xi}^2, \ldots, \boldsymbol{\xi}^M$ to obtain M outputs $X^1, X^2, \ldots, X^M$*

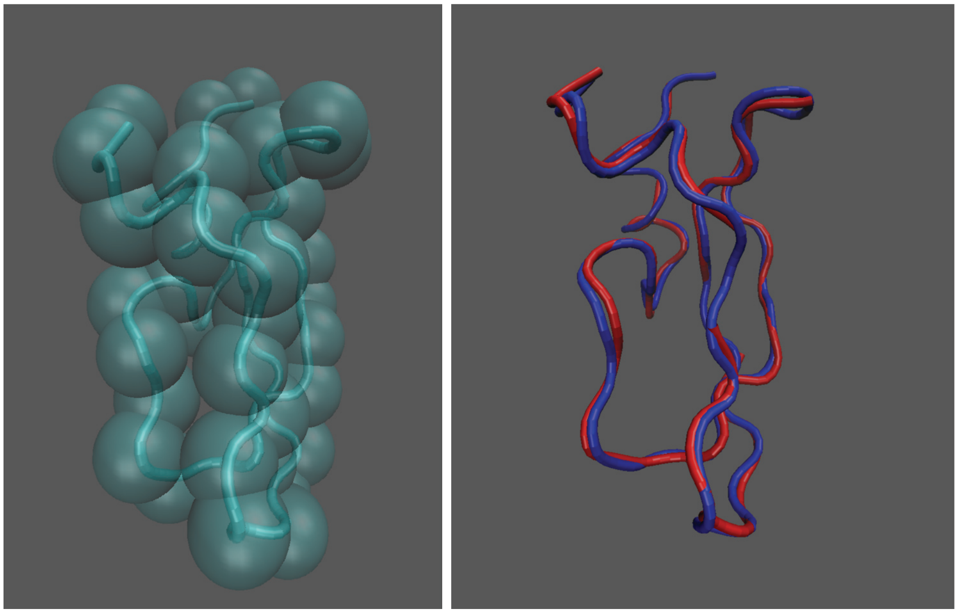


FIG. 2. (Left) *Tube diagram of the equilibrium structure of bovine pancreatic trypsin inhibitor (PDB code:* 5pti*) with spheres denoting the residue* C_α *positions.* (Right) *Tube diagrams of the molecule representing instantaneous conformational states under thermal fluctuation.*

(where $X^q = X(\boldsymbol{\xi}^q)$*). Denote* $\boldsymbol{X} = (X^1, X^2, \ldots, X^M)$ *as the "observation" in* $(P_{1,\epsilon})$*. The "measurement matrix"* $\boldsymbol{\Psi}$ *is constructed as* $\boldsymbol{\Psi}_{i,j} = \psi_{\boldsymbol{\alpha}_j}(\boldsymbol{\xi}^i)$*, where* $\psi_{\boldsymbol{\alpha}_j}$ *are the basis functions. The size of* $\boldsymbol{\Psi}$ *is* $M \times n$*, where* n *is the total number of basis functions depending on* P *in* (3.3)*.*

Step 3*. Set the tolerance* ϵ *in* $(P_{1,\epsilon})$ *by employing the cross-validation method.*

Step 4*. Solve the* ℓ_1 *minimization problem*

$$\arg\min_{\boldsymbol{c}^*} \|\boldsymbol{c}^*\|_1 \quad \textit{subject to} \quad \|\boldsymbol{\Psi}\boldsymbol{c}^* - \boldsymbol{X}\|_2 \le \epsilon$$

to obtain the gPC coefficients $\boldsymbol{c}^{\{\boldsymbol{\xi}\}}$*.*

Step 5*. Evaluate the gradient matrix* $\mathbf{G} = \mathbb{E}\big[\nabla X^{\mathrm{gPC}}(\boldsymbol{\xi})\nabla X^{\mathrm{gPC}}(\boldsymbol{\xi})^T\big]$*, given* $\boldsymbol{c}^{\{\boldsymbol{\xi}\}}$*, and define the random vector* $\boldsymbol{\chi}$ *by* (3.13)*. Compute the sample of* $\boldsymbol{\chi}$ *as* $\boldsymbol{\chi}^q = \mathbf{Q}^T\boldsymbol{\xi}^q, q = 1, \ldots, M$*.*

Step 6*. Construct new "measurement matrix"* $\tilde{\boldsymbol{\Psi}}$ *by setting* $\tilde{\boldsymbol{\Psi}}_{ij} = \psi_{\boldsymbol{\alpha}_j}(\chi^i)$*. Construct the gPC expansion of* $X(\boldsymbol{\chi})$ *by repeating Steps* 3–4 *on random vector* $\boldsymbol{\chi}$ *and using* $\boldsymbol{X} = (X^1, X^2, \ldots, X^M)^T$ *that have been determined in Step* 2*.*

4. Numerical results. As an example, we apply our method to quantify the uncertainty in the solvent-accessible surface area (SASA) caused by conformational fluctuations in the biomolecule bovine pancreatic trypsin inhibitor (PDB code: 5pti) [51], shown in Figure 2. The SASA is an essential element of numerous solvation models [5, 44, 40]. The SASA for the entire molecule can be decomposed into residue-specific contributions, allowing us to explore the influence of conformational fluctuations on local area uncertainty. The SASA is calculated following Shrake and Rupley [45], setting N^p nearly equidistant probing points on the solvent particle and determining the SASA value for each residue from the fraction of probing points that are not buried by any of the neighboring residues. In particular, we choose $N^p \approx 2.5 \times 10^5$ such that the numerical error ϕ satisfies $|\phi| / |X| \lesssim 1.0 \times 10^{-4}$; see section 4.2 for further discussion on sensitivity study on the accuracy of the constructed surrogate model by choosing different magnitudes of numerical error.

To demonstrate the applicability of our method in exploiting information from limited sampling data, we focus on the performance of our method when constructing a surrogate model using less than 2500 sample data. This performance is assessed relative to two reference systems: a direct Monte Carlo simulation of the conformational space with 10^6 sampling data as well as a system constructed by the standard sparse grid collocation method. We test our method by examining the L_2 error of the model as well as the Kullback–Leibler divergence between the PDFs obtained from this new approach and the reference data.

4.1. Surrogate model for SASA of individual residues.

Figure 2 shows a sketch of the CG biomolecular model under equilibrium and thermal-fluctuation states. Following [4], each residue is modeled as a single α-carbon particle, as shown in Figure 2(left). Due to thermal fluctuations, the molecule exhibits a distribution of conformation states where individual residues may deviate from the equilibrium positions, as shown in Figure 2(right). To model the fluctuation of individual residues, we construct the ANM correlation matrix $\mathbf{C}$ by (2.5) using a cutoff distance for the harmonic potential of $r_c = 9.8$Å. The radius values of the α-carbon residue and the solvent probe were set to 2.8 and 1.2 Å, respectively, for the SASA calculations.

We first consider local properties and study the SASA of residue P14. Starting with the full 168-dimensional random correlation matrix $\mathbf{C}$, we construct the local correlation matrix $\mathbf{C}'$ via (2.7) by setting the neighbor cutoff distance r_c^p to be 9.5 Å. This cutoff value yields eight neighboring residues and therefore a 27-dimensional random space $\mathbb{R}^{27}(\boldsymbol{\xi})$ by (2.8). As shown in Figure 3, the PDFs of the SASA of residue P14 extracted from the local and the full random conformation spaces agree well with each other, indicating that this particular property can be represented within a reduced space rather than the full 168-dimensional space. The dashed line in Figure 3 represents the PDF extracted from the local random space by neglecting the fluctuation correlation between different residues (e.g., setting the off-diagonal blocks to zero). The resulting distribution is wider than that predicted by the full correlation matrix. This is not surprising since the off-diagonal elements represent the harmonic potential contribution of molecular deformation in (2.1). Neglecting the off-diagonal block elements results in a more "flexible" molecule model which lacks the harmonic restraints and therefore exhibits a wider distribution of SASA values.

Next, we construct the surrogate model by computing the gPC coefficients within the reduced random space $\mathbb{R}^{27}(\boldsymbol{\xi})$ following the method presented in section 3. First, we calculate the gPC coefficients $c_{\boldsymbol{\alpha}}^{\{\boldsymbol{\xi}\}}$ up to order $P = 2$ (406 basis functions) by setting $M = 300$ in Algorithm 1 and applying Steps 1–4. Given $c_{\boldsymbol{\alpha}}^{\{\boldsymbol{\xi}\}}$, we next construct the approximate gradient matrix $\mathbf{G}$ by (3.14). Eigendecomposition of this matrix provides a set of rotated random variables $\boldsymbol{\chi}$ by (3.13) (Step 5 in Algorithm 1). Figure 4 shows the resulting normalized eigenvalues of $\mathbf{G}$ and the reduced correlation matrix $\mathbf{C}'$. We note that $\mathbf{C}'$ is independent of the target quantity X; it is completely determined by the molecular structure. The eigenvalues of $\mathbf{C}'$ decay slowly, at a rate similar to the full correlation matrix $\mathbf{C}$ (not shown in the plot), while the eigenvalues of the gradient matrix $\mathbf{G}$ decay much more quickly. This result indicates that, for a particular quantity X, the eigenvectors of $\mathbf{C}$ do not necessarily correspond to the directions with the steepest decay of variability in a target property.

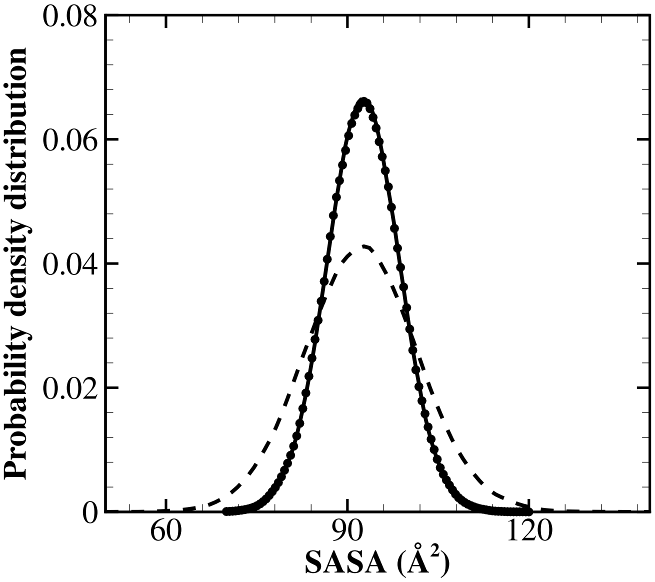


Fig. 3. *PDF of the SASA of the* 14*th residue obtained from the full correlation matrix* **C** *(solid line) and the local reduced correlation matrix* **C**′ *("●" symbol). The dashed line represents the distribution obtained from the reduced correlation matrix where off-diagonal elements are set to zero.*

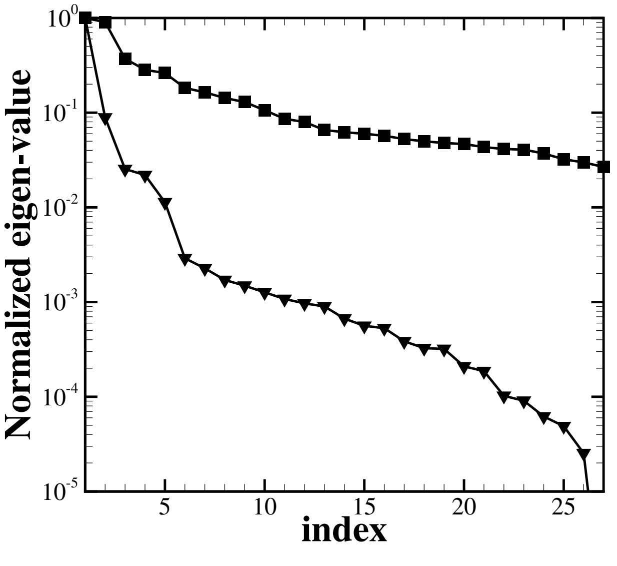


Fig. 4. *Normalized eigenvalues of the gradient matrix* **G** *("▼" symbol) and the correlation matrix* **C**′ *("■" symbol).*

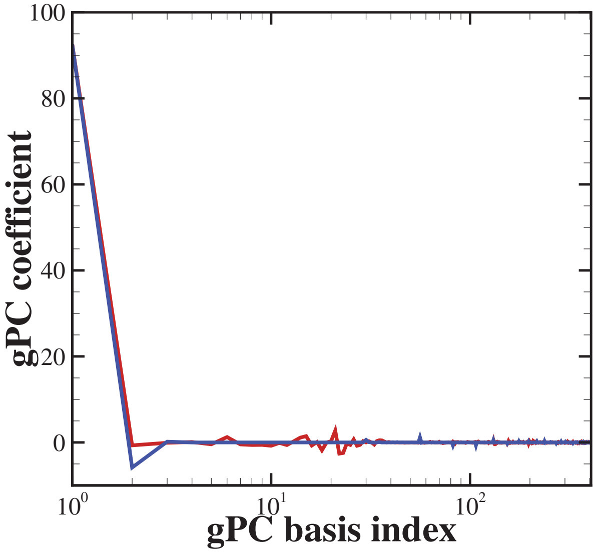


FIG. 5. *gPC coefficients (up to second order) for the SASA value on the* 14*th residue obtained from the compressive sensing method with respect to random vector* ξ *(red) and* χ *(blue) with dimension* $d = 27$*. Color is distinguishable only in the online version.*

Given the variables $\boldsymbol{\chi}$, we compute the corresponding gPC coefficients $c_{\boldsymbol{\alpha}}^{\{\boldsymbol{\chi}\}}$ with order $P = 2$ by applying Step 6 in Algorithm 1. The results are shown in Figure 5. Compared with $c_{\boldsymbol{\alpha}}^{\{\boldsymbol{\chi}\}}$, the spectrum of $c_{\boldsymbol{\alpha}}^{\{\boldsymbol{\chi}\}}$ exhibits a higher degree of sparsity, as expected. This result indicates that, with the same polynomial order, the target quantity X can be approximated using *fewer* gPC terms with respect to the set of random variables $\boldsymbol{\chi}$ than with $\boldsymbol{\xi}$.

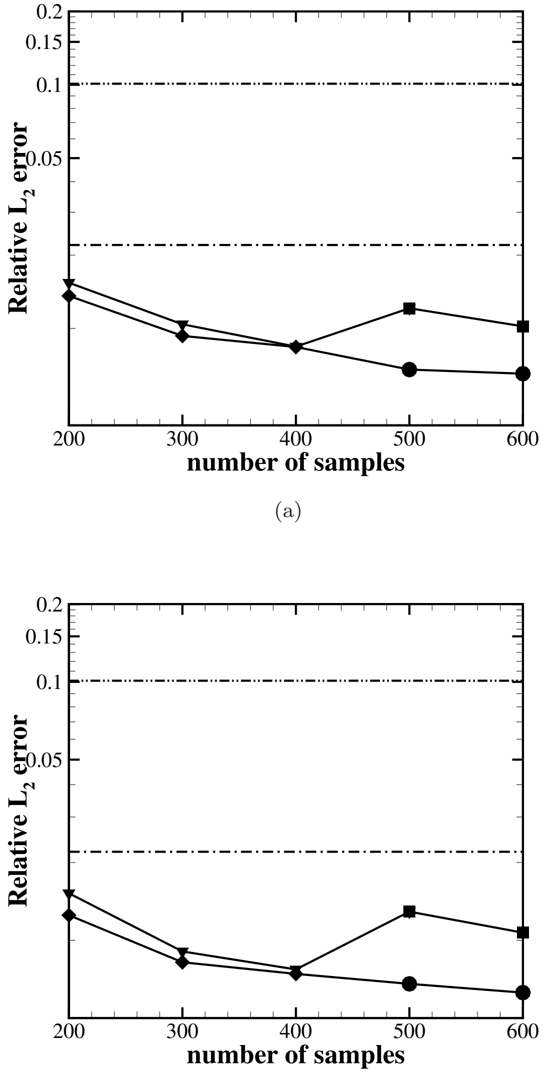
To examine the constructed surrogate model, we compute the relative L_2 error ϵ of the surrogate model by

$$\epsilon = \left(\frac{\int |X(\boldsymbol{\xi}) - \tilde{X}(\boldsymbol{\xi})|^2 \rho(\boldsymbol{\xi}) \mathrm{d}\boldsymbol{\xi}}{\int |X(\boldsymbol{\xi})|^2 \rho(\boldsymbol{\xi}) \mathrm{d}\boldsymbol{\xi}} \right)^{1/2}, \tag{4.1}$$

where $\tilde{X}$ is the gPC expansion of X by (3.3) with $c_{\boldsymbol{\alpha}}^{\{\boldsymbol{\xi}\}}$ and $c_{\boldsymbol{\alpha}}^{\{\boldsymbol{\chi}\}}$, respectively. As $X(\boldsymbol{\xi})$ is unknown in general, we use Monte Carlo sampling to approximate the integral in (4.2),

$$\epsilon \approx \left(\frac{\sum_{i=1}^{N_s} |X(\boldsymbol{\xi}^i) - \tilde{X}(\boldsymbol{\xi}^i)|^2}{\sum_{i=1}^{N_s} |X(\boldsymbol{\xi}^i)|^2} \right)^{1/2}, \tag{4.2}$$

where N_s is the number of sampling data. In this work, we choose $N_s = 10^6$.



(a)

(b)

Fig. 6. *Relative L_2 error of the SASA value on residue* 14 *predicted by the gPC expansions $\widetilde{X}(\boldsymbol{\xi})$ and $\widetilde{X}(\boldsymbol{\chi})$, where the gPC coefficients are obtained from two separate sets of sampling data, represented by* (a) *and* (b)*, respectively. The symbols "▼" and "■" denote the second- and third-order gPC expansions by $\boldsymbol{\xi}$. The symbols "◆" and "●" denote the second- and third-order gPC expansions by $\boldsymbol{\chi}$. The dash-dot and dash-dot-dot lines represent the relative L_2 error of the first- and second-order gPC expansions obtained from level-*1 *and level-*2 *sparse grid points, using* 55 *and* 1513 *sample points, respectively.*

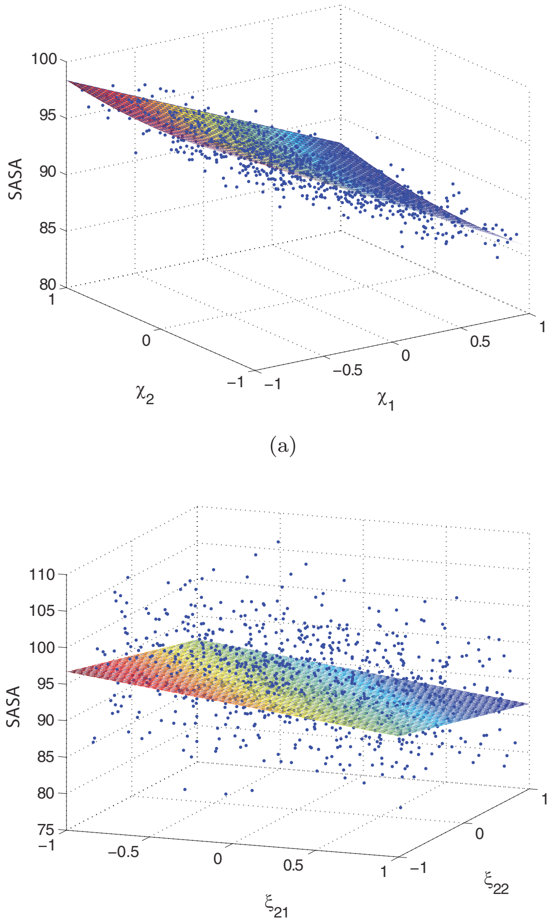
Figure 6 shows the relative L_2 error of the constructed surrogate model with gPC coefficients recovered from two independent sets of sample data. For each sample set, we use 200–400 points to construct the order $P = 2$ gPC expansion with 406 basis functions and 500–600 sample points to construct the order $P = 3$ of gPC expansion with 4060 basis functions. For each case, the L_2 error decreases as we increase the number of sampling points from 200 to 400. For the same number of sample points, the surrogate models constructed with respect to $\boldsymbol{\chi}$ exhibit a smaller L_2 error than those constructed with respect to $\boldsymbol{\xi}$. In particular, given the same number of sampling points, sparser gPC coefficients $\mathbf{c}$ in (3.8) lead to more accurate recovery of $\mathbf{c}$ from the compressive sensing method by (3.9). The accuracies of the compressive sensing methods based on $\boldsymbol{\xi}$ and $\boldsymbol{\chi}$ are comparable when the number of sampling points is close to the number of basis functions.

For random variables $\boldsymbol{\xi}$, the L_2 error changes nonmonotonically as we compute $c_{\boldsymbol{\alpha}}$ at order $P = 3$ by increasing numbers of sampling points. The error increases as we increase the number of sample points to 500 and then decreases as we increase the number of sample points to 600 (although it remains larger than the 400-point error). This behavior is primarily due to the fact that the number of basis functions for $P = 3$ is much larger than the number of sample points, and therefore $c_{\boldsymbol{\alpha}}^{\{\boldsymbol{\xi}\}}$ is poorly recovered due to insufficient sample points. However, for the transformed random variables $\boldsymbol{\chi}$, $c_{\boldsymbol{\alpha}}^{\{\boldsymbol{\chi}\}}$ can be accurately recovered due to the high sparsity of the gPC spectrum with a monotonic decrease in error with increasing numbers of sampling points. We note that ideally, the rotation process can be used iteratively. We leave this for future study.

We examined the surrogate model constructed by the sparse grid method based on Gaussian quadrature collocation points and Smolyak structure with gPC coefficients computed according to (3.5). Figure 6 shows the relative L_2 error of the surrogate model constructed by approximating the integral in (3.4) with level-1 and level-2 sparse grid methods using 55 and 1513 sample points, respectively. Note that the algebraic accuracies of the level-1 and level-2 sparse grid methods we use are 3 and 5, respectively. Therefore, we construct first- and second-order gPC expansions with level-1 and level-2 methods, respectively. The sparse grid results show systematically larger L_2 errors than in the compressive sensing approach. An unexpected phenomenon is that the error of the second-order expansion is larger than that of the first-order expansion. This behavior will be explained in section 4.2.

The differences between the models constructed by $c_{\boldsymbol{\alpha}}^{\{\boldsymbol{\xi}\}}$ and $c_{\boldsymbol{\alpha}}^{\{\boldsymbol{\chi}\}}$ can be further illustrated by examining the response surfaces in the *reduced* random space shown in Figure 7. This figure shows the response surfaces $\widetilde{X}(\boldsymbol{\xi})$ and $\widetilde{X}(\boldsymbol{\chi})$ with respect to two random variables with the remaining 25 random variables fixed. The gPC coefficients are computed using 300 sample points with the order $P = 2$ for both cases. For $\widetilde{X}(\boldsymbol{\chi})$, we only consider the first two random variables χ_1 and χ_2. For $\widetilde{X}(\boldsymbol{\xi})$, we consider the random variables ξ_{21} and ξ_{22} which are associated with the largest magnitudes of the first-order gPC coefficients. For each case, we fixed the remaining variables as constant values extracted from an i.i.d. standard normal distribution $\mathcal{N}(0, 1)$.

The behavior of the Monte Carlo data around the response surfaces in Figure 7 indicates that the variation of $\widetilde{X}(\boldsymbol{\chi})$ strongly depends on χ_1, χ_2 while the dependence is much weaker for ξ_{21} and ξ_{22}. Furthermore, most of the symbols generated by $\widetilde{X}(\boldsymbol{\chi})$ fall near the *reduced* response surface $\widetilde{X}(\boldsymbol{\chi})$ with small deviation while the deviations for $\widetilde{X}(\boldsymbol{\xi})$ are much larger around the response surface $\widetilde{X}(\boldsymbol{\xi})$. As expected with rotation of the space by (3.13), this result indicates that $\widetilde{X}(\boldsymbol{\chi})$ can be fitted fairly well by only using two variables. However, if we use the original random variables, the reduced response surface cannot be captured well even if we use the two most important variables associated with the first-order gPC expansion. Figure 7 clearly illustrates that the different sparsities of $\mathbf{c}$ result in different accuracies for the recovered response



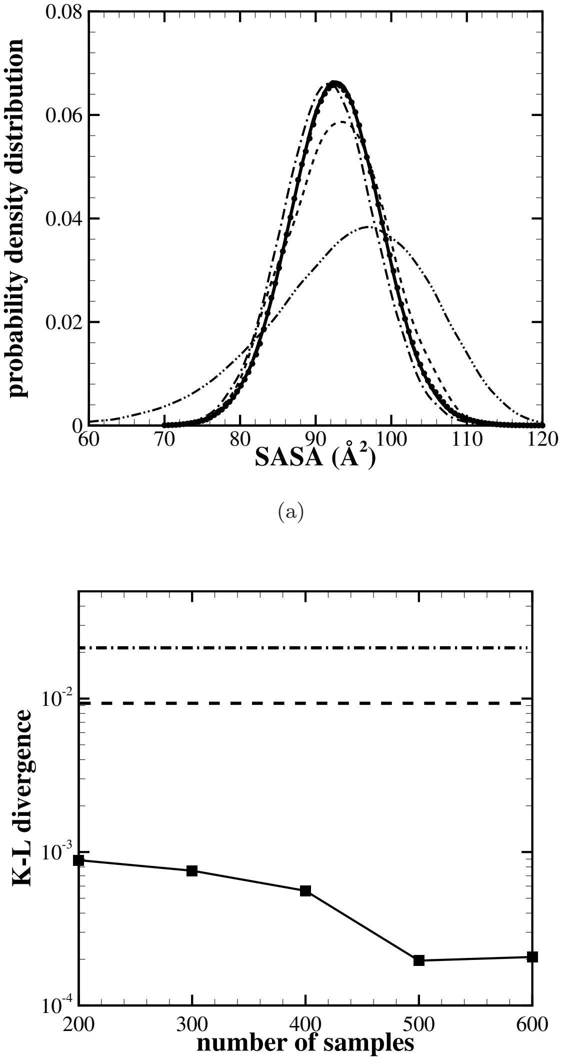
(a)

(b)

Fig. 7. (a) *The reduced response surface constructed by* $\widetilde{X}(\chi_1, \chi_2, \chi_3^0, \ldots, \chi_{27}^0)$*, where* $(\chi_3^0, \ldots, \chi_{27}^0)$ *are fixed values extracted from the i.i.d. normal distribution* $\mathcal{N}(0, 1)$*. The scattered blue (dotted) symbols are direct numerical simulation results on stochastic points* $(\chi_1, \chi_2, \ldots, \chi_{27})$ *in* $\mathbb{R}^{27}$ *following an i.i.d. normal distribution* $\mathcal{N}^{27}(0, 1)$*.* (b) *The reduced response surface constructed by* $\widetilde{X}(\xi_1^0, \ldots, \xi_{21}, \xi_{22}, \ldots, \xi_{27}^0)$ *where* $(\xi_1^0, \ldots, \xi_{20}^0, \xi_{23}^0, \ldots, \xi_{27}^0)$ *are fixed values extracted from the i.i.d. normal distribution* $\mathcal{N}(0, 1)$*. The scattered blue (dotted) symbols are direct numerical simulation results on points* $(\xi_1, \xi_2, \ldots, \xi_{27})$ *following an i.i.d. normal distribution* $\mathcal{N}^{27}(0, 1)$*.*

surfaces $\widetilde{X}(\boldsymbol{\chi})$ and $\widetilde{X}(\boldsymbol{\xi})$.

To evaluate the statistical information extracted from the surrogate model, we compute the SASA PDF for target residue P14 by evaluating 10^6 sampling data points with the constructed surrogate model. These results are shown in Figure 8(a) and compared with a reference solution based on the PDF computed from 10^6 direct Monte Carlo sample points. The compressive sensing method with 300 sample points yields the closest approximation of the reference solution. In contrast, the PDFs constructed by the direct Monte Carlo and sparse grid methods show significant deviation from the reference solution. To quantify the numerical error of the obtained PDFs, we



(a)

(b)

Fig. 8. (a) *PDF of the SASA values on residue P*14 *obtained from the gPC expansion* $\widetilde{X}(\boldsymbol{\chi})$ *using* 300 *sampling points ("•" symbol). Reference solution (solid line) is obtained from Monte Carlo sampling using* 10^6 *sample points. Results of the first- and second-order gPC expansions obtained from level-*1 *(dash-dot line,* 55 *sample points) and level-*2 *(dash-dot-dot line,* 1513 *sample points) sparse grid methods; results from direct Monte Carlo sampling methods (dashed line,* 300 *sample points) are also presented for comparison.* (b) *Kullback–Leibler divergence between the PDF of the reference solution and the PDFs obtained from gPC expansion* $\widetilde{X}(\boldsymbol{\chi})$ *("■" symbol) with varying numbers of sample points. Level-*1 *sparse grid (dash-dot line) and direct Monte Carlo (dashed line,* 300 *sample points) results are presented for comparison.*

computed the Kullback–Leibler divergence

$$D_{\mathrm{KL}} = \int_{-\infty}^{\infty} \ln\left(\frac{f^N(X)}{f^0(X)}\right) f^N(X)\,\mathrm{d}X \tag{4.3}$$

with the discrete form, where $f^N(X)$ and $f^0(X)$ represent the PDFs of the numerical and reference solutions, respectively. For the compressive sensing method, D_{KL} decreases as we increase the number of sampling points, which is consistent with the L_2 error of the surrogate model (Figure 6). The plateau value at 500–600 sampling points is primarily due to the finite resolution of the PDF: a sensitivity study shows that D_{KL} between two i.i.d. sets of 10^6 Monte Carlo sample points is on the order of 10^{-4}. In contrast, D_{KL} values of the PDFs obtained by the level-1 and level-2 sparse grid methods are about 20 and 450 times larger (respectively) than the results of the compressive sensing method.

4.2. Error sources and sensitivity analysis. To further investigate the applicability of the numerical methods for biomolecular systems, we quantified the SASA uncertainty for two other residues P11 and P20, which have 13 and 20 neighboring residues and correspond to random conformation spaces $\mathbb{R}^{42}$ and $\mathbb{R}^{63}$, respectively. For each case, we constructed the surrogate model by the compressive sensing method with respect to both $\boldsymbol{\xi}$ and $\boldsymbol{\chi}$, as well as by the level-1 (first-order gPC expansion) and level-2 (second-order gPC expansion) sparse grid methods. Figure 9 shows the relative L_2 error of the surrogate models, the PDFs of the SASA values, and the Kullback–Leibler divergence with respect to the reference solution.

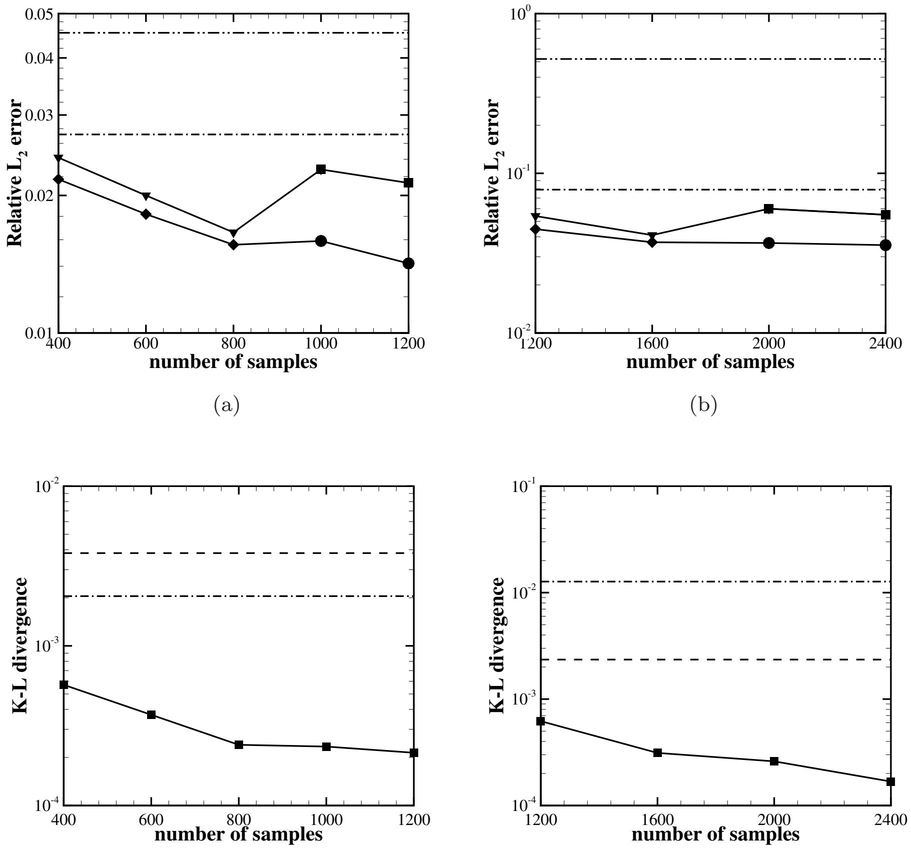
Similar to the results for residue P14, the surrogate models constructed with respect to $\boldsymbol{\chi}$ yield a smaller error than those constructed with respect to $\boldsymbol{\xi}$. The accuracy of the $\boldsymbol{\xi}$ and $\boldsymbol{\chi}$ compressive sensing methods is comparable when the number of sampling points is close to the number of basis functions. However, the surrogate model constructed with respect to $\boldsymbol{\chi}$ is more accurate than $\boldsymbol{\xi}$ when the number of sampling points is much less than the number of basis functions, e.g., when the third-order gPC terms are incorporated. In particular, the surrogate model (second-order gPC expansion) for residue 20 constructed by the level-2 sparse grid in random space $\mathbb{R}^{63}$ yields the largest deviation from the reference solution.

For the present system, the relatively large surrogate model error constructed by the sparse grid method (e.g., (3.5)) can be explained as follows. Given the target quantity X computed at collocation points, the gPC coefficient $c_{\boldsymbol{\alpha}}$ is computed by

$$\begin{aligned} c_{\boldsymbol{\alpha}} &= \sum_{i=1}^{N^{\mathrm{sp}}} w^i(\bar{X}(\boldsymbol{\xi}^i) + \phi(\boldsymbol{\xi}^i))\psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi}^i) \\ &= \sum_{i=1}^{N^{\mathrm{sp}}} w^i(\bar{X}^i + \phi^i)\psi_{\boldsymbol{\alpha}}^i, \end{aligned} \tag{4.4}$$

where $\bar{X}^i = \bar{X}(\boldsymbol{\xi}^i)$, $\phi^i = \phi(\boldsymbol{\xi}^i)$, and $\psi_{\boldsymbol{\alpha}}^i = \psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi}^i)$ represent the true solutions, numerical error, and Hermite basis function evaluated at the sparse grid collocation point $\boldsymbol{\xi}^i$, respectively; N^{sp} is the required number of sampling points with integral accuracy up to order $2P+1$; and ϕ is the associated numerical error accompanied with the computed value of $\bar{X}$. We assume that

$$|\phi^i| \ll |\bar{X}^i| \tag{4.5}$$



(a) (b)

(c) (d)

FIG. 9. (a)–(b) *Relative L_2 error of the SASA value on residue* 11 (a) *and* 20 (b) *predicted by the gPC expansions $\widetilde{X}(\boldsymbol{\xi})$ and $\widetilde{X}(\boldsymbol{\chi})$. The symbols "▼" and "■" denote the second- and third-order gPC expansions by $\boldsymbol{\xi}$. The symbols "◆" and "●" denote the second- and third-order gPC expansions by $\boldsymbol{\chi}$. The dash-dot lines represent the relative L_2 error of the first-order gPC expansion obtained from level-*1 *sparse grid methods using* 85 *and* 127 *sample points. The dash-dot-dot lines represent the relative L_2 error of the second-order gPC expansion obtained from level-*2 *sparse grid methods using* 3613 *and* 8065 *sample points.* (c)–(d) *Kullback–Leibler divergence between the PDF of the reference solution and the PDFs obtained from the constructed surrogate models ("■" symbol) for residues* 11 (c) *and* 20 (d)*. Level-*1 *sparse grid (dash-dot line) and direct Monte Carlo (dashed line,* 300 *sample points) results are presented for comparison.*

and that $c_{\boldsymbol{\alpha}}$ can be approximated by

$$
\begin{aligned}
c_{\boldsymbol{\alpha}} &= \sum_{i=1}^{N^{\mathrm{sp}}} w^i \bar{X}^i \psi_{\boldsymbol{\alpha}}^i + \sum_{i=1}^{N^{\mathrm{sp}}} w^i \phi^i \psi_{\boldsymbol{\alpha}}^i \\
&= \bar{c}_{\boldsymbol{\alpha}} + \sum_{\substack{|\boldsymbol{\alpha}+\boldsymbol{\beta}| \\ >2P+1}} \sum_{i=1}^{N^{\mathrm{sp}}} \bar{c}_{\boldsymbol{\beta}} w^i \psi_{\boldsymbol{\alpha}}^i \psi_{\boldsymbol{\beta}}^i + \sum_{i=1}^{N^{\mathrm{sp}}} w^i \phi^i \psi_{\boldsymbol{\alpha}}^i,
\end{aligned}
\tag{4.6}
$$

where $\bar{c}_{\boldsymbol{\alpha}} = \int \bar{X}(\boldsymbol{\xi}) \psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi}) \rho(\boldsymbol{\xi}) \mathrm{d}\boldsymbol{\xi}$ represents the true value of the gPC coefficient of

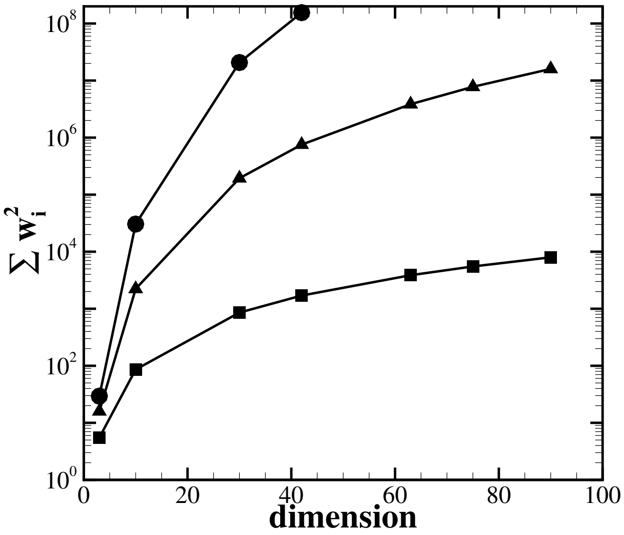


FIG. 10. *Variance of the numerical error term* $\sum_{i=1}^{N^{\mathrm{sp}}} \phi_i w_i$ *(normalized by* σ_ϕ^2*) for* c_0 *computed by level-*1 *("■"), level-*2 *("▲"), and level-*3 *("●") sparse grid points.*

index $\boldsymbol{\alpha}$ and $\bar{c}_{\boldsymbol{\beta}}$ represents the gPC coefficients with order $|\boldsymbol{\alpha}+\boldsymbol{\beta}| > 2P+1$. The second term on the right-hand side of (4.6) represents the aliasing error due to the sparse grid approximation. The third term $w^i\phi^i\psi^i_{\boldsymbol{\alpha}}$ represents the error due to the numerical error ϕ accompanied with the numerical computation of $\bar{X}$.

For systems of high dimensionality, the aliasing error $\sum_{\substack{|\boldsymbol{\alpha}+\boldsymbol{\beta}|\\ >2P+1}} \sum_{i=1}^{N^{\mathrm{sp}}} \bar{c}_{\boldsymbol{\beta}} w^i \psi^i_{\boldsymbol{\alpha}} \psi^i_{\boldsymbol{\beta}}$ and the numerical error $\sum_{i=1}^{N^{\mathrm{sp}}} w^i\phi^i\psi^i_{\boldsymbol{\alpha}}$ may both induce pronounced error to the numerical computation of $c_{\boldsymbol{\alpha}}$. Specifically, we assume that the numerical error ϕ_i superimposed on each collocation point is i.i.d. with zero mean and small variance $|\sigma_\phi^2| \ll |X|^2$. Given this assumption, the term $\sum_{i=1}^{N^{\mathrm{sp}}} w^i\phi^i\psi^i_{\boldsymbol{\alpha}}$ is zero mean with variance

$$\mathrm{Var}\left(\sum_{i=1}^{N^{\mathrm{sp}}} w^i\phi^i\psi^i_{\boldsymbol{\alpha}}\right) = \sum_{i=1}^{N^{\mathrm{sp}}} (w^i)^2(\psi^i_{\boldsymbol{\alpha}})^2\sigma_\phi^2. \tag{4.7}$$

When the dimension of $\boldsymbol{\xi}$ is large, we note that the weight distribution on sparse grid points is inhomogeneous, i.e.,

$$\sum_i w^i = 1, \qquad \exists k, |w^k| \gg 1. \tag{4.8}$$

Figure 10 plots the variance of the term $\sum_{i=1}^{N^{\mathrm{sp}}} \phi^i w^i$ (normalized by σ_ϕ^2) for c_0 (i.e., the coefficient of basis function $\psi_{\mathbf{0}}(\boldsymbol{\xi}) \equiv 1$) computed at different levels of sparse grid points. As the dimension increases, the variance of the error term increases rapidly, leading to nonnegligible errors in the computation of $c_{\boldsymbol{\alpha}}$.

For illustration purposes, we consider the following 27-dimensional function:

$$f(\boldsymbol{\xi}) = 305 + \sum_{k=1}^{27} \xi_k - 0.01 \left(\sum_{k=1}^{27} \xi_k^2 \right)^2 . \tag{4.9}$$

We first construct a first-order gPC expansion f_1 by using the level-1 sparse grid method (algebraic accuracy 3) to compute the coefficients:

$$c_{\boldsymbol{\alpha}} = \int_{\mathbb{R}^{27}} f(\boldsymbol{\xi})\psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi})\rho(\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi} \approx \sum_{i=1}^{N_1^{\mathrm{sp}}} f(\boldsymbol{\xi}^i)\psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi}^i)w^i, \tag{4.10}$$

where $\boldsymbol{\xi}^i, w^i$ are sparse grid points and corresponding weights and N_1^{sp} is the total number of level-1 sparse grid points. Next we construct a second-order gPC expansion f_2 by using the level-2 sparse grid method (algebraic accuracy 5) to compute the coefficients in the same manner. We compute the relative L_2 error of f_1 and f_2 as

$$\epsilon_k = \|f_k - f\|_2 / \|f\|_2 = \left(\sum_{q=1}^{N_4^{\mathrm{sp}}} (f_k(\boldsymbol{\xi}^q) - f(\boldsymbol{\xi}^q))^2 w^q \right)^{1/2} \Bigg/ \left(\sum_{q=1}^{N_4^{\mathrm{sp}}} f(\boldsymbol{\xi}^q)^2 w^q \right)^{1/2}, \; k = 1, 2, \tag{4.11}$$

where we use the level-4 sparse grid method (algebraic accuracy 9) so that the numerical integral gives an accurate result. The second-order expansion yields a larger L_2 error due to the aliasing error in numerical integration. In particular, $\epsilon_1 = 0.029$ while $\epsilon_2 = 0.100$.

We also constructed the gPC expansion of f by sample points superimposed with numerical error:

$$c_{\boldsymbol{\alpha}} = \int_{\mathbb{R}^{27}} f(\boldsymbol{\xi})\psi_i(\boldsymbol{\xi})\rho(\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi} \approx \sum_{q=1}^{N_1^{\mathrm{sp}}} f(\boldsymbol{\xi}^q)(1+\sigma\zeta)\psi_i(\boldsymbol{\xi}^q)w^q, \tag{4.12}$$

where ζ is a standard Gaussian random variable and σ is the magnitude of the noise. We repeat each test with 1000 independent sets of noise and present the mean and standard deviation of the L_2 error in Table 1. As σ increases from 10^{-4} to 10^{-3}, the relative L_2 error further increases. Moreover, the second-order expansion yields much larger error than the first-order expansion due to the more inhomogeneous weight distribution.

Table 1

L_2 error of the first- and second-order expansions of f constructed by level-1 (ϵ_1) and level-2 (ϵ_2) sparse grid methods with sampling data superimposed with different magnitudes of numerical error.

σ	ϵ_1	ϵ_2
1×10^{-3}	0.04 ± 0.02	1.1 ± 0.8
5×10^{-4}	0.03 ± 0.01	0.5 ± 0.4
1×10^{-4}	0.029 ± 0.002	0.14 ± 0.09

Similar to the simple numerical example presented above, the surrogate model error of the biomolecular system constructed by the sparse grid method is determined by both the aliasing error and the numerical error on the sampling point. Here we

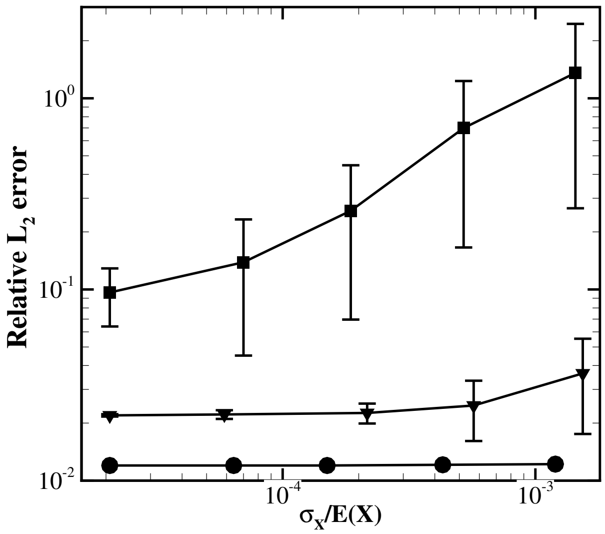


FIG. 11. *Relative L_2 error of the SASA value on residue* 14 *predicted by the first-order gPC expansion constructed by the level-*1 *sparse grid method (*"▼" *symbol), the second-order gPC expansion constructed by the level-*2 *sparse grid method (*"■" *symbol), and the second-order gPC expansion constructed by compressive sensing (*"●" *symbol using* 200 *sample points) under different accuracy levels $\sigma_X/\mathbb{E}(X)$. For each accuracy level,* 32 *sets of independent computations are conducted to compute the L_2 error of the constructed response surface.*

systematically investigate the L_2 error of the surrogate model of the target property X (e.g., the SASA value) on residue P14. The target quantity X on the sampling point is computed under various accuracy levels with relative error from approximately 10^{-5} to 10^{-3}. The different accuracy levels are achieved by choosing different numbers of probe points on the solvent particle when computing the SASA value of the target residue. For each accuracy level, we conduct a random three-dimensional rotation of the molecule and conduct 32 computations of the SASA value on residue P14. We approximate the relative error by $\sigma_X/\mathbb{E}(X)$ with σ_X and $\mathbb{E}(X)$ defined by

$$\sigma_X = \frac{1}{S}\sum_{i=1}^{S}\sigma_{X^i}, \quad \mathbb{E}(X) = \frac{1}{S}\sum_{i=1}^{S}X^i, \tag{4.13}$$

where σ_{X^i} is the standard deviation of 32 independent computation values of X on sample point $\boldsymbol{\xi}^i$ and S is the total number of sample points. We emphasize that σ_X defined by (4.13) is not equal to ϕ (e.g., the difference between the observation and true solution). However, σ_X provides a useful guide for understanding the magnitude of the disturbance on the sample data. We also note that all the numerical results presented in section 4.1 were computed using sampling data with relative error $\sigma_X/\mathbb{E}(X) \approx 5 \times 10^{-5}$.

Figure 11 shows the relative L_2 error of the gPC expansion using the compressive sensing, level-1 sparse grid, and level-2 sparse grid methods. The results from the sparse grid methods are very sensitive to the accuracy level of the sample point. For high accuracy levels, the L_2 error is mainly due to the aliasing error. Increasing $\sigma_X/\mathbb{E}(X)$ from 2×10^{-5} to 1.2×10^{-3}, the mean value of the relative L_2 error increases from 2.20% to 3.36% for the level-1 sparse grid method and from 9.66% to 135.13% for the level-2 sparse grid method. In contrast, the compressive sensing method is insensitive to the imposed error on X for the present system; the resulting error is nearly constant for $\delta \in \left[10^{-5}, 10^{-3}\right]$. This result suggests another advantage of the present method: the present method is more stable in the presence of limited accuracy in the computed target quantity. For high-dimensional systems, the performance of the sparse grid method strongly depends on the accuracy of the evaluation of X at collocation points. Similar phenomena have been reported previously [59]. In practice, it may be computationally infeasible to evaluate X at the accuracy level required for stable sparse grid results. However, our new method based on compressive sensing shows a much weaker dependence on accuracy at individual sample points.

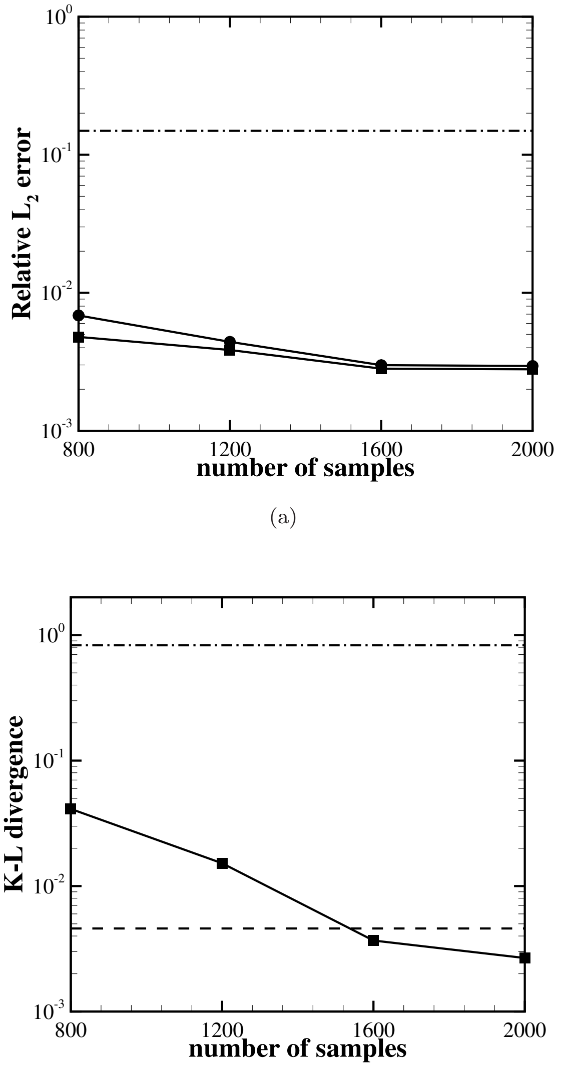
In order to explore the sensitivity of the accuracy level on the constructed gPC expansion, we conducted 32 independent computations on the target quantity X by randomly rotating the biomolecule 32 times on each sampling point. However, we are cautious to claim that numerical error ϕ^i superimposed on the X^i is i.i.d. among the sampling points. The i.i.d. assumptions adopted in (4.7) and (4.12) are used to demonstrate that the numerical error ϕ may further induce error to the constructed gPC expansion. The study presented in this section demonstrates that the sparse grid method may induce relatively large errors to the constructed surrogate model. Rigorous error analysis of the sparse grid method in high-dimensional/complex systems is beyond the scope of this work. However, there appear to be at least two important error sources (aliasing and numerical error on X) that could lead to erroneous results when applying the sparse grid method to high-dimensional systems such as biomolecules. We note that other specific structured or adaptive sparse grid methods [21, 30, 33, 34] may alleviate the instability issue in high-dimensional systems. However, these methods either have less flexibility (the required number of sampling points is fixed for each accuracy level) or require a specialized design for adaptivity criteria.

4.3. Surrogate model for total molecular SASA. Finally, we apply our method to quantify the uncertainty of the total SASA for the entire molecule. Unlike the previous local per-residue SASA, this target quantity depends on the conformation states of all residues. We construct the gPC expansion within the full random space $\mathbb{R}^{168}$. Due to the high dimensionality, we use a second-order gPC expansion with 14365 basis functions. Figure 12 shows the relative L_2 error of the surrogate model and the Kullback–Leibler divergence of the PDFs.

We note that the Hermite basis functions associated with the normal distribution are *unbounded*, which leads to inhomogeneous error distributions in the random space. Figure 13 shows the average error distribution of the surrogate model within different regimes of the SASA value. The average error of the surrogate model of X within $[x_1, x_2]$ is defined by

$$\mathbb{E}\left(\epsilon(x_1, x_2)\right) = \left(\frac{\displaystyle\sum_i (X^{\mathrm{gPC}}(\boldsymbol{\xi}^i) - X(\boldsymbol{\xi}^i))^2 I_{(x_1,x_2)}(X(\boldsymbol{\xi}^i))}{\displaystyle\sum_i I_{(x_1,x_2)}(X(\boldsymbol{\xi}^i))} \right)^{\frac{1}{2}}, \tag{4.14}$$

where $I_{(x_1,x_2)}(X(\boldsymbol{\xi}^i))$ is an indicator function which is 1 if $X(\boldsymbol{\xi}^i) \in [x_1, x_2]$ and 0 otherwise. As shown in Figure 13, the error exhibits a minimum value near the equilibrium state and increases as the X approaches the tails of the SASA PDF. This



(a)

(b)

Fig. 12. (a) *Relative L_2 error of the total molecular SASA by gPC expansion* $\widetilde{X}(\boldsymbol{\xi})$ *(“●”) and* $\widetilde{X}(\boldsymbol{\chi})$ $\widetilde{X}(\boldsymbol{\chi})$*. The dash-dot line represents the relative L_2 error obtained from sampling on level-*1 *sparse grid points. Sampling over the level-*2 *sparse grid points generates erroneous results, as discussed in the text.* (b) *Kullback–Leibler divergence between the PDF of the reference solution and the PDFs obtained from surrogate models* $\widetilde{X}(\boldsymbol{\chi})$ *(“■”), level-*1 *sparse grid (dash-dot line), and direct MD sampling (dash line,* 2400 *sample points).*

demonstrates that the constructed surrogate model is not a *global* approximation of the target quantity X over the *entire* random space. Instead, it provides an approximation of X with respect to the *local* points near equilibrium within the random space. Nevertheless, in practice, we are generally interested in the variation of X with response to conformation fluctuation near the equilibrium state, e.g., the relatively small thermally induced molecular fluctuations considered in the present work.

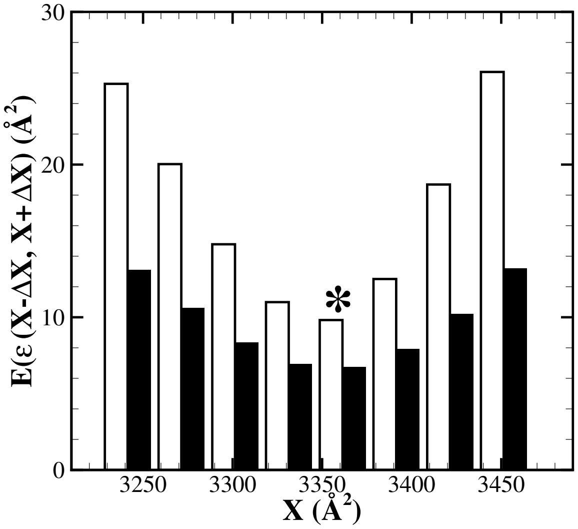


FIG. 13. *The average error distribution of the surrogate models of the total-SASA within different regimes* $[X - \Delta X, X + \Delta X]$. ΔX *is chosen as* 5 $\mathring{A}^2$. *The surrogate models are constructed using* 800 *(blank) and* 1600 *(filled) sample points, respectively. The mean value of total SASA is approximately* 3351 $\mathring{A}^2$ *(denoted by "⋆" symbol), corresponding to conformations near the equilibrium state with respect to the thermal fluctuation.*

Similar to the "local" properties discussed above, the gPC expansion recovered by our compressive sensing method yields the smallest error. However, the advantage of our new method over direct Monte Carlo sampling for the global SASA is not as large as in the case of local properties. By constructing multidimensional basis functions through tensor products of one-dimensional basis functions, the upper bound (here the upper bound exists because the sampling of the Gaussian random variables is truncated in practice) of the basis function becomes larger, which decreases the efficiency of the compressive sensing method. This is similar to the phenomenon observed by others [39, 54]. If only statistical information such as expectation values or PDFs is needed, other methods such as quasi–Monte Carlo [35, 46] may be suitable for high-dimensional systems.

Acknowledgments. We would like to thank Xiaoliang Wan, Wen Zhou, and Tom Goddard for fruitful discussions.

REFERENCES

[1] ℓ_1-*MAGIC*, http://statweb.stanford.edu/~candes/l1magic/.

[2] S. A. Adcock and J. A. McCammon, *Molecular dynamics: Survey of methods for simulating the activity of proteins*, Chem. Rev., 106 (2006), pp. 1589–1615.

[3] E. Alexov, E. L. Mehler, N. A. Baker, A. M. Baptista, Y. Huang, F. Milletti, J. E. Nielsen, D. Farrell, T. Carstensen, M. H. M. Olsson, J. K. Shen, J. Warwicker, S. Williams, and J. M. Word, *Progress in the prediction of pKa values in proteins*, Proteins, 79 (2011), pp. 3260–3275.

[4] A. R. Atilgan, S. R. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar, *Anisotropy of fluctuation dynamics of proteins with an elastic network model*, Biophys. J., 80 (2001), pp. 505–515.

[5] N. A. Baker, *Biomolecular applications of Poisson–Boltzmann methods*, Rev. Comput. Chem.,

21 (2005), pp. 349–379.

[6] B. Brooks and M. Karplus, *Harmonic dynamics of proteins: Normal modes and fluctuations in bovine pancreatic trypsin inhibitor*, Proc. Nat. Acad. Sci. U.S.A., 80 (1983), pp. 6571–6575.

[7] A. M. Bruckstein, D. L. Donoho, and M. Elad, *From sparse solutions of systems of equations to sparse modeling of signals and images*, SIAM Rev., 51 (2009), pp. 34–81.

[8] J. Cai, S. Osher, and Z. Shen, *Convergence of the linearized Bregman iteration for ℓ_1-norm minimization*, Math. Comp., 78 (2009), pp. 2127–2136.

[9] J. Cai, S. Osher, and Z. Shen, *Linearized Bregman iterations for compressed sensing*, Math. Comp., 78 (2009), pp. 1515–1536.

[10] E. J. Candès, *The restricted isometry property and its implications for compressed sensing*, C. R. Math. Acad. Sci. Paris, 346 (2008), pp. 589–592.

[11] E. J. Candès and T. Tao, *Decoding by linear programming*, IEEE Trans. Inform. Theory, 51 (2005), pp. 4203–4215.

[12] M. L. Connolly, *Solvent-accessible surfaces of proteins and nucleic acids*, Science, 221 (1983), pp. 709–713.

[13] P. G. Constantine, E. Dow, and Q. Wang, *Active subspace methods in theory and practice: Applications to kriging surfaces*, SIAM J. Sci. Comput., 36 (2014), pp. A1500–A1524.

[14] CVX Research, Inc., *CVX: Matlab Software for Disciplined Convex Programming*, version 2.0; http://cvxr.com/cvx/ (April 2011).

[15] D. L. Donoho, M. Elad, and V. N. Temlyakov, *Stable recovery of sparse overcomplete representations in the presence of noise*, IEEE Trans. Inform. Theory, 52 (2006), pp. 6–18.

[16] A. Doostan and H. Owhadi, *A non-adapted sparse approximation of PDEs with stochastic inputs*, J. Comput. Phys., 230 (2011), pp. 3015–3034.

[17] R. O. Dror, R. M. Dirks, J. P. Grossman, H. Xu, and D. E. Shaw, *Biomolecular simulation: A computational microscope for molecular biology*, Annu. Rev. Biophys., 41 (2012), pp. 429–452.

[18] J. Foo and G. E. Karniadakis, *Multi-element probabilistic collocation method in high dimensions*, J. Comput. Phys., 229 (2010), pp. 1536–1557.

[19] J. Foo, X. Wan, and G. E. Karniadakis, *The multi-element probabilistic collocation method (ME-PCM): Error analysis and applications*, J. Comput. Phys., 227 (2008), pp. 9572–9595.

[20] B. Ganapathysubramanian and N. Zabaras, *Sparse grid collocation schemes for stochastic natural convection problems*, J. Comput. Phys., 225 (2007), pp. 652–685.

[21] A. Genz and B. D. Keister, *Fully symmetric interpolatory rules for multiple integrals over infinite regions with Gaussian weight*, J. Comput. Appl. Math., 71 (1996), pp. 299–309.

[22] R. G. Ghanem and P. D. Spanos, *Stochastic Finite Elements: A Spectral Approach*, Springer-Verlag, New York, 1991.

[23] N. Go, T. Noguti, and T. Nishikawa, *Dynamics of a small globular protein in terms of low-frequency vibrational modes*, Proc. Nat. Acad. Sci. U.S.A., 80 (1983), pp. 3696–3700.

[24] T. Goldstein and S. Osher, *The split Bregman method for L1-regularized problems*, SIAM J. Imaging Sci., 2 (2009), pp. 323–343.

[25] T. Haliloglu, I. Bahar, and B. Erman, *Gaussian dynamics of folded proteins*, Phys. Rev. Lett., 79 (1997), pp. 3090–3093.

[26] R. C. Harris, A. H. Boschitsch, and M. O. Fenley, *Influence of grid spacing in Poisson–Boltzmann equation binding energy estimation*, J. Chem. Theory Comput., 9 (2013), pp. 3677–3685.

[27] B. Lee and F. M. Richards, *The interpretation of protein structures: Estimation of static accessibility*, J. Mol. Biol., 55 (1971), pp. 379–400.

[28] M. Levitt, C. Sander, and P. S. Stern, *The normal modes of a protein: Native bovine pancreatic trypsin inhibitor*, Int. J. Quant. Chem.: Quantum Biology Symposium, 10 (1983), pp. 181–199.

[29] X. Li, *Finding deterministic solution from underdetermined equation: Large-scale performance variability modeling of analog/RF circuits*, IEEE Trans. Comput.-Aided Design Integr. Circuits Syst., 29 (2010), pp. 1661–1668.

[30] X. Ma and N. Zabaras, *An adaptive hierarchical sparse grid collocation algorithm for the solution of stochastic differential equations*, J. Comput. Phys., 228 (2009), pp. 3084–3113.

[31] X. Ma and N. Zabaras, *An adaptive high-dimensional stochastic model representation technique for the solution of stochastic partial differential equations*, J. Comput. Phys., 229 (2010), pp. 3884–3915.

[32] A. McCammon and S. C. Harvey, *Dynamics of Protein and Nucleic Acids*, Cambridge University Press, Cambridge, UK, 1987.

[33] A. Narayan and D. Xiu, *Stochastic collocation methods on unstructured grids in high dimensions via interpolation*, SIAM J. Sci. Comput., 34 (2012), pp. A1729–A1752.

[34] A. Narayan and D. Xiu, *Constructing nested nodal sets for multivariate polynomial interpolation*, SIAM J. Sci. Comput., 35 (2013), pp. A2293–A2315.

[35] H. Niederreiter, *Random Number Generation and Quasi-Monte Carlo Methods*, SIAM, Philadelphia, 1992.

[36] F. Nobile, R. Tempone, and C. G. Webster, *An anisotropic sparse grid stochastic collocation method for partial differential equations with random input data*, SIAM J. Numer. Anal., 46 (2008), pp. 2411–2442.

[37] K. Petras, *Smolpack: A Software for Smolyak Quadrature with Clenshaw-Curtis Basis-Sequence*, http://people.sc.fsu.edu/~jburkardt/c_src/smolpack/smolpack.html (2003).

[38] J. W. Ponder and D. A. Case, *Force fields for protein simulations*, in Advances in Protein Chemistry, Vol. 66, Elsevier, New York, 2003, pp. 27–85.

[39] H. Rauhut and R. Ward, *Sparse Legendre expansions via l_1-minimization*, J. Approx. Theory, 164 (2012), pp. 517–533.

[40] P. Ren, J. Chun, D. G. Thomas, M. J. Schnieders, M. Marucho, J. Zhang, and N. A. Baker, *Biomolecular electrostatics and solvation: A computational perspective*, Q. Rev. Biophys., 45 (2012), pp. 427–491.

[41] T. J. Richmond, *Solvent accessible surface area and excluded volume in proteins*, J. Mol. Biol., 178 (1984), pp. 63–89.

[42] F. Rizzi, H. N. Najm, B. J. Debusschere, K. Sargsyan, M. Salloum, H. Adalsteinsson, and O. M. Knio, *Uncertainty quantification in MD simulations. Part* I: *Forward propagation*, Multiscale Model. Simul., 10 (2012), pp. 1428–1459.

[43] F. Rizzi, H. N. Najm, B. J. Debusschere, K. Sargsyan, M. Salloum, H. Adalsteinsson, and O. M. Knio, *Uncertainty quantification in MD simulations. Part* II: *Bayesian inference of force-field parameters*, Multiscale Model. Simul., 10 (2012), pp. 1460–1492.

[44] B. Roux and T. Simonson, *Implicit solvent models*, Biophys. Chem., 78 (1999), pp. 1–20.

[45] A. Shrake and J. A. Rupley, *Environment and exposure to solvent of protein atoms. Lysozyme and insulin*, J. Mol. Biol., 79 (1973), pp. 351–371.

[46] I. H. Sloan and S. Joe, *Lattice Methods for Multiple Integration*, Oxford University Press, New York, 1994.

[47] F. Tama and Y.-H. Sanejouand, *Conformational change of proteins arising from normal mode calculations*, Protein Eng., 14 (2001), pp. 1–6.

[48] R. Tipireddy and R. Ghanem, *Basis adaptation in homogeneous chaos spaces*, J. Comput. Phys., 259 (2014), pp. 304–317.

[49] M. M. Tirion, *Large amplitude elastic motions in proteins from a single-parameter, atomic analysis*, Phys. Rev. Lett., 77 (1996), pp. 1905–1908.

[50] E. van den Berg and M. P. Friedlander, *Probing the Pareto frontier for basis pursuit solutions*, SIAM J. Sci. Comput., 31 (2008), pp. 890–912.

[51] A. Wlodawer, J. Walter, R. Huber, and L. Sjölin, *Structure of bovine pancreatic trypsin inhibitor. Results of joint neutron and x-ray refinement of crystal form* II, J. Mol. Biol., 180 (1984), pp. 301–329.

[52] D. Xiu and J. S. Hesthaven, *High-order collocation methods for differential equations with random inputs*, SIAM J. Sci. Comput., 27 (2005), pp. 1118–1139.

[53] D. Xiu and G. E. Karniadakis, *The Wiener–Askey polynomial chaos for stochastic differential equations*, SIAM J. Sci. Comput., 24 (2002), pp. 619–644.

[54] L. Yan, L. Guo, and D. Xiu, *Stochastic collocation algorithms using l_1-minimization*, Int. J. Uncertain. Quantif., 2 (2012), pp. 279–293.

[55] X. Yang, M. Choi, G. Lin, and G. E. Karniadakis, *Adaptive ANOVA decomposition of stochastic incompressible and compressible flows*, J. Comput. Phys., 231 (2012), pp. 1587–1614.

[56] X. Yang and G. E. Karniadakis, *Reweighted ℓ_1 minimization method for stochastic elliptic differential equations*, J. Comput. Phys., 248 (2013), pp. 87–108.

[57] W. Yin, S. Osher, D. Goldfarb, and J. Darbon, *Bregman iterative algorithms for ℓ_1-minimization with applications to compressed sensing*, SIAM J. Imaging Sci., 1 (2008), pp. 143–168.

[58] Z. Zhang, M. Choi, and G. E. Karniadakis, *Error estimates for the ANOVA method with polynomial chaos interpolation: Tensor product functions*, SIAM J. Sci. Comput., 34 (2012), pp. A1165–A1186.

[59] Z. Zhang, M. V. Tretyakov, B. Rozovskii, and G. E. Karniadakis, *A recursive sparse grid collocation method for differential equations with white noise*, SIAM J. Sci. Comput., 36 (2014), pp. A1652–A1677.