QUANTIFYING THE INFLUENCE OF CONFORMATIONAL UNCERTAINTY IN BIOMOLECULAR SOLVATION *

HUAN LEI† XIU YANG‡ BIN ZHENG† GUANG LIN† AND NATHAN 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. We also 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. Further more, 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

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 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 [16, 2]. Over the past few decades, this approach has made great progress in the development of accurate empirical force field as well as efficient simulation algorithms [40]. 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 non-ergodic 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 [50, 27, 48]. In particular, by only modeling interaction between the neighboring α-carbon (C_α), ENMs are able to predict structural fluctuations (e.g., Debye-Waller or B-factors) with surprising accuracy [27, 4].

*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 DOE under Contract DE-AC05-76RL01830. We would like to thank Xiaoliang Wan, Wen Zhou and Tom Goddard for fruitful discussions. H. Lei acknowledge travel grant from IMA for workshop Uncertainty Quantification in Materials Modeling and travel grant from DOE for Conference on Data Analysis 2014.

†Pacific Northwest National Laboratory, Richland, Washington WA 99352, USA

‡Division of Applied Mathematics, Brown University, Providence, RI 02912, USA

§email: Nathan.Baker@pnnl.gov

The simplified potentials used by CG models such as ENM allow us to examine structural fluctuations in a semi-analytical manner. However, unlike MD simulations, these CG models do not provide a direct approach for calculating the uncertainty these fluctuations induce on various biomolecular properties of interest. This leads to an important practical question: how do we utilize the stochastic information inferred from these models to quantify uncertainty of the various biological related properties? In many applications, a single native conformation of a molecule is used when computing a range of properties such as molecular volume and area [29, 43, 11], electrostatic and solvation properties [42, 46], 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 MD simulations.

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 conformational state distribution. The distribution of values for the target property can be calculated directly from this conformational state distribution without the need for additional MD simulations. To the best of our knowledge, this is the first demonstration of how a target property response surface – including property uncertainty – can be directly calculated from the biomolecular conformation distribution.

To construct the surrogate model, we adopt the generalized polynomial chaos (gPC) [22, 54] and formulate the target property as an expansion of a set of gPC basis functions determined by the specific conformation states. Within this framework, we can formulate conformational uncertainty 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 (PCM) such as ANOVA [33, 17, 59, 56] and sparse grid methods [53, 19, 18, 38, 32] 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 [39]). The second barrier is the presence of the 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, 28]. The error between the true value and the computed value 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 non-collocation method based on compressive sensing [10, 14, 7] for high-dimensional biomolecular systems which reduces the influence of the limited accuracy of the target property while increasing 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 [31, 15, 55, 57].

2. Stochastic model. In this section, we introduce a semi-analytical stochastic model representing the thermal fluctuations of a biomolecule around its native state. Previous studies [50, 27, 48] have demonstrated that low-frequency conformational modes can be accurately reproduced by a coarse-grained ENM where neighboring particles interact via a simple harmonic potential. 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 coarse-grained ENM, we note that further dimensionality reduction can be obtained for biomolecular target properties that have local dependence on structure; i.e., the values associated with a particular property may depend only on a subset of atoms in the molecule.

2.1. Full stochastic model of conformation fluctuation. We construct the stochastic conformation space of the bio-molecular system based on the coarse-grained (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. ANM can be viewed as a simplified CG model of normal mode analysis [23, 6, 30, 34], where the model potential does not rely on the complex atomic-detail force field. Consider a bio-molecule of N residues with equilibrium positions $\overline{\mathbf{R}}^T = \left[\overline{\mathbf{r}}_1^T\ \overline{\mathbf{r}}_2^T \cdots\ \overline{\mathbf{r}}_N^T\right]$, where $\overline{\mathbf{r}}_i$ is a 3-dimensional vector representing the equilibrium position of residue i. The harmonic approximation for the potential energy V with respect to the instantaneous position $\mathbf{R}^T = \left[\mathbf{r}_1^T\ \mathbf{r}_2^T \cdots\ \mathbf{r}_N^T\right]$ is given by

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

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

Given the potential defined by Eq. (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, Y and Z represent the 3-dimensional Cartesian coordinates of residues i and j. This harmonic form for the potential leads to Gaussian statistics for the conformational probability distribution [4] such that the fluctuation correlation between individual residue i and j can be determined by the pseudo-inverse of the Hessian matrix $\mathbf{H}$ as

$$\left\langle \Delta\mathbf{R}_i \Delta\mathbf{R}_j^T \right\rangle = \frac{k_B T}{\gamma} \left[\mathbf{H}^{-1}\right]_{ij}, \tag{2.2}$$

where $\Delta\mathbf{R}_i = \mathbf{r}_i - \overline{\mathbf{r}}_i$, k_B is the Boltzmann constant, T is the temperature, and $\left[\mathbf{H}^{-1}\right]_{ij}$ is a 3×3 matrix with the elements corresponding to the submatrix of the pseudo-inverse of the Hessian matrix $\mathbf{H}$ with rows from $3i-2$ to $3i$ and columns from $3j-2$ to $3j$.

To construct the stochastic conformation space, we perform an eigendecomposition of $\mathbf{H}$

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

where λ_i is the *i-th* 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 *i-th* eigenvector of $\mathbf{H}$. Following Eq. (2.2), the fluctuation correlation matrix $\mathbf{C}$ can be determined by

$$\mathbf{C}_{ij} \equiv \left\langle \Delta\mathbf{R}_i \Delta\mathbf{R}_j^T \right\rangle \tag{2.5a}$$

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

where $\mathbf{U} = \left(\frac{k_B T}{\gamma}\right)^{\frac{1}{2}} \mathbf{W}\mathbf{\Lambda}^{-\frac{1}{2}}$. Given the fluctuation correlation matrix $\mathbf{C}$, we can construct the stochastic conformation space 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 $\boldsymbol{\xi} = (\xi_1, \xi_2, \cdots, \xi_{3N-6})$ is an independent and identically distributed (i.i.d.) standard Gaussian random vector. Given a value of $\boldsymbol{\xi}$, the corresponding conformation is fully determined (on the CG level), allowing us to calculate related target properties, denoted by $X(\boldsymbol{\xi})$.

2.2. Reduced-dimensionality stochastic model for conformational fluctuations. Some target properties of interest, particularly those related to a specific residue, may have only local dependence on the molecular conformation rather than depending on all degrees of freedom in the biomolecule. We denote such quantities as "local" properties $X^{\{p\}}$ if the target quantity $X^{\{p\}}$ on residue p is independent of the position of residue j, i.e., $\frac{\partial X^{\{p\}}}{\partial \mathbf{R}_j} = 0$. Moreover, we note that this locality implies $\nabla_{\mathbf{R}} X^{\{p\}}$ could be sparse and have reduced dimension if most of the entry of $\nabla_{\mathbf{R}} X^{\{p\}}$ is zero or close to zero. if $X^{\{p\}}$ on residue p is independent of the position of residue j. For example, the solvent accessible surface area (SASA) of a specific residue p should only depend on the positions of residue p as well as its neighboring residues within a certain cutoff distance r_c. Under such circumstances, the full position fluctuation correlation matrix $\mathbf{C}$ can be reduced to a $3N \times 3N$ matrix $\mathbf{C}'$ with larger sparsity, i.e., the (i, j) element (a 3×3 matrix) is given by

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

where $r_{ip} = |\mathbf{r}_p - \mathbf{r}_i|$, $r_{jp} = |\mathbf{r}_p - \mathbf{r}_j|$, and r_c^p is a cut-off distance of residue p beyond which $X^{\{p\}}$ is independent of the residue i.

Fig. 2.1 illustrates the dimensionality reduction procedure for local properties as discussed above. The lefthand matrix represents the full property correlation matrix $\mathbf{C}$ where each block element (i, j) represents a 3×3 property correlation matrix between individual residues i and j. We color the non-zero blocks in blue and set the rest of the elements to zero. For a specific residue, its local property only depends on neighboring residues within a cut-off distance r_c^p and obtain a correlation matrix $\mathbf{C}^{\{p\}}$ with reduced dimension $d = 3\sum_i^N h(r_c^p - r_{ip})$. Similar to Eq. (2.6), we 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. standard 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 analysis, as introduced in next section. Without loss of generality, we note that many properties of interest can be local and therefore omit the superscript $\{p\}$ in the rest of this manuscript, using $X(\boldsymbol{\xi})$ 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 generalized polynomial chaos (gPC) expansion with a brief discussion on possible difficulties with popular collocation methods.

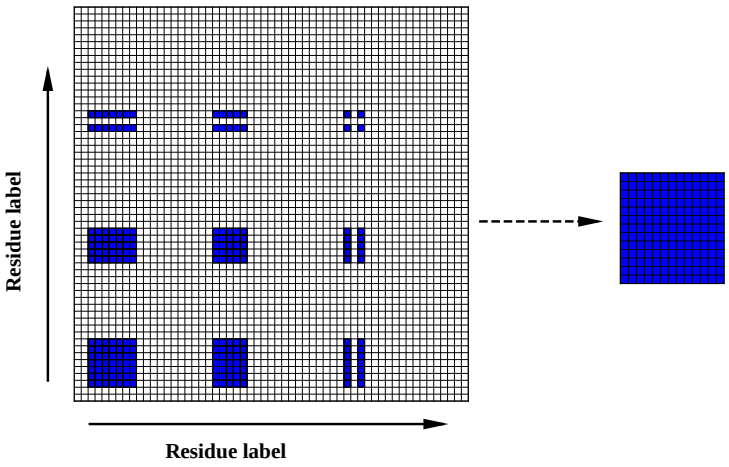


FIG. 2.1. *Sketch of a typical reduced property correlation matrix. The square on the left hand side represents the full correlation matrix. Each block represents* 9 *elements (in x, y and z directions) of a residue in the correlation matrix* $\mathbf{C}$*. The blocks in blue color represent the matrix 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.*

Next, we introduce a non-collocation method to construct the gPC expansion based on compressive sensing. We note that the performance of the compressive sensing method depends on the sparsity of the gPC coefficient. 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. We use the gPC expansion to construct the response surface of the target property X with respect to the stochastic model parameter $\boldsymbol{\xi}$, the molecular conformation. However, we note that the computated result of X is generally subject to finite precision and accuracy, e.g.,

$$X(\boldsymbol{\xi}) = \bar{X}(\boldsymbol{\xi}) + \phi, \tag{3.1}$$

where $\bar{X}(\boldsymbol{\xi})$ is the true value of the target property, $X(\boldsymbol{\xi})$ is the numerically computed value, and ϕ represents the difference between $\bar{X}(\boldsymbol{\xi})$ and $X(\boldsymbol{\xi})$. We assume that

$$|\phi| \ll |X| \tag{3.2}$$

and use a gPC expansion to represent the target property X by

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

$$\psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi}) = \psi_{\alpha^1}(\xi_1)\psi_{\alpha^2}(\xi_2)\cdots\psi_{\alpha^d}(\xi_d), \quad \alpha^i \in \mathbb{N} \cup \{0\}, \tag{3.3b}$$

where $\boldsymbol{\alpha}$ is a multi-index and $c_{\boldsymbol{\alpha}}$ is the gPC coefficient to be determined. We choose $\psi_{\boldsymbol{\alpha}}$ as the d-dimensional orthonormal Hermite polynomials, which is consistent with the d-dimensional Gaussian distribution used to model conformational fluctuations in ENM, as discussed in Sec. 2.1. We truncate the expression (3.3) up to polynomial order P, hence X is approximated as:

$$X(\boldsymbol{\xi}) \approx \widetilde{X}(\boldsymbol{\xi}) = \sum_{|\boldsymbol{\alpha}|=0}^{P} c_{\boldsymbol{\alpha}} \psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi}), \tag{3.4}$$

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

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

$$c_\alpha = \frac{\int X(\boldsymbol{\xi})\psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi})\mathrm{d}\mathcal{P}(\boldsymbol{\xi})}{\int \psi^2_{\boldsymbol{\alpha}}(\boldsymbol{\xi})\mathrm{d}\mathcal{P}(\boldsymbol{\xi})}, \tag{3.5}$$

where the integration can be accomplished by using probabilistic collocation approaches such as tensor product [44, 45] or sparse grid [53, 19, 18] methods. Numerical computation of X is performed on specific collocation sampling points $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \cdots, \boldsymbol{\xi}_Q$, e.g.,

$$\int X(\boldsymbol{\xi})\psi_\alpha(\boldsymbol{\xi})\mathrm{d}\mathcal{P}(\boldsymbol{\xi}) \approx \sum_{i=1}^{Q} X_i \psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi}_i) w_i, \tag{3.6}$$

where w_i is the corresponding weight associated with collocation point $\boldsymbol{\xi}_i$.

However, for the high-dimensional biomolecular systems considered the present work, the required number of collocation sampling points Q can be huge computationally intractable. For example, a small biomolecular system with a 27-dimensional reduced conformation random space would require $Q = 7.6\times10^{12}$ tensor product collocation points to construct a quadratic-order gPC expansion. Standard sparse grid method based on Gaussian quadrature and Smolyak construction reduce this space 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 3rd-order approximation is 27829) which makes it difficult to incorporate adaptive sampling strategies. Moreover, as will be shown in Sec. 4, the numerical error ϕ associated with X may lead to erroneous approximation of $c_{\boldsymbol{\alpha}}$ even if Eq. (3.2) is satisfied. To overcome the above difficulties, we compute the gPC expansion by applying compressive sensing as described in Sec. 3.2.

3.2. Compressive sensing method. To construct the gPC expansion in Eq. (3.4), we compute $X(\boldsymbol{\xi})$ on M sampling points $(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \cdots, \boldsymbol{\xi}_M)$ generated by probability measure $\mathcal{P}(\boldsymbol{\xi})$, which is the d-dimensional i.i.d. standard normal distribution in the present work. We discretize Eq. (3.4) as a linear system

$$\begin{pmatrix} \psi_{\boldsymbol{\alpha}_1}(\boldsymbol{\xi}_1) & \psi_{\boldsymbol{\alpha}_2}(\boldsymbol{\xi}_1) & \cdots \\ \psi_{\boldsymbol{\alpha}_1}(\boldsymbol{\xi}_2) & \psi_{\boldsymbol{\alpha}_2}(\boldsymbol{\xi}_2) & \cdots \\ & \vdots & \end{pmatrix} \begin{pmatrix} c_{\boldsymbol{\alpha}_1} \\ c_{\boldsymbol{\alpha}_2} \\ \vdots \end{pmatrix} = \begin{pmatrix} X(\boldsymbol{\xi}_1) \\ X(\boldsymbol{\xi}_2) \\ \vdots \end{pmatrix} + \boldsymbol{\varepsilon},$$

or equivalently,

$$\boldsymbol{\Psi c} = \boldsymbol{X} + \boldsymbol{\varepsilon}, \tag{3.7}$$

where $\boldsymbol{\Psi}$ is the "measurement matrix" with entries $\Psi_{i,j} = \psi_{\boldsymbol{\alpha}_j}(\boldsymbol{\xi}_i)$, $\boldsymbol{c}$ is the vector of the gPC coefficients, $\boldsymbol{X}$ is the vector consists of the outputs and $\boldsymbol{\varepsilon}$ is related to the truncation error and numerical error ϕ.

It has been proven [10, 13] that, given the knowledge $\boldsymbol{c}$ is sparse, it is possible to solve the underdetermined system $\boldsymbol{\Psi c} = \boldsymbol{y}$ via solution of the optimization problem (P_h)

$$(P_h): \qquad \min_{\boldsymbol{c}} \|\boldsymbol{c}\|_h \quad \text{subject to} \quad \boldsymbol{\Psi c} = \boldsymbol{u}, \tag{3.8}$$

where h is usually taken as 0 or 1. The "ℓ_0 norm" is defined as the non-zeros entries of a vector, $\|\boldsymbol{c}\|_0 \stackrel{\text{def}}{=} |\{i : c_i \neq 0\}|$, and the ℓ_1 norm is defined as $\|\boldsymbol{c}\|_1 \stackrel{\text{def}}{=} \sum_i |c_i|$. Solution of this problem by combinatorial search (i.e., sweeping exhaustively through all possible sparse subsets) has computational complexity that is exponential in m, the length of $\boldsymbol{c}$. Instead, the most popular method for solution is to relax the ℓ_0 minimization (P_0) to a ℓ_1 minimization (P_1), which is a convex optimization problem. If the observation is superimposed with numerical error/noise, i.e.,

$$\boldsymbol{y} = \boldsymbol{\Psi c} + \boldsymbol{\varepsilon},$$

then we need to modify (P_h) to be

$$(P_{h,\epsilon}): \qquad \min_c \|\boldsymbol{c}\|_h \quad \text{subject to} \quad \|\boldsymbol{\Psi c} - \boldsymbol{u}\|_2 \leq \delta, \tag{3.9}$$

where $\delta = \|\varepsilon\|_2$ is the truncation error tolerance. The $(P_{1,\epsilon})$ optimization problem can be solved using classical convex optimization solvers (e.g., `CVX` [25]), sparse recovery software packages (e.g., `SPGL1` package [51], ℓ_1-`MAGIC` [1]), or the split Bregman method [24, 58, 9, 8]. In this paper we use `SPGL1`.

To solve Eq. (3.9), we need the value of δ, which is generally not known *a priori*. In this work, we estimate δ using a cross-validation method [15] We first divide M sampling data into two parts denoted by M_r (white color) and M_v (black color). Second, $\boldsymbol{c}$ is computed with M_r sample points with a chosen tolerance error δ_r. Next, an optimized estimate $\hat{\delta}_r$ is determined such that $\|\boldsymbol{\Psi}^{M_v}\boldsymbol{c} - \boldsymbol{X}^{M_v}\|_2$ is minimized. Finally, we repeat the above process for different replicas of the sample points and determine a new optimized δ as $\delta = \sqrt{M/M_r}\hat{\delta}_r$. In this work, we set $M_r = 2M/3$ and performed the cross-validation for three replications.

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 $n+1$, as well as the sparsity of the linear system in Eq. (3.7). In general, accuracy improves with either larger M/n ratios or sparser target vectors $\boldsymbol{c}$. A straightforward way to increase M/n is to reduce the dimension of stochastic space. 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 et al. [12] have developed an alternative approach that focuses on increasing sparsity by analysis of variability in the target properties. For the target $X(\boldsymbol{\xi})$, we define gradient matrix $\mathbf{G}$ following [12] by

$$\mathbf{G} = \mathbb{E}\left[\nabla X(\boldsymbol{\xi})\nabla X(\boldsymbol{\xi})^T\right] \tag{3.10}$$

where $\nabla X(\boldsymbol{\xi})$ is the gradient vector defined by $\nabla X(\boldsymbol{\xi}) = \left[\frac{\partial X}{\partial \xi_1} \cdots \frac{\partial X}{\partial \xi_d}\right]$. We conduct the eigendecomposition

$$\mathbf{G} = \mathbf{Q}\mathbf{K}\mathbf{Q}^T, \qquad \mathbf{Q} = [\mathbf{q}_1\ \mathbf{q}_2 \cdots\ \mathbf{q}_d], \tag{3.11a}$$

$$\mathbf{K} = \mathrm{diag}(k_1, \cdots, k_d), \quad k_1 \geq \cdots \geq k_d \geq 0, \tag{3.11b}$$

where $\mathbf{q}_i$ is the i-th eigenvector of $\mathbf{G}$. Therefore, within the probability measure $\mathcal{P}(\boldsymbol{\xi})$, the target property X exhibits the largest variability along the direction $\mathbf{q}_1$ while it exhibits the least variability along the direction $\mathbf{q}_d$. This rotation motivates the definition of a new random vector

$$\boldsymbol{\chi} = \mathbf{Q}\boldsymbol{\xi}, \tag{3.12}$$

where Q is unitary and $\boldsymbol{\chi}$ is i.i.d. (also similar to Ref. [49]). 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 while the gPC coefficients associated with other variables exhibiting much smaller value (or even close to 0), yielding larger sparsity than the random vector $\boldsymbol{\xi}$ for the linear system defined in Eq. (3.7). Hence, if we recover the gPC coefficients with respect to $\boldsymbol{\chi}$ in Eq. (3.7) with the compressive sensing method, we expect more accurate result than directly recovering from $\boldsymbol{\xi}$.

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 via the gPC expansion recovered from $\boldsymbol{\xi}$, e.g.

$$\mathbf{G} \approx \mathbb{E}\left[\nabla X^{\mathrm{gPC}}(\boldsymbol{\xi})\nabla X^{\mathrm{gPC}}(\boldsymbol{\xi})^T\right] \tag{3.13a}$$

$$X^{\mathrm{gPC}}(\boldsymbol{\xi}) = \sum_{|\boldsymbol{\alpha}|=0}^{P} c_{\boldsymbol{\alpha}}^{\{\boldsymbol{\xi}\}} \psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi}), \tag{3.13b}$$

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 in $\mathcal{P}(\boldsymbol{\xi})$ and can be used to define $\mathbf{Q}$ and, therefore, the new random basis $\boldsymbol{\chi} = \mathbf{Q}\boldsymbol{\xi}$. Finally, this new basis can be used to re-construct the gPC expansion of X with respect to the $\boldsymbol{\chi}$ which, in general, yields an expansion of larger 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 vectors is not universal, it depends on the specific target property X.*

REMARK 3.2. *The gradient matrix $\mathbf{G}$ is approximated by Eq. (3.13). Therefore, eigenvectors $[\mathbf{r}_1\ \mathbf{r}_2 \cdots\ \mathbf{r}_d]$ may not correspond exactly to the steepest decay directions of variability for the target property X. Nevertheless, we adopt Eq. (3.13) to construct a "rotated" space that provides larger (if not optimal) sparsity.*

We summarize the entire procedure presented above (Sections 2 and 3) in Algorithm 1 and, in the next Section, 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, construct the full stochastic conformation space given (Eq. 2.6). For "local" target properties, further reduce the dimension of the stochastic conformation space (Eq. 2.8). Denote the parameter of the stochastic space as a d-dimensional i.i.d. standard normal random vector $\boldsymbol{\xi}$.

Step 2. Generate M sampling points $\boldsymbol{\xi}^1, \boldsymbol{\xi}^2, \cdots, \boldsymbol{\xi}^M$ based on the distribution of $\boldsymbol{\xi}$. Numerically compute X on $\boldsymbol{\xi}^1, \boldsymbol{\xi}^2, \cdots, \boldsymbol{\xi}^M$ to obtain M outputs $X^1, X^2, \cdots, X^M$. Denote $\boldsymbol{X} = (X^1, X^2, \cdots, X^M)$ as the "observation" in $(P_{1,\delta})$. 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 (3.4).

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

Step 4. Solve the weighted ℓ_1 minimization problem

$$\boldsymbol{c} = \arg\min \|\boldsymbol{c}\|_1 \quad \textit{subject to} \quad \|\boldsymbol{\Psi}\boldsymbol{c} - \boldsymbol{u}\|_2 \leq \delta.$$

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

Step 5. Evaluate the gradient matrix $\mathbb{E}\left[\nabla X^{\mathrm{gPC}}(\boldsymbol{\xi}) \nabla X^{\mathrm{gPC}}(\boldsymbol{\xi})^T\right]$, given $\boldsymbol{c}^{\{\boldsymbol{\xi}\}}$ and define the random vector $\boldsymbol{\chi}$ by Eq. (3.12).

Step 6. Construct the gPC expansion of $X(\boldsymbol{\chi})$ by repeating steps 3-4 on random vector $\boldsymbol{\chi}$ by utilizing the M sampling data $(X^1, X^2, \cdots, X^M)$ that have been determined in step 2.

4. Numerical Results. As an example, we apply our method to quantify the uncertainty in solvent-accessible surface area (SASA) caused by conformational fluctuations in the biomolecule bovine pancreatic trypsin inhibitor (PDB code: 5pti) [52], shown in Fig. 4.1 SASA is an essential element of numerous solvation models [5, 46, 42] and the total area can be decomposed into residue-specific contributions, allowing us to explore the influence of conformational fluctuations on local area uncertainty. 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 points. This performance is assessed relative to two reference systems: a direct Monte Carlo simulation of the conformational space with 10^6 sampling points as

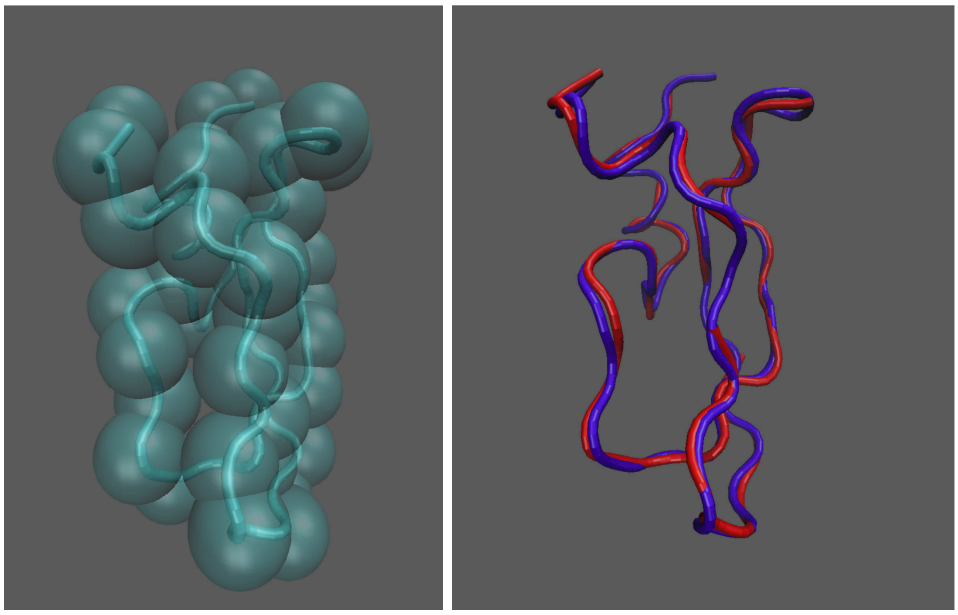


FIG. 4.1. *(a) Tube diagram of the equilibrium structure of bovine pancreatic trypsin inhibitor (PDB code: 5pti) with spheres denoting the residue C_α positions. (b) Tube diagrams of the molecule representing instantaneous conformational states under thermal fluctuation.*

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 probability density functions obtained from this new approach and the reference data.

4.1. Surrogate model for SASA of individual residues. Fig. 4.1 shows a sketch of the CG biomolecular model under equilibrium and thermal-fluctuation states. Following Ref. [4], each residue is modeled as a single α-carbon particle as shown in Fig. 4.1(a). Due to thermal fluctuations, the molecule exhibits a distribution of conformation states where individual residues may deviate from the equilibrium positions, as shown in Fig. 4.1(b). To model the fluctuation of individual residues, we construct the ANM correlation matrix $\mathbf{C}$ by Eq. (2.5) using a cut-off distance for the harmonic potential of $r_c = 9.8$Å. The radius values 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 Eq. (2.7) by setting the neighbor cut-off distance r_c^p to be 9.5 Å. This cutoff value yields 8 neighboring residues and therefore a $27-$dimensional random space $\mathbb{R}^{27}(\boldsymbol{\xi})$ by Eq. (2.8). We examine the constructed local random space through the probability density function (PDF) of the SASA. As shown in Fig. 4.2, 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 Fig. 4.2 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 predicted by the full correlation matrix which is not surprising since the off-diagonal elements represent the harmonic potential contribution of molecular deformation in Eq. (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 Sec. 3. First, we calculate the gPC coefficients $c_\alpha^{\{\boldsymbol{\xi}\}}$ up to order $\mathbf{P} = 2$ (406 basis functions) by using 300 points and applying setp 4 in Algorithm

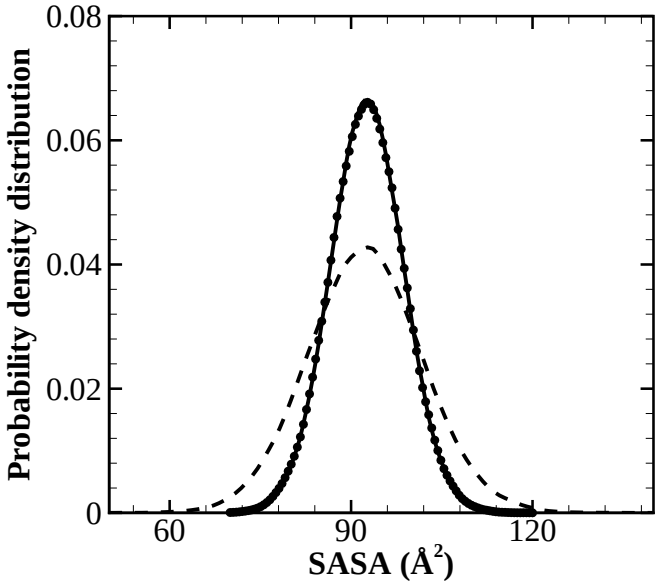


FIG. 4.2. *Probability density function of the SASA of the 14th residue obtained from the full correlation matrix* $\mathbf{C}$ *(solid line) and the local reduced correlation matrix* $\mathbf{C}'$ *("•" symbol). The dashed line represents the distribution obtained from the reduced correlation matrix where off-diagonal elements are set to zero.*

1. Given $c_\alpha^{\{\boldsymbol{\xi}\}}$, we next construct the approximate gradient matrix $\mathbf{G}$ by Eq. (3.13). Eigendecomposition of this matrix provides a set of rotated random variables $\boldsymbol{\chi}$ by Eq. (3.12). Fig. 4.3 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. Therefore, although the dimension of the random space is fixed by the molecular structure, we are still able to define a set of vectors (i.e., the eigenvectors of $\mathbf{G}$) such that the variability of X decays faster along these directions.

Given the variables $\boldsymbol{\chi}$, we compute the corresponding gPC coefficients $c_\alpha^{\{\boldsymbol{\chi}\}}$ with order $\mathbf{P} = 2$ on the same 300 points used for $c_\alpha^{\{\boldsymbol{\xi}\}}$. The results are shown in Fig. 4.4. Compared with $c_\alpha^{\{\boldsymbol{\chi}\}}$, the spectrum of $c_\alpha^{\{\boldsymbol{\chi}\}}$ exhibits a higher degree of sparsity, as expected. This result indicates that, with the same polynomial order, the target quantity X can be represented 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

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

where $\tilde{X}$ is the gPC expansion of X by Eq. (3.4) with $c_\alpha^{\{\boldsymbol{\xi}\}}$ and $c_\alpha^{\{\boldsymbol{\chi}\}}$, respectively. As $X(\boldsymbol{\xi})$ is

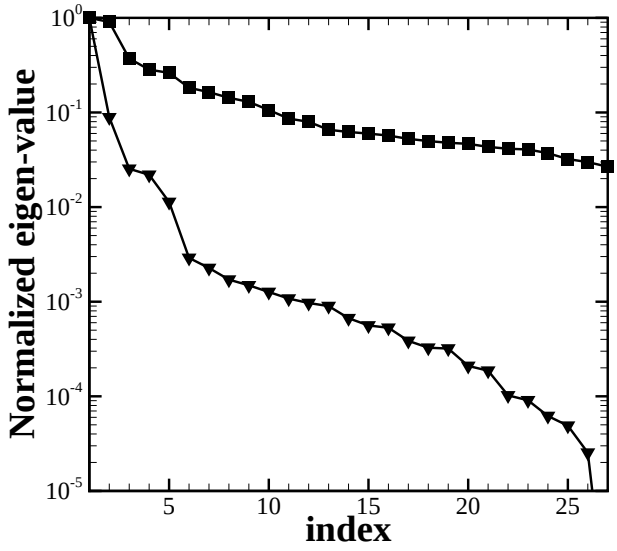


FIG. 4.3. *Normalized eigenvalues of the gradient matrix* $\mathbf{G}$ *("▼" symbol) and the correlation matrix* $\mathbf{C}'$ *("■" symbol).*

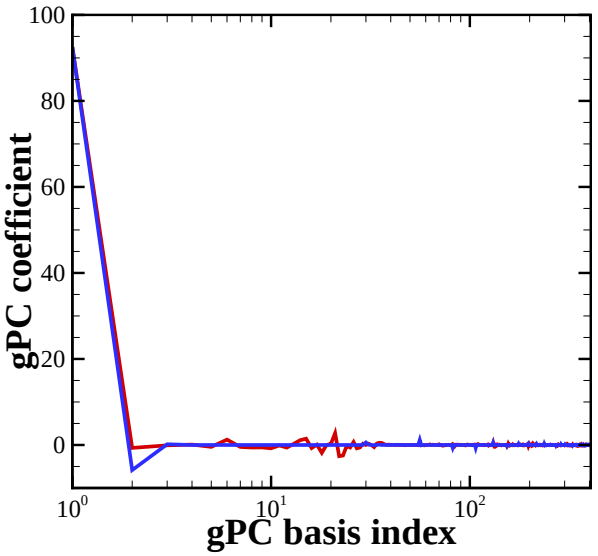
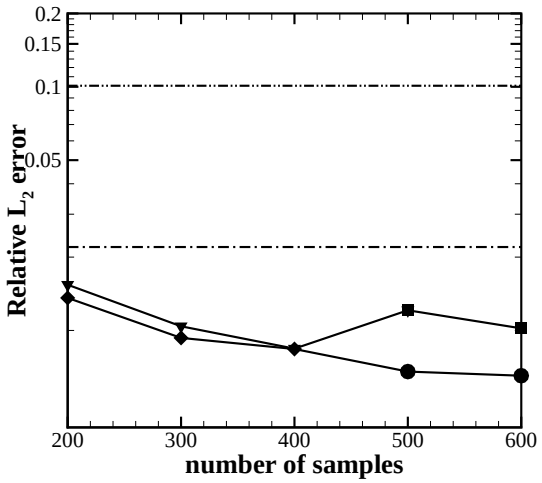


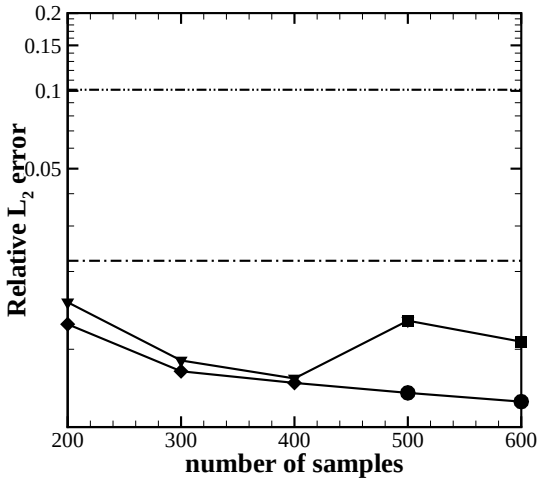
FIG. 4.4. *gPC coefficients (up to 2nd order) for the SASA value on the 14th residue obtained from CS method with respect to random vector* ξ *(red) and* χ *(blue) with dimension* $d = 27$.

unknown in general, we use Monte Carlo sampling to approximate the integral in Eq. (4.2)

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



(a)



(b)

FIG. 4.5. *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 2nd and 3rd order gPC expansion by $\boldsymbol{\xi}$. The symbols "◆" and "●" denote the 2nd and 3rd order gPC expansion by $\boldsymbol{\chi}$. The dash-dot and dash-dot-dot lines represent the relative L_2 error obtained from level-1 and level-2 sparse grid points, using 55 and 1513 sample points respectively.*

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

Fig. 4.5 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 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 Eq. (3.7) lead to more accurate recovery of $\mathbf{c}$ from the compressive sensing method by Eq. (3.8). 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.

For random variables $\boldsymbol{\xi}$, the L_2 error changes non-monotonically as we compute c_α 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 point to 600 (although it remains larger than the 400-point error). This behavior is 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 $c_\alpha^{\{\boldsymbol{\xi}\}}$ is poorly recovered due to insufficient sample points. However, for the transformed random variables $\boldsymbol{\chi}$, $c_\alpha^{\{\boldsymbol{\xi}\}}$ 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 examined the surrogate model constructed by the sparse grid method based on Gaussian quadrature collocation points. Fig. 4.5 shows the relative L_2 error of the surrogate model constructed by approximating the integral in Eq. (3.5) with level-1 and level-2 sparse grid methods using 55 and 1513 sample points, respectively. The sparse grid results show systematically larger L_2 errors than the compressive sensing approach.

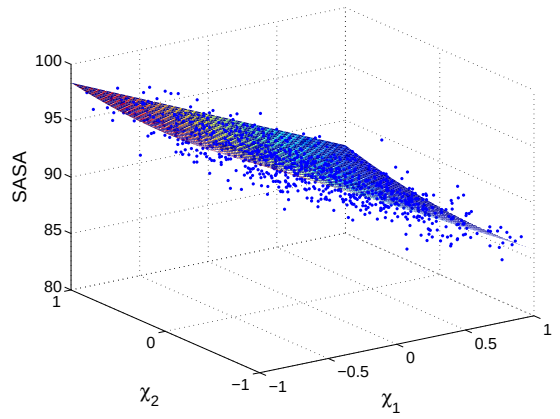
The differences between the models constructed by $c_\alpha^{\{\boldsymbol{\xi}\}}$ and $c_\alpha^{\{\boldsymbol{\chi}\}}$ can be further illustrated by examining the response surfaces in the *reduced* random space shown in Fig. 4.6. This figure shows the response surfaces $\widetilde{X}(\boldsymbol{\xi})$ 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 Fig. 4.6 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 Eq. (3.12), 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 can not be captured well even if we use the two most important variables associated with the first order gPC expansion. Fig, 4.6 clearly illustrates that the different sparsities of $\mathbf{c}$ result in different accuracies for the recovered response surfaces $\widetilde{X}(\boldsymbol{\chi})$ and $\widetilde{X}(\boldsymbol{\xi})$.

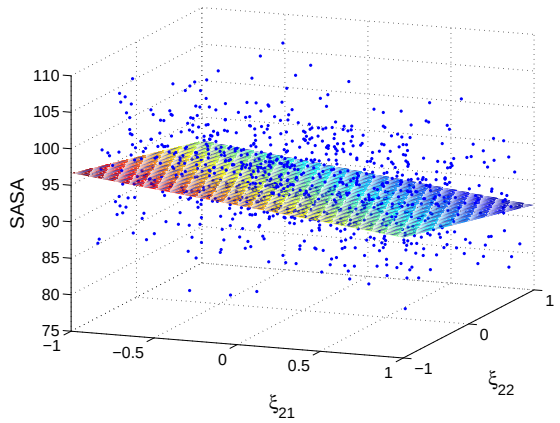
To evaluate the statistical information extracted from the surrogate model, we compute the SASA PDF for target residue P14 by evaluating 10^6 sampling data points with the constructed surrogate model. These results are shown in Fig. 4.7(a) and compared with a reference solution based on the PDF computed from 10^6 direct MC sample points as the reference solution. 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 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 solution, 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 (Fig. 4.5). 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 MC 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.



(a)

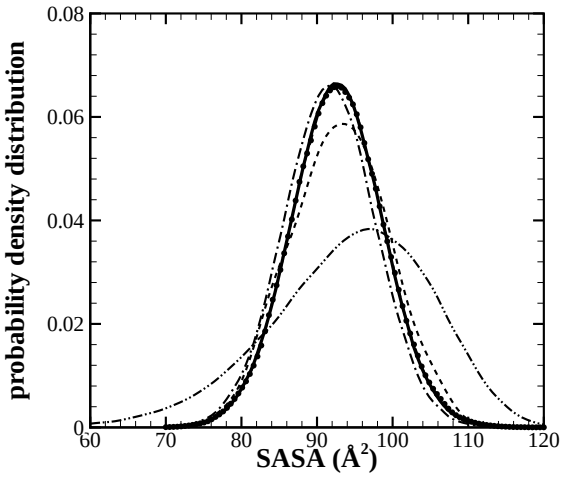


(b)

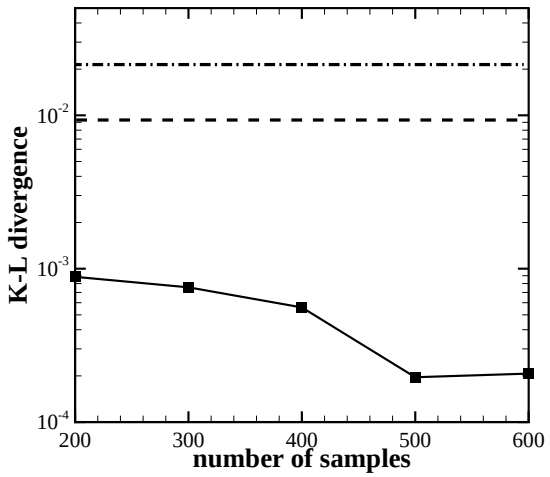
FIG. 4.6. *(a) The reduced response surface constructed by* $\widetilde{X}(\chi_1, \chi_2, \chi_3^0, \cdots, \chi_{27}^0)$*, where* $(\chi_3^0, \cdots, \chi_{27}^0)$ *are fixed values extracted from the i.i.d. normal distribution* $\mathcal{N}(0,1)$*. The scattered symbols (blue points) are direct numerical simulation results on stochastic points* $(\chi_1, \chi_2, \cdots, \chi_{27})$ *in* $\mathbb{R}^{27}$ *following an i.i.d. normal distribution* $\mathcal{N}^{27}(0,1)$*. (b) The reduced response surface constructed by* $\widetilde{X}(\xi_1^0, \cdots, \xi_{21}, \xi_{22}, \cdots, \xi_{27}^0)$ *where* $(\xi_1^0, ..., \xi_{20}^0, \xi_{23}^0, ..., \xi_{27}^0)$ *are fixed values extracted from the i.i.d. normal distribution* $\mathcal{N}(0,1)$*. The scattered symbols (blue points) are direct numerical simulation results on points* $(\xi_1, \xi_2, \cdots, \xi_{27})$ *following an i.i.d. normal distribution* $\mathcal{N}^{27}(0,1)$*.*

4.2. Error sources and sensitivity analysis. To further investigate the applicability of the numerical methods for biomolecular systems, we quantify 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 construct the surrogate model by the compressive sensing method with respect to both $\boldsymbol{\xi}$ and $\boldsymbol{\chi}$, as well as by the level-1 and 2 sparse grid methods. Fig. 4.8 shows the relative L_2 error of the surrogate models, the PDFs of the SASA values, and the K-L divergence with respect to the reference solution.

Similar to the results for residue P14, the surrogate models constructed with respect to $\boldsymbol{\chi}$ yield smaller error than the ones with 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. For both cases, the sparse grid methods yield larger errors due to the high dimensionality of the system [21, 26]. In particular, the surrogate model for residue 20 constructed by the level-2 sparse grid in random space $\mathbb{R}^{63}$ yields the largest



(a)



(b)

FIG. 4.7. *(a) Probability density function (PDF) of the SASA values on residue P14 obtained from the gPC expansion* $\widetilde{X}(\boldsymbol{\chi})$ *using* 300 *sampling points ("•" symbol). Reference solution (solid line) is obtained from MC sampling using* 10^6 *sample points. Results from level-1 (dash-dot line,* 55 *sample points) and level-2 (dash-dot-dot line,* 1513 *sample points) sparse grid method, as well as direct MC sampling methods (dash line,* 300 *sample points) are also presented for comparison. (b) Kullback-Leibler divergence between 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 (dash line,* 300 *sample points) results are presented for comparison.*

deviation from the reference solution.

We also examined the sensitivity of the gPC expansion with respect to numerical error ϕ associated with limited accuracy in the computed value of X as defined in Eq. (3.1). We construct the gPC expansion of the SASA of residue 11 by adding random noise to X at each collocation point

$$X' = X(1.0 + \delta\zeta) \tag{4.4}$$

where ζ is a random variable uniformly distributed between $[-1, 1]$ and the magnitude of the random noise is governed by $\delta \in \left[10^{-5}, 10^{-4}\right]$. For each δ, we construct the gPC coefficients by adopting

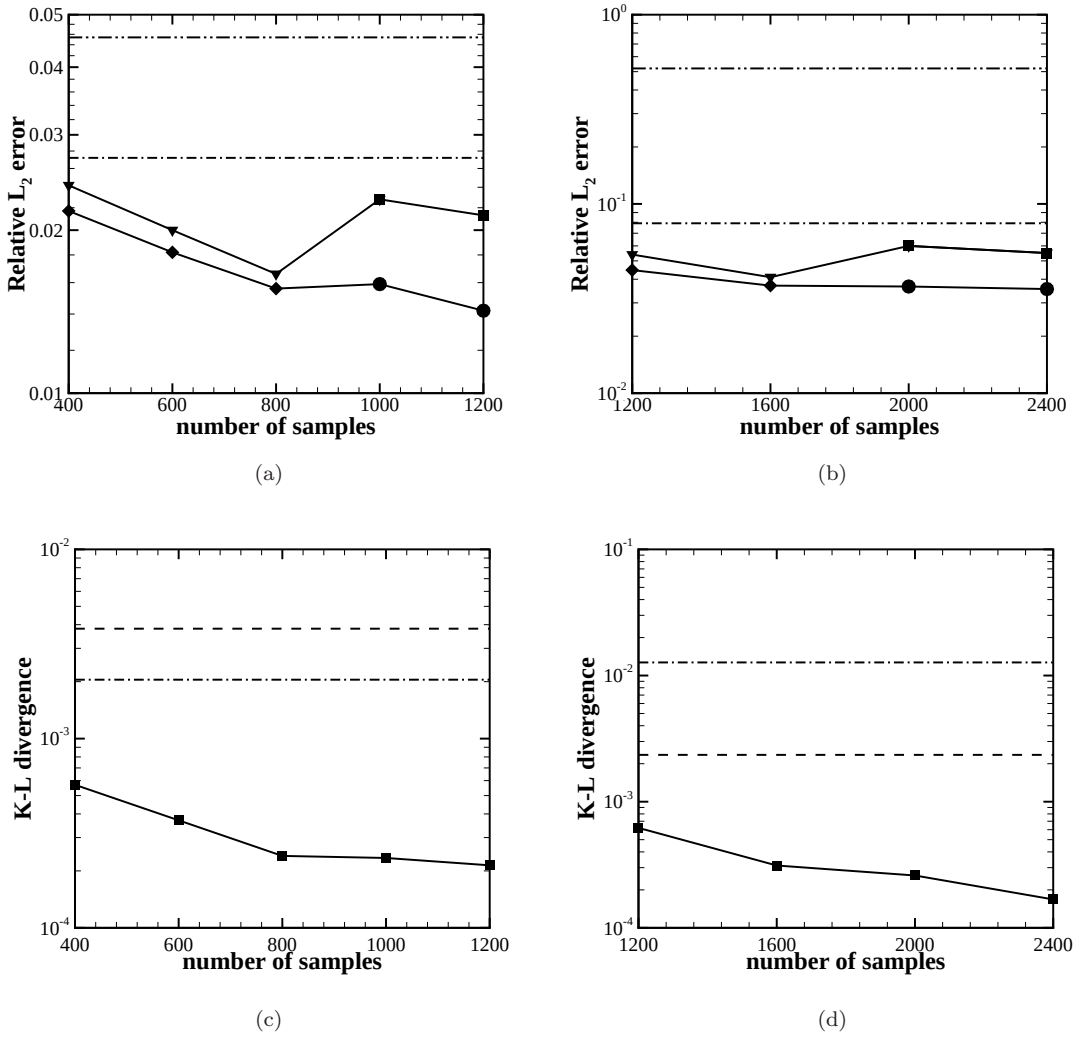


FIG. 4.8. *(a-b) Relative L_2 error of the SASA value on residue 11 (a) and 20 (b) predicted by the gPC expansions $\widetilde{X}(\boldsymbol{\xi})$ and $\widetilde{X}(\boldsymbol{\chi})$. The symbols "▼" and "■" denote the 2nd and 3rd order gPC expansion by $\boldsymbol{\xi}$. The symbols "◆" and "●" denote the 2nd and 3rd order gPC expansion by $\boldsymbol{\chi}$. The dash-dot lines represent the relative L_2 error obtained from level-1 sparse grid methods using 85 and 127 sample points. The dash-dot-dot lines represent the relative L_2 error obtained from level-2 sparse grind methods using 3613 and 8065 sample points. (c-d) Kullback-Leibler divergence between the PDFs obtained from the constructed surrogate models ("■" symbol) for residue 11 (c) and 20 (d). Level 1 sparse grid (dash-dot line) and direct Monte Carlo (dash line, 300 sample points) results are presented for comparison.*

8 sets of i.i.d. random variables ζ on the collocation points. Fig. 4.9 shows the relative L_2 error of the resulting gPC expansion using the compressive sensing method as well as the level-2 sparse grid method. The results from the sparse grid method are very sensitive to the random noise imposed on X. By increasing δ from 10^{-5} to 10^{-4}, the mean value of the relative L_2 error substantially increases from 6.3% to 24%. 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 $\delta \in \left[10^{-5}, 10^{-4}\right]$. This result suggests another advantage of the present method: stability in the presence of numerical error due to limited accuracy in the computed target quantity. For high dimensional systems, the performance of sparse grid method strongly depends on the accuracy of the evaluation of X at collocation points. In practice, it may be computationally infeasible to evaluate X at the accuracy required for stable sparse grid results. Instead, our new method based on compressive sensing shows a much weaker

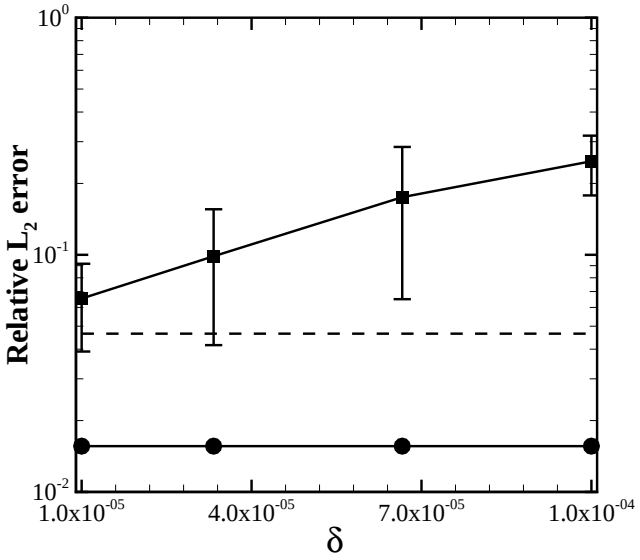


FIG. 4.9. *Relative L_2 error of the SASA value on residue 11 predicted by a gPC expansion constructed by the compressive sensing ("●" symbol) and level-2 sparse grid method ("■" symbol) in the presence of simulated random noise term. The dash line represents the L_2 error from sparse grid method without random noise term.*

dependence on accuracy at individual sample points.

The sparse grid method sensitivity of the constructed gPC expansion with respect to random noise can be understood as follows. Given the target quantity X computed at collocation points, we compute the gPC coefficient $c_{\boldsymbol{\alpha}}$ by

$$c_{\boldsymbol{\alpha}} = \sum_{i=1}^{N^{\text{sp}}} w_i (\bar{X}(\boldsymbol{\xi}_i) + \phi(\boldsymbol{\xi}_i)) \psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi}_i), \tag{4.5}$$

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

$$|\phi_(\boldsymbol{\xi}_i)| \ll |\bar{X}(\boldsymbol{\xi}_i)| \tag{4.6}$$

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

$$\begin{aligned} c_{\boldsymbol{\alpha}} &= \sum_{i=1}^{N^{\text{sp}}} w_i \bar{X}(\boldsymbol{\xi}_i) \psi_{\alpha i} + \sum_{i=1}^{N^{\text{sp}}} w_i \phi_(\boldsymbol{\xi}_i) \psi_{\alpha i} \\ &= \bar{c}_{\boldsymbol{\alpha}} + \sum_{\substack{|\boldsymbol{\alpha}+\boldsymbol{\beta}| \\ >2P+1}} \sum_{i=1}^{N^{\text{sp}}} \bar{c}_{\boldsymbol{\beta}} w_i \psi_{\boldsymbol{\beta} i} \psi_{\boldsymbol{\alpha} i} + \sum_{i=1}^{N^{\text{sp}}} w_i \phi_i \psi_{\alpha i}, \end{aligned} \tag{4.7}$$

where $\bar{c}_{\boldsymbol{\alpha}} = \int \bar{X}(\xi) \psi_\alpha(\xi) d\mathcal{P}(\xi)$ represents the true value of the gPC coefficient of 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 righthand side of Eq. (4.7) represents aliasing error due to the sparse grid approximation. The third term $w_i \phi_i \psi_{\alpha i}$ represents the error due to approximation of $\bar{X}$. We assume that the numerical error ϕ_i superimposed on each collocation point is i.i.d. with zero mean and small variance $|\sigma_\phi^2| \ll |X|^2$.

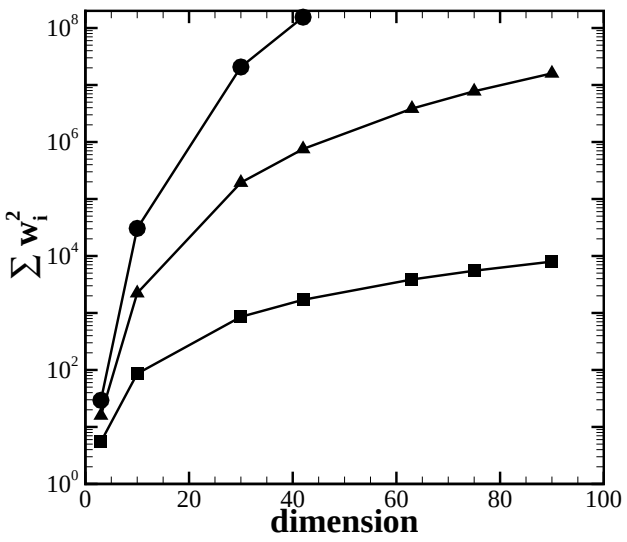


FIG. 4.10. *Variance of the numerical error term $\sum_{i=1}^{N^{\rm sp}} \phi_i w_i$ (normalized by σ_ϕ^2) for c_0 computed by level-1 ("■"), level-2 ("▲") and level-3 ("●") sparse grid points.*

Then, the term $\sum_{i=1}^{N^{\rm sp}} w_i \phi_i \psi_{\alpha i}$ is zero mean with variance

$$\mathrm{Var}\left(\sum_{i=1}^{N^{\rm sp}} w_i \phi_i \psi_{\alpha i}\right) = \sum_{i=1}^{N^{\rm sp}} w_i^2 \psi_{\alpha i}^2 \sigma_\phi^2. \tag{4.8}$$

For a high-dimensional random space $\mathbb{R}^d$, we note that the weight distribution on sparse grid points is inhomogeneous; i.e.,

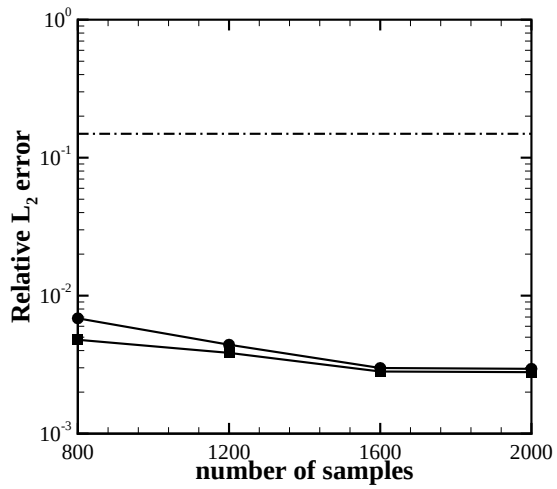
$$\sum_i w_i = 1, \qquad \exists k, |w_k| \gg 1. \tag{4.9}$$

For example, for a level-2 sparse grid based on Gaussian quadrature points, the largest weight is 8.0×10^2 and 1.9×10^3 for $d = 42$ and $d = 63$, respectively. Fig. 4.10 plots the variance of the term $\sum_{i=1}^{N^{\rm sp}} \phi_i w_i$ (normalized by σ_ϕ^2) for c_0 (i.e., $\psi_{\alpha i} \equiv 1$) computed by different levels of sparse grid points. As the dimension increases, the variance of the error term increases rapidly. Moreover, higher levels of sparse grid sampling points results in larger magnitudes of variance due to the increased inhomogeneity in weight distributions. For high-dimensional biomolecular systems, the variance term $\sum_{i=1}^{N^{\rm sp}} w_i \phi_i \psi_{\alpha i}$ could become comparable with $\bar{c}_{\boldsymbol{\alpha}}$ even if Eq. (4.6) is satisfied, leading to non-negligible errors in the computation of $c_{\boldsymbol{\alpha}}$.

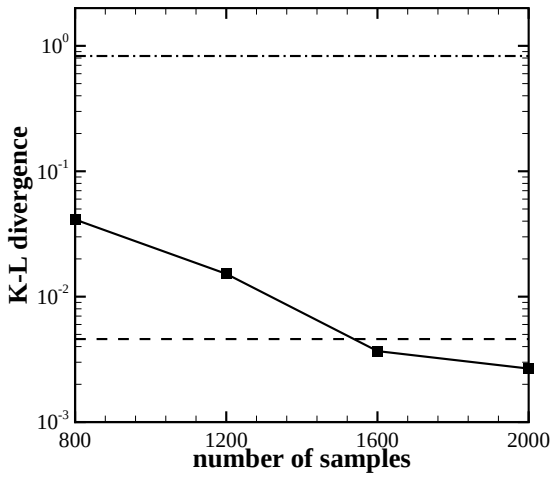
REMARK 4.1. *Rigorous error analysis of the sparse grid method in high-dimensional/complex systems is beyond the scope of this work. To this end, there may be two error sources (aliasing and numerical error on X) that could lead to erroneous results when directly applying the sparse grid method to high-dimensional systems such as biomolecules.*

REMARK 4.2. *We note that other specific structured or adaptive sparse grid methods [20, 32, 35, 36] 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



(a)



(b)

FIG. 4.11. *(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 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).*

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. Fig. 4.11 shows the relative L_2 error of the surrogate model and the K-L 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. Fig. 4.12 shows the average error distribution of the surrogate model within different regimes of the SASA value. The average

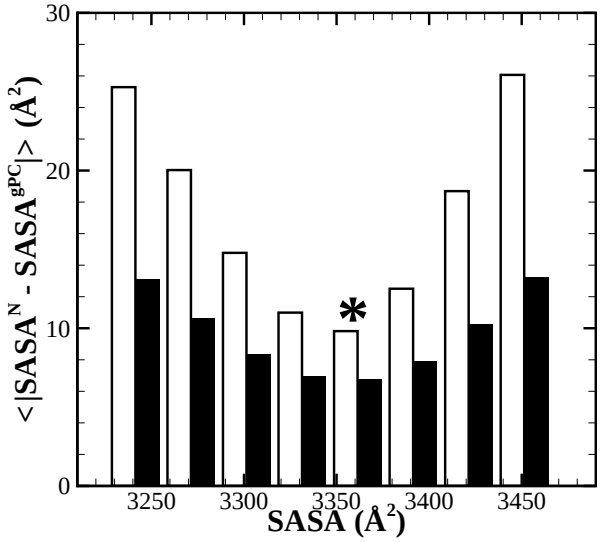


FIG. 4.12. *The average error distribution of the surrogate models of the total-SASA within different regimes. The surrogate models are constructed using 800 (blank) and 1600 (filled) sample points, respectively. The mean value of total-SASA is about* 3351 $\AA^2$ *(denoted by "⋆" symbol), corresponding to conformations near the equlibirium state with respect to the thermal fluctuation.*

error of the surrogate model of X within $[x_1, x_2]$ is defined by

$$\left\langle \epsilon(x_1, x_2)^2 \right\rangle = \frac{\sum_i (X^{\mathrm{gPC}}(\boldsymbol{\xi}_i) - X(\boldsymbol{\xi}_i))^2 I_{(x_1,x_2)}(X(\boldsymbol{\xi}_i))}{\sum_i I_{(x_1,x_2)}(X(\boldsymbol{\xi}_i))}, \tag{4.10}$$

where $I_{(x_1,x_2)}(X(\boldsymbol{\xi}_i))$ is an indicator function which is 1 if the $X(\boldsymbol{\xi}_i) \in [x_1, x_2]$ and 0 otherwise. As shown in Fig. 4.12, the error exhibits a minimum value near the equilibrium state while the value increases as the X approaches the tails of the SASA PDF. Therefore, we emphasize 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 fluctuation considered in the present work.

Similar to the "local" properties discussed above, the gPC expansion recovered by our compressive sensing method yields the smallest error. However, regarding the statistical information, we note that the advantage of our new method over direct Monte Carlo sampling for the global SASA quantity is not as large as in the cases of local properties. This result is because by constructing multi-D basis function through tensor product of one-dimensional basis functions, the upper bound (Here, the upper bond is due to 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 analyzed in Ref. [41, 55]. Alternatively, if only statistical information such as expectation values or PDFs are needed, other methods such as Monte Carlo, quasi-Monte Carlo [37, 47] may be suitable for high-dimensional systems. On the other hand, our method further provides an accurate and efficient approach to explore the response surface of target quantities in high-dimensional systems.

REFERENCES

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

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

[3] Emil Alexov, Ernest L. Mehler, Nathan Baker, António M. Baptista, Yong Huang, Francesca Milletti, Jens E. Nielsen, Damien Farrell, Tommy Carstensen, Mats H. M. Olsson, Jana K. Shen, Jim Warwicker, Sarah Williams, and J. Michael 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*, Biophysical Journal, 80 (2001), pp. 505–515.

[5] Nathan A. Baker, *Biomolecular applications of PoissonBoltzmann methods*, 21 (2005), pp. 349–379.

[6] B. Brooks and M. Karplus, *Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor*, Proceedings of the National Academy of Sciences, 80 (1983), pp. 6571–6575.

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

[8] Jianfeng Cai, Stanley Osher, and Zuowei Shen, *Convergence of the linearized bregman iteration for ℓ_1-norm minimization*, Math. Comput., 78 (2009), pp. 2127–2136.

[9] ———, *Linearized bregman iterations for compressed sensing*, Math. Comput., 78 (2009), pp. 1515–1536.

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

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

[12] Paul G. Constantine, Eric Dow, and Qiqi Wang, *Active subspace methods in theory and practice: applications to kriging surfaces*, 2013.

[13] David L Donoho, *Compressed sensing*, IEEE Trans. Inf. Theory, 52 (2006), pp. 1289–1306.

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

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

[16] Ron O. Dror, Robert M. Dirks, J. P. Grossman, Huafeng Xu, and David E. Shaw, *Biomolecular simulation: A computational microscope for molecular biology*, Annual Review of Biophysics, 41 (2012), pp. 429–452.

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

[18] Jasmine Foo, Xiaoliang Wan, and George Em Karniadakis, *The multi-element probabilistic collocation method (ME-PCM): error analysis and applications*, J. Comput. Phys., 227 (2008), pp. 9572–9595.

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

[20] Alan Genz and B.D. Keister, *Fully symmetric interpolatory rules for multiple integrals over infinite regions with gaussian weight*, Journal of Computational and Applied Mathematics, 71 (1996), pp. 299 – 309.

[21] Thomas Gerstner and Michael Griebel, *Numerical integration using sparse grids*, Numerical Algorithms, 18 (1998), pp. 209–232.

[22] Roger G. Ghanem and Pol 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*, Proceedings of the National Academy of Sciences, 80 (1983), pp. 3696–3700.

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

[25] M. Grant and S. Boyd, *CVX: Matlab software for disciplined convex programming*. http://cvxr.com/cvx/.

[26] Gerstner Griebel and Markus Holtz, *Dimension-wise integration of high-dimensional functions with applications to finance*, Journal of Complexity, 26 (2010), pp. 455 – 489.

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

[28] Robert C. Harris, Alexander H. Boschitsch, and Marcia O. Fenley, *Influence of grid spacing in PoissonBoltzmann equation binding energy estimation*, J. Chem. Theory Comput., 9 (2013), pp. 3677–3685.

[29] B. Lee and F. M. Richards, *The interpretation of protein structures: Estimation of static accessibility*, Journal of Molecular Biology, 55 (1971), pp. 379–IN4.

[30] 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.

[31] Xin Li, *Finding deterministic solution from underdetermined equation: large-scale performance variability modeling of analog/rf circuits*, Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on, 29 (2010), pp. 1661–1668.

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

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

[34] A. McCammon and S. C. Harvey, *Dynamics of protein and nucleic acids*, Cambridge University Press.,

Cambridge, 1987.
[35] A. Narayan and D. Xiu, *Stochastic collocation methods on unstructured grids in high dimensions via interpolation*, SIAM Journal on Scientific Computing, 34 (2012), pp. A1729–A1752.
[36] ———, *Constructing nested nodal sets for multivariate polynomial interpolation*, SIAM Journal on Scientific Computing, 35 (2013), pp. A2293–A2315.
[37] H. Niederreiter, *Random number generation and quasi-Monte Carlo methods*, SIAM, Philadelphia, PA, 1992.
[38] 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.
[39] 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.`
[40] Jay W. Ponder and David A. Case, *Force Fields for Protein Simulations*, vol. 66 of Advances in Protein Chemistry, Elsevier, 2003, pp. 27–85.
[41] Holger Rauhut and Rachel Ward, *Sparse legendre expansions via l_1-minimization*, J. Approx. Theory, 164 (2012), pp. 517–533.
[42] Pengyu Ren, Jaehun Chun, Dennis G. Thomas, Michael J. Schnieders, Marcelo Marucho, Jiajing Zhang, and Nathan A. Baker, *Biomolecular electrostatics and solvation: a computational perspective*, Quarterly Reviews of Biophysics, 45 (2012), pp. 427–491.
[43] Timothy J. Richmond, *Solvent accessible surface area and excluded volume in proteins*, Journal of Molecular Biology, 178 (1984), pp. 63–89.
[44] F. Rizzi, H. N. Najm, B. J. Debusschere, K. Sargsyan, M. Salloum, H. Adalsteinsson, and O. M. Knio, *Uncertainty quantification in md simulation. part i: Forward propagation*, Multiscale Modeling & Simulation, 10 (2012).
[45] ———, *Uncertainty quantification in md simulation. part ii: Bayesian inference of force-field parameters*, Multiscale Modeling & Simulation, 10 (2012), pp. 1460–1492.
[46] Benoit Roux and Thomas Simonson, *Implicit solvent models*, Biophysical Chemistry, 78 (1999), pp. 1–20.
[47] I. H. Sloan and Joe S., *Lattice Methods for Multiple Integration*, Oxford University Press, New York, 1994.
[48] F. Tama and Y.-H. Sanejouand, *Conformational change of proteins arising from normal mode calculations*, Protein Engineering, 14 (2001), pp. 1–6.
[49] Ramakrishna Tipireddy and Roger Ghanem, *Basis adaptation in homogeneous chaos spaces*, Journal of Computational Physics, 259 (2014), pp. 304 – 317.
[50] Monique M. Tirion, *Large amplitude elastic motions in proteins from a single-parameter, atomic analysis*, Phys. Rev. Lett., 77 (1996), pp. 1905–1908.
[51] E. van den Berg and M. P. Friedlander, *Probing the pareto frontier for basis pursuit solutions*, SIAM Journal on Scientific Computing, 31 (2008), pp. 890–912.
[52] 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.*, Journal of Molecular Biology, 180 (1984), pp. 301–329.
[53] Dongbin Xiu and Jan S. Hesthaven, *High-order collocation methods for differential equations with random inputs*, SIAM J. Sci. Comput., 27 (2005), pp. 1118–1139.
[54] Dongbin Xiu and George Em Karniadakis, *The Wiener-Askey polynomial chaos for stochastic differential equations*, SIAM J. Sci. Comput., 24 (2002), pp. 619–644.
[55] Liang Yan, Ling Guo, and Dongbin Xiu, *Stochastic collocation algorithms using l_1-minimization*, Int. J. Uncertainty Quantification, 2 (2012), pp. 279–293.
[56] Xiu Yang, Minseok Choi, Guang Lin, and George Em Karniadakis, *Adaptive ANOVA decomposition of stochastic incompressible and compressible flows*, J. Comput. Phys., 231 (2012), pp. 1587 – 1614.
[57] Xiu Yang and George Em Karniadakis, *Reweighted ℓ_1 minimization method for stochastic elliptic differential equations*, J. Comput. Phys., 248 (2013), pp. 87 – 108.
[58] Wotao Yin, Stanley Osher, Donald Goldfarb, and Jerome Darbon, *Bregman iterative algorithms for ℓ_1-minimization with applications to compressed sensing*, SIAM J. Imaging Sci., 1 (2008), pp. 143–168.
[59] Zhongqiang Zhang, Minseok Choi, and George Em Karniadakis, *Error estimates for the ANOVA method with polynomial chaos interpolation: Tensor product functions*, SIAM J. Sci. Comput., 34 (2012), pp. A1165–A1186.