Elsevier Editorial System(tm) for Computer Methods in Applied Mechanics and Engineering
Manuscript Draft

Manuscript Number: CMAME-D-18-01024R1

Title: A data-driven framework for sparsity-enhanced surrogates with arbitrary mutually dependent randomness

Article Type: Research Paper

Keywords: data driven; arbitrary randomness; mutual dependence; compressed sensing; sparsity enhancement; uncertainty quantification

Corresponding Author: Dr. Nathan Andrew Baker, Ph.D.

Corresponding Author's Institution: Pacific Northwest National Laboratory

First Author: Huan Lei

Order of Authors: Huan Lei; Jing Li; Peiyuan Gao; Panagiotis Stinis; Nathan Baker

Abstract: The challenge of quantifying uncertainty propagation in real-world systems is rooted in the high-dimensionality of the stochastic input and the frequent lack of explicit knowledge of its probability distribution. Traditional approaches show limitations for such problems, especially when the size of the training data is limited. To address these difficulties, we have developed a general framework of constructing surrogate models on spaces of stochastic input with arbitrary probability measure irrespective of the mutual dependencies between individual components of the random inputs and the analytical form. The present Data-driven Sparsity-enhancing Rotation for Arbitrary Randomness (DSRAR) framework includes a data-driven construction of multivariate polynomial basis for arbitrary mutually dependent probability measures and a sparsity enhancement rotation procedure. This sparsity-enhancing rotation method was initially proposed in our previous work for Gaussian density distributions, which may not be feasible for non-Gaussian distributions due to the loss of orthogonality after the rotation. To remedy such difficulties, we developed a new data-driven approach to construct orthonormal polynomials for arbitrary mutually dependent randomness, ensuring the constructed basis maintains the orthogonality/near-orthogonality with respect to the density of the rotated random vector, where directly applying the regular polynomial chaos including arbitrary polynomial chaos shows limitations due to the assumption of the mutual independence between the components of the random inputs. The developed DSRAR framework leads to accurate recovery, with only limited training data, of a sparse representation of the target functions. The effectiveness of our method is demonstrated in challenging problems such as partial differential equations and realistic molecular systems within high-dimensional conformational spaces where the underlying density is implicitly represented by a large collection of sample data, as well as systems with explicitly given non-Gaussian probabilistic measures.

Research Data Related to this Submission

There are no linked research data sets for this submission. The following reason is given:
Data will be made available on request

Response to referee's report for the manuscript #CMAME-D-18-01024

December 8, 2018

Manuscript: #CMAME-D-18-01024.
Title: A data-driven framework for sparsity-enhanced surrogates with arbitrary mutually dependent randomness
Previous Title: Data-driven approach of quantifying uncertainty in complex systems with arbitrary randomness.
Authors: H. Lei, J. Li, P. Gao, P. Stinis and N. A. .Baker.

Dear Editor:
We have addressed *all* of the points that the referees raised and also made all the recommended changes regarding our manuscript. We provide point-by-point answers below. In the revised manuscript, revisions are marked in blue color.

We thank the referees for their thorough review and constructive comments that helped us to improve the quality of our manuscript. We also thank you for your time and consideration of this paper.

Sincerely,
Huan Lei
Jing Li
Peiyuan Gao
Panos Stinis
Nathan Baker

Reviewer 1

Reviewer Comment: *This work is an extension of the authors' previous work on uncertainty quantification in complex systems from observation data by removing the assumption of Gaussianity. The novelty in this work includes the construction of an orthogonal basis in random space using the Gram-Schmidt procedure and application of the authors' work to deal with non-Gaussian processes. This work is an important and interesting exploration toward real applications using the state-of-the-art uncertainty quantification techniques. However, many places in this highly technical paper are unclear to me in both mathematics and descriptions. I would suggest a major revision.*

1. **Reviewer Comment:** *Page 8. Line 45-46. The claim that "Consequently, the orthogonal polynomials basis ..." is not true. Whether the involving random variables are independent or not, we can always represent the considered process with a tensor product of univariate basis functions.*

Author Response: We refer to the case where each component of $\boldsymbol{\xi}$ can be mutually dependent. We have clarified this and revised the sentence by "Consequently, the orthogonal polynomial basis $\psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi})$ cannot be directly constructed as a tensor product of univariate orthonormal basis functions in each component of $\boldsymbol{\xi}$."

2. **Reviewer Comment:** *Page 9. The Gram-Schmidt orthogonalization often becomes unstable when the orthogonality is lost after several steps. What do you do when you have this problem? Also, what are your requirements on your samples to assure the success of the orthogonalization.*

Author Response: The numerical instability can be addressed by employing modified Gram-Schmidt orthogonalization. For the current work, we checked the orthonormality after the basis was constructed in all examples we presented in this manuscript, the normalized Gram-Schmidt orthogonalization worked quite well, and a possible reason is that we implemented with the polynomials basis up to relatively low degree. We add the modified Gram-Schmidt orthogonalization as an alternative if there exists an issue with instability. There is no requirements on the samples to assure the success of the orthogonalization.

3. **Reviewer Comment:** *Page 9. Line 30-31. By "an i.i.d random vector", do you mean a random vector with i.i.d elements? Similar issues appear in several places in the paper.*

Author Response: Yes. We replaced all instances of " i.i.d. random vector" with "random vector with i.i.d. components" to clarify this point.

4. **Reviewer Comment:** *How (3.6) is derived from (3.4)? I don't see any connection between these two formulas.*

Author Response: We do not derive (3.6) from (3.4). We use the orthonormal conditions and the representations of (3.4) to calculate the coefficients f_{β}^{α}.

5. **Reviewer Comment:** *Page 11. Lines 50-51. If an ordered monomial basis is used, will there be a problem of ill-conditioning?*

Author Response: The ordered monomial basis is used for representation of our orthonormal basis. When we construct the measurement matrix for compressive sensing, we only use the constructed orthonormal basis or the near-orthonormal basis. There are no conditioning problems with these bases.

6. **Reviewer Comment:** *Page 13. Equation (3.14). Why is it important to consider the gradient matrix? For moderately nonlinear f and Gaussian $\boldsymbol{\xi}$, the gradient might be a good*

indicator of the covariance matrix. However, for non-Gaussian $\boldsymbol{\xi}$, this may not be the case. The gradient can be misleading for a non-Gaussian case.

Author Response: We evaluate the gradient matrix $\mathbf{G} = \mathbb{E}\left[\nabla f(\boldsymbol{\xi})\nabla f(\boldsymbol{\xi})^T\right]$ to find the sorted variability directions with respect to the QoI. Eigendecomposition $\mathbf{G} = \mathbf{Q}\mathbf{K}\mathbf{Q}^T$, $\mathbf{Q} = [\mathbf{q}_1\ \mathbf{q}_2 \cdots\ \mathbf{q}_d]$, $\mathbf{K} = \mathrm{diag}(k_1, \cdots, k_d)$ shows that the QoI f exhibits the largest variability along $\mathbf{q}_1$ and the least along $\mathbf{q}_d$ with respect to $\rho(\boldsymbol{\xi})$. We refer to Ref. [5, 2] for details.
This is true for nonlinear f and non-Gaussian $\boldsymbol{\xi}$. As a special case, if f is linear, $k_2 = \cdots = k_d \equiv 0$ and QoI only depends on the univariate $\mathbf{q}_1^T\boldsymbol{\xi}$.
We note that the gradient matrix $\mathbf{G}$ depends on the QoI f and does not serve as an approximation of the covariance matrix. Alternatively, we use $\mathbf{G}$ to define a new random vector $\boldsymbol{\chi} = \mathbf{Q}^T\boldsymbol{\xi}$ s.t. the QoI $g(\boldsymbol{\chi}) = f(\mathbf{Q}\boldsymbol{\chi})$ shows sparse coefficient (e.g., see Fig. 5). Accordingly, the surrogate model $\tilde{g}(\boldsymbol{\chi})$ constructed by compressed sensing is more accurate than $\tilde{f}(\boldsymbol{\xi})$. Since f is nonlinear in general, here we does not conduct dimension reduction, i.e., the dimension of $\boldsymbol{\chi}$ equals to dimension of $\boldsymbol{\xi}$.
To construct $\tilde{g}$, we note that for non-Gaussian $\boldsymbol{\xi}$, the rotated vector $\boldsymbol{\chi} = \mathbf{Q}^T\boldsymbol{\xi}$ does not retain the same density distribution $\rho(\cdot)$. This is different from the case where both $\boldsymbol{\xi}$ and $\boldsymbol{\chi}$ are Gaussian. The orthonormal basis $\psi(\boldsymbol{\xi})$ does not retain orthonormal with respect to $\boldsymbol{\chi}$. This is one of the major motivations for the current work, where we developed a general framework to construct orthonormal bases $\phi(\boldsymbol{\chi})$ with respect to arbitrary non-Gaussian density. The present method enables us to retain the orthonormal properties after the rotation and construct surrogate $\tilde{g}(\boldsymbol{\chi})$ with respect to the new random vector $\boldsymbol{\chi}$.

7. **Reviewer Comment:** *Page 13. Line 40. Why the model is $f(\chi)$ instead of $f \cdot Q(\chi)$?*

Author Response: To answer this question, we revised the document and replaced "where the surrogate model $f(\chi)$ can be represented by a sparser coefficient vector c as compared with $f(\xi)$" by "$f((\mathbf{Q}^T)^{-1}\boldsymbol{\chi}) = f(\mathbf{Q}\boldsymbol{\chi}) = f(\boldsymbol{\xi})$ can be approximated by an expansion of orthonormal polynomial basis of $\boldsymbol{\chi}$ with sparser coefficient vector $\boldsymbol{c}$ as compared to $f(\boldsymbol{\xi})$ being expanded by orthonormal basis of $\boldsymbol{\xi}$."

8. **Reviewer Comment:** *Numerical results. · It is assumed that ϵ is from a Gaussian mixture with its components whose mean, and covariance are randomly chosen. I don't understand why the Legendre polynomial chaos is chosen for comparison. · It seems that none of the experiments can be reproduced. No random generators are reported.*

Author Response: In the present study, we consider systems with arbitrary randomness (e.g., samples generated by Gaussian mixture, linear transformation from i.i.d. uniform variables, trajectory data set from Molecular Dynamics simulation). For such systems, standard polynomial bases such as Legendre polynomials and Hermite polynomials are commonly chosen for surrogate construction. However, as the reviewer suggests (see also the reviewer comment 15), directly applying such bases has limitations. Therefore, we proposed the present approach based on the data-driven construction of orthonormal basis. We compared our results with ones regressed with Legendre chaos and Hermite chaos. For the purposes of reproducibility, we have provided the seed numbers of the random number generators and name of generation function called to generate the Gaussian Mixture data set in Appendix D as suggested (all of the simulations are done in Matlab).

9. **Reviewer Comment:** *Page 16. Lines 44-45. Why the loss of orthonormality leads to worse performance?*

Author Response: There are 2 possible reasons. One is that in approximation theory,

the orthogonal basis expansion gives the best $l2$ approximation. Another reason is that in compressive sensing, the orthonormal basis gives better recovery.

10. **Reviewer Comment:** *Page 16. Lines 48-49. What is the condition? Can this condition be given briefly here?*

Author Response: The sufficient number stated in [52] ([30] in previous version) is given in Theorem 2.5 and we revised the sentence to " M suggested by the sufficient condition (Theorem 2.5) originally given " for clarification.

11. **Reviewer Comment:** *Page 17. Line 37. Why is the error evaluated in l2-norm? instead of l1 norm.*

Author Response: Due to the orthonormality of the basis functions in probability measure, the l_2 error of the surrogate model is equal to the l_2 error of the coefficient vector. Accordingly, we examined the l_2-error of the coefficient vector to measure the accuracy of the surrogates. Also the l_2-error is conventional in the community of compressive sensing.

12. **Reviewer Comment:** *Page 19. What is the level 4 sparse grid integration rule? What quadrature rule is used? Are they sufficient to give the desired accuracy? Also, I don't see anything plotted in Figure 3 on the integration of $\tilde{f}$ using a sparse grid.*

Author Response: We used Smolyak's sparse grid quadrature rule of level 4. It is exact for integration of polynomial of degree 7 and is sufficient to calculate numerical l_2 error of surrogate model (polynomial degree 3). Figure 3 presents the relative $l2$-errors of different regressed surrogates. The $l2$-errors, defined by Eq. (4.12), are evaluated by Smolyak's sparse grid quadrature rule of level 4.

13. **Reviewer Comment:** *Page 20. Why is the density function of the form (4.10) chosen? PDFs of this type are extremely difficult because of the singularity on the boundary.*

Author Response: We chose the density function of the form (4.10) as the random variables with this density function have been used a lot for benchmark problems in compressive sensing for UQ. And also as the reviewer suggested, PDFs of this type are extremely difficult, we would like to test our method for such difficult problems.

14. **Reviewer Comment:** *Page 21. Line 43-44. What is "each training sample size number M"?*

Author Response: For clarification, we added a sentence "This small set $\{\boldsymbol{z}^{(i)}\}_{i=1}^{M}$ is usually called the training sample set and M is the training sample size." in Section 3 Methods 3 lines below the section title.

15. **Reviewer Comment:** *Page 23. Section 4.3.2. Again why Legendre basis (and Hermite basis) is chosen? The sparsity here doesn't mean anything since these bases can be useless at all (no capability of representing the current function).*

Author Response: Both Legendre basis and Hermite basis up to degree P can span polynomial spaces of degree P. This means that whenever the target function can be approximated by polynomials of order P, then both Legendre basis and Hermite basis can be used. However, in the approach of compressive sensing, using both Legendre basis and Hermite basis may not give good representations. Currently, Legendre polynomials and Hermite polynomials are commonly chosen for surrogate construction in UQ studies in both benchmark test and real-world application. However, as the reviewer suggested, direct employing such bases may not be appropriate in practice. We totally agree on this point, which motivates us to develop the present method to handle such problems. Here,

we compare the surrogates constructed by our approach with the ones constructed with Legendre basis and/or Hermite basis to demonstrate effectiveness of our approach.

16. **Reviewer Comment:** *Page 42-43. What do you mean by "$\boldsymbol{\xi}$ ... represents* 99.99% *of the variance"?*

Author Response: Since there is no page 42-43 of this manuscript, we assume the reviewer meant the Page 24, line 42-43. This is a standard approach for dimensionality reduction in molecular systems. Here we employ PCA (principal component analysis) to construct the random (conformational) space of the molecule, and after the dimension reduction we kept 99.99% of the total variance of the original data set.

17. **Reviewer Comment:** · *Not sure whether arbitrary randomness is appropriate in the title and abstract. Observe that the authors have two examples: one is a randomized Gaussian mixture: the other also uses Gaussian mixtures.*

Author Response: In this manuscript, we implemented our approach for various cases. Some of the data sets are generated from Gaussian mixture distributions (example 4.1, 4.3.1, 4.3.2), some are generated linear transformation from other i.i.d. distributions such as uniform distribution and distribution with PDF (4.6) (example 4.2.1) and PDF (4.10) (example 4.2.2), and some are from MD simulations (example 4.3.3). Another reason we employ the Gaussian mixture model is that a wide range of arbitrary distributed data can be approximated by Gaussian mixture model.
Moreover, there is a major difference between the present study and the previous UQ studies on based gPC (and aPC). The DSRAR framework developed in the present study does not assume the mutual independence between the components of random inputs; and therefore can be more realistic and particularly well-suited for uncertainty quantification for complex systems, where the knowledge of the underlying randomness can be implicit. Accordingly, we have changed the title to "A data-driven framework for sparsity-enhanced surrogates with arbitrary mutually dependent randomness" and emphasized such difference in the abstract and Introduction of the revised manuscript.

18. **Reviewer Comment:** *Some details Here is a partial list I have when I read the paper. The authors should use more caution when proofreading the paper.*

Author Response: We thank the reviewer's careful reading for helping us to improve the quality of this manuscript. We have addressed the comments listed below and also carefully checked the revised manuscript.

19. **Reviewer Comment:** *Page 8. Line 43. What does "the dimension of $\boldsymbol{\xi}$ are independent " mean?*

Author Response: We replaced "the dimension of $\boldsymbol{\xi}$ are independent" by "each component of $\boldsymbol{\xi}$ is mutually independent".

20. **Reviewer Comment:** *Page 12. Line 52 Insert "a" before "smaller basis bound." This also applies to Line 42 on Page 27.*

Author Response: Fixed.

21. **Reviewer Comment:** *Page 13. Lines 51-52 what does it mean by saying "$\boldsymbol{\xi}$ are i.i.d random variables"?*

Author Response: We rewrite this sentence with "In particular, if $\boldsymbol{\xi}$ is a random vector with i.i.d. Gaussian components, $\boldsymbol{\chi}$ is also a random vector with i.i.d. Gaussian components" for clarification.

22. **Reviewer Comment:** *Page 15. Line 42. An article "a" is missing before "uniform distribution".*

Author Response: Fixed.

23. **Reviewer Comment:** *Appendix B.2. What is the basis bound K I checked throughout the paper and didn't find where it is defined.*

Author Response: K is defined in (2.9). For clarification, we revised that part by adding "where K is called the basis bound. "

Reviewer 2

Reviewer Comment: *This paper presents an interesting strategy to construct a surrogate model using gPC in order to perform Uncertainty Quantification. The novelty of the research lies in the ability of the method to construct a surrogate model with arbitrary probability density functions for the input random variables. Thanks to their sparsity-enhancing procedure, the method is able to recover the quantity of interest with a limited number of samples. The overall quality of the manuscript is good. The method is presented in detailed and the authors tested their approach using various test cases. I find that while the article is, in essence, worthy of publication, I would request the authors to please respond to my questions and kindly address some specific corrections that I have noted.*

Author Response: We thank the referee for the detailed reading of the manuscript and constructive suggestions that help improve our manuscript. In particular, we appreciate the referee's suggestions on computing the predictive coefficient Q_2 and the Sobol indices on the test systems, which provide complementary information for the tested systems and demonstrate the effectiveness of our approach.

1. **Reviewer Comment:** *From the title and abstract, an important information is missing. It is unclear that you are using gPC as a surrogate model. This is also the case in your highlights were it should be clear. Also, the first highlight is not very informative as this is the essence of UQ as each parameter is naturally described by a PDF (being known or not)..*

Author Response: We have changed the title to "A data-driven framework for sparsity-enhanced surrogates with arbitrary mutually dependent randomness". We have also revised the abstract and high-light accordingly to specify the motivation and the approach of the present study.

2. **Reviewer Comment:** *In the introduction, the authors start by clearly stating their UQ problem. The constrains are: (i) high dimensions with (ii) dependent random variables and (iii) unknown probability density functions. Following, the attention is directly shifted to gPC methods and some strategies to overcome the three constrains are given. But the bibliographical review lack information about other surrogate strategies such as Gaussian Processes that do not intrinsically suffer from these problems. As the main opponent method, this could be mentioned through classical references (Rasmussen, Gaussian processes for machine learning, 2006) and comparative studies (P.Roy, Comparison of polynomial chaos and Gaussian process surrogates for uncertainty quantification and correlation estimation of spatially distributed open-channel steady flows, Stochastic Environmental Research and Risk Assessment, 2017). Also, sparse grid methods are commonly used in high-dimensional problems.*

Author Response: We appreciate the referee's suggestions and have included a brief literature review of the Gaussian Process and sparsity grid method in the introduction of the revised manuscript.

3. **Reviewer Comment:** *In the explanation of the CS method, it is stated that the method is suited for linear functions. Does it mean that the overall method presented in this work is not suited for non-linear and complex system? This is not clear as the method is presented to be a general framework.*

Author Response: The method presented in the present study works for non-linear and complex systems. The surrogate model is represented as the expansion of a set of multivariate polynomial basis, i.e., $\hat{f}(\boldsymbol{\xi}) = \sum_{i=1}^{N} c_i \psi_i(\boldsymbol{\xi})$, where $\psi_i(\cdot)$ is the basis and c_i is the

coefficient. To determine the coefficients with training set $\left\{\boldsymbol{\xi}^{(i)}, f(\boldsymbol{\xi}^{(i)})\right\}_{i=1}^{M}$, we employ CS to solve the linear system $\boldsymbol{A}\boldsymbol{c} = \boldsymbol{b}$, where $\boldsymbol{A} \in \mathbb{R}^{M\times N}$ with $[\boldsymbol{A}]_{ij} = \psi_j(\boldsymbol{\xi}^{(i)})$ and $\boldsymbol{b} \in \mathbb{R}^M$ with $b_i = f(\boldsymbol{\xi}^{(i)})$. This does not mean the obtained surrogate function is linear. In this study, we considered the scenario when the available training points is limited (i.e., $M < N$), and we employ the CS method to determine $\boldsymbol{c}$.

4. **Reviewer Comment:** *May I suggest the use of classical test functions used in the field? g-function and Ishigami are notably used. g-function is particularly interesting due to the ability to set the dimensionality and the influent parameters (example here: Kucherenko, Exploring multidimensional spaces: a Comparison of Latin Hypercube and Quasi Monte Carlo Sampling Techniques,2015).I am concerned about the ability of the method to really perform well with complex, non-linear, non-monotonous and high-dimensional functions.*

Author Response: The present work is based on the assumption of compressed sensing, which assumes that there exists a sparse representation of the multivariate basis for the QoI. Accordingly, it is not suited for g-functions or Ishigami functions, which are not in the scope of CS. We have added a clarification in the introduction and summary of revised manuscript.

5. **Reviewer Comment:** *Complementary to the classical l2 error, the predictivity coefficient Q2 is most of the time used (Marrel, Calculations of Sobol indices for the Gaussian process meta model. Reliability Engineering & System Safety. 2009).*

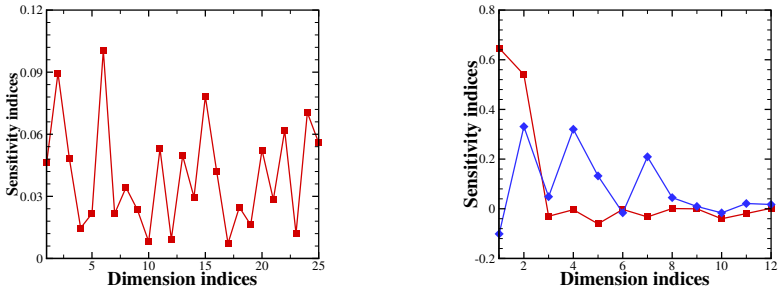
Author Response: We have computed the predictivity coefficient Q_2 (following the definition in Ref. [3]) for the test cases of Gaussian Mixture systems (with $d = 25$ and $p = 3$) and the biomolecular systems. The results are listed in Tab. 1, where the surrogate models are constructed by the present data-driven basis approach. We have included the such results in the Appendix of the revised manuscript.

Table 1: The predictivity coefficient Q_2 for polynomial function with Gaussian Mixture measure ($d = 25$ and $p = 3$) and the molecular system for solvation energy and SASA of atom H9.

molecule solvation	M	80	160	240	320	400
	Q_2	0.995715	0.999132	0.999731	0.999864	0.999911
molecule SASA	M	200	300	400	500	600
	Q_2	0.988675	0.996069	0.998272	0.998709	0.999027
Gaussian Mixture	M	200	300	400	500	600
	Q_2	0.998372	0.999347	0.999844	0.999892	0.999941

6. **Reviewer Comment:** *It would be interesting to add information about the effective number of important variables in each case. $d = 25$ in most cases but what are the Sobol indices? Are all these dimensions as important? Analysis using Type A, B and C functions as proposed in Kucherenko2015 is quite informative and allows to really grasp the complexity of the cases. Especially for the motivation case, it lacks a UQ analysis.*

Author Response: We would like to thank the referee's suggestion on studying the Sobol indices. The present method allows us to construct surrogate models for systems with dependent random variables. Accordingly, we have further computed the Sobol sensitivity indices for the test cases of Gaussian Mixture systems (with $d = 25$ and $p = 3$) and the molecular systems with dependent random variables following Ref. [1]. The results



(a) (b)

Figure 1: The first-order sobol sensitivity indices for (a) polynomial function with Gaussian Mixture mesure ($d = 25$, $p = 3$) (b) molecular system for solvation energy ("—■—") and SASA of atom H9("-·-◆-·-").

are presented in Fig. 1, where the surrogate models are constructed by the present data-driven basis approach using $M = 800$, $M = 240$ and $M = 600$ training points, respectively. Based on the analysis, it is shown the dominant components are on the dimensions $(1, 2, 3, 6, 11, 13, 14, 15, 16, 20, 22, 24, 25)$, $(1, 2, 5)$ and $(1, 2, 4, 5, 7)$ (90% of total variance). We have also included these results in the Appendix of the revised manuscript.

7. **Reviewer Comment:** *Regarding the results. I found the sizes of the samples quite large. The classical rule of thumb to construct a surrogate model is 10n (Jones, Efficient global optimization of expensive black-box functions, J Glob Optim, 1998 and here is a comprehensive review Liu, A survey of adaptive sampling for global meta modeling in support of simulation based complex engineering design, Structural and Multidisciplinary Optimization, 2018). This may be linked to my previous comment. Using classical functions would allow fair comparisons between different surrogate strategies.*

Author Response: In this study, we utilize a relatively large size of sample set S to approximate the underlying probability measure $\rho(\cdot)$ and construct the orthonormal basis. On the other hand, the size of the training set M is still limited. We have examined the error of the surrogate model constructed using a broad range of training sample size similar to Ref. [4]. As shown in Tab. 1, the predictivity coefficient Q_2 is larger than 0.98 when the training size is on order $O(10d)$, which means that our approach still produced good approximations.

8. **Reviewer Comment:** *In all the test cases, have you used QMC or other sampling technic other than crude MC in order to generate the samples from the sought PDFs? This is known to impact a lot the quality of the surrogates. Furthermore, when constructing a surrogate model, the design of experiments does not have to follow any specific distribution. Once you obtained the surrogate, you can evaluate it using these distributions. This should be discussed even though you are in a data-driven approach as you could choose to filter the data or some time you are able to fetch new data.*

Author Response: In the present study, we focus on the scenario where the underlying measure is implicitly represented sample set and the training set is randomly collected from

the sample set and paired with the values of QoI. Accordingly, the surrogate model is constructed in a non-intrusive way; and we do not have the samples on the pre-selected points. Indeed, to construct the surrogate model, the measure of the sampling points does not need to follow any specific distributions (e.g., uniform, Gaussian, Gamma, etc.) We choose the training sample points and the data-driven orthonormal basis following the underlying measure mainly for two reasons (I) the constructed surrogate model is l_2 optimal with respect to the underlying measure (II) for some systems such as molecular system, it may not be straightforward to sample the QoI following a specific distribution; it would be more natural and convenient to collect samples directly following the underlying distribution (i.e., molecular dynamics trajectories).

9. **Reviewer Comment:** *Lastly, the introduction lack a concise explanation of the previously introduced method*

Author Response: We have included a brief explanation of the previous method in the revised manuscript.

10. **Reviewer Comment:** *A reference to classical UQ work can be added for the unfamiliar reader such as: Saltelli, Global Sensitivity Analysis. The Primer, 2007.*

Author Response: We have added the reference.

11. **Reviewer Comment:** *Two ideas: numerical simulations and experimental simulations. Revise to make it clear that the tools are the same and specify that you will focus on numerical experiments.*

Author Response: We have revised the Introduction accordingly.

12. **Reviewer Comment:** *Lines 10-11. Trying to reduce the dimensionality is a consequence of the numerical cost. The two sentences are confusing. It is suggested that the objective is to reduce the dimensionality. This should be revised. Also add a reference to the dimensionality issue.*

Author Response: We have revised the Introduction and added a reference.

13. **Reviewer Comment:** *Line 14 Add a reference of such realistic systems with not independent variables.*

Author Response: We have added a reference with a brief explanation.

14. **Reviewer Comment:** *Page 2, line 15: You are pointing out 3 challenges as opposed to 2 on line 7. You are focussing on these aspects but there are other challenges. Revise your statement about the number of important challenge by pointing out that these are your challenge of interest.*

Author Response: This refers to the part of the second challenge. We have modified the introduction to clarify this.

15. **Reviewer Comment:** *Page 2, lines 21-22: you referred to UQ, talk about sensitivity indices.*

Author Response: We have added sensitivity indices.

16. **Reviewer Comment:** *A better reference can be used in an UQ context such as: Kucherenko, Exploring multidimensional spaces: a Comparison of Latin Hypercube and Quasi Monte Carlo Sampling Techniques,2015.*

Author Response: We have added this reference.

17. **Reviewer Comment:** *page2, line25 : inefficient is not the word. It is the numerical cost.*

Author Response: We have modified the sentence.

18. **Reviewer Comment:** *page2, line26 : Remove etc..*

Author Response: We have removed the word.

19. **Reviewer Comment:** *Page 2, line 26: Preferably refer to the original papers instead of 6-7: Mckay, A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code, 1979*

Author Response: We have added the above reference.

20. **Reviewer Comment:** *Page 2, lines 26-28: explain why QMC or LHS are not suited with unknown distributions. This statement is strong and should be backed up.*

Author Response: We have modified this sentence.

21. **Reviewer Comment:** *Page 3, lines 41-50: For someone new to gPC, it will be hard to understand the curse of dimensionality from this.*

Author Response: We have modified this paragraph.

22. **Reviewer Comment:** *Page3, lines 54-61: Missing reference about the biomolecular example. Also, this part is hard to follow without prior knowledge about the physics. For instance, what is a conformational state? This should be revised.*

Author Response: We have modified this paragraph and added a relevant reference.

23. **Reviewer Comment:** *Page 4, line 77: Reference 47 does not seem relevant for copula*

Author Response: Copulas are discussed in Sec. 3.4 of Reference 47.

24. **Reviewer Comment:** *general UQ framework for.*

Author Response: We have modified this sentence.

25. **Reviewer Comment:** *Page 4, line 98: it is a strong statement to make saying that you have access to the analytical form of the PDF in real-world applications*

Author Response: We may not know the analytical form of PDF in real-world applications; and this scenario is discussed as the situation (I) in the paragraph; while there are also systems where the underlying distribution is accessible from physical principle (e.g., Boltzmann distribution of Hamiltonian systems), or the randomness is specified by users as direct input.

26. **Reviewer Comment:** *page 4, line 105: Remove rest.*

Author Response: We have removed the word.

27. **Reviewer Comment:** *Page 5, line 112: biomolecular.*

Author Response: We have changed the word.

28. **Reviewer Comment:** *Page 6, Eq 2.4: c is not defined clearly and it is central in the following.*

Author Response: We have modified this part.

29. **Reviewer Comment:** *Explain what is N, this is important.*

Author Response: We have modified this part.

30. **Reviewer Comment:** *You clearly introduce an notation for $l1$ norm before. Be consistent. Also explain what are A and b.*

Author Response: We have specified a previous definition $|\cdot|$ for l_1 norm on multi-index. We have added the explanation of A and b.

31. **Reviewer Comment:** *page 6: Missing equation number after 2.5.*

Author Response: We appreciate the suggestion. As this equation is not referred elsewhere in the manuscript, we skipped the equation label.

32. **Reviewer Comment:** *Page 7, Def 2.2: s = 1,2, ..., up to ?*

Author Response: We have modified this part.

33. **Reviewer Comment:** *Page 7, line 139: correct ϵ.*

Author Response: We have corrected this typo.

34. **Reviewer Comment:** *Page 7, Th 2.3: Define M and is N same as page 6?*

Author Response: M refers to the row of matrix $\boldsymbol{B}$. This theorem holds under the conditions specified and N is not necessarily any specific value.

35. **Reviewer Comment:** *Page 7, Def 2.4: n defined after.*

Author Response: We have modified this part.

36. **Reviewer Comment:** *page7: Missing two equations number before 2.8 and one after 2.9.*

Author Response: We appreciate the suggestion. As this equation is not referred elsewhere in the manuscript, we skipped the equation label.

37. **Reviewer Comment:** *Page 6-7: Introduce beforehand what you will express mathematically. The list of definition and theorem as it appears unclear. Also, what is the conclusion of this subsection(CS)?*

Author Response: We have modified this part with a few transition sentences.

38. **Reviewer Comment:** *Page 8, lines 150-155: The definition of A should appear in the CS subsection. Some elements found here could serve as an introduction to CS subsection.*

Author Response: The $\boldsymbol{A}$ in CS subsection refers to a matrix considered in the linear system $\boldsymbol{Ac} = \boldsymbol{b}$. The theory of CS discussed in Section 2 holds in general. As an application, we specify $\boldsymbol{A}$ as the measurement matrix and $\boldsymbol{b}$ as the samples of QoI to recover the coefficient $\boldsymbol{c}$. Therefore, we specified the definition of $\boldsymbol{A}$ in Section 3.

39. **Reviewer Comment:** *Page 8, line 155: our approach, which is? From the text before you presented previous work and your contribution does not appear.*

Author Response: We have modified this sentence.

40. **Reviewer Comment:** *page 8, line 159: Clarify notations of the random vector. Here ξ and in page 6 line 128 it is z. Same goes with N sand M.*

Author Response: ξ in Section 3 is related to $\boldsymbol{z}$ in Section 2. $\boldsymbol{z}$ in Section 2 represents a random vector as the input of a multivariate polynomial function, and ξ in Section 3 refers to the random vector of the surrogate model we aim to construct. Later on, we constructed the surrogate model with respect to the rotated random vector χ. Under such consideration,

we employed different notations in section 2 and 3. Here we use N_s with a subscript "s" to represent the total number of samples in set S.

41. **Reviewer Comment:** *Page 10: see previous remark, N is not the same as Ns but this is really confusing as there is no clear explanation of N*

Author Response: N is the number of basis specified in Eq. (2.5) and text following in the revised manuscript.

42. **Reviewer Comment:** *page 10: Missing equation number before 3.8.*

Author Response: We appreciate the suggestion. As this equation is not referred elsewhere in the manuscript, we skipped the equation label.

43. **Reviewer Comment:** *Page 13, Eq 3.14: By using gradient information, you are not ensured to have the global interaction of the parameters. Only global sensitivity indices such as Sobol can guarantee it. Especially in high dimensions with a low number of samples, if the function is not smooth enough you can be assured that gradient information will be erroneous.*

Author Response: The gradient information alone is insufficient to characterize the global interaction of the parameters, and the accuracy of the gradient matrix is not guaranteed if QoI is non-smooth. This is the major reason we did not conduct further dimension reduction following the evaluation of the gradient matrix. Instead, we conducted eigendecomposition and establish a linear transform of the random vector $\boldsymbol{\xi}$ into $\boldsymbol{\chi}$ (with the same dimension) such that the coefficient vector $\boldsymbol{c}$ with respect to $\boldsymbol{\chi}$ is sparser (therefore more favorable for CS), if the gradient matrix can partially capture the variability of QoI along the d directions. There could be alternative approaches to establish such transformation (i.e., via nonlinear transformation), which admit sparse representation of QoI. This direction is worth further exploration in future studies.

44. **Reviewer Comment:** *Page 15, line 254: Better define ω*

Author Response: We have added a definition to the paragraph.

45. **Reviewer Comment:** *page 15: Missing equation number after 4.3.*

Author Response: We appreciate the suggestion. As this equation is not referred elsewhere in the manuscript, we skipped the equation label.

46. **Reviewer Comment:** *page17, fig1: The mean is interesting but figure would benefit from information about the variance also.*

Author Response: We have added error bars.

47. **Reviewer Comment:** *Page 24, line 360: The motivation example should be a separated subsection to separate analytical tests from it.*

Author Response: We have separated this example as a new subsection.

48. **Reviewer Comment:** *Page27, lines 407-408: To be correct, you have done an uncertainty propagation study. UQ is a more general term encompassing sensitivity analysis as well.* **Author Response:** We have modified this sentence.

49. **Reviewer Comment:** *Ref 56 missing year.* **Author Response:** We have included the year.

50. **Reviewer Comment:** *Ref 60 solver.* **Author Response:** We have fixed the typo.

Reviewer 4

Reviewer Comment: *The manuscript deals with an approach designed to enhance the performance of forward Uncertainty Quantification analysis in the presence of high-dimensionality and arbitrary probability density distributions of the inputs. The two main components of the framework are (i) the possibility to construct the optimal (orthogonal/near-orthogonal) polynomial basis for the compressive sensing and (ii) the use of a Active Subspace based rotation to enhance the sparsity of the coefficient vector. The manuscript consists of a theoretical introduction followed by a set of numerical results where the proposed approach is compared against other related strategies.*
I found the manuscript overall carefully written, the idea interesting, the numerical examples relevant and I think it perfectly fits the scope of the Journal. However, I would recommend some modifications before accepting it for publication. In the following I report my suggestions without ranking them in a particular order of importance.

1. **Reviewer Comment:** *The introduction presents a relevant literature survey. I found also interesting and appropriate that the authors include a reference to Monte Carlo sampling methods as a robust and alternative way to deal with the same problem. My suggestion here is to extend this reference to Monte Carlo by adding some of the most recent directions in this context. I am referring to the large recent literature based on multilevel and multifidelity sampling strategies. The reason for that is that I believe the approach proposed in this work is somehow complementary with respect to sampling based approach because it is more effective when the dimensionality of the input space is not extremely large and the solution is regular. If those conditions are not met, it is still mandatory to rely on sampling based approach. I think that adding a comment along those lines might help the reader.*

Author Response: Recent work on multilevel-MC and multifidelity-MC have been developed to improve sampling efficiency and are relevant within the scope of Monte Carlo sampling. We appreciate the referee pointed this out and we have added this discussion to the revised manuscript.

2. **Reviewer Comment:** *Please check page 7 line 14*

Author Response: We would like to thank the referee for careful reading of our manuscript. We have fixed this typo in the revised manuscript.

3. **Reviewer Comment:** *I really found the discussion carried over at pages 11 and 12 about the near-orthonormal basis very interesting. However, only the case of two disjoint sets is considered, whereas in many cases one would like to add new samples to a previous computation. This might happen for instance if new data become available after the first regression is performed. What does it happen in this case? Is the orthonormality retained? Should Algorithm 3 be modified? I think it is fair to mention this issue without fully addressing it and, for instance, add it as one of the future directions.*

Author Response: In this study, we construct the near-orthonormal basis in a heuristic way by directly splitting the sampling set S into two disjoint sets. In practice, if the new available data set shares the same underlying probability $\rho(\cdot)$ with the previous sample set S, then constructed bases will retain proper orthonormality. However, it is worth exploring how to utilize the new data set to design some more sophisticated (cross-validation) strategies to further optimize the orthonormal threshold values and the near-orthonormal basis construction procedure. We have added a brief discussion in the revised manuscript.

4. **Reviewer Comment:** *I would suggest to be more explicit about the amount of data*

required for the identification of the probability measure. In the summary the authors clearly state that this is possible under the assumption that a large collection of samples is available. Even if this might be the case in some applications, in many other cases the problem is that a limited number of samples are available. I would recommend to include a comment on that already in the abstract/introduction.

Author Response: For the case of implicit distributions, the current study assumes that a large collection of samples is available, which represents the underlying probability measure (while the size of the training set can still be small). We have included this comment in the abstract and introduction of the revised manuscript.

5. **Reviewer Comment:** *I found the numerical results interesting and relevant. However, I found the introduction of Section 4 (Results) a little bit too concise. In particular the strategy adopted for the numerical tests is not completely clear to me. The data set S is divided (evenly) in two subsets (I assume with the intersection set empty) and the first is used for constructing the basis, while the second is used to construct and test the surrogate. We have here two training sets S_1 and part of $S_{2,1}$ which are used to build the basis and to compute the coefficients, and a set $S_{2,2} = S_2 \ S_{2,1}$ which should be used for the test of the surrogate. My main point is that the surrogate should be built on a training data set which need to be disjoint from the test dataset. I think the authors need to clarify this point and make it more explicit in the text (i.e. how many training points? how many test points? etc.).*

Author Response: The numerical tests were conducted exactly as the referee described above. The training set $S_{2,1}$ and the test set $S_{2,2}$ are disjointly chosen from S_2. The size of the training set is $O(10^2) - O(10^3)$ and size of the test set is $O(10^5)$ in the present study. We have clarified this in the revised manuscript.

6. **Reviewer Comment:** *Related to the previous point, for instance if one looks at Section 4.1 the presentation of the numerical experiment gets a little bit confusing (at least for me). $S = 2 \times 10^5$, so S_1 and S_2 are 10^5, however the minimization problem is solved for increasing M (number of measurements) which I think it is the right thing to do. However, it is not clear what happens to S_1, is the cardinality of S_1 also equal to M? It would be inconsistent in my opinion to use the full set of samples S_1 to build the base and M samples for the output. I suggest to clarify this point.*

Author Response: In this numerical test, the cardinality of S_1 is 10^5. $M \sim O(10)$ represents the size of the training set $S_{2,1}$ chosen from S_2. S_1 and $S_{2,1}$ are disjoint sets (please also refer to the above response.) The motivation is to examine, for various bases, how many training points (e.g., M) is needed to accurately recover the vector $\mathbf{c}$. We have clarified this in the revised manuscript.

Reviewer 5 Reviewer Comment: *This paper addresses some challenging problems in real-life applications that exhibit randomness with imprecise and non-standard input distributions often represented by a finite set of samples. I think this paper is worth to be published, but please note the following comments and suggestions.*

1. **Reviewer Comment:** *First of all, the authors are commended for a relatively comprehensive survey of the current state-of-the-art methods setting a good stage for subsequent sections.*

Author Response: We appreciate the referee's comment on the literature review of our manuscript. Following the suggestions of Review 4, we have also added the relevant literatures of recent work on multilevel-MC and multifidelity-MC in the scope of Monte Carlo sampling in the revised manuscript.

2. **Reviewer Comment:** *The reference to 'high-dimensionality' in the paper seems to be somewhat misleading. The highest dimension studied in the numerical experiments is about 20. Typically, high-dimensionality refers to hundreds to millions of dimensions such as those corresponding to random fields. Sparsification and active subspace were mentioned as ways to reduce the number of effective basis functions, but they may not be able to handle really 'high' dimensions. Claiming to deal with high dimensionality may invite many questions about why this or that machine learning method is not considered. So it is better to clarify what type of high-dimensional problems you are trying to address.*

Author Response: Indeed we address the problems with arbitrary randomness in dimension of $O(10)$ in the present study. In the original version of the manuscript, we referred to "high-dimensional" in the scope of stochastic modeling of molecular systems (within the conformational space). In this field, the number of collective variables (i.e., the dimension of randomness) is often limited to 5 or less due to the numerical challenges in accurate quantification of free energy landscape (e.g., please refer to molecular system where the underlying density is approximated by Gaussian mixture in Sec. 4.3.3) The data-driven and sparsity enhancement approach developed in the present study provides a work-around approach to accurately quantify the uncertainty of such systems.
We appreciate the referee for pointing out this possible confusion. In the revised manuscript, we have specified the dimensionality of problems we aim to address.

3. **Reviewer Comment:** *The authors should spell out what is new in the orthonormalization procedure, i.e. what is new compared to Dunkl and Xu's approach? I think the orthonormalization procedure is somewhat basic. The potentially interesting and useful idea is near-orthonormalization because in practice the data set may not be complete. I suggest more focus be put on demonstrating the effectiveness and limitations of near-orthogonality. For example, In Remark 3.2 you mentioned that Eqn 3.9 "provides a heuristic approach to construct near-orthonormal basis." While this heuristic currently is not accompanied by any rigorous error analysis but it seems to outperform the exact orthonormal basis, it is desirable to add some numerical experiments to study how performance degrades when the relaxation parameters varies (that is, how —c-c'— deteriorates with respect to loss of orthonormality?) The authors do state in the Conclusion section that this is something they plan to investigate, but the inclusion of a small numerical study in this paper will make the paper more complete.*

Author Response: The orthonormal basis is constructed following the Dunkl and Xu's approach. We highlighted this in the revised manuscript. In the current study, we did not include a systematic study of the near-orthonormal basis. In our understanding, the differ-

ent numerical performance between the orthonormal and near-orthonormal bases arises from the different null spaces of the measurement matrices constructed from the two bases. Accordingly, there might be alternative approach to optimize the basis construction; however, we feel that it is best to defer such investigation to a separate study.

4. **Reviewer Comment:** *It will be more informative if the authors include a cost comparison between Algorithm 4 and Algorithm 4 without the use of different bases (χ) - that is, Step 4-6. Numerically, the examples show better accuracy of the former but for what additional costs*

Author Response: The additional cost of Step 4 - 6 mainly consists of two parts: construction of the basis with respect to χ and construction of the surrogate model $\tilde{f}(\chi)$. For the current numerical examples considered in the present study, this additional cost is less than 0.6 CPU (3.7 GHz Quad-Core Intel Xeon E5) hour in general. For realistic applications, the overhead of Step 4 - 6 could be relatively small if sampling of QoI is expensive or the available training set is limited. We have included this discussion in the revised manuscript.

5. **Reviewer Comment:** *Minor: on p. 21 line 34: What is Λ_p^d?*

Author Response: Λ_p^d is a multi-indices set defined in Eq. (2.4). We have clarified this in the revised manuscript.

References

[1] G. Chastaing, F. Gamboa, and C. Prieur. Generalized sobol sensitivity indices for dependent variables: numerical methods. *Journal of Statistical Computation and Simulation*, 85(7):1306–1333, 2015.

[2] Paul G Constantine, Eric Dow, and Qiqi Wang. Active subspace methods in theory and practice: Applications to kriging surfaces. *SIAM J. Sci. Comput.*, 36(4):A1500–A1524, 2014.

[3] Amandine Marrel, Bertrand Iooss, Béatrice Laurent, and Olivier Roustant. Calculations of sobol indices for the gaussian process metamodel. *Reliability Engineering & System Safety*, 94(3):742 – 751, 2009.

[4] Pamphile T. Roy, Nabil El Moçayd, Sophie Ricci, Jean-Christophe Jouhaud, Nicole Goutal, Matthias De Lozzo, and Mélanie C. Rochoux. Comparison of polynomial chaos and gaussian process surrogates for uncertainty quantification and correlation estimation of spatially distributed open-channel steady flows. *Stochastic Environmental Research and Risk Assessment*, 32(6):1723–1741, Jun 2018.

[5] Trent Michael Russi. *Uncertainty Quantification with Experimental Data and Complex System Models*. PhD thesis, University of California, Berkeley, 2001.

Highlights

- Surrogate construction for arbitrary mutually dependent probability measures
- A general framework for both explicit non-Gaussian densities and implicit probability measures
- Data-driven orthonormal-basis construction coupled with sparsity-enhancing rotations
- Application to molecular systems with non-Gaussian mutually dependent randomness

A data-driven framework for sparsity-enhanced surrogates with arbitrary mutually dependent randomness

Huan Lei[a,1,*], Jing Li[a,1], Peiyuan Gao[a], Panagiotis Stinis[a,b], Nathan A. Baker[a,c,*]

[a] *Pacific Northwest National Laboratory, Richland, WA 99352.*
[b] *Department of Applied Mathematics, University of Washington, Seattle, WA 98195.*
[c] *Division of Applied Mathematics, Brown University, Providence, RI 02912.*

Abstract

The challenge of quantifying uncertainty propagation in real-world systems is rooted in the high-dimensionality of the stochastic input and the frequent lack of explicit knowledge of its probability distribution. Traditional approaches show limitations for such problems, especially when the size of the training data is limited. To address these difficulties, we have developed a general framework of constructing surrogate models on spaces of stochastic input with arbitrary probability measure irrespective of the mutual dependencies between individual components of the random inputs and the analytical form. The present *Data-driven Sparsity-enhancing Rotation for Arbitrary Randomness* (DSRAR) framework includes a data-driven construction of multivariate polynomial basis for arbitrary mutually dependent probability measures and a sparsity enhancement rotation procedure. This sparsity-enhancing rotation method was initially proposed in our previous work [1] for Gaussian density distributions, which may not be feasible for non-Gaussian distributions due to the loss of orthogonality after the rotation. To remedy such difficulties, we developed a new data-driven approach to construct orthonormal polynomials for arbitrary mutually dependent randomness, ensuring the constructed basis maintains the orthogonality/near-orthogonality with respect to the density of the rotated random vector, where directly applying the regular polynomial chaos including arbitrary polynomial chaos (aPC) [2] shows limitations due to the assumption of the mutual independence between the components of the random inputs. The developed DSRAR framework leads to accurate recovery, with only limited training data, of a sparse representation of the target functions. The effectiveness of our method is demonstrated in challenging problems such as partial differential equations and realistic molecular systems within high-dimensional ($O(10)$) conformational spaces where the underlying density is implicitly represented by a large collection of sample data, as well as systems with explicitly given non-Gaussian probabilistic measures.

Keywords: data-driven, arbitrary randomness, mutual dependence, compressed sensing, sparsity enhancement, uncertainty quantification

*Corresponding author
Email addresses: huan.lei@pnnl.gov (Huan Lei), nathan.baker@pnnl.gov (Nathan A. Baker)
[1] The first two authors contributed equally

1. Introduction

A fundamental problem in uncertainty quantification (UQ) [3] is to calculate the statistical properties of a quantity of interest (QoI) due to various sources of randomness, e.g., numerical simulations subject to uncertain parameters, initial conditions and/or boundary conditions, as well as experimental measurements in the presence of material heterogeneity, thermal fluctuations. Such sources of uncertainty are usually characterized by high-dimensional random variables whose probability measures can be either discrete or continuous. In real-world systems, there are usually two crucial challenges to accurately quantify the propagation of the randomness from the input to the system response. The first challenge comes from the high-dimensionality of the random inputs. For such systems, limited computational resources often motivates further dimensionality reduction [4]. However, it is often non-trivial to accurately transfer the high-dimensional random space into a low-dimensional random space. This results in the numerical intractability of quantifying the uncertainty of the QoI from training data of limited size. The second challenge arises from frequent dependencies and arbitrary distribution of the random inputs. Typically, random inputs are represented by random vectors with mutually independent components. For realistic systems, the underlying distribution of the inputs can often involve dependencies that cannot be ignored (e.g., see molecule systems in Ref. [5] and Sec. 4.4). Moreover, the input distribution could be even unknown and thus we may only have access to it implicitly through a collection of samples. This creates further numerical obstacles in characterizing the random inputs as well as their effect on the system response. In the current work, we present a *Data-driven Sparsity-enhancing Rotation for Arbitrary Randomness* (DSRAR) framework for dealing with all of the aforementioned challenges. While we focused on numerical experiments in the present study, the developed framework can be also applied to UQ in experimental studies.

In practice, a straightforward and robust approach is the Monte Carlo (MC) method, which involves collecting a large number of samples of the random inputs from their distribution, evaluating the QoI at each sample point, and then obtaining the statistical properties (mean, variance, sensitivity indices, probability density function, probability of a certain event etc.) of the QoI. Unfortunately, to get an accurate estimate, the MC method requires a large number of simulations due to its slow convergence rate [6, 7]. Furthermore, for large or complex systems, even a single instance of these simulations may require very large computational resources. Under such circumstances, the computational cost of MC method can become extremely large. Several approaches have been developed to alleviate such difficulties. For instance, sampling approaches such as multilevel-MC [8, 9, 10] and multifidelity-MC [11, 12] have been designed to optimize the computational load when samples of the QoI are available at hierarchical levels of accuracy; sampling approaches like quasi-MC [13, 14, 15] and Latin Hypercube sampling [16, 17, 18], have been designed to accelerate convergence. However, when the underlying distribution of the inputs is arbitrary and not explicitly given, the latter two sampling strategies may lose their advantage if it is not straightforward to generate quasi-random sequences

following the underlying distribution.

An alternative approach approximates the QoI via constructing the surrogate model of the random inputs and then calculates the statistics of the QoI analytically or numerically. Among such approaches, the most popular are the Gaussian Process [19, 20, 21], and the polynomial chaos expansion originally introduced by Wiener [22], applied to UQ by Ghanem [23] and extended to the generalized polynomial chaos (gPC) expansion by Xiu [24]. The Gaussian Process (GP) is a stochastic process which approximates the values of the QoI at every finite sets of sample point as multivariate Gaussian random vectors. The flexibility of the mean and covariance functions enables GP to characterize a wide range of function behavior with broad applications on UQ [21, 25, 26, 27, 28]. The gPC expansion approximates the QoI by a set of simple basis functions. It is known to be a *mathematically optimal* approximation of the QoI when the basis functions are chosen to be orthogonal with respect to the probability measure of the random inputs. This approach has been demonstrated for diverse applications in UQ [29, 30, 31, 32, 33, 34, 35, 36] due to its spectral convergence under certain situations. In this study, we focus on the approach developed based on gPC and we refer to previous publications [37, 38, 39, 40] (and the references therein) for comparative studies of the two approaches.

In principle, if the orthogonal polynomial type and the corresponding random variables are determined, both intrusive and non intrusive methods can be used to evaluate the coefficients of the expansion. For example, stochastic collocation, based on tensor products of one-dimensional quadrature rules, is often employed when dimensionality is small [41, 42, 43], with the number of basis functions given by $(p+d)!/p!d!$, where p is polynomial order and d is the dimension. However, as the dimension increases, the number of quadrature points needed for the tensor product rule increases exponentially. To mitigate this issue, sparse grid and adaptive collocation methods have been proposed to deal with moderate dimensionality [42, 44, 45, 46, 47, 48, 49]. When the dimension of the random inputs is large, none of the above collocation methods is feasible. In the case of a limited number of available simulations and large dimensionality, compressed sensing (CS) approaches have been used to construct sparse polynomial approximations of the QoI [50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62]. Finally, we note that gPC (including extensions such as arbitrary polynomial chaos [2]), in its current form, can only handle random vector with independent identically distributed (i.i.d.) components in standard types (uniform, Gaussian, gamma, beta, etc.). For other distributions, a pre-processing step is required to transform the original random variables into i.i.d. random variables of standard types. In general, these transformations are highly nonlinear which result in the final QoI function approximation to be a high-degree polynomial in order to maintain accuracy.

The methods discussed above rely on the *explicit* knowledge of the underlying probability measures and/or the assumption of mutual independence between the components of the random inputs. However, such assumptions on the random inputs can be quite restrictive for realistic applications. One such example

is the UQ for molecular system properties QoIs due to conformational fluctuations [63]. For such systems, the random inputs are the various conformational states (i.e., the instantaneous structure) of the molecule. The underlying distribution is determined by the free energy function of the system, which is essentially the multi-dimensional marginal density distribution with respect to the (Boltzmann) distribution of the full Hamiltonian system. Unfortunately, numerical evaluation of the free energy function is a well-known challenging problem. Although various sampling strategies have been developed [64, 5, 65], the explicit free energy function is usually unknown for dimensions greater than 4. In practice, the underlying density is only known implicitly through a large collection of the molecule conformational states obtained from experiments or simulated trajectories. Another commonly encountered example arises in our recent work [1] on constructing sparse representations of a QoI based on CS. Inspired by the active subspace method [66], we proposed a method to enhance the sparsity of polynomial expansion in terms of a new random vector via unitary rotation of the original random vector. For i.i.d. Gaussian random inputs, the new random vector retains the same distribution. However, for non-Gaussian random inputs, which are more realistic for applications, the new random vector does not retain the mutual independence even if the original random vector elements are i.i.d.

For problems with non-Gaussian random inputs, the traditional approach is to cast the available statistics into a family of standard distributions and then to apply the gPC techniques discussed above. Gaussian mixture models, due to their flexibility, are broadly employed to approximate the distribution of the data. With the distribution approximated, a gPC expansion of the QoI can be constructed for each Gaussian component. The statistical properties of the QoI are derived by combining the statistical properties of all components [67, 68]. However, there are two drawbacks of the Gaussian mixture approach: (i) it lacks one-to-one correspondence between one instance of random inputs and the approximated function evaluation, (ii) it is difficult to determine an appropriate and accurate probability density approximation when the dimension is larger than one. Copulas have been employed to treat dependent probabilistic models for surrogate construction in [69]. Zabaras [70] has established a graph-based approach to factorize the joint distribution into a set of conditional distributions based on the dependence structure of the variables. Alternatively, several studies have been devoted to constructing orthogonal polynomial bases using the moments of the random variables. Orthogonal polynomial chaos for random vectors with independent components of arbitrary measure was proposed in [2, 71, 72, 73, 74]. Ahlfeld investigated the quadrature rule of this arbitrary polynomial chaos (aPC) and proposed a sparse quadrature rule for the integration which can facilitate the evaluation of the expansion coefficients [75]. However, those quadrature rules of arbitrary polynomial chaos again assume *the components of the random inputs are mutually independent.*

In this paper, we develop a general UQ framework for constructing surrogate models via DSRAR *irrespective of possible mutual dependencies between the random input components.* This approach is different

from the aforementioned studies based on polynomial chaos expansions and, therefore, can be particularly useful for realistic systems where the input distributions can be non-standard or unknown analytically. The key idea is a data-driven approach for basis construction, consisting of multivariate orthonormal *polynomials for arbitrary mutually dependent* (amdP), coupled with the previously developed rotation-based sparsity enhancement approach [1]. This can be viewed a special case of the present method when the random inputs are from a Gaussian distribution. When the size of the training set is limited, the method can recover the expansion coefficients by CS, under the assumption that there exists a sparse representation of surrogate model. As we will show, directly employing a regular polynomial basis and/or the sparsity enhancement rotation on the random input may result in large recovery error due to the violation of orthogonality for non-standard density distributions. The procedure of data-driven basis construction described in the present study retains proper orthogonality with respect to the associated random inputs and therefore ensures more accurate recovery. In this sense, the present method takes advantage of both the orthonormal basis expansion and the enhanced sparsity of the expansion coefficients. The method deals with two situations widely encountered in real-world applications: (I) probability measures that are implicitly represented by a large collection of samples and (II) non-Gaussian probability measures with explicit (analytical) forms. For the first situation, we construct orthonormal polynomial bases with respect to discrete measures on the sample set. Besides the exact orthonormal basis, we also propose a heuristic method to construct a *near-orthonormal* basis, which yields a smaller basis bound than the exact orthonormal basis and results in more accurate recovery of the sparse representation. For the second situation, we construct the orthonormal basis when the quadrature rules for polynomial integration are known. This construction is especially well suited to random variables obtained from sparsity enhancement of non-Gaussian distributions.

The paper is organized as follows. In Section 2, we present the problem setup and briefly review preliminary background on multivariate orthogonal polynomials and compressed sensing. In Section 3, we present the DSRAR framework by first introducing the methods to construct data-driven orthonormal amdP basis. When the underlying density is implicitly represented by a large collection of random input samples, we propose a heuristic approach to construct a near-orthonormal basis along with some heuristics on the advantage over an exactly orthonormal basis. Then we introduce the rotation-based sparsity enhancement method and provide algorithmic details on how to combine the data-driven basis construction and sparsity enhancement rotation. In Section 4, we demonstrate the developed framework in a realistic molecular system fluctuating in a high-dimensional conformational space ($O(10)$) as well as partial differential equations (PDEs) with arbitrary randomness where the underlying distributions are either explicitly known or implicitly represented by a large collection of samples. Concluding remarks and directions for future work are provided in Section 5.

2. Background

2.1. Approximation with orthogonal polynomials

We begin with a few facts about multivariate orthogonal polynomials [76]. Let Π^d be the set of polynomials in d variables on $\mathbb{R}^d$. Polynomials in Π^d are naturally indexed by the multi-indices set $\mathbb{N}_0^d$. For $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$ and $\boldsymbol{z} = (z_1, \ldots, z_d)$, a monomial z^α is defined by $\boldsymbol{z}^{\boldsymbol{\alpha}} = z_1^{\alpha_1} \cdots z_d^{\alpha_d}$ and the degree of $\boldsymbol{z}^{\boldsymbol{\alpha}}$ is defined by $|\boldsymbol{\alpha}| = a_1 + \cdots + a_d$. From now on, without confusion, $|\cdot|$ operating on a multi-index $\boldsymbol{\alpha}$ denotes the ℓ_1 norm of $\boldsymbol{\alpha}$ while $|\cdot|$ operating on a set T denotes the cardinality of T. The degree of a polynomial is defined by the largest degree of its monomial terms. Then the space of polynomials of degree at most p is defined by

$$\Pi_p^d := \operatorname{span}\{\boldsymbol{z}^{\boldsymbol{\alpha}} : |\boldsymbol{\alpha}| \leq p, \boldsymbol{\alpha} \in \mathbb{N}_0^d\} \text{ and } \dim \Pi_p^d = \begin{pmatrix} p+d \\ p \end{pmatrix}. \tag{2.1}$$

If we equip $\mathbb{R}^d$ with a probability measure ρ, then we can define an inner product on Π^d,

$$\langle f, g \rangle_\rho = \int_{\mathbb{R}^d} fg \, \mathrm{d}\rho \qquad f, g \in \Pi^d. \tag{2.2}$$

f and g are said to be orthogonal with respect to ρ if $\langle f, g \rangle_\rho = 0$. Given such an inner product, and an order of the set $\mathbb{N}_0^d$, we can apply the Gram-Schmidt process on the ordered set $\{\boldsymbol{z}^{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \mathbb{N}_0^d\}$ to generate a sequence of orthogonal polynomials. We will revisit this construction in Section 3.1. When $d > 1$, there is no natural order among monomials. As a result, multivariate orthogonal polynomials are, in general, not uniquely determined. In this paper, we choose the *graded lexicographic order* when applying the Gram-Schmidt process, that is, $\boldsymbol{z}^{\boldsymbol{\alpha}} \succ \boldsymbol{z}^{\boldsymbol{\beta}}$ if $|\boldsymbol{\alpha}| > |\boldsymbol{\beta}|$ or if $|\boldsymbol{\alpha}| = |\boldsymbol{\beta}|$ and the first nonzero entry in the difference $\boldsymbol{\alpha} - \boldsymbol{\beta}$ is positive.

When a simulation model is expensive to run, building an approximation of the response of the model output with respect to the variations in the model input can often be an efficient approach to quantify uncertainty propagation. The polynomial approximation of a function (model) $f(\boldsymbol{z}) : \mathbb{R}^d \to \mathbb{R}, d \geq 1$ where $\boldsymbol{z} = (z_1, \ldots, z_d) : \Omega \to \mathbb{R}^d$ is a d-dimensional random variable with associated probability measure $\rho(\boldsymbol{z})$, which is widely used due to its fast convergence when $f(\boldsymbol{z})$ is analytic. In this paper, we will approximate f using an orthogonal polynomial basis. It is a generalization of the gPC expansion which usually deals with i.i.d. random variables.

Let $\Psi = \{\psi_{\boldsymbol{\alpha}}(\boldsymbol{z}) : \boldsymbol{\alpha} \in \mathbb{N}_0^d\}$ be a set of orthonormal polynomial basis of Π^d associated with the measure $\rho(\boldsymbol{z})$, that is,

$$\int \psi_{\boldsymbol{\alpha}}(\boldsymbol{z}) \psi_{\boldsymbol{\beta}}(\boldsymbol{z}) \, \mathrm{d}\rho(\boldsymbol{z}) = \delta_{\boldsymbol{\alpha}\boldsymbol{\beta}}, \quad \boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{N}_0^d, \tag{2.3}$$

where $\delta_{\boldsymbol{\alpha\beta}} := \prod_{i=1}^{d} \delta_{\alpha_i,\beta_i}$ to be the multi-index Kronecker delta. Then the pth-degree arbitrary orthogonal polynomial expansion $f_p(\boldsymbol{z})$ of function $f(\boldsymbol{z})$ associated with ψ is defined as,

$$f(\boldsymbol{z}) \approx f_p(\boldsymbol{z}) := \sum_{\boldsymbol{\alpha}\in\Lambda_p^d} c_{\boldsymbol{\alpha}}\psi_{\boldsymbol{\alpha}}(\boldsymbol{z}), \quad \Lambda_p^d = \left\{\boldsymbol{\alpha}\in\mathbb{N}_0^d : |\boldsymbol{\alpha}| \le p\right\}, \tag{2.4}$$

where $c_{\boldsymbol{\alpha}}$ is the coefficient to be evaluated. Using an ordering of the orthonormal polynomial basis, we can change (2.4) into the following single index version

$$f_p(\boldsymbol{z}) = \sum_{\boldsymbol{\alpha}\in\Lambda_p^d} c_{\boldsymbol{\alpha}}\psi_{\boldsymbol{\alpha}}(\boldsymbol{z}) = \sum_{n=1}^{N} c_n\psi_n(\boldsymbol{z}), \tag{2.5}$$

where N is the total number of basis and is given by

$$N = \dim \Pi_p^d = |\Lambda_p^d| = \begin{pmatrix} d+p \\ p \end{pmatrix}.$$

2.2. Compressed sensing

Compressed sensing is a well-studied and popular approach to find sparse solutions to linear equations [77, 78, 79, 80]. In this subsection, we briefly review the theory of CS and discuss the conditions which allow accurate recovery of solutions to underdetermined linear system.

Under certain assumptions, the solution—or its approximation—can be found by the well-studied ℓ_1 minimization, i.e., finding the minimizer

$$\min \|\boldsymbol{c}\|_1 \quad \text{subject to } \boldsymbol{Ac} = \boldsymbol{b}, \tag{2.6}$$

where $\boldsymbol{A} \in \mathbb{R}^{M\times N}$, $\boldsymbol{b} \in \mathbb{R}^M$ and $\|\boldsymbol{c}\|_1 = \sum_{i=1}^{N} |c_i|$ is the ℓ_1 norm of the vector $\boldsymbol{c}$.

When the data $\boldsymbol{b}$ is contaminated by noise, the constraint in (2.6) is relaxed to obtain the basis pursuit denoising problem,

$$\min \|\boldsymbol{c}\|_1 \quad \text{subject to } \|\boldsymbol{Ac} - \boldsymbol{b}\|_2 \le \sigma, \tag{2.7}$$

where σ is an estimate of the ℓ_2 norm of the noise. The optimization problems (2.6) and (2.7) can be solved with efficient algorithms from convex optimization [81].

Next we discuss the conditions for the sparse recovery of $\boldsymbol{c}$.

Definition 2.1. A vector $\boldsymbol{c}$ is said to be s-sparse if it has at most s nonzero entries, i.e., $\boldsymbol{c}$ is supported on $T \subset \{1,\dots,N\}$ with $|T| \le s$.

Definition 2.2 (Restricted isometry constant [82, 83])**.** For each integer $s = 1, 2, \dots, N$ define the isometry constant δ_s of a matrix $\boldsymbol{A}$ as the smallest number such that

$$(1-\delta_s)\|\boldsymbol{c}\|_2^2 \le \|\boldsymbol{Ac}\|_2^2 \le (1+\delta_s)\|\boldsymbol{c}\|_2^2$$

holds for any s-sparse vector $\boldsymbol{c} \in \mathbb{R}^N$.

The restricted isometry constants (RICs) characterizes matrices that are nearly orthonormal. The spare recovery is established by the following theorem.

Theorem 2.3 (Sparse Recovery for restricted isometry property (RIP)-Matrices)**.** *Let* $\boldsymbol{A} \in \mathbb{R}^{M\times N}$*. Assume that its isometry constant* δ_{2s} *satisfies* $\delta_{2s} < 0.4931$*. Let* $\boldsymbol{c} \in \mathbb{R}^N$*, and assume noisy measurements* $\boldsymbol{b} = \boldsymbol{A}\boldsymbol{c}+\boldsymbol{\eta}$ *are given with* $\|\boldsymbol{\eta}\|_2 \le \sigma$*, then the minimizer* $\boldsymbol{c}^*$ *of*

$$\min \|\boldsymbol{c}\|_1 \quad \textit{subject to } \|\boldsymbol{A}\boldsymbol{c}-\boldsymbol{b}\|_2 \le \sigma,$$

satisfies

$$\begin{aligned} \|\boldsymbol{c}-\boldsymbol{c}^*\|_2 &\le C_1 \frac{\sigma_s(\boldsymbol{c})}{\sqrt{s}} + C_2\sigma, \\ \|\boldsymbol{c}-\boldsymbol{c}^*\|_1 &\le C_3\sigma_s(\boldsymbol{c}) + C_4\sqrt{s}\sigma. \end{aligned} \tag{2.8}$$

where constants C_1*,* C_2*,* C_3 *and* C_4 *depend only on* δ_{2s}*, and* $\sigma_s(\boldsymbol{c}) = \inf_{\boldsymbol{c}_s:\|\boldsymbol{c}_s\|_0\le s} \|\boldsymbol{c}-\boldsymbol{c}_s\|_1$ *with* $\|\boldsymbol{c}_s\|_0$ *indicates the number of nonzero entries of* $\boldsymbol{c}_s$ *In particular, if* $\boldsymbol{c}$ *is s-sparse, then the reconstruction is exact.*

Proof. See Rauhut and Ward [52]. □

A bounded orthonormal system has the following definition.

Definition 2.4. $\{\psi_n\}, n = 1,\dots,N$ is a bounded orthonormal system, if

$$K := \sup_n \|\psi_n\|_\infty = \sup_n \sup_{\boldsymbol{z}} |\psi_n(\boldsymbol{z})| < \infty, \tag{2.9}$$

where K is called the basis bound.

These definitions allow us to establish the recoverability of (2.6) based on the RIP.

Theorem 2.5 (RIP for bounded orthonormal systems)**.** *Let* $\boldsymbol{A} \in \mathbb{R}^{M\times N}$ *be the interpolation matrix with entries* $\{a_{j,n} = \psi_n(\boldsymbol{z}^{(j)})\}_{1\le n\le N, 1\le j\le M}$ *(see* (3.2)*), where* $\{\psi_n\}$ *is a bounded orthonormal system satisfying* (2.9)*. Assume that*

$$M \ge C\delta^{-2}K^2 s \log^3(s)\log(N),$$

then with probability at least $1 - N^{-\gamma\log^3(s)}$*, the RIC* δ_s *of* $1/\sqrt{M}\boldsymbol{A}$ *satisfies* $\delta_s \le \delta$*. Here,* $C, \gamma > 0$ *are universal constants.*

Proof. See Rauhut and Ward [52]. □

Theorem 2.3 and Theorem 2.5 establish the sparse recoverability of the bounded orthonormal systems.

3. Methods

In this section, we introduce the DSRAR framework to construct surrogate model. The goal of this study is to determine, given a small set of $M \ll N$ unstructured realizations $\{\boldsymbol{z}^{(i)}\}_{i=1}^{M}$ and the corresponding outputs $b = (f(\boldsymbol{z}^{(1)}), ...f(\boldsymbol{z}^{(M)}))^T$, the polynomial approximation in (2.4) or (2.5) when $f(\boldsymbol{z})$ has a sparse representation. This small set $\{\boldsymbol{z}^{(i)}\}_{i=1}^{M}$ is usually called *training set* and M is the *training sample size.* There are two quantities we need to compute: (i) an appropriate orthonormal polynomial basis ψ and (ii) an interpolation-type sparse solution $\boldsymbol{c} = (c_1, \ldots, c_N)^T \in \mathbb{R}^N$ such that $f_p(\boldsymbol{z}^{(i)}) = f(\boldsymbol{z}^{(i)})$ for $i = 1, \ldots, M$ with the smallest possible number nonzero $\boldsymbol{c}$. The basis construction, step (i), will be discussed in detail in Section 3.1. We can reformulate the second part as the following constrained optimization problem,

$$\min \|\boldsymbol{c}\|_0 \quad \text{subject to } \boldsymbol{A}\boldsymbol{c} = \boldsymbol{b}, \tag{3.1}$$

where $\|\boldsymbol{c}\|_0$ indicates the number of nonzero entries of $\boldsymbol{c}$ and $\boldsymbol{A} \in \mathbb{R}^{M\times N}$ (usually called the measurement matrix) is written as

$$\boldsymbol{A} = (a_{ij})_{1\le i\le M, 1\le j\le N}, \quad a_{ij} = \psi_j(\boldsymbol{z}^{(i)}). \tag{3.2}$$

It is well known that this ℓ_0 minimization problem (3.1) is NP-hard [84]. As mentioned in Section 2.2, CS is a well-studied and popular approach to find sparse solutions to (3.1) through ℓ_1-minimization shown in (2.6) (no noise) or (2.7) (with noise). Therefore, the approach introduced below can be viewed as a method for data-driven construction of bases that allow sparse representation and accurate recovery for QoIs in UQ applications.

3.1. Data-driven construction of the amdP basis

Let us start with a set of samples of d-dimensional random vector $\boldsymbol{\xi} \in \mathbb{R}^d$, i.e., $S := \{\boldsymbol{\xi}^{(k)}\}_{k=1}^{N_s}$ with the underlying probability measure $\rho(\boldsymbol{\xi})$. S is usually called the *sample set.* We aim to construct a set of orthonormal polynomial basis functions $\{\psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi})\}_{|\boldsymbol{\alpha}|=0}^{p}$ with respect to $\rho(\boldsymbol{\xi})$ in Π_p^d, the space of polynomials up to degree p. Since $\rho(\boldsymbol{\xi})$ can be non-Gaussian or even unknown, *we do not make the assumption that each component of $\boldsymbol{\xi}$ is mutually independent, even under a linear transformation such as those based on principal component analysis (PCA).* Consequently, the orthogonal polynomial basis $\psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi})$ cannot be directly constructed as a tensor product of univariate orthonormal basis functions in each component of $\boldsymbol{\xi}$. Below, we introduce a data-driven approach to construct multivariate amdP randomness.

3.1.1. Orthonormal basis

When we have a collection of random samples S, and the underlying probability measure $\rho(\boldsymbol{\xi})$ can be approximated by the discrete measure $\nu_S(\boldsymbol{\xi})$

$$\rho(\boldsymbol{\xi}) \approx \nu_S(\boldsymbol{\xi}) := \frac{1}{N_s} \sum_{\boldsymbol{\xi}^{(k)}\in S} \delta_{\boldsymbol{\xi}^{(k)}}(\boldsymbol{\xi}), \tag{3.3}$$

where $\delta_{\boldsymbol{\xi}^{(k)}}$ is the Dirac measure, that is $\delta_{\boldsymbol{\xi}^{(k)}}(\boldsymbol{\xi})$ is equal to 1 when $\boldsymbol{\xi} = \boldsymbol{\xi}^{(k)}$ and 0 otherwise. Given the inner product defined as in (2.2) with ρ replaced by the discrete measure ν_S, we can construct a set of orthonormal multivariate polynomial basis functions $\{\psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi})\}_{|\boldsymbol{\alpha}|=0}^{p}$ via the Gram-Schmidt orthogonalization process on an ordered monomial basis $\{\hat{\psi}_{\boldsymbol{\alpha}}(\boldsymbol{\xi})\}_{|\boldsymbol{\alpha}|=0}^{p}$. Here, we use the aforementioned graded lexicographic ordering of the multi-index.

Similar to Dunkl and Xu [76], $\psi_{\boldsymbol{\alpha}}$ can be constructed using the recursive formulation

$$\psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi}) = f_{\boldsymbol{\alpha}}^{\boldsymbol{\alpha}}\hat{\psi}_{\boldsymbol{\alpha}}(\boldsymbol{\xi}) - \sum_{\boldsymbol{\beta}\prec\boldsymbol{\alpha}} f_{\boldsymbol{\beta}}^{\boldsymbol{\alpha}}\psi_{\boldsymbol{\beta}}(\boldsymbol{\xi}), \tag{3.4}$$

where $\hat{\psi}_{\boldsymbol{\alpha}}(\boldsymbol{\xi}) := \prod_{k=1}^{d}\xi_k^{\alpha_k}$ represents the multivariate monomial basis function. The expression $\boldsymbol{\beta}\prec\boldsymbol{\alpha}$ means that the multi-index $\boldsymbol{\beta}$ comes before $\boldsymbol{\alpha}$ under the chosen ordering. The coefficients $f_{\boldsymbol{\beta}}^{\boldsymbol{\alpha}}$ are determined by imposing an orthonormal condition with respect to the discrete measure ν_S, i.e.,

$$\begin{aligned}\int \psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi})\psi_{\boldsymbol{\beta}}(\boldsymbol{\xi})\,\mathrm{d}\rho(\boldsymbol{\xi}) &\approx \int \psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi})\psi_{\boldsymbol{\beta}}(\boldsymbol{\xi})\,\mathrm{d}\nu_S(\boldsymbol{\xi})\\ &= \frac{1}{N_s}\sum_{k=1}^{N_s}\psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi}^{(k)})\psi_{\boldsymbol{\beta}}(\boldsymbol{\xi}^{(k)})\\ &\equiv \delta_{\boldsymbol{\alpha},\boldsymbol{\beta}}, \qquad \boldsymbol{\beta}\preceq\boldsymbol{\alpha}.\end{aligned} \tag{3.5}$$

Equations (3.4) and (3.5) generate a set of orthonormal basis functions on the discrete measure ν_S irrespective of the mutual dependence between the components of $\boldsymbol{\xi}$. We employ $\{\psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi})\}_{|\boldsymbol{\alpha}|=0}^{p}$ as the amdP basis on $\rho(\boldsymbol{\xi})$. Numerically, the modified Gram-Schmidt orthogonalization can be used as an alternative approach when the number of basis is too large and there exists instability in the standard Gram-Schmidt orthogonalization.

When $\rho(\boldsymbol{\xi})$ is known explicitly, orthonormal basis functions can also be constructed by taking the general formulation in Equation (3.4) and imposing the inner product in Equation (2.2) with respect to ρ. Here we will also consider a special case when $\boldsymbol{\xi}$ is a random vector that is linearly transformed from a random vector $\boldsymbol{z}$ with i.i.d. components via $\boldsymbol{\xi} = \mathbf{Q}\boldsymbol{z}$. This case is motivated by the sparsity enhancement approach discussed in Sec. 3.2. In particular, we assume that the quadrature rule of polynomial integration with respect to the probability measure of $\boldsymbol{z}$ is explicitly known.

Given these assumptions, $f_{\boldsymbol{\beta}}^{\boldsymbol{\alpha}}$ can be determined by

$$\begin{aligned}\int \psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi})\psi_{\boldsymbol{\beta}}(\boldsymbol{\xi})\,\mathrm{d}\rho(\boldsymbol{\xi}) &= \sum_{k=1}^{N_Q}\psi_{\boldsymbol{\alpha}}\left(\mathbf{Q}\boldsymbol{z}_Q^{(k)}\right)\psi_{\boldsymbol{\beta}}\left(\mathbf{Q}\boldsymbol{z}_Q^{(k)}\right)w_k\\ &\equiv \delta_{\boldsymbol{\alpha},\boldsymbol{\beta}}, \qquad \boldsymbol{\beta}\preceq\boldsymbol{\alpha}.\end{aligned} \tag{3.6}$$

where $\left\{\boldsymbol{z}_Q^{(k)}\right\}_{k=1}^{N_Q}$ and $\{w_k\}_{k=1}^{N_Q}$ represent the quadrature points and weights constructed to yield an exact integration with probability measure of $\boldsymbol{z}$ for polynomials of degree $|\boldsymbol{\alpha}| + |\boldsymbol{\beta}|$ or less.

Algorithms 1 and 2 summarize the procedure of orthonormal basis construction when $\rho(\boldsymbol{\xi})$ is implicitly represented by a sample set S and known explicitly, respectively. There is no unique system of orthogonal

ALGORITHM 1: Construct the orthonormal amdP basis $\{\psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi})\}_{|\boldsymbol{\alpha}|=0}^{p}$ on discrete sample set S.

1: Given sample set $S = \left\{\boldsymbol{\xi}^{(k)}\right\}_{k=1}^{N_s}$.
2: Given a fixed multi-index order $\left\{\boldsymbol{\alpha}^{(l)}\right\}_{l=1}^{N}$.
3: **for** $l = 1$ to N **do**
4: Let $\boldsymbol{\alpha} = \boldsymbol{\alpha}^{(l)}$, construct $\psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi}) = f_{\boldsymbol{\alpha}}^{\boldsymbol{\alpha}}\hat{\psi}_{\boldsymbol{\alpha}}(\boldsymbol{\xi}) - \sum_{\boldsymbol{\beta}\prec\boldsymbol{\alpha}} f_{\boldsymbol{\beta}}^{\boldsymbol{\alpha}}\psi_{\boldsymbol{\beta}}(\boldsymbol{\xi})$ subject to Equation (3.5).
5: **end for**

ALGORITHM 2: Construct the orthonormal amdP basis $\{\psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi})\}_{|\boldsymbol{\alpha}|=0}^{p}$ with probability measure $\rho(\boldsymbol{\alpha})$.

1: Given a multi-index order $\left\{\boldsymbol{\alpha}^{(l)}\right\}_{l=1}^{N}$.
2: **for** $l = 1$ to N **do**
3: Let $\boldsymbol{\alpha} = \boldsymbol{\alpha}^{(l)}$, construct $\psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi}) = f_{\boldsymbol{\alpha}}^{\boldsymbol{\alpha}}\hat{\psi}_{\boldsymbol{\alpha}}(\boldsymbol{\xi}) - \sum_{\boldsymbol{\beta}\prec\boldsymbol{\alpha}} f_{\boldsymbol{\beta}}^{\boldsymbol{\alpha}}\psi_{\boldsymbol{\beta}}(\boldsymbol{\xi})$ by evaluating the basis inner product using existing quadrature rule or Equation (3.6) if $\boldsymbol{\xi}$ can be linearly transformed from a random vector with i.i.d. components $\boldsymbol{z}$ with an explicitly known quadrature rule.
4: **end for**

polynomial basis functions for both scenarios if $d > 1$; different orderings of $\boldsymbol{\alpha}$ lead to different orthogonal basis [76]. On the other hand, the constructed orthonormal basis is unique up to unitary transformations as we prove in Theorem 3.1.

Theorem 3.1. *Let $\{\psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi})\}_{|\boldsymbol{\alpha}|=0}^{p}$ be a set of orthonormal polynomial basis with respect to the measure $\rho(\boldsymbol{\xi})$, $\boldsymbol{\xi} \in \mathbb{R}^d$. Denote by $\boldsymbol{\Psi}(\boldsymbol{\xi})$ the polynomial basis vector*

$$\boldsymbol{\Psi}(\boldsymbol{\xi}) := (\psi_{\boldsymbol{\alpha}^{(1)}}, \cdots, \psi_{\boldsymbol{\alpha}^{(N)}})^T, \tag{3.7}$$

where $\boldsymbol{\alpha}^{(1)}, \cdots, \boldsymbol{\alpha}^{(N)}$ is the arrangement of multi-index $\boldsymbol{\alpha}$ according to a fixed multi-index order. Let $\boldsymbol{\chi} = \mathbf{Q}\boldsymbol{\xi}$, where $\mathbf{Q} \in \mathbb{R}^{d\times d}$ is invertible. Let $\{\phi_{\boldsymbol{\beta}}(\boldsymbol{\chi})\}_{|\boldsymbol{\beta}|=0}^{p}$ be a set of orthonormal polynomial basis functions with respect to a measure $\rho'(\boldsymbol{\chi})$ constructed with order $\boldsymbol{\beta}^{(1)}, \cdots, \boldsymbol{\beta}^{(N)}$, where $\rho'(\boldsymbol{\chi})$ is induced from $\rho(\boldsymbol{\xi})$. Then there exists a unitary matrix $\mathbf{U}$ such that $\boldsymbol{\Phi}(\boldsymbol{\chi}) = \mathbf{U}\boldsymbol{\Psi}(\boldsymbol{\xi})$, where $\boldsymbol{\Phi}(\boldsymbol{\chi}) := (\phi_{\boldsymbol{\beta}^{(1)}}, \cdots, \phi_{\boldsymbol{\beta}^{(N)}})^T$ denotes the corresponding polynomial basis vector.

Proof. Let $\hat{\boldsymbol{\Psi}}(\boldsymbol{\xi})$ be the monomial basis vector. Note that $\{\psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi})\}_{|\boldsymbol{\alpha}|=0}^{p}$ and $\{\phi_{\boldsymbol{\beta}}(\boldsymbol{\xi})\}_{|\boldsymbol{\beta}|=0}^{p}$ are two sets of basis in Π_p^d. There exists transfer matrices $\mathcal{M}_\psi$ and $\mathcal{M}_\phi \in \mathbb{R}^{N\times N}$ such that

$$\boldsymbol{\Psi}(\boldsymbol{\xi}) = \mathcal{M}_\psi\hat{\boldsymbol{\Psi}}(\boldsymbol{\xi}), \quad \boldsymbol{\Phi}(\boldsymbol{\xi}) = \mathcal{M}_\phi\hat{\boldsymbol{\Psi}}(\boldsymbol{\xi}).$$

With $\boldsymbol{\chi} = \mathbf{Q}\boldsymbol{\xi}$, $\boldsymbol{\Phi}(\mathbf{Q}\boldsymbol{\xi})$ is also a basis in Π_p^d. Then there exists an invertible matrix $\mathcal{T} \in \mathbb{R}^{N\times N}$ such that

$$\boldsymbol{\Phi}(\boldsymbol{\chi}) = \boldsymbol{\Phi}(\mathbf{Q}\boldsymbol{\xi}) = \mathcal{T}\hat{\boldsymbol{\Psi}}(\boldsymbol{\xi}),$$

which gives $\boldsymbol{\Phi}(\boldsymbol{\chi}) = \mathcal{U}\boldsymbol{\Psi}(\boldsymbol{\xi})$, where $\mathcal{U} = \mathcal{T}\mathcal{M}_{\psi}^{-1}$. Recall $\{\psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi})\}_{|\boldsymbol{\alpha}|=0}^{p}$ and $\{\phi_{\boldsymbol{\beta}}(\boldsymbol{\chi})\}_{|\boldsymbol{\beta}|=0}^{p}$ are orthonormal basis with respect to $\rho(\boldsymbol{\xi})$ and $\rho'(\boldsymbol{\chi})$, we have

$$\mathbf{I} = \int \boldsymbol{\Phi}(\boldsymbol{\chi})\boldsymbol{\Phi}(\boldsymbol{\chi})^T \,\mathrm{d}\rho'(\boldsymbol{\chi}) = \int \mathcal{U}\boldsymbol{\Psi}(\boldsymbol{\xi})\boldsymbol{\Psi}(\boldsymbol{\xi})^T\mathcal{U}^T \,\mathrm{d}\rho(\boldsymbol{\xi}) = \mathcal{U}\mathcal{U}^T \tag{3.8}$$

□

We do not need further assumptions on $\rho(\boldsymbol{\xi})$ because Theorem 3.1 holds both when $\rho(\boldsymbol{\xi})$ is a measure on the continuous random vector $\boldsymbol{\xi}$ (with probability density function $\omega(\boldsymbol{\xi})$) or a discrete measure $\nu_S(\boldsymbol{\xi})$ on a sample set S. Furthermore, it is straightforward to show the following Corollary.

Corollary 1. *Let* $S_1 := \left\{\boldsymbol{\xi}^{(k)}\right\}_{k=1}^{M}$ *and* $S_2 := \left\{\boldsymbol{\chi}^{(k)}\right\}_{k=1}^{M}$ *be two sets of random sampling points where* $\boldsymbol{\chi}^{(k)} = \mathbf{Q}\boldsymbol{\xi}^{(k)}$ *with invertible* $\mathbf{Q}$. *Let* $\mathbf{G}_{\boldsymbol{\xi}}$ *and* $\mathbf{G}_{\boldsymbol{\chi}}$ *be the Gram matrix constructed by* $\boldsymbol{\Psi}(\boldsymbol{\xi})$ *and* $\boldsymbol{\Phi}(\boldsymbol{\chi})$ *defined in Theorem 3.1, i.e.,* $\mathbf{G}_{\boldsymbol{\xi}} := \sum_{k=1}^{M} \boldsymbol{\Psi}(\boldsymbol{\xi}^{(k)})\boldsymbol{\Psi}(\boldsymbol{\xi}^{(k)})^T/M$ *and* $\mathbf{G}_{\boldsymbol{\chi}} := \sum_{k=1}^{M} \boldsymbol{\Phi}(\boldsymbol{\chi}^{(k)})\boldsymbol{\Phi}(\boldsymbol{\chi}^{(k)})^T/M$. *Then* $\mathbf{G}$. *has invariant* l_2 *norm, that is,* $\|\mathbf{G}_{\boldsymbol{\xi}}\|_2 = \|\mathbf{G}_{\boldsymbol{\chi}}\|_2$. *Moreover,* $\|\mathbf{G}_{\boldsymbol{\xi}} - \mathbf{I}\|_2 = \|\mathbf{G}_{\boldsymbol{\chi}} - \mathbf{I}\|_2$.

In general, the l_2 norm of $\|\mathbf{G}_{\boldsymbol{\xi}} - \mathbf{I}\|_2$ is independent of specific monomial order of $\boldsymbol{\alpha}$ and invariant under linear transformations of the random vector. The basis functions $\{\psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi})\}_{|\boldsymbol{\alpha}|=0}^{p}$ constructed by Equations (3.4) and (3.5) provide an appropriate candidate for representing the surrogate model $f(\boldsymbol{\xi})$ via CS.

3.1.2. Near-orthonormal basis

When $\rho(\boldsymbol{\xi})$ is implicitly represented by a sample set S, we employ the discrete measure ν_S to construct $\{\psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi})\}_{|\boldsymbol{\beta}|=0}^{p}$. However, we note that the training set that queries $f(\cdot)$, denoted by S_f, may not be a subset of S. In practice, the sample set S and the training set S_f are usually collected in a sequential manner or directly from different experiments, although individual sampling points of both S and S_f follow the same distribution. Since S only contains a finite number of samples of $\boldsymbol{\xi}$, basis $\{\psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi})\}_{|\boldsymbol{\alpha}|=0}^{p}$ constructed by (3.4) and (3.5) is not the "exact orthonormal" basis with respect to $\rho(\boldsymbol{\xi})$. Especially, let $S' = \{\boldsymbol{\xi'}_k\}_{k=1}^{N_s}$ be another sample set following the same distribution $\rho(\cdot)$ and $\nu_{S'}(\cdot)$ be the discrete measure defined on S'. For the orthonormal amdP basis functions $\{\psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi})\}_{|\boldsymbol{\alpha}|=0}^{p}$ constructed on S, we have $\mathbb{E}\left[\boldsymbol{\Psi}(\boldsymbol{\xi})\boldsymbol{\Psi}(\boldsymbol{\xi})^T\right] \neq \mathbf{I}$ under the discrete measure $\nu_{S'}(\boldsymbol{\xi})$ and vice versa.

The above observation forces us to re-examine the orthonormal condition imposed by (3.5). Since the pre-constructed basis $\psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi})$ does not retain the exact orthonormal condition when later being applied to approximate $f(\boldsymbol{\xi})$, we may relax the condition when determining the coefficients $f_{\boldsymbol{\beta}}^{\boldsymbol{\alpha}}$ in (3.4). In the present study, we propose the following heuristic criterion

$$\arg\min_{\hat{\mathbf{f}}^{\boldsymbol{\alpha}}} \|\hat{\mathbf{f}}^{\boldsymbol{\alpha}}\|_2 \;\text{ subject to }\; \left|\int \psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi})\psi_{\boldsymbol{\beta}}(\boldsymbol{\xi}) \,\mathrm{d}\nu_S(\boldsymbol{\xi}) - \delta_{\boldsymbol{\alpha},\boldsymbol{\beta}}\right| < \zeta_{\boldsymbol{\alpha},\boldsymbol{\beta}}, \quad \boldsymbol{\beta} \leq \boldsymbol{\alpha}, \tag{3.9}$$

where $\hat{\mathbf{f}}^{\boldsymbol{\alpha}}$ is the coefficient vector of $\psi_{\boldsymbol{\alpha}}$ when represented using monomial basis functions, i.e., $\psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi}) = \sum_{\boldsymbol{\beta}\preceq\boldsymbol{\alpha}} \hat{f}^{\boldsymbol{\alpha}}_{\boldsymbol{\beta}}\hat{\psi}_{\boldsymbol{\beta}}(\boldsymbol{\xi})$. $\hat{\mathbf{f}}^{\boldsymbol{\alpha}}$ is related to $\mathbf{f}^{\boldsymbol{\alpha}}$ through the linear transformation

$$\hat{\mathbf{f}}^{\boldsymbol{\alpha}} = \begin{pmatrix} \mathbf{F} & 0 \\ 0 & 1 \end{pmatrix} \mathbf{f}^{\boldsymbol{\alpha}}, \tag{3.10}$$

where $\mathbf{F}$ is an upper triangle matrix determined by pre-computed $\hat{\mathbf{f}}^{\boldsymbol{\beta}}, \boldsymbol{\beta} \prec \boldsymbol{\alpha}$, i.e.,

$$[\mathbf{F}]_{I_{\boldsymbol{\beta}'} I_{\boldsymbol{\beta}}} = \begin{cases} \hat{f}^{\boldsymbol{\beta}}_{\boldsymbol{\beta}'} & \boldsymbol{\beta}' \preceq \boldsymbol{\beta} \\ 0 & \boldsymbol{\beta}' \succ \boldsymbol{\beta}, \end{cases} \tag{3.11}$$

where $I_{\boldsymbol{\beta}}$ represents the mapping from multi-index to single index.

The parameter $\zeta_{\boldsymbol{\alpha},\boldsymbol{\beta}}$ quantifies the relaxation of the orthonormal condition. We split the sample set S equally into two parts $S := S_1 \cup S_2$. Denote $\left\{\psi^{(1)}_{\boldsymbol{\alpha}}(\boldsymbol{\xi})\right\}^{p}_{|\boldsymbol{\alpha}|=0}$ and $\left\{\psi^{(2)}_{\boldsymbol{\alpha}}(\boldsymbol{\xi})\right\}^{p}_{|\boldsymbol{\alpha}|=0}$ the orthonormal bases constructed by Equations (3.4) and (3.5) on the discrete measures $\nu_{S_1}(\boldsymbol{\xi})$ and $\nu_{S_2}(\boldsymbol{\xi})$, respectively. Inspired by cross-validation, we have chosen $\zeta_{\boldsymbol{\alpha},\boldsymbol{\beta}} = \dfrac{|\zeta_1| + |\zeta_2|}{2\sqrt{2}}$

$$\zeta_1 = \int \psi^{(1)}_{\boldsymbol{\alpha}}(\boldsymbol{\xi})\psi^{(1)}_{\boldsymbol{\beta}}(\boldsymbol{\xi})\,\mathrm{d}\nu_{S_2}(\boldsymbol{\xi}), \quad \zeta_2 = \int \psi^{(2)}_{\boldsymbol{\alpha}}(\boldsymbol{\xi})\psi^{(2)}_{\boldsymbol{\beta}}(\boldsymbol{\xi})\,\mathrm{d}\nu_{S_1}(\boldsymbol{\xi}). \tag{3.12}$$

ALGORITHM 3: Construct the near-orthonormal amdP basis $\{\psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi})\}^{p}_{|\boldsymbol{\alpha}|=0}$ on discrete sample set S.

1: Collect samples of $\boldsymbol{\xi}$ from sample set $S = \left\{\boldsymbol{\xi}^{(k)}\right\}_{k=1}^{N_s}$, split S equally into two disjoint subsets, i.e., $S = S_1 \cup S_2$, $S_1 \cap S_2 = \emptyset$.

2: Given fixed monomial index order $\left\{\boldsymbol{\alpha}^{(l)}\right\}_{l=1}^{N}$, construct the orthonormal amdP basis $\left\{\psi^{(1)}_{\boldsymbol{\alpha}}(\boldsymbol{\xi})\right\}^{p}_{|\boldsymbol{\alpha}|=0}$ and $\left\{\psi^{(2)}_{\boldsymbol{\alpha}}(\boldsymbol{\xi})\right\}^{p}_{|\boldsymbol{\alpha}|=0}$ on set S_1 and S_2 by **Algorithm 1**.

3: **for** $l = 1$ to N **do**

4: Let $\boldsymbol{\alpha} = \boldsymbol{\alpha}^{(l)}$, construct $\psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi}) = f^{\boldsymbol{\alpha}}_{\boldsymbol{\alpha}}\hat{\psi}_{\boldsymbol{\alpha}}(\boldsymbol{\xi}) - \sum_{\boldsymbol{\beta}\prec\boldsymbol{\alpha}} f^{\boldsymbol{\alpha}}_{\boldsymbol{\beta}}\psi_{\boldsymbol{\beta}}(\boldsymbol{\xi})$ on by Equations (3.9), (3.10), and (3.12).

5: **end for**

Algorithm 3 describes construction for a set of near-orthonormal amdP basis functions on the sample set S. When applied to the sample set S' to approximate $f(\boldsymbol{\xi})$, the basis shows comparable orthonormal conditions with the basis constructed by (3.5). Such results can be partially understood by the theoretical bound from Theorem 2.5 on the number of samples M for exact recovery in orthonormal polynomial systems, $M \geq C_1 K^2 s \log^3(s)\log(N)$, where $s = \|\boldsymbol{c}\|_0$ and $K = \sup_{\boldsymbol{\alpha}} \|\psi_{\boldsymbol{\alpha}}\|_\infty$. Theoretical analysis of the recovery error under different basis functions is out of the scope of the present work and is left for future investigation. However, we note that the accuracy of the surrogate model $f(\boldsymbol{\xi})$ can be further improved by enhancing

the sparsity of $\boldsymbol{c}$. This can be achieved through the ideas presented in our previous work [1] which will be extended to general distributions below.

Remark 3.2. We emphasize that (3.9) provides a heuristic approach to construct the near-orthonormal amdP basis functions $\psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi})$ with a smaller basis bound. In practice, (3.9) can be further relaxed to

$$\begin{aligned} \arg\min_{\hat{\mathbf{f}}^{\boldsymbol{\alpha}}} \|\hat{\mathbf{f}}^{\boldsymbol{\alpha}}\|_2 \ \text{ subject to } & \sum_{|\boldsymbol{\beta}|=r,\boldsymbol{\beta}<\boldsymbol{\alpha}} \left| \int \psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi})\psi_{\boldsymbol{\beta}}(\boldsymbol{\xi})\,\mathrm{d}\nu_S(\boldsymbol{\xi}) \right|^2 < \sum_{|\boldsymbol{\beta}|=r,\boldsymbol{\beta}<\boldsymbol{\alpha}} \zeta_{\boldsymbol{\alpha},\boldsymbol{\beta}}^2, \\ & \left| \int \psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi})\psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi})\,\mathrm{d}\nu_S(\boldsymbol{\xi}) - 1 \right| < \zeta_{\boldsymbol{\alpha},\boldsymbol{\alpha}}, \quad r = 0, \cdots, |\boldsymbol{\alpha}|, \end{aligned} \tag{3.13}$$

which shows similar numerical performance. There is no theoretical guarantee yet that Equations (3.9) and (3.13) yield a smaller basis bound than (3.5) on S_f, S or the entire domain of $\boldsymbol{\xi}$. We numerically compare some properties of different bases in Section 4.1, which illustrate the performance of the near-orthonormal amdP basis constructed above. There may exist other numerical approaches to optimize $\psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi})$ that can lead to an even smaller basis bound. We also note that the threshold $\zeta_{\boldsymbol{\alpha},\boldsymbol{\beta}}$ is determined by directly splitting S into two disjoint sets. In practice, it is possible to design more sophisticated strategies to optimize the choice of $\zeta_{\boldsymbol{\alpha},\boldsymbol{\beta}}$ and the basis construction procedure. We leave such studies for future work.

3.2. Sparsity enhancement

For the linear system in (2.6), the numerical accuracy of the recovered $\tilde{c}$ via l_1-minimization depends on the sparsity of $\boldsymbol{c}$. This dependence motivates us to develop a numerical approach to further enhance the sparsity of $\boldsymbol{c}$ through the variability analysis of $f(\boldsymbol{\xi})$ [1]. If we know $f(\boldsymbol{\xi})$ explicitly, the (sorted) directions of variance in $f(\boldsymbol{\xi})$ under the distribution of $\boldsymbol{\xi}$ can be found based on the active subspace method [85, 66]. In particular, we define the gradient matrix $\mathbf{G}$ by

$$\mathbf{G} = \mathbb{E}\left[\nabla f(\boldsymbol{\xi}) \nabla f(\boldsymbol{\xi})^T\right] \tag{3.14}$$

where $\nabla f(\boldsymbol{\xi})$ is the gradient vector defined by $\nabla f(\boldsymbol{\xi}) = \left(\frac{\partial f}{\partial \xi_1}, \frac{\partial f}{\partial \xi_2}, \cdots \frac{\partial f}{\partial \xi_d}\right)^T$. Eigendecomposition of $\mathbf{G}$,

$$\mathbf{G} = \mathbf{Q}\mathbf{K}\mathbf{Q}^T, \qquad \mathbf{Q} = [\mathbf{q}_1\ \mathbf{q}_2 \cdots\ \mathbf{q}_d], \tag{3.15a}$$

$$\mathbf{K} = \mathrm{diag}(k_1, \cdots, k_d), \quad k_1 \geq \cdots \geq k_d \geq 0, \tag{3.15b}$$

yields the sorted variability directions $\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_d$. Accordingly, we may define a new random vector $\boldsymbol{\chi}$ following the sorted variability directions via linear transformation

$$\boldsymbol{\chi} = \mathbf{Q}^T\boldsymbol{\xi}. \tag{3.16}$$

$f(\boldsymbol{\xi}) = f((\mathbf{Q}^T)^{-1}\boldsymbol{\chi}) = f(\mathbf{Q}\boldsymbol{\chi})$ can be approximated by expansion in an orthonormal polynomial basis $\boldsymbol{\chi}$ with a coefficient vector $\boldsymbol{c}$ which is sparser than the $f(\boldsymbol{\xi})$ being expanded by orthonormal basis of $\boldsymbol{\xi}$. For the

remainder of this paper, we use $\mathbf{Q}$ to denote the rotation matrix to transform $\boldsymbol{\xi}$ to $\boldsymbol{\chi}$ and $g(\boldsymbol{\chi})$ to represent $f(\mathbf{Q}\boldsymbol{\chi})$.

In practice, $f(\boldsymbol{\xi})$ is usually not explicitly known. We may numerically approximate $\mathbf{G}$ by

$$\mathbf{G} \approx \mathbb{E}\left[\nabla \tilde{f}(\boldsymbol{\xi}) \nabla \tilde{f}(\boldsymbol{\xi})^T\right], \tag{3.17}$$

where $\tilde{f}(\boldsymbol{\xi})$ represents the approximation of $f(\boldsymbol{\xi})$ by the orthonormal polynomial basis functions $\psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi})$ as proposed in [1] or obtained via solving (2.6) with the data-driven basis approach (i.e., basis functions constructed with respect to an arbitrary measure) described in Section 3.1. In particular, if $\boldsymbol{\xi}$ is a random vector with i.i.d. Gaussian components, $\boldsymbol{\chi}$ is also a random vector with i.i.d. Gaussian components. Thus, $\tilde{f}(\boldsymbol{\xi})$ and $\tilde{g}(\boldsymbol{\chi}) := \tilde{f}(\mathbf{Q}\boldsymbol{\chi})$ can be represented by the orthonormal basis functions of the same form, e.g., tensor products of univariate Hermite polynomials. Without of lost of generality, from now on, we use $\tilde{g}(\boldsymbol{\chi})$ to represent $\tilde{f}(\mathbf{Q}\boldsymbol{\chi})$.

However, if $\rho(\boldsymbol{\xi})$ is not i.i.d. Gaussian, $\boldsymbol{\chi}$ and $\boldsymbol{\xi}$ do not generally have the same distribution. Therefore, an orthonormal polynomial basis $\psi(\cdot)$ with respect to $\boldsymbol{\xi}$ cannot *be directly applied to $\boldsymbol{\chi}$.* The general approach presented in Section 3.1 enables us to construct the amdP basis with respect to the probability measure of the rotated vector $\boldsymbol{\chi}$. The two orthonormal bases associated with $\boldsymbol{\xi}$ and $\boldsymbol{\chi}$ respectively are related to each other via a unitary transformation as shown in Theorem 3.1. In particular, if $\rho(\boldsymbol{\xi})$ is implicitly described by a sample set $S = \left\{\boldsymbol{\xi}^{(k)}\right\}_{k=1}^{N_s}$, $\mathbf{G}$ can be easily evaluated by representing $\psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi})$ via the monomial basis, i.e., $\psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi}) = \sum_{\boldsymbol{\beta} \preceq \boldsymbol{\alpha}} \hat{f}_{\boldsymbol{\beta}}^{\boldsymbol{\alpha}} \hat{\psi}_{\boldsymbol{\beta}}(\boldsymbol{\xi})$ via Equation (3.10) and then integrating with discrete measure ν_S. By transforming S and S_f into $\left\{\boldsymbol{\chi}^{(k)}\right\}_{k=1}^{N_s}$ and $\left\{\boldsymbol{\chi}'^{(k)}\right\}_{k=1}^{M}$, the orthonormal and near-orthonormal amdP basis functions with respect to $\boldsymbol{\chi}$ can be constructed by Eqs. (3.5) (3.13). The surrogate model $\tilde{g}(\boldsymbol{\chi})$ can then be constructed by solving (2.6).

The entire DSRAR procedure is presented in **Algorithm 4**. Compared with $\tilde{f}(\boldsymbol{\xi})$, $\tilde{g}(\boldsymbol{\chi})$ shows smaller numerical error in general. The additional cost of sparsity enhancement procedure in Step 4 - 6 is less than 0.6 CPU (3.7 GHz Quad-Core Intel Xeon E5) hour for the numerical examples considered in this study. For realistic applications, the overhead of Step 4 - 6 could be relatively small if sampling of QoI is expensive or the available training set is limited.

The DSRAR framework described above is also applicable to systems with standard density distributions, where $\rho(\boldsymbol{\xi})$ is known explicitly. Without loss of generality, we assume that an orthonormal polynomial basis $\{\psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi})\}_{|\boldsymbol{\alpha}|=0}^{p}$ is known. Evaluation of $\mathbf{G}$ by (3.17) on $\rho(\boldsymbol{\xi})$ is straightforward. The surrogate model of f can be constructed via l_1 minimization with enhanced sparsity through Algorithm 5.

The procedures for random vector rotation and surrogate construction presented in Algorithms 4 and 5 can be conducted in an iterative manner. We have investigated this issue [86] by applying a previously developed rotation procedure [1] successively to systems with underlying Gaussian distributions. For the

ALGORITHM 4: DSRAR: Surrogate model construction with discrete sample set S and training set S_f.

1: Collect the sample set within the random space $S = \{\boldsymbol{\xi}^{(k)}\}_{k=1}^{N_s}$.

2: Generate evaluations of f on training set $S_f = \left\{\boldsymbol{\xi}'^{(k)}\right\}_{k=1}^{M}$ with M outputs $f_1, f_2, \cdots, f_M$.

3: Construct the data-driven amdP basis $\{\psi_i(\boldsymbol{\xi})\}_{i=1}^{N}$ on discrete measure $\nu_S(\boldsymbol{\xi})$ as the exact orthonormal basis by Algorithm 1 or the near orthonormal basis by Algorithm 3.

4: Evaluate the measurement matrix $\boldsymbol{A}_{ij} = \psi_j(\boldsymbol{\xi}'^{(i)})$, $1 \le i \le M$, $1 \le j \le N$; construct surrogate model $\tilde{f}(\boldsymbol{\xi}) = \sum_{|\alpha|=0}^{p} c_{\boldsymbol{\alpha}} \psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi})$ by solving the l_1-minimization problem.

5: Evaluate the gradient matrix $\mathbf{G} \approx \mathbb{E}\left[\nabla \tilde{f}(\boldsymbol{\xi}) \nabla \tilde{f}(\boldsymbol{\xi})^T\right]$ on $\nu_S(\boldsymbol{\xi})$. Find the eigendecomposition $\mathbf{G} = \mathbf{Q}\mathbf{K}\mathbf{Q}^T$, define sample set $\{\boldsymbol{\chi}^{(k)}\}_{k=1}^{N_s}$ and training set $\left\{\boldsymbol{\chi}'^{(k)}\right\}_{k=1}^{M}$ by $\boldsymbol{\chi}^{(k)} = \mathbf{Q}^T \boldsymbol{\xi}^{(k)}$, $\boldsymbol{\chi}'^{(k)} = \mathbf{Q}^T \boldsymbol{\xi}'^{(k)}$.

6: Reconstruct the data-driven amdP basis $\{\phi_{\boldsymbol{\alpha}}(\boldsymbol{\chi})\}_{|\boldsymbol{\alpha}|=0}^{p}$ by Algorithm 3 and surrogate model $\tilde{g}(\boldsymbol{\chi})$ with enhanced sparsity following Step 3 and Step 4.

ALGORITHM 5: DSRAR: Surrogate model construction with training set S_f and probability measure $\rho(\boldsymbol{\xi})$.

1: Evaluate f on training set $S_f = \left\{\boldsymbol{\xi}'^{(k)}\right\}_{k=1}^{M}$ with M outputs $f_1, f_2, \cdots, f_M$.

2: Evaluate the measurement matrix $\boldsymbol{A}_{ij} = \psi_j(\boldsymbol{\xi}'^{(i)})$, $1 \le i \le M$, $1 \le j \le N$; construct surrogate model $\tilde{f}(\boldsymbol{\xi}) = \sum_{|\alpha|=0}^{p} c_{\boldsymbol{\alpha}} \psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi})$ by solving l_1 minimization problem.

3: Evaluate the gradient matrix $\mathbf{G} = \mathbb{E}\left[\nabla \tilde{f}(\boldsymbol{\xi}) \nabla \tilde{f}(\boldsymbol{\xi})^T\right]$ on $\rho(\boldsymbol{\xi})$. Conduct eigendecomposition $\mathbf{G} = \mathbf{Q}\mathbf{K}\mathbf{Q}^T$ and define training set $\{\boldsymbol{\chi}'^{(k)}\}_{k=1}^{M}$, $\boldsymbol{\chi}'^{(k)} = \mathbf{Q}^T \boldsymbol{\xi}'^{(k)}$.

4: Re-construct the orthonormal amdP basis $\{\phi_{\boldsymbol{\alpha}}(\boldsymbol{\chi})\}_{|\boldsymbol{\alpha}|=0}^{p}$ with respect to $\rho'(\boldsymbol{\chi})$ by **Algorithm 2**. Construct the surrogate model $\tilde{g}(\boldsymbol{\chi})$ with enhanced sparsity following Step 3.

systems studied in the present work, the improvement of the numerical accuracy is marginal after the first rotation procedure. Therefore, the numerical results with only one rotation procedure will be presented in this manuscript.

4. Results

This section presents the numerical results of the present DSRAR framework for surrogate model construction with arbitrary underlying distributions. For numerical examples where the probability measure $\rho(\boldsymbol{\xi})$ (with density function $\omega(\boldsymbol{\xi})$) is not known explicitly and is represented by a discrete data set $S = \{\boldsymbol{\xi}^{(k)}\}_{k=1}^{N_s}$, we split S equally into two subsets $S = S_1 \cup S_2$. We use S_1 to construct the data-driven amdP basis and split S_2 into two disjoint subset $S_2 = S_{2,1} \cup S_{2,2}$, where $S_{2,1}$ is the training set for surrogate model construction and $S_{2,2}$ is the test set to evaluate the accuracy of the constructed surrogate model. The size of the training set is $O(10^2) - O(10^3)$ and size of the test set is $O(10^5)$.

4.1. Accurate recovery of linear systems with data-driven bases

In this test, we collected a sample set $S = \{\boldsymbol{\xi}^{(k)}\}_{k=1}^{N_s}$ with $N_s = 2 \times 10^5$. The random vector $\boldsymbol{\xi}$ followed the Gaussian mixture distribution

$$\omega(\boldsymbol{\xi}) = \sum_{i=1}^{N_m} a_i \mathcal{N}(\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}) \tag{4.1}$$

where N_m is the number of Gaussian modes. We set $N_m = 3$, $a_i > 0$ for $i = 1, 2, 3$ and $\sum_{i=1}^{3} a_i = 1$. For each Gaussian mode, $\boldsymbol{\mu}_i$ is a 25-dimensional i.i.d. random vector with uniform distribution $\mathcal{U}[-2.5, 2.5]$ on each dimension and then shifted such that $\sum_{i=1}^{3} a_i \boldsymbol{\mu}_i = 0$. The matrices $\boldsymbol{\Sigma}_i$ were chosen such that

$$\boldsymbol{\Sigma}_i = (\boldsymbol{\Upsilon}_i {\boldsymbol{\Upsilon}_1}^T + \mathbf{I})/4, \tag{4.2}$$

where $\boldsymbol{\Upsilon}_i$ is a random matrix with i.i.d. entries from $\mathcal{U}[0, 1]$ for $i = 1, 2, 3$.

We considered a linear system

$$\boldsymbol{A}\boldsymbol{c} = \boldsymbol{b} + \boldsymbol{\varepsilon}$$

and recovered $\boldsymbol{c}$ using M training points by solving the l_1 minimization problem defined by (2.6) where

$$[\boldsymbol{A}]_{i,j} = \psi_j(\boldsymbol{\xi}^{(i)}), \quad b_i = \sum_{k=1}^{N} c_k \psi_k(\boldsymbol{\xi}^{(i)}), \tag{4.3}$$

with $1 \le i \le M$, $1 \le j \le N$, and $\boldsymbol{\varepsilon}$ is noise with $\|\boldsymbol{\varepsilon}\|_2 \le 10^{-7}$. We set $d = 25$, $p = 2$ and $N = \begin{pmatrix} d+p \\ p \end{pmatrix} = 351$. The basis functions $\psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi})$ were constructed on the set S_1 by the following approaches:

1. the orthonormal amdP basis subject to Equations (3.4) and (3.5);
2. the near-orthonormal amdP basis subject to Equation (3.9);

3. tensor product of univariate normalized Legendre polynomials (both sampling points and training points are scaled to lie in $[-1, 1]$ on each dimension accordingly).

Training points from set S_2 were used to examine the recovery accuracy of $\boldsymbol{c}$.

4.1.1. Sparse linear systems

First, we considered the scenario where $\boldsymbol{c}$ is a s-sparse vector and employed the following theoretical bound to examine the recovery accuracy via l_1-minimization.

Theorem 4.1. *Given a matrix $\boldsymbol{\Psi} \in \mathbb{R}^{M\times N}$ and set $T_{\boldsymbol{\alpha}}$ with $s = |T_{\boldsymbol{\alpha}}|$, a s-sparse vector $\boldsymbol{c}$ with non-zero entries on $T_{\boldsymbol{\alpha}}$ can be exactly recovered via l_1-minimization if $\frac{\theta_s}{1-\delta_s} < 0.5$, where δ_s and θ_s are defined by*

$$\begin{aligned} \delta_s &:= \inf\left[\delta : (1-\delta)\|\boldsymbol{y}\|_2^2 \le \|\boldsymbol{\Psi}_t \boldsymbol{y}\|_2^2 \le (1+\delta)\|\boldsymbol{y}\|_2^2\right],\ \forall t \subseteq \mathbf{T}, \forall \boldsymbol{y} \in \mathbb{R}^{|t|} \\ \theta_s &:= \inf\left[\theta : |\langle \boldsymbol{\Psi}_{t'}\boldsymbol{y}', \boldsymbol{\Psi}_t \boldsymbol{y}\rangle| \le \theta \|\boldsymbol{y}'\|_2 \|\boldsymbol{y}\|_2\right],\ \forall t \subseteq \mathbf{T}, t' \nsubseteq \mathbf{T}, |t'| \le s, \forall \boldsymbol{y} \in \mathbb{R}^{|t|}, \boldsymbol{y}' \in \mathbb{R}^{|t'|} \end{aligned} \tag{4.4}$$

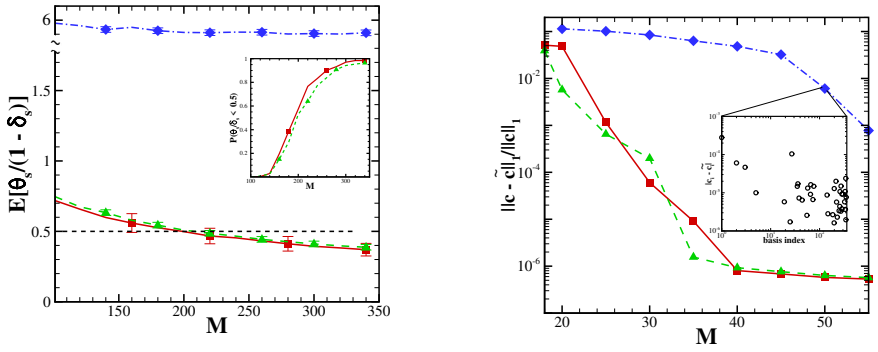
where $\boldsymbol{\Psi}_t$ and $\boldsymbol{\Psi}_{t'}$ denote the sub-matrices of $\boldsymbol{\Psi}$ with column indices in t and t' respectively.

Theorem 4.1 (see Appendix A for proof) provides a sufficient condition to exactly recover $\boldsymbol{c}$ with non-zero entries on index set $T_{\boldsymbol{\alpha}}$. For numerical study, we randomly chose an index set $T_{\boldsymbol{\alpha}}$ from Λ_p^d with $|T_{\boldsymbol{\alpha}}| = 3$, where Λ_p^d is defined by (2.4). For each training set, we constructed the measurement matrix $\boldsymbol{A}$ with different bases and computed $\theta_s/(1-\delta_s)$ by (4.4). Figure 1(a) shows the mean value $\mathbb{E}[\theta_s/(1-\delta_s)]$ on 200 independent training sets chosen from S_2 for each M. The exact and near-orthonormal bases yield similar results: $\mathbb{E}[\theta_s/(1-\delta_s)]$ becomes smaller than 0.5 as M approaches 210, which is also shown in the inset plot of Figure 1(a). In contrast, $\mathbb{E}[\theta_s/(1-\delta_s)]$ obtained from Legendre polynomial basis shows worse performance due to the loss of orthonormality.

In our numerical experiments, we were able to recover $\boldsymbol{c}$ using fewer samples than the number M—as suggested by the sufficient condition (Theorem 2.5) originally given by Rauhut [52]—since this number is based on the worst case scenario and is not, in general, a sharp bound. Figure 1(b) shows the numerical results of a test case with $\boldsymbol{c}_{T_{\boldsymbol{\alpha}}} = 1$, $\boldsymbol{c}_{T_{\boldsymbol{\alpha}}^c} = 0$, $|T_{\boldsymbol{\alpha}}| = 5$. For each M, 200 CS implementations were conducted to compute the average of the relative error $\|\boldsymbol{c} - \tilde{\boldsymbol{c}}\|_1/\|\boldsymbol{c}\|_1$. The exact and near-orthonormal amdP bases show similar performance, where $\boldsymbol{c}$ can be accurately recovered (up to $\|\boldsymbol{\varepsilon}\|_2$) using $M = 45$ training points. In contrast, the Legendre basis yields larger relative error in ℓ_1-norm. The relative error of the recovered coefficients from one CS implementation with Legendre basis is shown in the inset plot of Figure 1(b).

4.1.2. Non-sparse linear systems

We also tested the recovery performance when the exact representation is not sparse. The vector $\boldsymbol{c}$ is chosen with a random non-zero index set $T_{\boldsymbol{\alpha}}$ with $|T_{\boldsymbol{\alpha}}| = 120$. Individual components of $\boldsymbol{c}_{T_{\boldsymbol{\alpha}}}$ are i.i.d. log-normal, such that $\log \boldsymbol{c}_{T_{\boldsymbol{\alpha}}} \sim \mathcal{N}(0, 2)$. For each size ($M$) of the training set, 200 CS implementations were



(a) (b)

Figure 1: The measurement matrices constructed by the exact and near-orthonormal bases exhibit similar performance in the theoretical (sufficient) bound and numerical results for recovery of sparse vector. Both bases outperform the Legendre basis. "—■—": the exact orthonormal amdP basis; "- -▲- -": the near-orthonormal amdP basis; "-·-◆-·-": Legendre basis. (a) Mean value of the theoretical bound $\mathbb{E}\left[\theta_s/\left(1-\delta_s\right)\right]$ of exact recovery for measurement matrices $\boldsymbol{A}$ constructed by various bases for the chosen non-zero index $T_{\boldsymbol{\alpha}}$ with $s = 3$. The error bar represents the standard deviation. The inset plot shows the theoretical prediction of the exact recovery probability. (b) Relative l_1 error of the recovered sparse vector ($s = 5$) using different training set size M. The inset plot shows the recovery error $\|\boldsymbol{c}-\tilde{\boldsymbol{c}}\|_1$ of one training set for the Legendre basis system.

conducted to compute the average of the numerical error $\|\boldsymbol{c}-\tilde{\boldsymbol{c}}\|_2$, as shown in Figure 2(a). Similar to the previous example, the Legendre basis exhibits the largest approximation error. The near-orthonormal basis shows smaller error than the exact orthonormal basis.

We also computed the density distribution of individual component $|\boldsymbol{c}_{i'} - \tilde{\boldsymbol{c}}_{i'}|$, where i' refers to single index sorted by the magnitude in descending order. Figure 2(b-d) shows that, compared with the exact orthonormal basis and the Legendre basis, the distribution of $\log|\boldsymbol{c}_{i'} - \tilde{\boldsymbol{c}}_{i'}|$ obtained from near-orthonormal basis is biased toward the smallest magnitudes for error of individual i'. This result can be interpreted as that the average of $\|\boldsymbol{c}-\tilde{\boldsymbol{c}}\|_2$ of the near orthogonal basis is smaller than that of the exact orthogonal basis and also outperforms the Legendre basis.

4.2. Systems with explicit knowledge of density function

In this subsection, we demonstrate the proposed method in systems with common non-Gaussian randomness with analytical density function $\omega(\boldsymbol{\xi})$. We show that the present method based on orthonormal basis construction and rotation of the random variables exploits the sparser representation of QoI while retaining proper orthogonality with respect to rotated variables. Therefore, it yields more accurate surrogate models than other approaches based on the direct recovery of $\boldsymbol{c}$ without the sparsity enhancement rotation procedure and/or directly applying the rotation procedure without reconstruction of the orthonormal amdP basis.

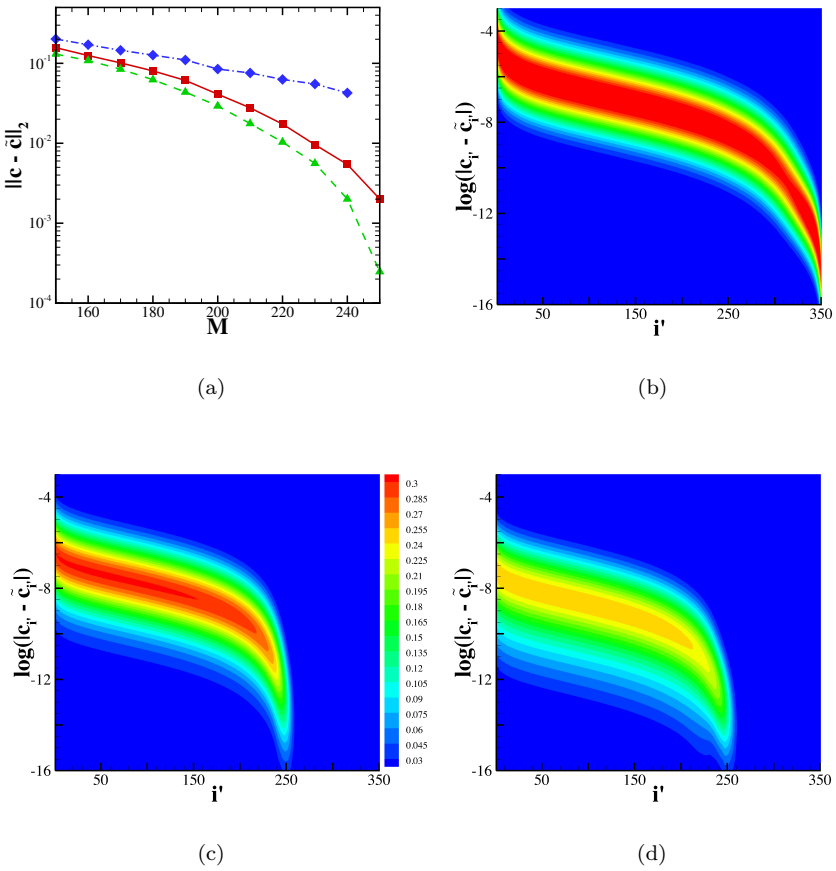


Figure 2: The measurement matrices constructed by different bases show different numerical performance for the recovery of non-sparse vector. The near-orthonormal basis shows the most accurate result. (a) l_1 error of the recovered vector $\mathbf{c}$ with different bases. "—■—": the exact orthonormal amdP basis; "- -▲- -": the near-orthonormal amdP basis; "-·-◆-·-": Legendre basis. (b-d) Contours of $|\mathbf{c}_{\boldsymbol{\alpha}'} - \tilde{\mathbf{c}}_{\boldsymbol{\alpha}'}|$ (sorted by magnitude) from training sets of size $M = 230$ with Legendre (top right), orthonormal (bottom left) and near-orthonormal bases (bottom right).

4.2.1. High-dimensional polynomial

For the first numerical example, we consider a high-dimensional polynomial function

$$f(\boldsymbol{\xi}) = \sum_{|\boldsymbol{\alpha}|\leq 3} \hat{c}_{\boldsymbol{\alpha}} \hat{\psi}_{\boldsymbol{\alpha}}(\boldsymbol{\xi}) = \sum_{i=1}^{N} \frac{\eta_i}{|i|^{1.5}} \hat{\psi}_i(\boldsymbol{\xi}), \tag{4.5}$$

where $\hat{\psi}_{\boldsymbol{\alpha}}$ and $\hat{\psi}_i$ represent monomial basis functions, η_i represents uniform random variables $\mathcal{U}[0,1]$. We employed this polynomial function with sparse coefficients as a benchmark problem to examine the recovery accuracy of the present method. $\boldsymbol{\xi}$ is a random vector consisting of 20 i.i.d. random variables. The density function of the i-th variable ξ_i is given by

$$\omega(\xi_i) = e^{-\xi_i}, \tag{4.6}$$

where the corresponding orthonormal basis are given by the Laguerre polynomials. Accordingly, we construct a 3^{rd}-order polynomial expansion $\tilde{f}(\boldsymbol{\xi})$ with $N = 1771$ multivariate basis functions, which are the tensor product of the univariate Laguerre polynomials. Figure 3 shows the relative l_2 error of $\tilde{f}$ computed by level 4 sparse grid integration. Similar to the previous example, the probability density function (PDF) of $\boldsymbol{\chi}$ does not retain the form $\omega'(\boldsymbol{\chi}) = \prod_{i=1}^{d} \exp(-\chi_i)$ after the rotation. Iteratively employing the multivariate Laguerre polynomials to represent $\tilde{g}(\boldsymbol{\chi})$ may result in erroneous prediction (the red dash-dotted curve). Alternatively, such a problem can be addressed by using the reconstructed orthonormal amdP basis with respect to $\boldsymbol{\chi}$, which yields a smaller error than $\tilde{f}(\boldsymbol{\xi})$ (the blue dashed curve).

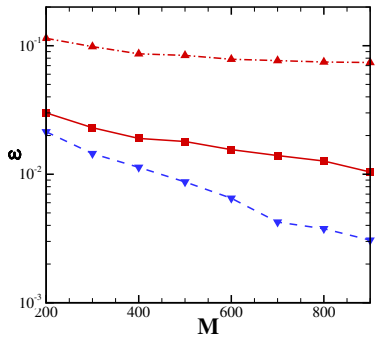


Figure 3: Sparsity-enhancing rotation with the reconstructed orthonormal amdP basis yields the most accurate recovery of a high-dimensional polynomial function of random vectors following density function given by Equation (4.6). Directly applying the rotation procedure without reconstructing the orthonormal basis yields errorneous prediction. "—■—": Laguerre polynomial basis with respect to $\boldsymbol{\xi}$; "-·-▲-·-": Laguerre polynomial basis with respect to rotated vector $\boldsymbol{\chi}$; "- - ▼- - ": the reconstructed amdP orthonormal basis with respect to rotated vector $\boldsymbol{\chi}$.

4.2.2. One-dimensional elliptic PDEs with high-dimensional random inputs

We applied the proposed method to model the solution to a one-dimensional (1D) elliptic PDE with high dimensional random input

$$-\frac{d}{dx}\left(D(x;\boldsymbol{\xi})\frac{du(x;\boldsymbol{\xi})}{dx}\right)=1,\quad x\in(0,1)$$
$$u(0)=u(1)=0, \tag{4.7}$$

where $a(x;\boldsymbol{\xi}):=\log D(x;\boldsymbol{\xi})$ is the stochastic input and $a(x;\boldsymbol{\xi})$ was a stationary process with correlation function

$$K(x,x')=\exp\left(\frac{|x-x'|}{l_c}\right), \tag{4.8}$$

where l_c is the correlation length. We constructed $a(x;\boldsymbol{\xi})$ by the Karhunen-Loève (KL) expansion:

$$a(x;\boldsymbol{\xi})=a_0(x)+\sigma\sum_{i=1}^{d}\sqrt{\lambda_i}\phi_i(x)\xi_i, \tag{4.9}$$

where $\{\lambda_i\}_{i=1}^d$, and $\{\phi_i(x)\}_{i=1}^d$ are the d largest eigenvalues and the corresponding eigenfunctions of $K(x,x')$. The values of λ_i and the analytical expressions for ϕ_i were available from the literature [87]. The ξ_i are i.i.d. random variables on $[-1,1]$. The density function of ξ_i is given by

$$\omega(\xi_i)=\frac{1}{\pi\sqrt{1-\xi_i^2}}, \tag{4.10}$$

where the corresponding orthonormal basis consists of Chebyshev polynomials of the first kind. For this example, we set $a_0(x)\equiv 1$, $\sigma=0.8$, $l_c=0.14$ and $d=16$. We chose the quantity of interest as $u(x;\boldsymbol{\xi})$ at $x=0.45$ and constructed a 3^{rd}-order polynomial expansion with $N=969$ basis functions. Figure 4 shows the relative l_2 error of the constructed $\tilde{f}(\boldsymbol{\xi})$ and $\tilde{g}(\boldsymbol{\chi})$. For the density function $\omega(\boldsymbol{\xi}_i)$ given by (4.10), $\tilde{f}(\boldsymbol{\xi})$ can be represented by a multivariate basis constructed by the tensor products of univariate Chebyshev polynomials. However, in general, the PDF of $\boldsymbol{\chi}$ does not retain the form $\omega'(\boldsymbol{\chi})=\prod_{i=1}^d\frac{1}{\pi\sqrt{1-\chi_i^2}}$. As shown in Figure 4, iteratively employing the multivariate Chebyshev polynomials to represent $\tilde{g}(\boldsymbol{\chi})$ (the red dash-dotted curve)—as done in previous studies [88]—resulted in a larger error than $\tilde{f}(\boldsymbol{\xi})$. Representing $\tilde{g}(\boldsymbol{\chi})$ by the reconstructed orthonormal amdP basis (the blue dashed curve) further decreases the numerical error compared to $\tilde{f}(\boldsymbol{\xi})$ (the solid red curve).

4.3. Systems with implicit knowledge of density function

In this suite of benchmark examples, we investigated the applicability and efficiency of the developed DSRAR framework based on data-driven orthonormal bases construction and sparsity enhanced rotation.

4.3.1. High-dimensional polynomials

We studied the ability of the data-driven method to recover a high-dimensional polynomial function

$$f(\boldsymbol{\xi})=\sum_{\boldsymbol{\alpha}\in T_{\boldsymbol{\alpha}}}\hat{\psi}_{\boldsymbol{\alpha}}(\boldsymbol{\xi}), \tag{4.11}$$

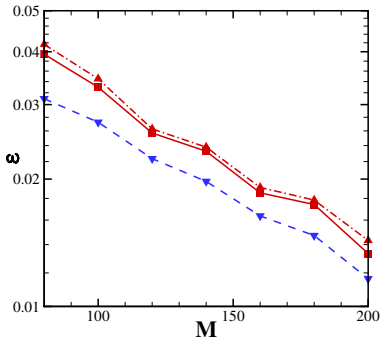


Figure 4: Sparsity-enhancing rotation with reconstructed orthonormal basis yield the most accurate surrogate models for a 1D elliptical PDE with random permeability coefficient modeled by Equations (4.9) and (4.10). Directly applying the rotation procedure without reconstructing the orthonormal basis yields increased numerical error. "—■—": Chebyshev polynomial basis with respect to $\boldsymbol{\xi}$; "-·-▲-·-": Chebyshev polynomial basis with respect to rotated vector $\boldsymbol{\chi}$; "- - ▼- -": the reconstructed orthonormal amdP basis with respect to rotated vector $\boldsymbol{\chi}$.

where $\hat{\psi}_{\boldsymbol{\alpha}}$ represents the monomial basis function, $T_{\boldsymbol{\alpha}}$ represents a set containing 50 indices randomly chosen from Λ_p^d with $d = 25$ and $p = 3$. The sample set S of random vector $\boldsymbol{\xi}$ for basis construction was generated from the Gaussian mixture model specified in (4.1) with $|S| = 2 \times 10^5$.

We approximated $f(\boldsymbol{\xi})$ by a 3rd-order polynomial expansion $\tilde{f}(\boldsymbol{\xi}) = \sum_{i=1}^{N} \tilde{c}_i \psi_i(\boldsymbol{\xi})$ with $N = 3276$. Figure 5(a) shows the relative l_2 error of the constructed surrogate model $\tilde{f}$ defined by

$$\epsilon = \left(\int (f(\boldsymbol{\xi}) - \tilde{f}(\boldsymbol{\xi}))^2 \, \mathrm{d}\nu_{S_2}(\boldsymbol{\xi}) \Big/ \int f(\boldsymbol{\xi})^2 \, \mathrm{d}\nu_{S_2}(\boldsymbol{\xi}) \right)^{\frac{1}{2}}, \tag{4.12}$$

where 20 implementations were utilized for each training sample size number M. As shown in Figure 5(a), $\tilde{f}(\boldsymbol{\xi})$ constructed by the near-orthonormal amdP basis yielded the smallest error while the tensor product of Legendre basis functions yielded the largest error. Accordingly, the magnitudes of the recovered coefficients $|\tilde{c}_i|$ by the exact and near-orthonormal bases decayed more quickly than those recovered using the Legendre basis functions, as shown in Figure 5(b). Furthermore, $\tilde{f}(\boldsymbol{\xi})$ allowed us to define a new random vector $\boldsymbol{\chi}$, which further enhanced the sparsity of $\boldsymbol{c}$, as shown in Figures 5(c) and (d). Following Step 5 in Algorithm 4, we defined a new random $\boldsymbol{\chi}$ through rotation. The associated representation coefficient vector $\boldsymbol{c}$ has enhanced sparsity.

However, for the exact and near-orthonormal basis, the $\tilde{g}(\boldsymbol{\chi})$ gave smaller errors (the dashed curve) than $\tilde{f}(\boldsymbol{\xi})$ (the solid curve), as shown in Figure 5(a). Thus, enhancing the sparsity of $\boldsymbol{c}$ alone does not guarantee enhanced accuracy of $\tilde{f}$. In particular, $\tilde{g}(\boldsymbol{\chi})$ constructed by the Legendre basis yielded larger error than $\tilde{f}(\boldsymbol{\xi})$ as demonstrated in Figure 5(a); although, the sparsity of $\boldsymbol{c}$ was greater, as seen in Figure 5(d). This behavior indicates that retaining the orthonormal condition can be crucial for the accurate construction of $\tilde{f}$. The basis bound (see Table B.2 in Appendix B) provides a metric to understand why the near-orthonormal basis

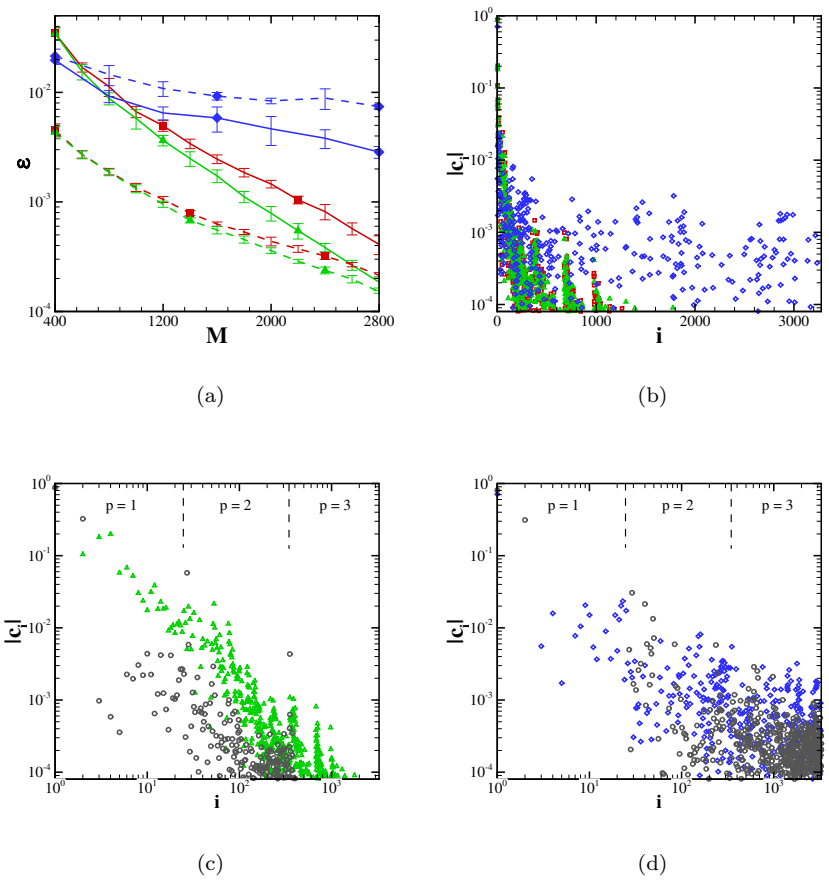


Figure 5: Numerical results for recovery of a high-dimensional polynomial function. The combination of near-orthonormal basis construction with the sparsity enhancement rotation procedure yields the most accurate results. Directly applying the rotation procedure to the Legendre basis may lead to increased error despite increased sparsity in $\boldsymbol{c}$. (a) Relative l_2 error of the recovered polynomial function with different bases: the exact orthonormal amdP basis with respect to $\boldsymbol{\xi}$ ("—■—") and $\boldsymbol{\chi}$ ("- -■- -"); the near-orthonormal amdP basis with respect to $\boldsymbol{\xi}$ ("—▲—") and $\boldsymbol{\chi}$ ("- -▲- -"); Legendre basis with respect to $\boldsymbol{\xi}$ (" —◆—") and $\boldsymbol{\chi}$(" - -◆- -"). (b) Coefficients magnitude $|c_i|$ recovered using different bases. "□": the exact orthonormal amdP basis with respect to $\boldsymbol{\xi}$; "△": the near-orthonormal amdP basis with respect to $\boldsymbol{\xi}$; " ◇": Legendre basis with respect to $\boldsymbol{\xi}$. (c) Recovered coefficient magnitude $|\mathbf{c}_i|$ using the near orthogonal basis with respect to $\boldsymbol{\xi}$ ("△") and $\boldsymbol{\chi}$ ("○"). The dashed vertical lines indicate the separation between different polynomial orders p. (d) Recovered coefficient magnitude $|\mathbf{c}_i|$ using the Legendre basis with respect to $\boldsymbol{\xi}$ (" ◇") and $\boldsymbol{\chi}$ ("○").

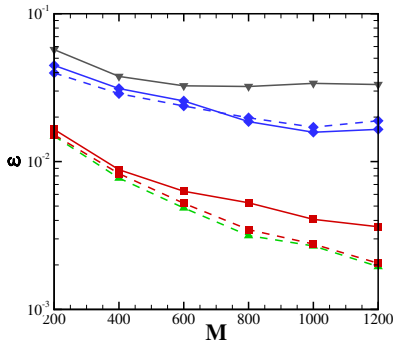


Figure 6: The combination of near-orthonormal basis construction and sparsity enhancement rotation yields the most accurate results, as shown through the relative l_2 error of the constructed surrogate model for the 1D elliptic PDE with random permeability coefficient: the exact orthonormal amdP basis with respect to $\boldsymbol{\xi}$ ("—■—") and $\boldsymbol{\chi}$ ("- - ■ - -"); Legendre basis with respect to $\boldsymbol{\xi}$ ("—◆—") and $\boldsymbol{\chi}$("- - ◆ - -"); Hermite basis with respect to $\boldsymbol{\xi}$ ("—▼—"); the near-orthornormal amdP basis with respect to $\boldsymbol{\chi}$ ("- - ▲ - -").

performs better than the exact orthonormal basis.

4.3.2. 1D elliptic PDEs with high-dimensional random inputs

In this example, we revisited the 1D elliptic PDE (4.7) with random coefficient given by Equation (4.9). Here we set $a_0(x) \equiv 1$, $\sigma = 1$, $l_c = 0.12$ and $d = 20$ such that $\sum_{i=1}^{d} \lambda_i > 0.91 \sum_{i=1}^{\infty} \lambda_i$.

Similar to the work by Zabaras et al. [70], a non-Gaussian multivariate distribution was used for $\boldsymbol{\xi} = (\xi_1, \xi_2, \cdots, \xi_d)$. We generated a sample set $\left\{\tilde{\boldsymbol{\xi}}^{(k)}\right\}_{k=1}^{Ns}$, where $N_s = 2 \times 10^5$ and $\tilde{\boldsymbol{\xi}}$ came from the Gaussian mixture distribution specified in (4.1). We used PCA to transform $\tilde{\boldsymbol{\xi}}$ to $\boldsymbol{\xi}$ such that $\mathbb{E}[\boldsymbol{\xi}_i] = 0$ and $\mathbb{E}[\boldsymbol{\xi}_i \boldsymbol{\xi}_j] = \delta_{ij}$. For each input sample $\boldsymbol{\xi}^{(k)}$, a and u only depended on x and the solution of the deterministic elliptic equation is given by [54]

$$\begin{aligned} u(x) &= u(0) + \int_0^x \frac{a(0)u(0)' - y}{a(y)}\,\mathrm{d}y \\ a(0)u(0)' &= \left(\int_0^1 \frac{y}{a(y)}\,\mathrm{d}y\right) \Big/ \left(\int_0^1 \frac{1}{a(y)}\,\mathrm{d}y\right). \end{aligned} \tag{4.13}$$

We chose the QoI to be $u(x;\boldsymbol{\xi})$ at $x = 0.35$ and constructed a 3^{rd}-order polynomial expansion with $N = 1771$ basis functions. Figure 6 shows the relative l_2 error of $\tilde{f}(\boldsymbol{\xi})$ (solid curve) and $\tilde{g}(\boldsymbol{\chi})$ (dashed curve) constructed by different bases. The data-driven bases (both exact orthonormal basis and near-orthonormal basis) showed more accurate results than the Legendre basis and the Hermite basis. In particular, the near-orthonormal basis with respect to the rotated variable $\boldsymbol{\chi}$ yielded the most accurate result (the green dashed curve). In contrast, directly employing the Legendre basis to the rotated variable $\boldsymbol{\chi}$ without reconstructing the basis function led to increased l_2 error, although $\boldsymbol{c}$ shows more sparsity in terms of $\boldsymbol{\chi}$ (the gray dashed curve) than $\boldsymbol{\xi}$ (the gray solid curve).

4.4. UQ study of a molecule system under Non-Gaussian conformational distributions

We demonstrated the proposed method on a physical system exploring conformational uncertainty in a small molecule system. Molecular properties, such as solvation energies or solvent-accessible surface areas (SASAs), are often calculated using single molecular conformations. However, due to thermal energy, a molecule undergoes conformational fluctuations which can induce significant uncertainty in properties calculated from single structures. Our previous work [1] was focused on quantifying this uncertainty using a simple multivariate Gaussian model for conformational fluctuations: the elastic network model [89]. However, it is well known that the conformational fluctuations are often non-Gaussian due to the complicated structure of the underlying energy landscape. Therefore, in the current study, we construct the data-driven basis *directly* from the samples of molecular trajectories collected from molecular dynamics (MD) simulations, thus eliminating the *over-simplified* Gaussian assumption.

We simulated the dynamics of the small molecule benzyl bromide under equilibrium (see Appendix E for details) and collected a sample set of the instantaneous molecular structure $\left\{\mathbf{r}^{(k)}\right\}_{k=1}^{N_s}$ from MD simulation trajectories over 20μs. In what follows, $N_s = 2 \times 10^5$ and $\mathbf{r}$ represent the positions of individual atoms. As a pre-processing step, we transformed $\left\{\mathbf{r}^{(k)}\right\}_{k=1}^{N_s}$ into a set of uncorrelated random vectors $S = \left\{\boldsymbol{\xi}^{(k)}\right\}_{k=1}^{N_s}$ via PCA:

$$\begin{aligned} \boldsymbol{\Sigma} &= \mathbb{E}\left[(\mathbf{r} - \bar{\mathbf{r}})\,(\mathbf{r} - \bar{\mathbf{r}})^T\right] \\ \boldsymbol{\Sigma} &= \mathbf{Q}\boldsymbol{\Gamma}\mathbf{Q}^T \quad \boldsymbol{\xi} = \boldsymbol{\Gamma}^{-1/2}\mathbf{Q}^T\mathbf{r}, \end{aligned} \tag{4.14}$$

where the average $\mathbb{E}[\cdot]$ is taken over the entire sample set and $\boldsymbol{\xi} \in \mathbb{R}^{12}$ is the normalized random vector that represents 99.99% of the observed variance. Figures 7(a) and (b) show the joint distributions of (ξ_1, ξ_2) and (ξ_1, ξ_3). Although the individual components of $\boldsymbol{\xi}$ are uncorrelated, the joint density distributions are mutually dependent and deviate from the standard Gaussian distributions.

We chose the polar solvation energy and SASA as the target QoIs for this system. The polar solvation energy was modeled by the Poisson-Boltzmann equation [90, 91]

$$-\nabla \cdot (\epsilon_f(\boldsymbol{x};\boldsymbol{\xi})\nabla\varphi(\boldsymbol{x};\boldsymbol{\xi})) = \rho_f(\boldsymbol{x};\boldsymbol{\xi}) \tag{4.15}$$

which relates the electrostatic potential φ to a dielectric coefficient ϵ_f and a fixed charge distribution ρ_f. Equation (4.15) is typically solved with Dirichlet boundary conditions set to an analytical asymptotic solution of the equation for an infinite domain. The dielectric coefficient ϵ_f implicitly represents the boundary between the atoms of the molecule and the surrounding solvent: the coefficient changes rapidly across this boundary from a low dielectric value in the molecular interior to a high dielectric value in the solvent. The charge distribution ρ_f is generally modeled as a collection of δ-like functions centered on the atoms of the molecule with magnitudes proportional to the atomic partial charges. Both ϵ_f and ρ_f are dependent on the

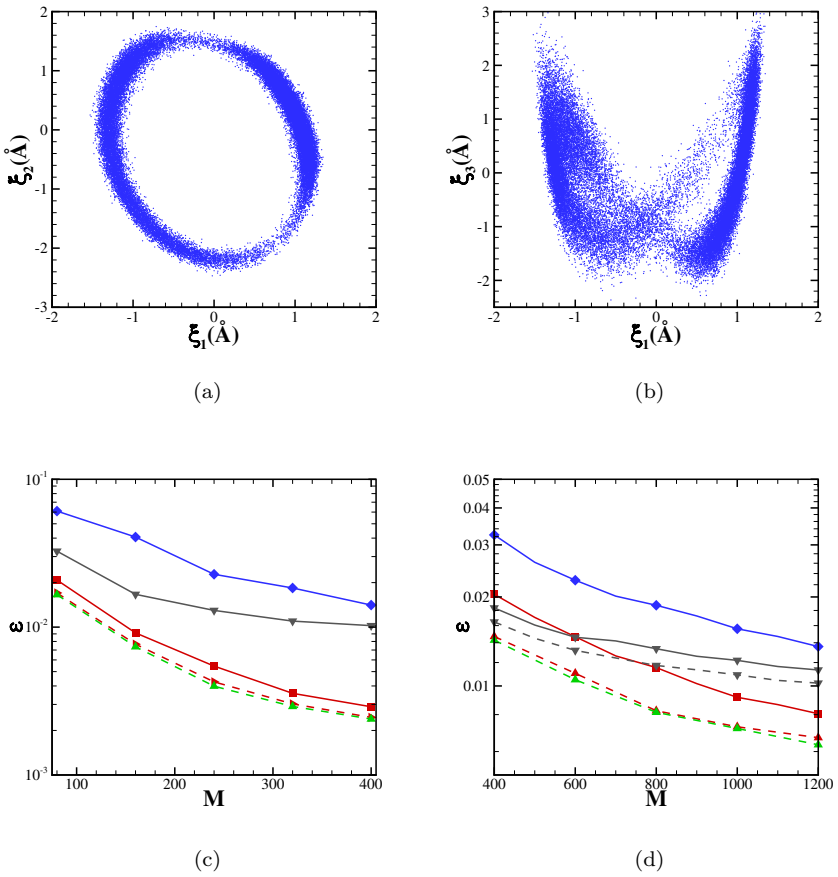
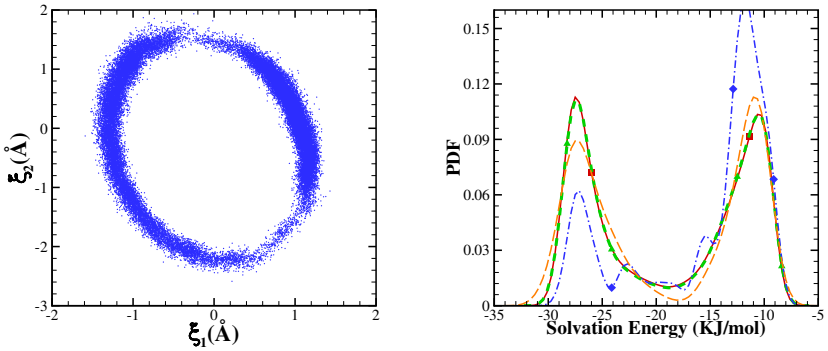


Figure 7: The present method based on data-driven basis construction and sparsity enhancement rotation yields the most accurate surrogate model for molecular systems with mutually dependent non-Gaussian density distributions. (a-b) Sampling points representing the joint distributions (ξ_1, ξ_2) (left) and (ξ_1, ξ_3) (right). (c-d) Relative l_2 error of the polar solvation energy (left) and the local SASA (right) of an individual atom (the H9 atom attached to the ortho-carbon atom) obtained with different numbers of training data M: the exact amdP orthonormal basis with respect to $\boldsymbol{\xi}$ ("—■—") and $\boldsymbol{\chi}$ ("- - ■ - -"); Hermite basis with respect to $\boldsymbol{\xi}$ ("—▼—") and $\boldsymbol{\chi}$ ("- - ▼ - -"); Legendre basis with respect to $\boldsymbol{\xi}$ (" —◆— "); the near-orthornormal amdP basis with respect to $\boldsymbol{\chi}$ ("- - ▲ - -").



(a) (b)

Figure 8: The present method yields the most accurate prediction on the PDF of the QoI for the molecular systems. Direct fitting of the underlying density $\omega(\boldsymbol{\xi})$ using Gaussian Mixture model may induce biased error to the PDF prediction. (a) Fitted random variables (ξ_1, ξ_2) with Gaussian mixture models. (b) PDF of the solvation energy obtained with the Gaussian Mixture model and the present data-driven approach. "—■—": reference solution obtained from 2×10^5 MC samples; "-·-◆-·-": direct MC sampling using the same set of 200 samples; "- - ▲ - -": present method using the same set of 200 samples; "_ _ _ _": fitting Gaussian Mixture model using 800 samples.

instantaneous molecular structure (i.e., $\boldsymbol{\xi}$). The polar solvation energy was calculated from

$$G_p(\boldsymbol{\xi}) = \int \rho_f(\boldsymbol{x};\boldsymbol{\xi}) \left(\varphi(\boldsymbol{x};\boldsymbol{\xi})) - \varphi_h(\boldsymbol{x};\boldsymbol{\xi}))\right) d\boldsymbol{x} \tag{4.16}$$

where φ_h is a reference potential obtained from solution of

$$-\epsilon_h \nabla^2 \varphi_h(\boldsymbol{x};\boldsymbol{\xi}) = \rho_f(\boldsymbol{x};\boldsymbol{\xi}) \tag{4.17}$$

where ϵ_h is a constant reference dielectric value. We used the Adaptive Poisson-Boltzmann Solver (APBS) software to solve the equations above [92]. Besides the solvation energy of the whole molecule, we also studied a local property like the SASA of an individual atom (the H9 atom attached to the ortho-carbon atom of the benzyl bromide molecule, see Figure E.11) by the Shake-Rupley algorithm [93] using APBS. Details of the APBS calculations are presented in Appendix E.

Figures 7(c) and (d) show the relative l_2 error of the constructed surrogate model $\tilde{f}(\boldsymbol{\xi})$ for the solvation energy and SASA using a 4^{th}-order gPC expansion with $N = 1820$ basis functions. For both QoIs, the near-orthonormal and orthonormal bases with respect to the rotated variable $\boldsymbol{\chi}$ (dashed curves) yield similar error which is much smaller than the error of Legendre and Hermite bases. A possible explanation for the similar performance of the near-orthonormal and orthonormal bases is the closeness of the basis bound estimates for these two bases (see Table B.3 in Appendix B).

Instead of the direct construction of $\tilde{f}(\boldsymbol{\xi})$ using data-driven basis functions, another possible approach to characterize the uncertainty of the molecular system is to fit the distribution density $\omega(\boldsymbol{\xi})$ with a distribution

model such as a Gaussian Mixture model. Figure 8 (a) shows a scatter plot of the joint distribution (ξ_1, ξ_2) extracted from the fitted Gaussian mixture distribution $\tilde{\omega}(\boldsymbol{\xi})$ using 7 Gaussian modes. Accordingly, we can construct the surrogate model for each Gaussian mode using standard Hermite basis function. However, it is well-known that accurate construction of $\omega(\boldsymbol{\xi})$ is a numerically challenging problem for $d > 4$. As shown in Figure 8(b), direct fitting $\omega(\boldsymbol{\xi})$ by $\tilde{\omega}(\boldsymbol{\xi})$ induces non-negligible error and leads to biased prediction of the PDF of the solvation energy. Furthermore, we lose the one-to-one mapping between the individual conformation state $\boldsymbol{\xi}$ and the QoIs through the constructed surrogate model $\tilde{f}(\boldsymbol{\xi})$.

5. Summary

In this study, we have developed a DSRAR framework for constructing surrogate models irrespective of the mutual dependence between the components of random inputs using limited training points. To the best of our knowledge, this problem has not been addressed by previous UQ studies based on polynomial chaos expansions. The DSRAR framework does not assume mutual independence between the components of random inputs and therefore can be applied to UQ in complex systems where information about the underlying random distribution can be implicit. To construct the surrogate model, this framework uses data-driven amdP basis construction and a sparsity-enhancing rotation procedure which leads to more accurate recovery of the sparse representation of the target function. The method benefits from both the orthonormal basis expansion and the enhanced sparsity of the expansion coefficients. With the assumption that there exists a sparse representation of the surrogate model, the DSRAR approach can be applied to challenging UQ problems under two widely encountered situations: (I) probability measure implicitly represented by a large collection of samples and (II) non-Gaussian probability measures with explicit (analytical) forms. For systems with explicit knowledge of the probability measure, our method exploits sparser representations of QoIs while retaining proper orthogonality with respect to rotated variables. For systems with randomness implicitly represented by a large collection of random samples, we also proposed a heuristic method to construct a *near-orthonormal* basis in addition to the exact orthonormal basis with respect to the discrete measure. The near-orthonormal basis shows a smaller basis bound and empirically yields more accurate representations. The numerical examples show the effectiveness of our method for realistic problems on quantifying uncertainty propagation in molecular system under conformational fluctuations as well as PDEs with arbitrary underlying probability measures.

For future study, we note that several issues not considered in the present work could further improve the performance of the present DSRAR framework. The heuristic approach to constructing near-orthonormal basis introduced in this study yields smaller basis bounds and more accurate representations than existing methods. However, we do not have the theoretical analysis to formally show that the near-orthonormal basis is optimal and to establish the conditions under which it outperforms the exact orthonormal basis. It

would be interesting to investigate different approaches of data-driven basis construction to further improve the properties of measurement matrix for CS purposes. For instance, if new data becomes available after the surrogate construction, it is worth exploring how to use the new information to design more sophisticated (cross-validation) strategies to optimize the orthonormal threshold values and the basis construction procedure. Furthermore, our study used a standard ℓ_1 minimization approach for relaxing the CS problem and recovering a sparse solution of the under-determined system. However, other optimization approaches can be employed when the measurement matrix is highly coherent when ℓ_1 minimization is not necessarily optimal. Finally, it would be interesting to employ the developed DSRAR approach for UQ study in other complex biological systems [94, 95]. Such results will be presented in a future publication.

Appendix A. Proof of Theorem 4.1

Proof. Let $\boldsymbol{v} \in$ Ker A and $\boldsymbol{x} \neq \boldsymbol{c}$ another solution of $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$. To show that $\boldsymbol{c}$ is the unique l_1 minimizer of $\boldsymbol{A}\boldsymbol{c} = \boldsymbol{b}$, it is sufficient if

$$\|\boldsymbol{v}_{T_{\boldsymbol{\alpha}}}\|_1 < \|\boldsymbol{v}_{T_{\boldsymbol{\alpha}}^c}\|_1, \tag{A.1}$$

which gives

$$\begin{aligned} \|\boldsymbol{c}\|_1 &\leq \|\boldsymbol{c} - \boldsymbol{x}_{T_{\boldsymbol{\alpha}}}\|_1 + \|\boldsymbol{x}_{T_{\boldsymbol{\alpha}}}\|_1 = \|\boldsymbol{c}_{T_{\boldsymbol{\alpha}}} - \boldsymbol{x}_{T_{\boldsymbol{\alpha}}}\|_1 + \|\boldsymbol{x}_{T_{\boldsymbol{\alpha}}}\|_1 = \|\boldsymbol{v}_{T_{\boldsymbol{\alpha}}}\|_1 + \|\boldsymbol{x}_{T_{\boldsymbol{\alpha}}}\|_1 \\ &< \|\boldsymbol{v}_{T_{\boldsymbol{\alpha}}^c}\|_1 + \|\boldsymbol{x}_{T_{\boldsymbol{\alpha}}}\|_1 = \|\boldsymbol{x}\|_1. \end{aligned} \tag{A.2}$$

To satisfy (A.1), we partition $T_{\boldsymbol{\alpha}}^c$ into $T_{\boldsymbol{\alpha}}^c = T_{\boldsymbol{\alpha},1}^c \bigcup T_{\boldsymbol{\alpha},2}^c \bigcup \cdots$, where $T_{\boldsymbol{\alpha},1}^c$ is the index set of s largest absolute entries of $\boldsymbol{v}$ in $T_{\boldsymbol{\alpha}}^c$, $T_{\boldsymbol{\alpha},2}^c$ is the index set of s largest absolute entries of $\boldsymbol{v}$ in $T_{\boldsymbol{\alpha}}^c T_{\boldsymbol{\alpha},1}^c$. Accordingly,

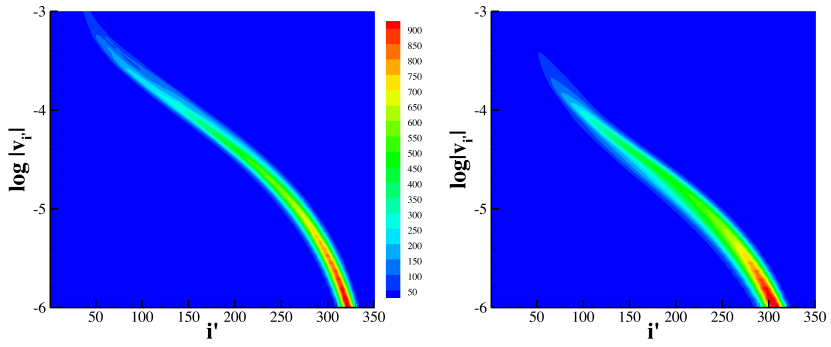
$$\|\boldsymbol{v}_{T_{\boldsymbol{\alpha}}}\|_2^2 \leq \frac{1}{1-\delta_s}\|\boldsymbol{A}\boldsymbol{v}_{T_{\boldsymbol{\alpha}}}\|_2 = \frac{1}{1-\delta_s}\sum_{k=1}\left\langle \boldsymbol{A}\boldsymbol{v}_{T_{\boldsymbol{\alpha}}}, \boldsymbol{A}(-\boldsymbol{v}_{T_{\boldsymbol{\alpha},k}^c})\right\rangle \leq \frac{\theta_s}{1-\delta_s}\sum_{k=1}\|\boldsymbol{v}_{T_{\boldsymbol{\alpha}}}\|_2\|\boldsymbol{v}_{T_{\boldsymbol{\alpha},k}^c}\|_2, \tag{A.3}$$

which gives $\|\boldsymbol{v}_{T_{\boldsymbol{\alpha}}}\|_2 \leq \frac{\theta_s}{1-\delta_s}\sum_{k=1}\|\boldsymbol{v}_{T_{\boldsymbol{\alpha},k}^c}\|_2$. The remaining of the proof is straightforward and follows Theorem 2.6 of Rauhut [96]. By the Cauchy-Schwarz inequality, we obtain

$$\|\boldsymbol{v}_{T_{\boldsymbol{\alpha}}}\|_1 \leq \frac{\theta_s}{1-\delta_s}\left(\|\boldsymbol{v}_{T_{\boldsymbol{\alpha}}}\|_1 + \|\boldsymbol{v}_{T_{\boldsymbol{\alpha}}^c}\|_1\right). \tag{A.4}$$

Equation (A.1) follows if $\frac{\theta_s}{1-\delta_s} < 0.5$. □

Remark Appendix A.1. We emphasize that Theorem 4.1 holds only for the given index set $T_{\boldsymbol{\alpha}}$; it provides a metric to examine the recovery accuracy with respect to measurement matrix $\boldsymbol{A}$ and should not be viewed as the sufficient condition for exact recovery of *arbitrary s-sparse vector* via l_1-minimization (see canonical references [82, 83, 96] for details). Theorem 4.1 also indicates that, for the given index set $T_{\boldsymbol{\alpha}}$, small $\|\boldsymbol{A}_{T_{\boldsymbol{\alpha}}}^* \boldsymbol{A}_{T_{\boldsymbol{\alpha}}} - I\|_2$ will promote the recover of $\boldsymbol{v}_{T_{\boldsymbol{\alpha}}}$.



(a) (b)

Figure B.9: The null spaces of measurement matrices constructed by the exact and near-orthonormal bases are different under $\|\boldsymbol{v}_{T_{\boldsymbol{\alpha}}}\|_1 > \left\|\boldsymbol{v}_{T_{\boldsymbol{\alpha}}^c}\right\|_1$—a necessary condition for $\boldsymbol{c}$ not being recoverable exactly. Density contour of the normalized null space vector component $\log|\mathrm{v}_{i'}|$ (sorted by magnitude) of the measurement matrix $\boldsymbol{A}$ constructed by orthogonal (a) and near-orthogonal basis functions (b) that satisfy $\|\boldsymbol{v}_{T_{\boldsymbol{\alpha}}}\|_1 > \left\|\boldsymbol{v}_{T_{\boldsymbol{\alpha}}^c}\right\|_1$ and $\|\boldsymbol{v}\|_2 = 1$.

Appendix B. Measurement matrix and basis bounds

Appendix B.1. Null space of measurement matrix from Section 4.1

Let $\tilde{\boldsymbol{c}} = \boldsymbol{c} + \boldsymbol{v}$, $\boldsymbol{v} \in$ Ker $\boldsymbol{A}$ where $\boldsymbol{A}$ is the measurement matrix defined in (4.3). From the null space property [96], $\tilde{\boldsymbol{c}}$ does not fully recover $\boldsymbol{c}$ by ℓ_1 minimization (i.e., equation (2.6)) only if $\|\tilde{\boldsymbol{c}}\|_1 < \|\boldsymbol{c}\|_1$. As a *necessary condition* for the failure of recovery, it requires

$$\|\boldsymbol{v}_{T_{\boldsymbol{\alpha}}}\|_1 > \left\|\boldsymbol{v}_{T_{\boldsymbol{\alpha}}^c}\right\|_1, \tag{B.1}$$

where $T_{\boldsymbol{\alpha}}^c$ refers to the complement of $T_{\boldsymbol{\alpha}}$. Accordingly, different null space of measurement matrix $\boldsymbol{A}$ generally leads to different recovery error.

We examined the above necessary condition (B.1) for different measurement matrices by randomly choosing a non-zero index set $T_{\boldsymbol{\alpha}}$ with $|T_{\boldsymbol{\alpha}}| = 50$ and $M = 180$. For $\boldsymbol{A}$ constructed by both basis sets, we collected 1000 normalized $\boldsymbol{v} \in$ Ker A that satisfy $\|\boldsymbol{v}_{T_{\boldsymbol{\alpha}}}\|_1 > \left\|\boldsymbol{v}_{T_{\boldsymbol{\alpha}}^c}\right\|_1$. Figure B.9 shows the density contour of individual component $|\boldsymbol{v}_{i'}|$ in log-scale, where i' refers to the index sorted by magnitude in descending order. The two basis sets demonstrate different distributions of $\log|\boldsymbol{v}_{i'}|$, which likely contribute to the different recovery errors shown in Figure 2.

Appendix B.2. Basis bounds

The lower bound of the required number of samples M given in Theorem 2.5 suggests that bases with smaller basis bounds K are preferred. We expect that smaller basis bounds will correlate with higher

accuracy representations. For the constructed basis set $\psi_i(\boldsymbol{\xi}), i = 1, \cdots, N$, we define the basis bound $\tilde{K}$ on the given data set S by

$$\tilde{K} := \frac{1}{|S_{M_\sigma}|} \sum_{\boldsymbol{\xi} \in S_{M_\sigma}} |k(\boldsymbol{\xi})|, \tag{B.2}$$

where the set S_{M_σ} is defined by $S_{M_\sigma} = \left\{ \boldsymbol{\xi} \middle| |k(\boldsymbol{\xi}) - \mathbb{E}[k]| > M_\sigma \sigma[k], \boldsymbol{\xi} \in S \right\}$. Here $k(\boldsymbol{\xi}) := \max_i |\psi_i(\boldsymbol{\xi})|$ denotes the maximum magnitude for an individual sampling point $\boldsymbol{\xi}$, $\mathbb{E}[k]$ and $\sigma[k]$ represent the mean and the standard deviation of $k(\boldsymbol{\xi})$ on S with respect to the discrete measure ν_S. In this study, we present $\tilde{K}$ as an indication of the difference between the exact and near-orthonormal basis function. In compressive sensing, the measurement matrix only consists of limited number of samples. Therefore, we employ the mean of the tails in the basis bounds as an indicator of the upper bound of the largest entry values from the measurement matrix. M_σ defines the range of this tail set. We choose $M_\sigma = 5$ if not specified otherwise.

Table B.1: $\tilde{K}$ of constructed basis set for Gaussian mixture system $d = 25$, $p = 2$ and $N_s = 1 \times 10^5$.

M_σ	3	4	5	6	$\max_{\boldsymbol{\xi} \in S} k(\boldsymbol{\xi})$
$\tilde{K}_{\text{orth}}$	10.359	12.048	13.895	15.513	22.208
$\tilde{K}_{\text{near-orth}}$	9.622	11.196	12.867	14.448	18.790

Following the definition by Equation (B.2), we examine the basis bound $\tilde{K}$ of the numerical examples presented in this study. Table B.1 shows the results of Gaussian mixture system $\left\{\boldsymbol{\xi}^{(i)}\right\}, i = 1, \cdots, N_s$ with $N_s = 1 \times 10^5$, $d = 25$ and $p = 2$ which is defined in Section 4.1. For different values of M_σ, $\tilde{K}$ of the near orthogonal basis shows consistently smaller values than the values of the exact orthogonal basis set.

Table B.2 shows the basis bound $\tilde{K}$ of the Gaussian mixture system which is studied in Section 4.3.1 with $N_s = 2 \times 10^5$, $d = 25$ and $p = 3$. The values of $\tilde{K}$ for the near orthogonal basis are consistently smaller than the value for the exact orthogonal basis set no matter on the original random sample set or the rotated sample set. Furthermore, we present the basis bounds on the rotated sampling set $\left\{\boldsymbol{\chi}_M^{(i)}\right\}_{i=1}^{N_s}$, where the subscript "$M$" refers to the different number of training points utilized to construct the surrogate model $X(\boldsymbol{\xi})$. The near-orthogonal basis yields smaller $\tilde{K}$ than the exact orthogonal basis in each case.

Similarly, Table B.3 shows $\tilde{K}$ of the constructed basis for uncertainty quantification of the molecular solvation energy ($d = 12$, $p = 4$ and $N_s = 2 \times 10^5$), which is studied in Section 4.4. The near-orthogonal basis yields smaller values consistently for different number ($\boldsymbol{\chi}_M$) of training points.

Appendix C. Other metrics for the surrogate model

Besides the relative l_2 error, we have also computed the predictivity coefficients Q_2 for the test cases of Gaussian mixture (with $d = 25$ and $p = 3$) and molecular systems. Similar to Marrel et al. [97], Q_2 is defined

Table B.2: $\tilde{K}$ of constructed basis set for Gaussian mixture system $d = 25$, $p = 3$ and $N_s = 2 \times 10^5$.

	$\boldsymbol{\xi}$	$\boldsymbol{\chi}_{M=400}$	$\boldsymbol{\chi}_{M=1200}$	$\boldsymbol{\chi}_{M=1600}$	$\boldsymbol{\chi}_{M=2400}$
$\tilde{K}_{\text{orth}}$	32.497	32.522	32.079	33.142	32.308
$\tilde{K}_{\text{near-orth}}$	28.320	29.811	29.407	29.512	29.192

Table B.3: $\tilde{K}$ of constructed basis set for molecular system $d = 12$, $p = 4$ and $N_s = 2 \times 10^5$.

	$\boldsymbol{\chi}_{M=80}$	$\boldsymbol{\chi}_{M=160}$	$\boldsymbol{\chi}_{M=240}$	$\boldsymbol{\chi}_{M=320}$	$\boldsymbol{\chi}_{M=400}$
$\tilde{K}_{\text{orth}}$	40.596	39.914	39.789	39.218	39.142
$\tilde{K}_{\text{near-orth}}$	39.970	39.278	39.290	38.528	38.631

by

$$Q_2 = 1 - \int (f(\boldsymbol{\xi}) - \tilde{f}(\boldsymbol{\xi}))^2 \, \mathrm{d}\nu_{S_2}(\boldsymbol{\xi}) / \int \left(f(\boldsymbol{\xi}) - \bar{f}\right)^2 \, \mathrm{d}\nu_{S_2}(\boldsymbol{\xi}), \tag{C.1}$$

where $\bar{f}$ represents the mean of QoI on S_2. The results are shown in Tab. C.4, where the surrogate models are constructed by the our data-driven basis approach.

Table C.4: The predictivity coefficient Q_2 for polynomial function with Gaussian Mixture measure ($d = 25$ and $p = 3$) and the molecular system for solvation energy and SASA of atom H9.

molecule solvation	M	80	160	240	320	400
	Q_2	0.995715	0.999132	0.999731	0.999864	0.999911
molecule SASA	M	200	300	400	500	600
	Q_2	0.988675	0.996069	0.998272	0.998709	0.999027
Gaussian Mixture	M	200	300	400	500	600
	Q_2	0.998372	0.999347	0.999844	0.999892	0.999941

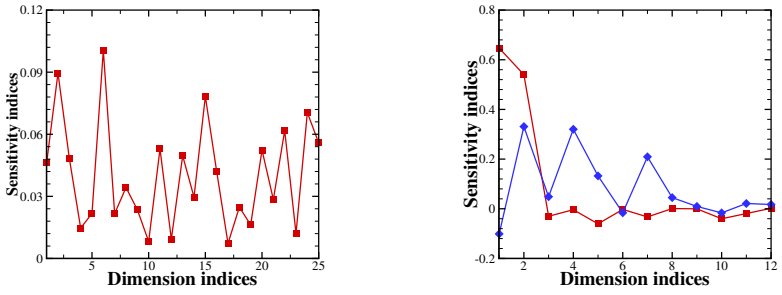
With the constructed surrogate model, we can compute the Sobol sensitivity indices for QoI with dependent random variables. In brief, $f(\boldsymbol{\xi})$ is expanded by

$$f(\boldsymbol{\xi}) = \eta_0(\boldsymbol{\xi}) + \sum_{\boldsymbol{\beta} \in \Theta^d} \eta_{\boldsymbol{\beta}}(\boldsymbol{\xi}), \tag{C.2}$$

where Θ^d represents the collection of all subsets of $[1 \ : \ d]$ and $\eta_{\boldsymbol{\beta}}(\boldsymbol{\xi})$ satisfies $\mathbb{E}\left[\eta_{\boldsymbol{\alpha}}, \eta_{\boldsymbol{\beta}}\right] = 0$, if $\boldsymbol{\alpha} \subset \boldsymbol{\beta}$. The sensitivity index $S_{\boldsymbol{\beta}}$ is given by

$$S_{\boldsymbol{\beta}} = \frac{\mathbb{V}(\eta_{\boldsymbol{\beta}}) + \sum_{\boldsymbol{\alpha} \cap \boldsymbol{\beta} \neq \boldsymbol{\alpha}, \boldsymbol{\beta}} \mathrm{Cov}(\eta_{\boldsymbol{\alpha}}, \beta_{\boldsymbol{\beta}})}{\mathbb{V}(f)} \tag{C.3}$$

where $\mathbb{V}(\cdot)$ refers to the variance on ν_S. We refer to Chastaing et al. [98] for the details. Fig. C.4 shows the first-order sensitivity indices for the test cases of Gaussian mixture systems ($d = 25$, $p = 3$) and the



(a) (b)

Figure C.10: The first-order sobol sensitivity indices for (a) polynomial function with Gaussian mixture measure ($d = 25$, $p = 3$) (b) molecular system for solvation energy ("—■—") and SASA of atom H9("-·-◆-·-").

molecular systems, where the surrogate models are constructed by the data-driven basis approach using $M = 800$, $M = 240$ and $M = 600$ training points, respectively. The dominant components are on the dimensions $(1, 2, 3, 6, 11, 13, 14, 15, 16, 20, 22, 24, 25)$, $(1, 2, 5)$ and $(1, 2, 4, 5, 7)$ (90% of total variance).

Appendix D. Generation of the Gaussian mixture data set

We used Matlab to generate the Gaussian mixture data set in Sec. 4.1 by calling the function `gmdistribution(`$\boldsymbol{\mu}$, $\{\Sigma_i\}_{i=1}^3$, $\boldsymbol{a}$`)` with $\boldsymbol{a} = (0.5358, 0.1281, 0.3361)$. $\boldsymbol{\mu}$ is a 25×3 random matrix with i.i.d. entries on $U[-2.5, 2.5]$. $\{\Sigma_i\}_{i=1}^3$ is a $25 \times 25 \times 3$ array where Σ_i is defined by

$$\boldsymbol{\Sigma}_i = (\boldsymbol{\Upsilon}_i \boldsymbol{\Upsilon}_i{}^T + \mathbf{I})/4, \tag{D.1}$$

where $\boldsymbol{\Upsilon}_i$ is a random matrix with i.i.d. entries from $\mathcal{U}[0, 1]$ for $i = 1, 2, 3$. $\boldsymbol{\mu}$ and Υ_i are generated by calling the Matlab function `rand()` with random number seed 200.

Appendix E. Molecular Dynamics simulation and calculation details

We performed all-atom MD simulation of benzyl bromide in water using GROMACS 5.1.2 [99]. The simulation system included a benzyl bromide molecule (see Figure E.11 for the molecular structure) and 1011 water molecules. The General AMBER Force Field (GAFF) [100] was used for the benzyl bromide parmameters. The partial charges of benzyl bromide molecule were calculated by RESP method [101]. Bond lengths of benzyl bromide were constrained using the LINCS algorithm [102]. The water molecule was modeled with the rigid TIP3P water model [103]. The bond lengths and angles were held constant through the SETTLE algorithm [104]. The system was equilibrated in the isothermal-isobaric ensemble for 10 ns at 300K and 1 bar after energy minimization. The van der Waals cut-off radii was 1.0 nm. Long-range

electrostatics were calculated using a Particle Mesh Ewald (PME) summation with grid spacing of 0.12 nm. The time step was 2 fs. Isobaric-isothermal simulations were equilibrated using a V-rescale thermostat and Berendsen barostat. Following equilibration, the simulation was run for a production period of 20 μs in a NVT ensemble with a Nosé-Hoover thermostat. The trajectory was stored every 10000 time steps.

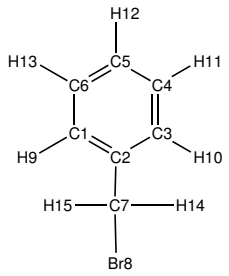


Figure E.11: Sketch of the molecule benzyl bromide with labeled atoms.

APBS calculations [105, 92] were performed with 129^3 grid points over a $40 \times 40 \times 40$ Å^3 coarse grid domain with focusing to a $14 \times 14 \times 14$ Å^3 fine grid domain with the grid origin located at the geometric center of the molecule. The Poisson equation was solved with Dirichlet boundary conditions based on the asymptotic behavior of multiple point charges in a homogeneous dielectric medium. The dielectric coefficient inside the domain used a van der Waals molecular volume definition with a dielectric value of 2.0 inside the molecule and 78.0 outside the molecule. Charges were modeled by Dirac delta functions but discretized to the finite difference grid points using a cubic spline approximation.

Acknowledgements

We thank Ling Guo (Shanghai Normal University), Lei Wu (Princeton University), Wen Zhou (Colorado State University), and David Sept (University of Michigan, ORCID:0000-0003-3719-2483) for helpful discussions. This work was supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research as part of the Collaboratory on Mathematics for Mesoscopic Modeling of Materials (CM4) and by the National Institutes of Health grant R01 GM069702. The research was performed using resources available through Research Computing at Pacific Northwest National Laboratory HL acknowledges grant support from AMS Simons Post-doctoral Travel Grant and PNNL Laboratory Directed Research & Development (LDRD) under project "Development of physics-compatible stochastic models for multiphysics systems".

References

[1] H. Lei, X. Yang, B. Zheng, G. Lin, N. A. Baker, Constructing surrogate models of complex systems with enhanced sparsity: Quantifying the influence of conformational uncertainty in biomolecular solvation,

SIAM Multiscale Model. Simul. 13 (4) (2015) 1327–1353.

[2] S. Oladyshkin, W. Nowak, Data-driven uncertainty quantification using the arbitrary polynomial chaos expansion, Reliability Engineering & System Safety 106 (2012) 179 – 190.

[3] A. Saltelli, Global sensitivity analysis: the primer, John Wiley, 2008.

[4] C. M. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics), Springer-Verlag, Berlin, Heidelberg, 2006.

[5] A. Laio, M. Parrinello, Escaping free-energy minima, Proceedings of the National Academy of Sciences 99 (20) (2002) 12562–12566.

[6] G. Fishman, Monte Carlo: Concepts, Algorithms, and Applications, Springer-Verlag New York, Inc., 1996.

[7] S. Kucherenko, D. Albrecht, A. Saltelli, Exploring multi-dimensional spaces: a Comparison of Latin Hypercube and Quasi Monte Carlo Sampling Techniques (2015). arXiv:arXiv:1505.02350.

[8] M. B. Giles, Multilevel Monte Carlo methods, Acta Numerica 24 (2015) 259 – 328.

[9] S. Heinrich, Multilevel Monte Carlo Methods, in: S. Margenov, J. Waśniewski, P. Yalamov (Eds.), Large-Scale Scientific Computing, Springer Berlin Heidelberg, Berlin, Heidelberg, 2001, pp. 58–67.

[10] M. Pisaroni, F. Nobile, P. Leyland, A Continuation Multi Level Monte Carlo (C-MLMC) method for uncertainty quantification in compressible inviscid aerodynamics, Computer Methods in Applied Mechanics and Engineering 326 (2017) 20 – 50.

[11] P. Koutsourelakis, Accurate Uncertainty Quantification Using Inaccurate Computational Models, SIAM Journal on Scientific Computing 31 (5) (2009) 3274–3300.

[12] B. Peherstorfer, K. Willcox, M. Gunzburger, Optimal Model Management for Multifidelity Monte Carlo Estimation, SIAM Journal on Scientific Computing 38 (5) (2016) A3163–A3194.

[13] B. Fox, Strategies for Quasi-Monte Carlo, Kluwer Academic Pub., 1999.

[14] H. Niederreiter, Random number generation and Quasi-Monte Carlo methods, SIAM, 1992.

[15] H. Niederreiter, P. Hellekalek, G. Larcher, P. Zinterhof, Monte Carlo and Quasi-Monte Carlo methods 1996, Springer-Verlag, 1998.

[16] M. D. McKay, R. J. Beckman, W. J. Conover, Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code, Technometrics 21 (2) (1979) 239–245.

[17] M. Stein, Large sample properties of simulations using Latin Hypercube Sampling, Technometrics 29 (2) (1987) 143–151.

[18] W. Loh, On Latin hypercube sampling, Ann. Stat. 24 (5) (1996) 2058–2080.

[19] J. Sacks, W. J. Welch, T. J. Mitchell, H. P. Wynn, Design and Analysis of Computer Experiments, Statist. Sci. 4 (4) (1989) 409–423.

[20] M. C. Kennedy, A. O'Hagan, Bayesian calibration of computer models, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63 (3) (2001) 425–464.

[21] C. Rasmussen, C. Williams, Gaussian Processes for Machine Learning, Adaptive Computation and Machine Learning, MIT Press, Cambridge, MA, USA, 2006.

[22] N. Wiener, The homogeneous chaos, Amer. J. Math. 60 (1938) 897–936.

[23] R. Ghanem, P. Spanos, Stochastic Finite Elements: A Spectral Approach, Springer-Verlag, 1991.

[24] D. Xiu, G. E. Karniadakis, The wiener-askey polynomial chaos for stochastic differential equations, SIAM J. Sci. Comput. 24 (2002) 619–644.

[25] P. Z. G. Qian, C. F. J. Wu, Bayesian Hierarchical Modeling for Integrating Low-Accuracy and High-Accuracy Experiments, Technometrics 50 (2) (2008) 192–204.

[26] B. Williams, D. Higdon, J. Gattiker, L. Moore, M. McKay, S. Keller-McNulty, Combining experimental data and computer simulations, with an application to flyer plate experiments, Bayesian Anal. 1 (4) (2006) 765–792. doi:10.1214/06-BA125.

[27] J. Oakley, A. O'Hagan, Bayesian inference for the uncertainty distribution of computer model outputs, Biometrika 89 (4) (2002) 769–784.

[28] B. A. Lockwood, M. Anitescu, Gradient-Enhanced Universal Kriging for Uncertainty Propagation, Nuclear Science and Engineering 170 (2) (2012) 168–195. doi:10.13182/NSE10-86.

[29] D. Xiu, G. Karniadakis, Modeling uncertainty in flow simulations via generalized polynomial chaos, J. Comput. Phys. 187 (2003) 137–167.

[30] R. Ghanem, S. Masri, M. Pellissetti, R. Wolfe, Identification and prediction of stochastic dynamical systems in a polynomial chaos basis, Comput. Meth. Appl. Math. Engrg. 194 (2005) 1641–1654.

[31] O. Knio, O. Le Maitre, Uncertainty propagation in CFD using polynomial chaos decomposition, Fluid Dyn. Res. 38 (9) (2006) 616–640.

[32] B. Sudret, Global sensitivity analysis using polynomial chaos expansions, Reliability Engineering & System Safety 93 (7) (2008) 964 – 979.

[33] J. Li, D. Xiu, A generalized polynomial chaos based ensemble Kalman filter with high accuracy, J. Comput. Phys. 228 (2009) 5454–5469.

[34] Y. Marzouk, D. Xiu, A stochastic collocation approach to bayesian inference in inverse problems, Communications in Computational Physics 6 (4) (2009) 826–847. doi:10.4208/cicp.2009.v6.p826.

[35] J. Li, D. Xiu, Evaluation of failure probability via surrogate models, J. Comput. Phys. 229 (2010) 8966–8980.

[36] J. Li, P. Stinis, Mori-zwanzig reduced models for uncertainty quantification, arXiv:1803.02826.

[37] R. Schobi, B. Sudret, J. Wiart, Polynomial-chaos-based Kriging, International Journal for Uncertainty Quantification 5 (2) (2015) 171–193.

[38] L. L. Gratiet, S. Marelli, B. Sudret, Metamodel-Based Sensitivity Analysis: Polynomial Chaos Expansions and Gaussian Processes, Springer International Publishing, Cham, 2016, pp. 1–37.

[39] N. Owen, P. Challenor, P. Menon, S. Bennani, Comparison of Surrogate-Based Uncertainty Quantification Methods for Computationally Expensive Simulators, SIAM/ASA Journal on Uncertainty Quantification 5 (1) (2017) 403–435. doi:10.1137/15M1046812.

[40] P. T. Roy, N. El Moçayd, S. Ricci, J.-C. Jouhaud, N. Goutal, M. De Lozzo, M. C. Rochoux, Comparison of polynomial chaos and Gaussian process surrogates for uncertainty quantification and correlation estimation of spatially distributed open-channel steady flows, Stochastic Environmental Research and Risk Assessment 32 (6) (2018) 1723–1741.

[41] L. Mathelin, M. Hussaini, A stochastic collocation algorithm for uncertainty analysis, Tech. Rep. NASA/CR-2003-212153, NASA Langley Research Center (2003).

[42] D. Xiu, J. Hesthaven, High-order collocation methods for differential equations with random inputs, SIAM J. Sci. Comput. 27 (3) (2005) 1118–1139.

[43] I. Babuška, F. Nobile, R. Tempone, A stochastic collocation method for elliptic partial differential equations with random input data, SIAM J. Numer. Anal. 45 (3) (2007) 1005–1034.

[44] F. Nobile, R. Tempone, C. G. Webster, A sparse grid stochastic collocation method for partial differential equations with random input data, SIAM J. Numer. Anal. 46 (5) (2008) 2309–2345.

[45] X. Ma, N. Zabaras, An adaptive hierarchical sparse grid collocation algorithm for the solution of stochastic differential equations, J. Comput. Phys. 228 (8) (2009) 3084–3113.

[46] J. Foo, G. E. Karniadakis, Multi-element probabilistic collocation method in high dimensions, J. Comput. Phys. 229 (5) (2010) 1536 – 1557.

[47] P. G. Constantine, M. S. Eldred, E. T. Phipps, Sparse pseudospectral approximation method, Computer Methods in Applied Mechanics and Engineering 229-232 (2012) 1 – 12. doi:https://doi.org/10.1016/j.cma.2012.03.019.

[48] J. D. Jakeman, S. G. Roberts, Local and dimension adaptive stochastic collocation for uncertainty quantification, in: Sparse grids and applications, Springer, 2013, pp. 181–203.

[49] J. Li, P. Stinis, A unified framework for mesh refinement in random and physical space, Journal of Computational Physics 323 (2016) 243 – 264. doi:https://doi.org/10.1016/j.jcp.2016.07.027.

[50] A. Doostan, H. Owhadi, A non-adapted sparse approximation of pdes with stochastic inputs, J. Comput. Phys 230 (2011) 3015–3034.

[51] L. Yan, L. Guo, D. Xiu, Stochastic collocation algorithms using l^1 minimization, Inter. J. Uncertain Quantification 2 (2012) 279–293.

[52] H. Rauhut, R. Ward, Sparse legendre expansions via ℓ_1-minimization, J. Approx. Theory 164 (2012) 517–533.

[53] L. Mathelin, K. A. Gallivan, A compressed sensing approach for partial differential equations with random input data, Communications in Computational Physics 12 (4) (2012) 919?954. doi:10.4208/cicp.151110.090911a.

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

[55] J. Hampton, A. Doostan, Compressive sampling of polynomial chaos expansions: Convergence analysis and sampling strategies., J. Comput. Phys 280 (2015) 363–386.

[56] J. Peng, J. Hampton, A. Doostan, On polynomial chaos expansion via gradient-enhanced ℓ_1-minimization, J. Comput. Phys 310 (2016) 440–458.

[57] L. Yan, Y. Shin, D. Xiu, Sparse approximation using $\ell_1 - \ell_2$ minimization and its application to stochastic collocation, SIAM Journal on Scientific Computing 39 (1) (2017) A229–A254. doi:10.1137/15M103947X.

[58] Y. L. Liu, L. Guo, Stochastic collocation via l1-minimisation on low discrepancy point sets with application to uncertainty quantification, EAJAM 6 (2016) 171–191.

[59] H. Lei, X. Yang, Z. Li, G. E. Karniadakis, Systematic parameter inference in stochastic mesoscopic modeling, J. Comput. Phys. 330 (4) (2017) 571–593.

[60] N. Alemazkoor, H. Meidani, Divide and conquer: An incremental sparsity promoting compressive sampling approach for polynomial chaos expansions, Computer Methods in Applied Mechanics and Engineering 318 (2017) 937 – 956. doi:https://doi.org/10.1016/j.cma.2017.01.039.

[61] P. Diaz, A. Doostan, J. Hampton, Sparse polynomial chaos expansions via compressed sensing and d-optimal design, Computer Methods in Applied Mechanics and Engineering 336 (2018) 640 – 666. doi:https://doi.org/10.1016/j.cma.2018.03.020.

[62] P. Rai, K. Sargsyan, H. Najm, Compressed sparse tensor based quadrature for vibrational quantum mechanics integrals, Computer Methods in Applied Mechanics and Engineering 336 (2018) 471 – 484. doi:https://doi.org/10.1016/j.cma.2018.02.026.

[63] K. Huang, Lectures on Statistical Physics and Protein Folding, WORLD SCIENTIFIC, 2005. doi:10.1142/5741.

[64] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, P. A. Kollman, The weighted histogram analysis method for free-energy calculations on biomolecules. i. the method, Journal of Computational Chemistry 13 (8) (1992) 1011–1021.

[65] L. Maragliano, E. Vanden-Eijnden, Single-sweep methods for free energy calculations, The Journal of Chemical Physics 128 (18) (2008) 184110. doi:10.1063/1.2907241.

[66] P. G. Constantine, E. Dow, Q. Wang, Active subspace methods in theory and practice: Applications to kriging surfaces, SIAM J. Sci. Comput. 36 (4) (2014) A1500–A1524.

[67] W. Li, G. Lin, An adaptive importance sampling algorithm for bayesian inversion with multimodal distributions, Journal of Computational Physics 294 (2015) 173 – 190.

[68] V. Vittaldev, R. P. Russell, R. Linares, Spacecraft uncertainty propagation using gaussian mixture models and polynomial chaos expansions, Journal of Guidance, Control, and Dynamics 39 (12) (2016) 2615–2626. doi:10.2514/1.g001571.

[69] J. Feinberg, H. P. Langtangen, Chaospy: An open source tool for designing methods of uncertainty quantification, Journal of Computational Science 11 (2015) 46 – 57. doi:https://doi.org/10.1016/j.jocs.2015.08.008.

[70] J. Wan, N. Zabaras, A probabilistic graphical model based stochastic input model construction, Journal of Computational Physics 272 (2014) 664 – 685. doi:https://doi.org/10.1016/j.jcp.2014.05.002.

[71] X. Wan, G. Karniadakis, Multi-element generalized polynomial chaos for arbitrary probability measures, SIAM J. Sci. Comput. 28 (2006) 901–928.

[72] J. A. S. Witteveen, H. Bijl, Modeling arbitrary uncertainties using gram-schmidt polynomial chaos, in: 44th AIAA Aerospace Sciences Meeting and Exhibit, American Institute of Aeronautics and Astronautics, 2006, pp. 1706–1713. doi:doi:10.2514/6.2006-896.

[73] M. Zheng, X. Wan, G. E. Karniadakis, Adaptive multi-element polynomial chaos with discrete measure: Algorithms and application to spdes, Applied Numerical Mathematics 90 (2015) 91–110.

[74] S. Yin, D. Yu, Z. Luo, B. Xia, An arbitrary polynomial chaos expansion approach for response analysis of acoustic systems with epistemic uncertainty, Computer Methods in Applied Mechanics and Engineering 332 (2018) 280 – 302. doi:https://doi.org/10.1016/j.cma.2017.12.025.

[75] R. Ahlfeld, B. Belkouchi, F. Montomoli, Samba: Sparse approximation of moment-based arbitrary polynomial chaos, Journal of Computational Physics 320 (2016) 1 – 16. doi:https://doi.org/10.1016/j.jcp.2016.05.014.

[76] C. F. Dunkl, Y. Xu, Orthogonal Polynomials of Several Variables, 2nd Edition, Encyclopedia of Mathematics and its Applications, Cambridge University Press, 2014. doi:10.1017/CBO9781107786134.

[77] E. Candès, M. Rudelson, T. Tao, R. Vershynin, Error correction via linear programming, in: 46th Annual IEEE Symposium on Foundations of Computer Science (FOCS'05), 2005, pp. 668–681.

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

[79] M. E. Davies, R. Gribonval, Restricted isometry constants where ℓ_p sparse recovery can fail for $0 < p \leq 1$, IEEE Trans. Inf. Theory 55 (2010) 2203–2214.

[80] D. Donoho, M. Elad, V. Temlyakov, Stable recovery of sparse overcomplete representations in the presence of noise, Information Theory, IEEE Transactions on 52 (1) (2006) 6–18.

[81] E. V. D. Berg, M. Friedlander, Spgl1: A solver for large-scale sparse reconstruction, http://www.cs.ubc.ca/labs/scl/spgl.

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

[83] E. J. Candès, J. Romberg, T. Tao, Stable signal recovery from incomplete and inaccurate measurements, Comm. Pure Appl. Math. 56 (2006) 1207–1223.

[84] B. K. Natarajan, Sparse approximate solutions to linear systems, SIAM J. Sci. Comput. 2 (1995) 227–234.

[85] T. M. Russi, Uncertainty Quantification with Experimental Data and Complex System Models, Ph.D. thesis, University of California, Berkeley (2001).

[86] X. Yang, H. Lei, N. A. Baker, G. Lin, Enhancing sparsity of hermite polynomial expansions by iterative rotations, J. Comput. Phys. 307 (2016) 94 – 109.

[87] M. Jardak, C.-H. Su, G. E. Karniadakis, Spectral polynomial chaos solutions of the stochastic advection equation, J. Sci. Comput. 17 (1-4) (2002) 319–338.

[88] X. Yang, X. Wan, L. Lin, A general framework of enhancing sparsity of generalized polynomial chaos expansions (2017). arXiv:arXiv:1707.02688.

[89] A. R. Atilgan, S. R. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin, I. Bahar, Anisotropy of fluctuation dynamics of proteins with an elastic network model, Biophysical Journal 80 (1) (2001) 505–515.

[90] P. Ren, J. Chun, D. G. Thomas, M. J. Schnieders, M. Marucho, J. Zhang, N. A. Baker, Biomolecular electrostatics and solvation: a computational perspective, Quarterly reviews of biophysics 45 (4) (2012) 427–491. doi:10.1017/S003358351200011X.

[91] N. A. Baker, Biomolecular Applications of Poisson?Boltzmann Methods, in: K. B. Lipkowitz, R. Larter, T. R. Cundari (Eds.), Reviews in Computational Chemistry, John Wiley & Sons, Inc., 2005, pp. 349–379. doi:10.1002/0471720895.ch5.

[92] E. Jurrus, D. Engel, K. Star, K. Monson, J. Brandi, L. E. Felberg, D. H. Brookes, L. Wilson, J. Chen, K. Liles, M. Chun, P. Li, D. W. Gohara, T. Dolinsky, R. Konecny, D. R. Koes, J. E. Nielsen, T. Head-Gordon, W. Geng, R. Krasny, G.-W. Wei, M. J. Holst, J. A. McCammon, N. A. Baker, Improvements to the APBS biomolecular solvation software suite, Protein Science 27 (1) (2018) 112–128. doi:10.1002/pro.3280.

[93] A. Shrake, J. Rupley, Environment and exposure to solvent of protein atoms. lysozyme and insulin, Journal of Molecular Biology 79 (2) (1973) 351 – 371.

[94] M. Rasheed, N. Clement, A. Bhowmick, C. Bajaj, Statistical Framework for Uncertainty Quantification in Computational Molecular Modeling, in: Proceedings of the 7th ACM International Conference on

Bioinformatics, Computational Biology, and Health Informatics, ACM, New York, NY, USA, 2016, pp. 146–155.

[95] N. Clement, M. Rasheed, C. L. Bajaj, Viral Capsid Assembly: A Quantified Uncertainty Approach, Journal of Computational Biology 25 (1) (2018) 51–71. doi:10.1089/cmb.2017.0218.

[96] H. Rauhut, Compressive sensing and structured random matrices, Radon Series Comp. Appl. Math. 9 (2010) 1–92.

[97] A. Marrel, B. Iooss, B. Laurent, O. Roustant, Calculations of Sobol indices for the Gaussian process metamodel, Reliability Engineering & System Safety 94 (3) (2009) 742 – 751.

[98] G. Chastaing, F. Gamboa, C. Prieur, Generalized Sobol sensitivity indices for dependent variables: numerical methods, Journal of Statistical Computation and Simulation 85 (7) (2015) 1306–1333. doi:10.1080/00949655.2014.960415.

[99] H. J. C. Berendsen, D. van der Spoel, R. van Drunen, GROMACS: A message-passing parallel molecular dynamics implementation, Computer Physics Communications 91 (1) (1995) 43–56. doi:10.1016/0010-4655(95)00042-E.

[100] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, D. A. Case, Development and testing of a general amber force field, Journal of Computational Chemistry 25 (9) (2004) 1157–1174.

[101] C. I. Bayly, P. Cieplak, W. Cornell, P. A. Kollman, A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the resp model, The Journal of Physical Chemistry 97 (40) (1993) 10269–10280.

[102] B. Hess, H. Bekker, J. C. Berendsen Herman, G. E. M. Fraaije Johannes, Lincs: A linear constraint solver for molecular simulations, Journal of Computational Chemistry 18 (12) (1998) 1463–1472.

[103] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein, Comparison of simple potential functions for simulating liquid water, The Journal of Chemical Physics 79 (2) (1983) 926–935.

[104] S. Miyamoto, A. Kollman Peter, Settle: An analytical version of the shake and rattle algorithm for rigid water models, Journal of Computational Chemistry 13 (8) (2004) 952–962.

[105] N. A. Baker, D. Sept, S. Joseph, M. J. Holst, J. A. McCammon, Electrostatics of nanosystems: application to microtubules and the ribosome, Proceedings of the National Academy of Sciences 98 (18) (2001) 10037–10041.

LaTeX Source Files

Click here to download LaTeX Source Files: revision.zip