Bypassing Backmapping: Coarse-Grained Electronic Property Distributions Using Heteroscedastic Gaussian Processes

J. Charlie Maier[1] and Nicholas E. Jackson[2]

[1)]Department of Physics, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA.

[2)]Department of Chemistry, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA.

(*Electronic mail: jacksonn@illinois.edu)

We employ deep kernel learning electronic coarse-graining (DKL-ECG) with approximate Gaussian Processes as a flexible and scalable framework for learning heteroscedastic electronic property distributions as a smooth function of coarse-grained (CG) configuration. The appropriateness of the Gaussian prior on the predictive CG property distributions is justified as a function of CG model resolution by examining the statistics of the target distributions. The certainties of the predictive CG distributions are shown to be limited by CG model resolution, with DKL-ECG predictive noise converging to the intrinsic physical noise induced by the CG mapping operator for multiple chemistries. Further analysis of the resolution dependence of the learned CG property distributions allows for the identification of CG mapping operators that capture CG degrees of freedom with strong electron-phonon coupling. We further demonstrate the ability to construct the exact quantum chemical valence electronic density of states (EDOS), including behavior in the tails of the EDOS, from an entirely CG model by combining iterative boltzmann inversion and DKL-ECG. DKL-ECG provides a means of learning CG distributions of all-atom properties that are traditionally “lost" in CG model development, introducing a promising methodological alternative to backmapping algorithms commonly employed to recover all-atom property distributions from CG simulations.

I. INTRODUCTION

Coarse-grained (CG) modeling[1–7] is an essential sub-field of molecular simulation aimed at understanding the mesoscopic spatiotemporal scales dictating function in soft materials. By averaging collections of all-atom (AA) degrees of freedom into CG pseudoatoms, the number of configurational degrees of freedom is reduced and computational efficiency is improved. These lower-dimensional CG models exhibit smoother potential energy surfaces than their AA counterparts, accelerating the sampling of long spatiotemporal scales. The effective potentials governing the thermodynamics of CG models can be inferred by sampling trajectories of AA systems using rigorous statistical mechanical techniques,[8–14] with recent approaches benefiting from advances in machine learning (ML).[15–19] Furthermore, removing “irrelevant" degrees of freedom often leads to more interpretable models that directly highlight the underlying structure-function relationships, rendering CG modeling both a computationally and conceptually efficient strategy.

Coincident with the advantages of CG models are the limitations imposed by marginalizing over subsets of AA degrees of freedom. CG models are inherently limited in their representation of AA systems as a result of the marginzalization procedure; to retrieve this lost information, the decimated AA degrees of freedom must be reimposed onto the CG representation, followed by resampling at the AA resolution. This challenge of retrieving lost information from CG models has produced a variety of algorithmic solutions referred to collectively as “backmapping",[20–29] that are designed to recover AA detail from thermodynamically sampled CG configurations. One implication of the information loss in CG models is that backmapping often requires assumptions that inhibit transferability. The non-invertibility of the CG map (i.e. one to many) further necessitates the inclusion of multi-step cycles of energy minimization and configurational sampling[30] to produce thermodynamically consistent AA configurations. These complications negate many of the ostensible computational cost reductions of CG modeling, as one must re-employ the AA model to obtain the “lost" AA properties. Recovery of this lost information is critical to multiscale simulations and represents an important fundamental research topic in CG modeling.

Coupling electronic structure analysis to CG modeling represents an illustrative case of the computational burdens of information transfer between CG and AA resolutions. In soft materials simulations, one often computes the electronic properties (e.g. electronic conductivity, optical absorption, reactivity, dielectric breakdown) as a function of thermodynamic and morphological parameters. The heterogeneity of soft materials requires simulations spanning mesoscopic length (10 nm - 1 mm) and time (1 ms) scales, but the electronic degrees of freedom necessitate quantum chemical (QC) methods. Consequently, AA configurations must be backmapped onto simulated CG configurations, followed by QC evaluations of the local electronic structure (note that due to the intrinsic structural heterogeneity of soft materials Bloch’s theorem cannot be employed). These calculations must then be repeated, ad nauseum, for every thermodynamic and morphological variation of interest, with requisite statistical sampling. Utilizing modern density functional theory (DFT) methods, this effort necessitates high-performance computing resources to predict the electronic properties of even a small portion of the morphological or thermodynamic design space. To tackle these categories of in silico design tasks that underscore modern soft materials challenges (e.g. chemical degradation, sustainability, flexible electronics), a computational paradigm shift is required.

The utilization of ML to augment CG molecular modeling is increasingly common. Using frameworks such as deep neural networks (DNNs) and Gaussian Process regression (GPR), multiple research groups have parameterized CG force fields,[15,18,31,32] that accurately reproduce the thermodynamics of underlying AA systems. Methods that incorporate physical symmetries such as rotation and permutation invariance directly into the structure of the ML framework have been particularly effective in reducing cost and improving accuracy.[16,18] ML has found interesting uses in solving traditionally subjective tasks essential to CG modeling: generating "good" CG mapping operators[24,33] or collective variables[34] and performing AA backmapping.[24–29] A particularly interesting thrust of research has employed ML to learn the underlying thermodynamic distribution of configurations, allowing for efficient sampling of a complex configurational space after training.[19,35,36] In aggregate, these advances motivate the potential efficacy of augmenting CG simulations using ML to improve information transfer between CG and AA resolutions in multiscale simulations.

Electronic Coarse-Graining (ECG)[37–39] has been introduced as a method for performing QC predictions using only a molecule's CG representation. In ECG, the conditional distribution of AA configurations mapped to a single CG configuration results in a thermodynamic distribution of electronic properties consistent with that CG configuration. By treating the AA property distribution as noisy observations of an underlying "true" function, GPR provides a framework for learning the CG-projected mean AA electronic structure, as well as a Gaussian fit on the conditional AA electronic property distribution.[39] This approach resembles CG methods for structural predictions[8] in which the mean projected AA force is used to construct a pairwise CG free energy function that reproduces the equilibrium distribution of the underlying AA model. Of critical note is that there is no a priori guaran-tee that the width of the AA distribution associated with a single CG configuration should be independent of CG configuration (i.e. homoscedastic). Consequently, GPR methods using homoscedastic noise kernels learn a single AA Gaussian distribution width averaged over all CG configurations.[39] To tackle the complexities of predicting AA distributions as a function of CG configuration, a more flexible heteroscedastic prediction of the AA distribution function is required. Such a method would represent a generalizable path forward for learning arbitrary AA property distributions directly from the CG representation, without recourse to backmapping.

Our previous work on ECG combined the representational power of DNNs with exact GPR within a Deep Kernel Learning (DKL) framework.[39] Exact GPR[40] assumes that all observations of the target variable are uniformly noisy, whereas a CG map may introduce a CG configuration-dependent level of noise. Additionally, due to inversion of the kernel matrix over the training set, exact GPR training scales as $O(N^3)$ with the number of training points. These two limitations can be addressed via the introduction of inducing point methods,[41–43] in which inference is performed through a smaller set of inducing variables, the size of which is fixed independently of the training set. This effort reduces training to O(N) and prediction to a O(1) in the number of training points. The use of inducing points can be further leveraged by treating them as a variational hyperparameter that improve performance[44] and provide a richer predictive noise that incorporates heteroscedasticity.

In this work, we extend ECG using a stochastic DKL framework with approximate GPR to predict DFT-calculated electronic properties at CG resolutions. This approach provides (i) improved scaling of ECG model training and (ii) flexibility to handle heteroscedastic data sets. First, we describe the method of CG mapping, data set generation. and the stochastic DKL framework developed in this work (termed DKL-ECG). We then estimate the ground truth CG distributions, providing quantitative validations of DKL-ECG predictions for both the predictive mean and noise. DKL-ECG is then applied to three canonical chemistries of electron transporting polymers, bithiophene (BT), TEMPO, and sexi(3-methyl)thiophene (S3MT), at a variety of CG resolutions to assess model performance. Finally, DKL-ECG is combined with Iterative Boltzmann Inversion (IBI) to simulate the full electronic density of states (EDOS), including accurate estimates in the tails of the EDOS, at a purely CG resolution without the use of backmapping. We then discuss the implications of this work and summarize our conclusions.

II. METHODS

A. Coarse-Grained Mapping

This paper studies the prediction of AA-derived QC properties using only a lower-dimensional CG representation of the molecule's configuration. While the methods outlined can be applied to any AA-derived target variable that follows a sufficiently well-behaved function of AA configuration, we use the highest occupied molecule orbital (HOMO) energy of molecules as the target AA property. We focus on three canonical electron transporting chemistries: BT (hole-transporting), TEMPO (radical-transporting), and S3MT (hole-transporting). Throughout this work a variety of terms are used in developing DKL-ECG; Table I provides a reference of the common notation used in this work.

A CG mapping is a linear projection from coordinates, r, for the set of all N atoms within a molecule, to the CG coordinates, R, for a predetermined set of M beads. We assume that each CG bead is located at the center of mass of the atoms belonging to the bead, and that each atom belongs to precisely one CG bead. This projection is non-invertible, and there exist infinitely many unique AA configurations that map to the same CG configuration. The exact nature of this distribution, p(rjR), is defined by molecular features (e.g. topology, nature of interatomic bonds) as well as environmental factors (e.g. temperature, pressure). Due to these features, predicting a single scalar value of chemical property (e.g. a molecular orbital energy) is not sufficient to characterize the distributional nature of the AA property at the CG resolution.

For each molecule in this work, the target AA distribution, p(r), is the equilibrium distribution for a single molecule in

Symbol	Description
$p(E\|R)$	Conditional distribution of HOMO energies, E, belonging to the same CG configuration, R, sampled over the AA equilibrium distribution. The standard deviation of this distribution is the CG noise.
$p(r\|R)$	Conditional distribution of AA configurations, r, that are mapped to the same CG configuration, sampled over the AA equilibrium distribution.
$p(E\|R)$	DKL-ECG predictive distribution approximating $p(E\|R)$. The standard deviation of this distribution is the predictive noise.
$p(E\|z)$	DKL-ECG predictive distribution at a given point in the DKL latent space, z. The standard deviation of this distribution is the DKL noise.
$Var(E\|R)$	The variance of the predictive distribution, $p(E\|R)$.
$p(E\|R;U+V)$	The estimate of the CG energy distribution at R sampled over the AA equilibrium distribution with an additional biasing potential, V.
$p(E\|R;U)$	The estimate of the CG energy distribution at R after unbiasing the additional applied potential.
$Var^C(E\|R)$	The estimated variance of the energy, for a cluster of points localized around R.
$Var^C(E\|R)$	The predictive variance over a cluster of points localized around R. This estimate takes into account the fluctuation of $p(\langle E\rangle\|R)$ over the cluster.
$p(E)$	The normalized sum of predictive distributions $p(E\|R)$ over a trajectory of CG configurations.

TABLE I: Table of terms and definitions used throughout this work.

vacuum. To generate a training set of molecular configurations for each studied molecule, AA MD simulations of a single molecule are performed in LAMMPS[45] using the optimized potentials for liquid simulation (OPLS)[46] force field (See SM for additional details). MD sampled configurations are then piped to a wB97X-D3/def2-SVP calculation using Orca[47] to generate all electronic molecular orbital energy levels, and HOMO energy was extracted. Data generation for S3MT is detailed in previous work,[39] and all S3MT data here are fully-flexible (no intramolecular constraints are applied).

For a given molecule and CG resolution, a choice of linear CG mapping operator was made based on symmetries of the molecule and preservation of the degrees of freedom suspected to be strongly correlated with the target variable. While the choice of CG mapping operator affects the level of CG noise in the target variable, this should not be interpreted as affecting the ultimate performance of the DKL-ECG regression model. In contrast to a single-valued regression model, here DKL-ECG is developed to learn an accurate estimate of the conditional distribution $p(E|R)$, where E represents the HOMO energy of the molecule, at a wide range of relative noise levels within a Gaussian approximation; reducing the noise present at any single CG resolution is a separate question related to the details of CG mapping operator selection. Details of the CG mapping operators are provided in Supplementary Materials Figure S1 and Table S1.

The CG resolution dependence of $p(E|R)$ can be intuited for certain limits of CG resolution. In the limit of M=N, the CG property distribution converges to a delta function consistent with the original one-to-one mapping between AA coordinates and the associated AA electronic structure (not accounting for electronic degeneracy). In the opposite limit (M = 1), the CG model has no internal degrees of freedom and the CG property distribution function is simply the full thermodynamic distribution function of the AA property (here, the thermodynamically averaged EDOS), the structure of which can assume any non-gaussian functional form (e.g. multimodal, long-tailed). In regimes where $N > M \gg 1$, single-valued regression models predicting the expectation value of the target distribution are likely reasonable approximations for the narrow distributions,[37,38] with recent work introducing homoscedastic GPR as a means to learn the noisy AA property distribution at intermediate CG resolutions.[39] Here, we extend this development to predict the variance of $p(E|R)$, and approximate the full conditional target variable distribution within a Gaussian approximation.

$$p(E|R) \quad N \qquad Exp(E)_{p(r|R)}; Var(E)_{p(r|R)} \tag{1}$$

We refer to the standard deviation of $p(E|R)$ as the CG noise. To train ML approximations to $p(E|R)$, CG configurations must be featurized into machine readable formats. Here this is accomplished using the flattened upper triangle of the CG distance matrix. Feature vectors are standardized such that each dimension of the training data has zero mean and unit variance.

With the CG mapping operator defined, a CG potential energy function can also be parameterized to facilitate sampling of CG configuration space. We use IBI[10,48] to learn a set of bond, angle, and dihedral potentials that reproduce the CG-mapped structural distribution functions. Following IBI convergence, the use of the CG potentials allows for CG configurational sampling, without further reference to the underlying AA model, at significantly reduced computational cost. Specifically, in conjunction with DKL-ECG, this allows for an entirely CG prediction of the structural and electronic properties of each chemical system without future reference to the original AA models. Details of the IBI procedure, including its convergence, are provided in the SM Section VIII.

B. Gaussian Process Regression

Exact GPR is a kernel method for estimating the value of a target function, f, when the given data consists of samples

of the function observed with a constant, or homoscedastic, noise (ϵ).[40]

$$y_i = f(x_i) + \epsilon_i; \epsilon_i \sim N(0; \sigma^2) \quad (2)$$

Provided a set of N_{tr} observations $\{(x_i; y_i) | 1 \leq i \leq N_{tr}\}$, a Gaussian Process (GP) is the further prior assumption on the auto-correlation of f,

$$p(f|x) = N(0; K_{xx}) \quad (3)$$

$$K(x; x') = \exp\left(-\frac{1}{2}(x - x')^T \Theta^2 (x - x')\right) \quad (4)$$

where K_{xx} is the matrix formed by evaluating a chosen kernel function over the given points. Here, we use a radial basis function (RBF) kernel, where Θ is a learned hyperparameter of the model that defines a characteristic length scale in each feature dimension. The bold $\mathbf{f}$ indicates a vector of observations of the single-valued function f, for a vector of points $\mathbf{x}$. Conditioning on the N_{tr} training points, a prediction for a set of test points, $\mathbf{x}$, can be exactly calculated.

$$p(\mathbf{y}|\mathbf{x}) = N(K_f (K_{ff} + \sigma^2 I)^{-1} \mathbf{y}; \sigma^2 + K - K_f (K_{ff} + \sigma^2 I)^{-1} K_f) \quad (5)$$

The hyperparameters of the kernel and σ, a scalar characterizing the homoscedastic noise, can be learned to minimize the log likelihood, evaluated over the training set. The previous result requires inverting the kernel matrix over the entire training set, which scales like $O(N_{tr}^3)$, and limits the use of exact GPR on large data sets.

Approximate GPR methods[43,44] provide a means of bypassing the limitations of exact GPR by using a set of variational hyperparameters called the inducing points, $\mathbf{u}$, to learn an approximation of the distribution that the training set is sampling. Each point in the set is parameterized by a tunable location in the training domain, $\mathbf{x}_u$, and the inducing point distribution

$$p(\mathbf{u}) = N(\mathbf{m}; S) \quad (6)$$

in which the vector $\mathbf{m}$ and matrix S are further weights of GPR. Inference on test points can be performed by integration over the N_{ind}-dimensional inducing point distribution, rather than the GP prior on the training points. The loss function used for approximate GPR is the predictive log likelihood for its performance on heteroscedastic data sets.[44]

$$p(\mathbf{y}|\mathbf{y}) = N(K_u K_{uu}^{-1} \mathbf{m}; K - K_u K_{uu}^{-1} K_u + K_u K_{uu}^{-1} S K_{uu}^{-1} K_u + \sigma^2) \quad (7)$$

We refer to the standard deviation of $p(\mathbf{y}|\mathbf{y})$ as the predictive noise. The scaling of approximate GPR training is $O(N_{tr} N_{ind}^2 + N_{ind}^3)$, with N_{ind} being set at model creation, so computations are not limited by N_{tr}. This independence from the training set size allows for the use of much larger training data sets for more complex systems. It has been shown that the necessary number of inducing points scales like $O(\log^{N_{lat}}(N_{tr}))$, for normally distributed N_{lat}-dimensional data.[49] For this work, we fixed the number of inducing points in DKL-ECG at 1000. This choice did not appear to limit the expressibility of the approximate GPR; determining the effect of reduced inducing points, in the case of a smaller computational budget, would require further analysis.

Another feature of the approximate GPR model is a useful decomposition of the predictive variance. For data sets with highly-varying local noise, the homoscedastic term, σ^2, is not the dominant contributor to the overall variance. The first two terms of the approximate GPR predictive variance, $K - K_u K_{uu}^{-1} K_u$, are a contribution from the uncertainty in the fit of the latent mean function. In low-data regions, this term reduces to the original GPR prior kernel of Equation 4. In high-data regions, this term reduces to 0, as the predictive mean is able to converge with minimal uncertainty. The third term in the predictive variance, $K_u K_{uu}^{-1} S K_{uu}^{-1} K_u$, expresses the variance associated with that point in the inducing approximation of the training distribution. In a high-data region, this term plus σ^2 can be roughly associated with the "physical" noise present due to CG information loss. Further discussion of the approximate GPR technique and an illustrative example of the varying contributions to a heteroscedastic predictive variance are presented in Appendix A.

C. Deep Kernel Learning

The distance matrix featurization of the CG coordinate vector, R_i, contains many highly correlated dimensions, which interferes with the practical ability of the kernel hyperparameters to converge. We employ a modification of DKL[50] to learn an effective feature representation for the GP kernel. In DKL, featurized data is first input to a feedforward DNN, g, parameterized by a set of weights, w_0, and encoded to a point, $\mathbf{z}_i$, in a lower-dimensional latent space.

$$g(R; w_0) : R^{M(M-1)/2} \to R^{N_f} \quad (8)$$

$$\mathbf{z}_i = g(R_i; w_0) \quad (9)$$

The length of this feature vector, N_f, is a hyperparameter set before training. We add a further variational layer which encodes the initial latent space projection to a normal distribution of the same dimension, similar to the encoding layer of a variational autoencoder.[51] By initializing the variational layer with a high encoding variance, we unbias the initial DKL encoding and make more potential projections more accessible to DKL-ECG. This layer is generated by two feedforward DNN, e_1 and e_2, with sets of parameters w_1 and w_2 that define the mean and logarithm of the variance of the encoded distribution. A stochastic sample, z_i, from this encoded distribution is then used as input to the GPR.

$$e_1(\tilde{z}; w_1) : R^{N_f} \to R^{N_f} \quad (10)$$

$$e_2(\tilde{z}; w_2) : R^{N_f} \to R \quad (11)$$

$$p(z|R_i; w_1; w_2) = N(e_1(\tilde{z}_i); \exp(e_2(\tilde{z}_i))I) \quad (12)$$

The GPR kernel is now a function of the lower-dimensional latent coordinates, rather than the original CG coordinates, which simplifies the task of fitting an accurate predictive function.

$$K(R_i; R_j; w_0; w_1; w_2) = K(z_i; z_j; w_0; w_1; w_2) \quad (13)$$

As the GPR loss function is expressed in terms of the kernel, and the transformed inputs are smooth functions of the DNN parameters, gradient descent on the GPR loss function simultaneously optimizes GPR kernel parameters and DKL parameters. This allows the DNN to learn a representation specifically suited to the task of predicting the target variable, increasing the representational power of DKL relative to traditional GPR.

By encoding each input configuration to a distribution in latent space, the overall projection of the training data can be regularized to any desired distribution by approximating the KL divergence, L_{KL}, of the target distribution and the combined distribution of the points in the training batch, $p_b(z)$. Given a specific prior distribution in latent space $p_0(z)$ and a batch of N_b training points, we define the approximation

$$p_b(z) = \frac{1}{N_b Z} \sum_{k=1}^{N_b} p_e(z|R_k) \quad (14)$$

$$L_{KL}(p_b \| p_0) = \int dz\, p_b(z) \log \frac{p_b(z)}{p^0(z)} \quad (15)$$

$$\approx \sum_{k=1}^{N_b} \log \frac{p_b(z_k)}{p^0(z^k)} \quad (16)$$

This formulation differs from pointwise latent space regularization functions by comparing the full batch distribution to the desired prior, rather than each training point individually. While this leads to a more complicated calculation of the regularization loss, it can be readily extended to other desired prior distributions, beyond a normal distribution in latent space. We incorporate this regularization loss to limit model overfitting as a result of the DKL projection and enforce specific structures on the latent projection of the data set.

The overall workflow is presented in Figure 1. We start with an AA MD trajectory which samples configurations from the canonical ensemble, and use those configurations for QC calculations to generate individual samples of the target HOMO energy distribution $p(E|R)$. AA configurations are mapped to a CG resolution by a defined linear CG mapping operator and featurized for input to DKL-ECG. Approximate GPR utilizing variational DKL is then used to produce heteroscedastic predictions of $p(E|R)$. This trained DKL-ECG model is then able to predict HOMO energy distributions for any other point in CG configuration space of the target molecule.

Implementation of DKL-ECG is performed through the GPyTorch package.[52,53] Details of DKL-ECG training are provided in the SM, with an implementation of the method provided at https://github.com/TheJacksonLab/DKL_ECG.

D. Estimates of Target Distributions

1. Additional Equilibrium Samples for Test Sets

To validate DKL-ECG, two types of data sets are constructed for validation using (i) equilibrium and (ii) constrained sampling techniques. Equilibrium sampling involves drawing statistically uncorrelated additional samples from the canonical ensemble using MD. For BT/TEMPO, the test set is generated by an additional 400/40 ns MD simulation. For S3MT the test set was generated by randomly splitting the data set into training and test sets. The reason for the order of magnitude difference between the BT and TEMPO data sets is that accurate evaluation of DKL-ECG predictions made on localized regions of CG space requires significantly more data than the amount needed to train DKL-ECG on the equilibrium distribution. Isolating only the subset of the sample of the equilibrium distribution in the neighborhood of a single CG configuration either requires a much higher data density, or results in high sampling uncertainty when estimating CG noise. BT exhibits a richer dependence of its electronic structure on configuration than TEMPO; the main influence on TEMPO's HOMO energy level is the N-O bond length, whereas the interplay S-C-C-S dihedral and intraring vibrations mediate electronic structure variations in BT. Therefore, we focus all data-intensive validations of DKL-ECG (i.e. noise estimates) on BT throughout the work.

2. Estimates of Local Energy Fluctuations in Test Sets

The test sets generated in the previous section are used as independent estimates of the local mean and noise of the conditional distribution $p(E|z)$. We refer to the standard deviation of $p(E|z)$ as the DKL noise. We take the mean values $e_1(\tilde{z}_i)$ of the latent projections of the test set, and use Gaussian Mixture Modeling (GMM) to cluster the overall set into samples of the local energy distribution. GMM is performed five times, with N = 20, 40, 60, 80, and 100 (low population clusters ($n_{pop} < 15$) are discarded). For each configuration in a remaining cluster, the predictive distribution is calculated. These test estimates of the local energy mean and noise are then compared to the corresponding predictions by DKL-ECG to give root-mean-squared error (RMSE) values for the mean and noise. All cluster distribution properties are evaluated as weighted sums according to the GMM-provided class probability over the elements of the test set whose highest classification probability is in that cluster; details of these calculations are provided in SM Section VII. GMM is implemented using the scikit-learn package.[54] When calculating local DKL noise values in clustered test data, the DKL-ECG predictive mean is

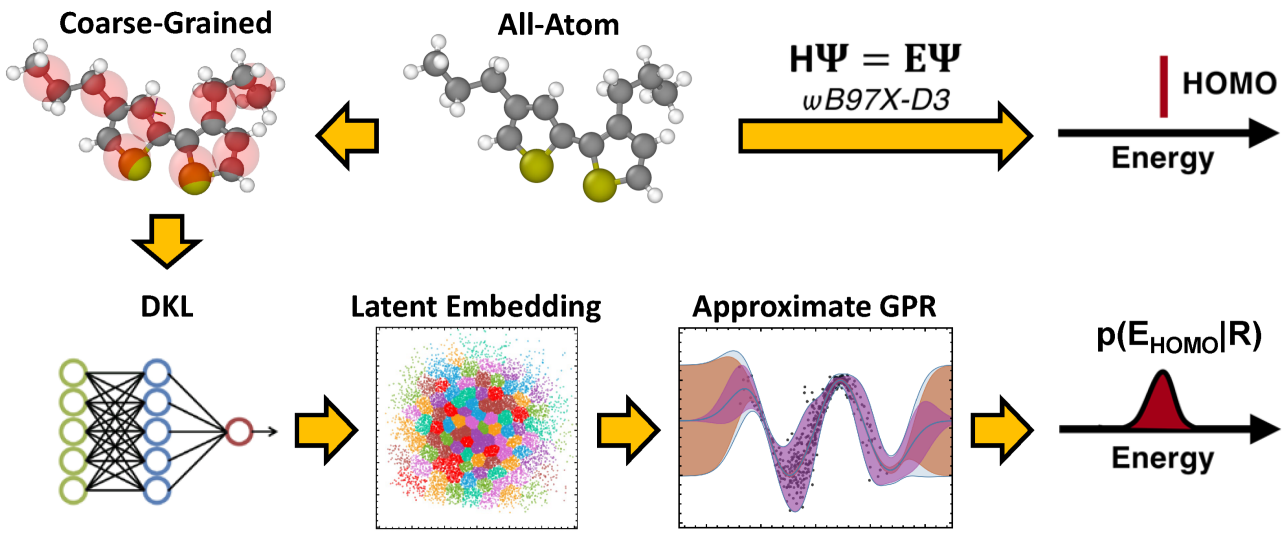

FIG. 1: Schematic of the DKL-ECG method. Predictions of all-atom fluctuations in a target electronic property are performed using only the CG representation of the molecule. This is compared to canonical individual samples from QC calculations. While individual QC calculations produce samples of the CG energy distribution, DKL-ECG predicts the full distribution. E_{HOMO} is the HOMO energy level, and throughout the text is expressed as simply E.

first subtracted from the test data energy value. This step facilitates the generation of higher-population clusters, and thus lower sampling uncertainty in the DKL noise. Sampling uncertainty can strongly affect the calculation of the global heteroscedasticity, especially when the heteroscedasticity present in the data is low. We also note that the process of encod-ing CG configurations to a lower-dimensional latent space via DKL is similar to the CG map, the initial source of noise in our target variable. We separately need to verify that the DKL noise matches the CG noise, and the DNN is only discarding information irrelevant to the prediction of the target variable.

3. Generation of CG Distributions through Backmapping

To determine the true CG noise at a given configuration, R, we sample p(rjR) through a constrained MD simulation, similar to other works backmapping CG configurations to AA coordinates.[22] This procedure is performed for only BT and TEMPO due to the more tractable CG configurational space for constrained sampling relative to S3MT. To generate constrained AA configurations, a CG configuration is drawn from the training distribution. The training configurations were first sorted according to the degree of freedom known to strongly influence the value of the HOMO energy level; ten configurations were then sampled from those nearest to a desired value of the DOF. For BT, the value of the S-C-C-S dihedral an-gle was used, and eight samples were taken uniformly from 0 to 315 degrees. For TEMPO, the length of the N-O bond was used, and nine samples were taken across four total stan-dard deviations of the training distribution. Sampling across CG configurations tests the ability of DKL-ECG to reproduce electronic distributions across the full range of the thermal energy distribution.

After choosing a CG configuration, a series of harmonic potentials (k = 5 10^4 kcal/mol Å ^{2}), denoted V, are placed at the initial center of mass of each CG bead, constraining the positions of the underlying AA coordinates to that point. The slope of these potentials bias the atoms to fluctuate only in the degrees of freedom that do not affect its CG represen-tation. The MD simulation timestep was changed to 0.05 fs due to the strong harmonic constraints, but no other environ-mental procedures were changed from the MD simulation for the training and test sets. For TEMPO, a 125 ps MD simu-lation was run, with a sampling interval of 2.5 fs. For BT, the length of the MD simulation was 600 ps, to reduce sam-pling uncertainty in the distributions. After sampling these constrained distributions, further ωB97X-D3/def2-SVP calculations were performed to generate the sampled distribution p(EjR). The constrained trajectory follows a biased distribution as the molecules experience the sum of the original atomic potential energies functions, U, and the newly applied harmonic bias V. To generate an unbiased estimate of the local mean and CG noise of the energy, we re-weight the expectation value according to

$$\langle A \rangle_U = \frac{\langle A e^{\beta V} \rangle_{U+V}}{\langle e^{\beta V} \rangle_{U+V}} \tag{17}$$

when calculating the mean and standard deviation of the energies for the constrained trajectory.

Due to the finite nature of the applied potentials, each sample of the constrained distribution, C, has some spread in both CG space and latent space. Fluctuations of the underlying mean energy function $\langle p(E|R) \rangle$ over the cluster domain can cause the sampled noise to be higher than the true noise at the

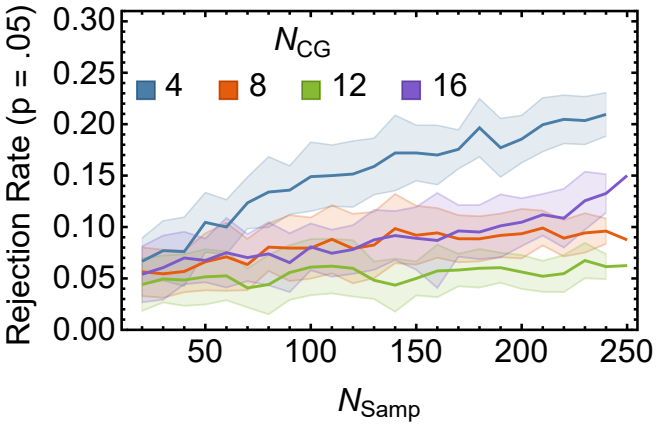

FIG. 2: Fraction of constrained BT energy distributions in which the null hypothesis of normality has been rejected by the Shapiro-Wilk test. Results are shown for different CG resolutions (N_{CG}) as a function of the number of samples (N_{Samp}) at each constrained configuration. Uncertainty bands indicate one standard deviation of 25 different samples.

target CG coordinate. We incorporate this into the model's prediction for the cluster by also sampling the predictive noise over the full domain of C, rather than just the target CG coordinate.

$$^{G}\mathrm{Var}\,(y|R) = \langle \mathrm{Var}(y|m_i)\rangle_{m_i|2C} \quad (18)$$

$$+\,\mathrm{Var}\,(\mathrm{Exp}(y|m_i))_{m_i|2C} \quad (19)$$

III. RESULTS

We first test the assumption that p(E|R) can be well-approximated by the normal distribution within DKL-ECG. A Shapiro-Wilk normality test[55] is applied to the constrained BT energy distributions. Each trajectory contains 250 total samples of the full distribution, p(E|R), drawn from constrained MD simulations. We draw N_{samp} samples without replacement from the distribution, and apply the test with a set significance value of 0.05 and a null hypothesis of a normal distribution. Figure 2 shows the fraction of distributions for which the null hypothesis was rejected as a function of the number of samples, N_{Samp}, and CG model resolution (N_{CG}). Finer resolution (larger N_{CG}) distributions are close to 5%, as expected by the choice of significance threshold. Small N_{CG} show higher rejection fractions, with the 4-resolution conditional distribution showing rejection as much as 20% of the time. This higher rejection rate reinforces that there is no guarantee for the molecule averaged CG conditional distribution (i.e. $N_{CG} = 1$, the electronic density of states of the system) to be normally distributed. Taken together, these results support the idea that the Gaussian assumption of the conditional distribution inherent to DKL is appropriate for our data sets, though may deteriorate at the coarsest model resolutions.

With p(E|R) being sufficiently approximated by a normal distribution at most CG resolutions, we next examine the ability of trained DKL-ECG models to reproduce the CG mean and noise of the target distribution throughout CG configuration space. Figure 3 illustrates the latent projection, $e_1 g(R)$, for the training set and constrained trajectories of BT at an 8-bead CG resolution. To aid in visualization, the latent space was regularized against the prior distribution N (0; 1), with regularization parameter $l = 10^{-4}$, and a latent feature dimension of two. First, Figure 3 visually corroborates that the projected training distributions match the Gaussian prior, in agreement with Figure 2. Second, Figure 3 proves that lo-cal regions in CG configuration space provided by constrained dynamics are mapped through the DNN representation of the kernel to local neighborhoods of latent space. It is generally observed that regions of comparable planarity of the S-C-C-S dihedral angle map to similar regions of the regularized latent space, with 90/270, 45/135/225/315, and 0/180 being approximately sorted according to z_2. A physical interpretation of z_1 is less forthcoming, but the general structuring of latent space according to known geometric descriptors relevant to large electron phonon couplings in BT is an encouraging feature of the learned latent space. Quantitatively, the DKL-ECG predictive distribution, p(E|R), over the GMM clustered data agrees well with the target distributions sampled by both explicitly constrained MD and GMM clustering over the canonically sampled data.

Next, we assess DKL-ECG's performance as a conventional regression model that outputs a single-valued prediction of the distribution mean for a given CG configuration. Figure 4 shows the RMSE on the training set, the average predictive noise predicted by DKL-ECG, the average CG noise derived from constrained trajectory data, and the average DKL noise derived from GMM clustering over canonically sampled trajectories of BT, TEMPO, and S3MT. The RMSE of the predictive mean converges to both the mean of the predictive noise and the CCG and GMM estimated noise values, for all molecules and CG resolutions. These facts indicate that there is no significant source of error in the DKL-ECG prediction beyond the intrinsic CG noise resulting from application of the CG mapping operator to the AA data - the predictive accuracy of DKL-ECG is CG mapping operator limited. This is particularly important as these results indicate that the noise value reached by the predictive distribution matches the physical CG noise in the data set, in spite of the dimensionality reduction of DKL. These results demonstrate that DKL-ECG can converge accurate prediction models of HOMO energies across multiple chemistries and CG resolutions, with nearly all prediction error can be associated with the intrinsic noise resulting from definition of the CG mapping operator.

An additional implication of the convergence of RMSE, predictive noise, and CCG sampled noise in Figure 4 is that DKL-ECG can be used to screen for "optimal" CG mapping operators. Sharp increases in the average predictive noise correspond to a sudden increase in the amount of information loss due to the CG mapping operator. For example, in TEMPO, the important difference between the 5-bead and 4-bead CG representations used in Figure 4b is the decimation of the N-O

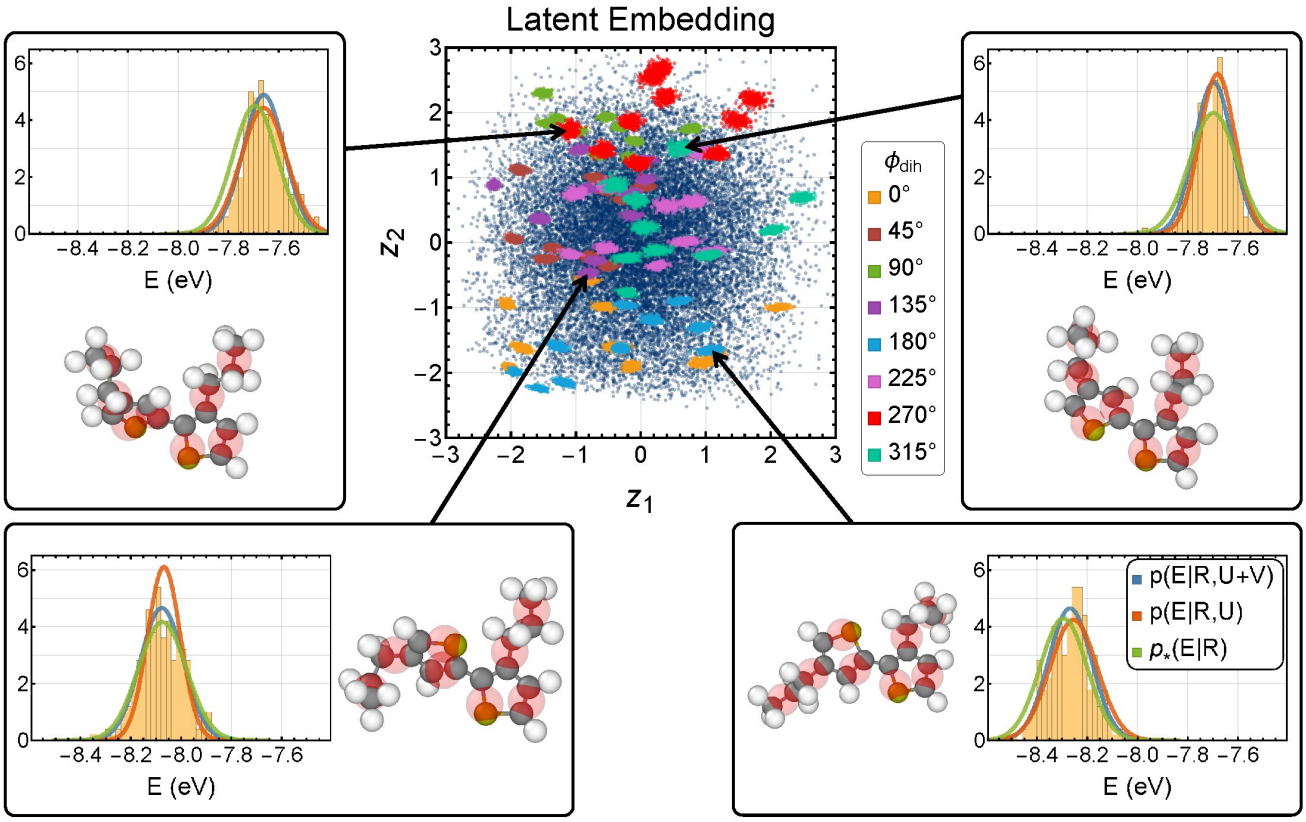

FIG. 3: Projection of AA BT MD trajectories (gray, white, and yellow atoms) constrained to regions of CG configuration space (red beads) onto a regularized DKL-ECG latent space. f_{dih} denotes the S-C-C-S interring dihedral angle of BT. p(EjR;U+V), p(EjR;U), and p(EjR) denote the conditional distribution for HOMO energy (E) predictions associated with constrained MD simulations (U+V), canonically sampled MD simulations (U), and the learned DKL-ECG model applied to GMM clusters from canonically sampled data, respectively.

bond degree of freedom. The length of this bond is strongly associated with the molecular orbital energy of the radical,[56] and consequently choosing a CG mapping operator that decimates this bond information introduces a large amount of CG noise (Figure 4b). Further coarsening of the CG resolution produces less significant changes in the noise. Similarly, BT exhibits a sharp increase in noise in going from four beads to six beads (Figure 4a), which is attributable to a deteriorated description of the intermonomer dihedral when averaging over both the conjugated rings and side chains. For S3MT, a steep increase in noise is observed in moving from 12 beads to 6 beads (Figure 4c), consistent with a loss in the ability to capture interring dihedral angles. Provided a single training data set, many different CG maps can be trialed by training DKL-ECG on each CG representation and evaluating the mean predictive noise. These results underscore the potential for DKL-ECG to identify important collective degrees of freedom to preserve when choosing a CG mapping operator.

Figure 5 shows the results of the component RMSE values for the GMM and CCG data sets. As these are errors in the parameters of a distribution, rather than the error of elements of a distribution measured against the mean, there is no longer a predefined lower bound. All error components generally decrease as a function of resolution, as expected due to the overall decrease in average distribution width at higher CG resolution. By scaling each element of the sum involved in the RMSE calculation by the reference DKL noise, we show that DKL-ECG achieves a consistent relative accuracy at all CG resolutions; this is discussed in further detail in SM Figures S2-3. The CCG component error follows the same trend as the GMM error, but at a consistently higher value. In particular, the coarsest resolution of BT has a larger gap between GMM and CCG errors than finer resolutions. It is unclear whether this is due to an overfitting of the sample of the equilibrium distribution, or due to the CCG samples not accurately replicating the conditional equilibrium distribution despite the reweighting scheme. Accurate sampling of the conditional distribution p(rjR) may break at coarser CG resolutions, when a single potential is biasing the motion of many more atoms.

The level of CG noise in the observations of the target variable is not only a function of the choice of molecule and CG map, but also thermodynamic parameters. Figure 6 shows the mean predictive noise, averaged over the testing points, for DKL-ECG trained on S3MT at multiple CG resolutions and temperatures from canonically sampled MD. Increasing

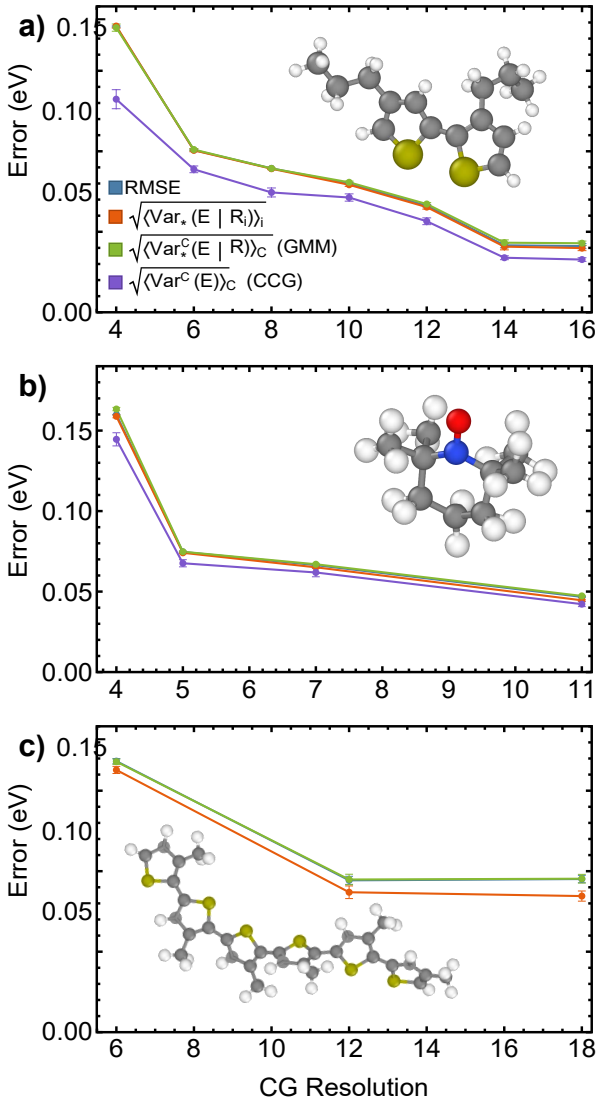

FIG. 4: RMSE of mean predictions (blue curve) and average CG noise estimates for (a) BT, (b) TEMPO, and (c) S3MT at 300 K. The orange curve is the GPR predictive variance, averaged pointwise over the test set. The green and purple curves are sampled noise estimates averaged over all configurational clusters, where each cluster is generated by GMM and CCG, respectively. Error bars of CCG curves are the standard error of the mean; other error bars are one standard deviation of 25 repeated trials.

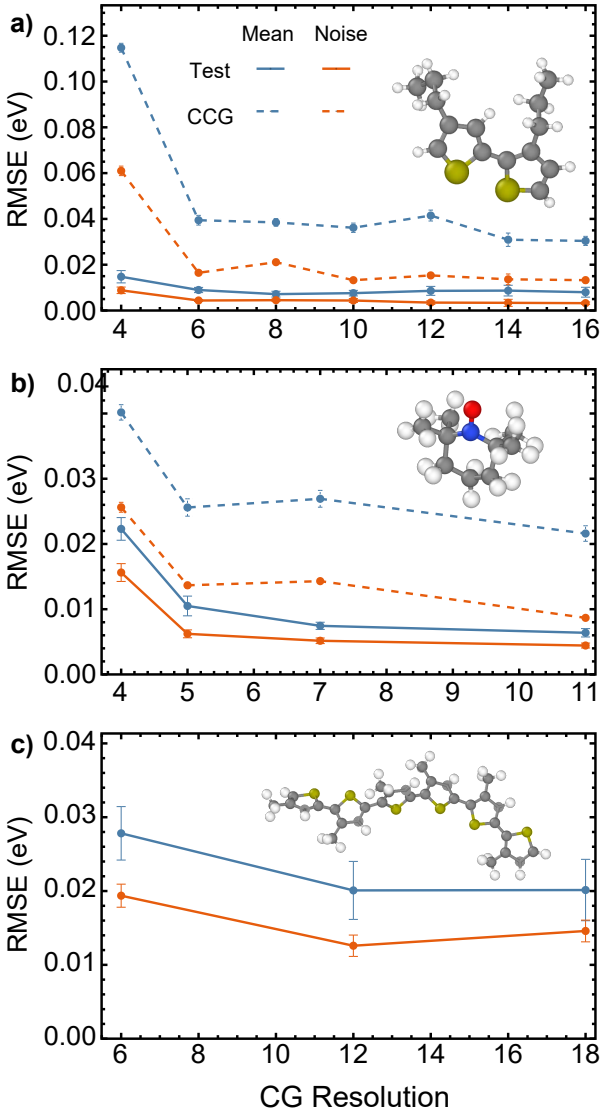

FIG. 5: RMSE for mean and standard deviation predictions on the GMM/CCG-clustered test sets (solid/dashed lines) for (a) BT, (b) TEMPO, and (c) S3MT at 300 K. Error bars are the sum in quadrature of the standard deviation of 25 repeated trials and mean sampling uncertainty of GMM estimates.

the temperature increases thermal fluctuations within each CG bead, leading to wider distributions of accessible AA configurations and larger fluctuations in the target variable. Figure 6 further supports the difference in predictive noise between the 6 and 12 bead CG resolutions of S3MT observed in Figure 4c. Decreasing the temperature leads to smaller fluctuations until, at zero temperature, one reaches the limit of quantum mechanical zero-point motions of the bonded topology not captured within classical MD sampling schemes.

DKL-ECG provides a means of predicting AA distribution functions using only CG models without re-referencing the AA resolution, dramatically improving computational efficiency. After training DKL-ECG on an AA data set, further exploration of configurational space can be performed without any future reference to the AA models. We use our training set of AA configurations to train effective CG bonded potentials via the IBI method. Using the converged IBI-generated potentials, a CG MD trajectory through configuration space was produced. As shown in SM Figure S13, this CG trajectory is projected by DKL to the same distribution in latent space as the DKL-ECG training set, confirming the validity of the IBI procedure. Using this IBI generated trajectory of CG configu-

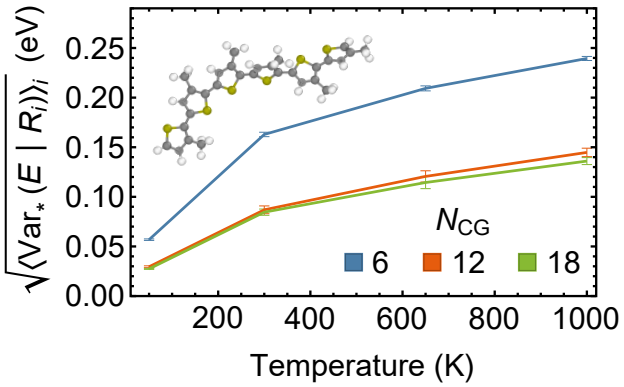

FIG. 6: Mean predictive noise of the S3MT testing set predicted by DKL-ECG as a function of simulation temperature and CG resolution. Error bars indicate one standard deviation of 25 repeated trials.

rations, we leverage the trained DKL-ECG model to generate predictions of the AA HOMO energy distribution at all points in the CG trajectory. A discrete set of sampled CG configurations, and single-valued estimations of their energies (e.g. the mean), produces only a discrete estimate of the overall density of states. However, DKL-ECG is capable of estimating a continuous energy distribution for each individual CG configuration within a Gaussian approximation, of which the average of these predictions over a sampled trajectory T can recreate a smooth estimation of the full DOS.

$$p(E) = \frac{1}{N_T} \sum_{i \in T} p(E \mid R_i) \qquad (20)$$

Figure 7 shows the originally sampled DOS of the HOMO energy level at the AA resolution, compared to predictions generated by DKL-ECG, for BT, TEMPO, and S3MT. A DOS generated by the distribution of predictions of the mean, as shown by $p^{IBI}(hEi)$, underestimates contributions of extreme energies and overestimates the middle of the distribution. In comparison, $p^{Te}(E)$ is the result of Equation 20 sampled over the test data set, and $p^{IBI}(E)$ is sampled over the IBI-generated trajectory. Both distributions exhibit quantitative agreement with the DOS of the underlying AA model. This result is critically important for CG modeling, as extreme values of the AA distribution (here, HOMO energies) often dictate phenomenology that controls ultimate material performance. In charge transporting polymers, this fact manifests particularly strongly as the tail of the density of states controls trap formation which dictates numerous device properties.[57] Figure 7 shows that extreme values of the AA distribution are effectively sampled by DKL-ECG models by virtue of incorporating the width of the distribution, without the need for backmapping and AA resampling paradigms. It should be further noted that the distribution being reconstructed, the full thermodynamic DOS, is also p(EjR) for a 1-bead CG mapping. While a trivial map eliminates the usefulness of a CG model, the full DOS clearly deviates from the assumption of normality made in the creation of DKL-ECG.

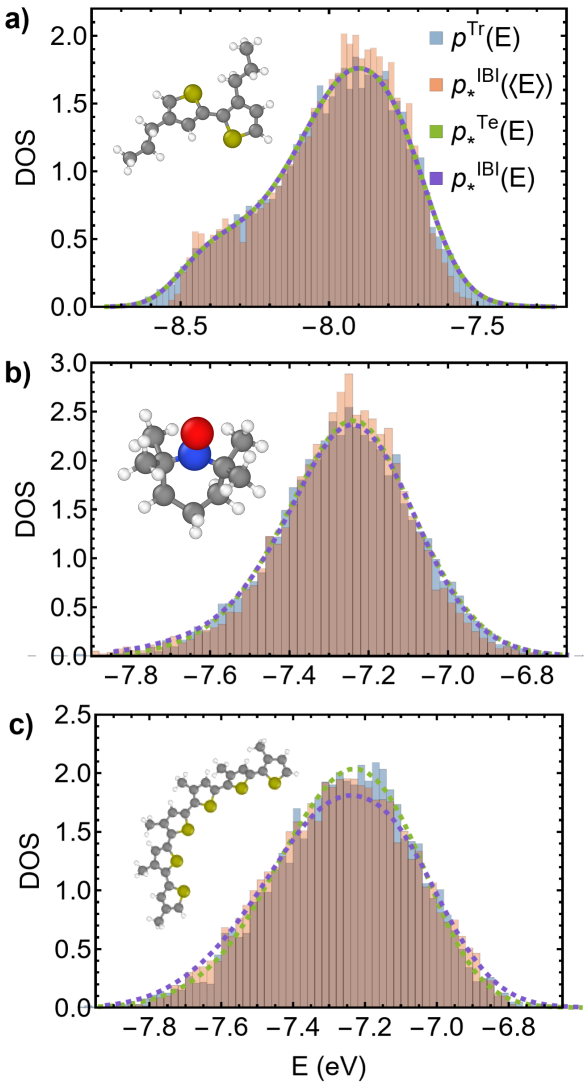

FIG. 7: Approximation of the 300 K (a) BT, (b) TEMPO, and (c) S3MT, HOMO DOS via four methods. The histogram $p^{Tr}(E)$ is the discrete sample of the DOS provided by the training set, while the histogram $p^{IBI}(hEi)$ is the prediction of the CG mean energy from DKL-ECG on the IBI trajectory. $p^{Te}(E)$ and $p^{IBI}(E)$ are the continuous DOS generated by the sum of DKL-ECG predictions on the testing set and IBI-generated CG configurations, respectively. All configurations used to generate p^{IBI} are drawn from IBI-derived CG potentials, and include no further reference to AA configurations.

IV. DISCUSSION

A. Bypassing Backmapping in Multiscale Modeling

Projecting the equilibrium distribution of a molecule to a CG resolution results in an information loss in the inference of chemical properties defined within the AA representation. While CG models are an important part of research into com-

plex molecular and polymeric systems, a single-valued prediction of an AA property as a function of CG configuration discards the full detail of the AA configuration belonging to the original equilibrium distribution around a point in CG configuration space. Generally, the process of backmapping is required to sample this conditional distribution, which requires complex computational workflows and many iterative layers of repeated AA simulations or ML models. As a significant merit of CG modeling is the reduction in computational cost relative to AA systems, the process of backmapping is a considerable concern with regards to the scalable screening of properties in complex morphologies, particularly those requiring expensive QC calculations on the AA resolution.

By explicitly treating the AA target variable as an intrinsically noisy probability distribution within CG configuration space, we have shown that DKL-ECG learns electronic structure predictions up to the limit of information loss from the CG mapping. The ability to learn the heteroscedastic noise values across CG configurations allows comparison of the levels of information loss across choices of map, and bypasses the need for backmapping to recreate the full atomistic variable distribution. Specifically, one can generate stochastic samples of the AA observable consistent with a point in CG configuration space by sampling a normal distribution with a learned mean and width as a function of CG configuration space (Figure 7). Consequently, in multiscale workflows there is considerable potential to learn information “lost" by application of the CG mapping operator to within a Gaussian approximation, and to use the learned distributions of DKL to directly sample the thermodynamic values of the property, without any future recourse to AA models. This represents a powerful general paradigm for reducing information loss and increasing computational efficiency in multiscale workflows utilizing CG models.

B. Generalization to Other All-Atom Property Distributions

While in this work DKL-ECG was used to learn the CG configuration dependence of the QC-derived HOMO energies of molecules, its applications are general across a breadth of AA-derived properties. Obvious future electronic property applications are optical spectra, charge density fluctuations, NMR chemical shifts, and multipolar moments over which thermodynamic averaging of QC calculations often occurs. It is likely that a DKL-ECG framework could also be employed to learn any arbitrary AA property distribution that is lost via the process of coarse-graining. Interesting future targets include vibrational spectra, hydrogen bonding, solvation shell environments, and general AA structural distributions.

DKL-ECG methodologies capable of learning both the mean and noise of the AA property distributions also possess interesting applications in the field of intermolecular potential fitting. As the derivative of a GP is another GP, it provides a natural method for making stochastic force predictions. In particular, learning not only the potential of mean force as a function of CG configuration, but also the local distribution of forces as a function of CG configuration may present a means of reincorporating fast, local dynamics lost in the smoother CG-mapped free energy surface under assumptions of appropriate decorrelation of force-force and velocity-force correlation functions. These considerations may potentially remedy some aspects of well-known dynamical deficiencies of CG models. Physically based models to this effect have managed to recapture lost dynamical information in CG models developed via force-matching,[58] and could find application to existing CG ML potentials using DNN or GP like frameworks.[16,17]

C. Limitations and Future Improvements

A clear limitation of the current DKL-ECG model is the choice of data featurization, the full CG distance matrix. While it provides a convenient way of enforcing rotational and translational symmetries, applying this featurization to a much larger system would result in an excessively large input vector due to the N^2 scaling of the distance matrix. Future work will explore modifying DKL with reduced distance matrices, or replacing the feed forward neural network with graph neural network or graph kernel methods capable of learning more robust three-dimensional representations with improved scaling.

The process of hyperparameter optimization introduces a computational overhead when fitting a single iteration of DKL-ECG. In addition, while our testing indicated full-batch training improved model performance, this may not be generally transferable. In some cases,[59] DKL has been shown to be susceptible to overfitting on data sets, with minibatching being an important factor in regularizing the DKL projection. This difference might be accounted for by recognizing that our training and testing sets are intended to be thorough samples of the full molecular equilibrium distribution, limiting the potential for overfitting. A more general understanding of the effects of data set and model hyperparameters should be reached in order to streamline this framework.

Convergence of the learned underlying AA distribution functions with respect to training data is also an important consideration. Here, we observe that $O(10^5)$ samples are needed to develop confident estimates of local AA distribution widths for BT when sampling is performed over the full equilibrium distribution. It appears that reasonable approximations of the mean and noise of the Gaussian distributions consistent with $O(10^5)$ samples are learned by DKL-ECG using $O(10^4)$ training samples. However, for QC data sets, particularly at levels of electronic structure theory surpassing DFT, this presents a considerable barrier to model training. Consequently, active learning strategies can be explored in future efforts.[39] However, thoughtful approaches to incorporating the correct thermodynamic distribution of AA training data, while simultaneously performing active sampling, will be required. Individual estimates of CG energy distributions can be accurately achieved with $O(10^3)$ samples per CG configuration, but this requires the assumption that $p(rjR)$ is accurately sampled. Deviations in this assumption result in a mismatch between the training and evaluation configuration distribution, and invalidating attempts at model evaluation.

Another point of consideration involves incorporating the variance decomposition described in Appendix A more systematically into the DKL-ECG method. For BT, TEMPO, and S3MT we are confident that we have obtained enough training data to sufficiently converge the estimate of the distribution widths. Consequently, the decomposition into s^A_K and s^A_S was not explicitly utilized in this work. However, in other systems where convergence has not been sufficiently achieved, such a decomposition into the uncertainty of the mean prediction and the expected CG noise could be more informative, but considerable future work needs to be done to formalize this strategy. Moreover, the basis of structural CG modeling is that sufficient AA data exists to converge all estimates of structural distribution functions in training, and consequently it is sensible to necessitate a similar requirement with respect to DKL-ECG model training.

A necessary aspect of predicting CG electronic variables is the ability to make heteroscedastic noise predictions; there is no reason to believe that fluctuations in the target variable are completely uniform across CG configurational space. Direct measurement of the level of heteroscedasticity in the data, calculated as the standard deviation of the local CG noise, is limited due to the sampling uncertainty of these local estimates. DKL-ECG often appears to show an underestimation of the standard deviation of predictive variance, both compared to DKL and CCG estimates of CG noise. Both of these sets suffer from artificial increases in the observed heteroscedasticity, due to the sampling uncertainty in the calculation of the standard deviation of a relatively small sample. These errors, as well as errors in the prediction of the local mean and standard deviation, improve as the size of the test set increases and sampling uncertainty is reduced. The effects of increasing testing set sizes on DKL-ECG evaluation accuracy is further detailed in SM Figure S12. The remaining gap between DKL-ECG predictive and CCG observed heteroscedasticity, requires further work in generating unbiased estimates of p(rjR), in order to more clearly determine the degree to which the DKL projection overfits when simplifying the training set.

V. CONCLUSIONS

We have developed a computational framework for bypassing backmapping that learns the AA property distribution functions associated with CG model representations within a heteroscedastic Gaussian approximation. We have applied this framework to learn the configuration dependence of QC-derived properties of multiple electron transporting chemistries at CG resolutions. This development enables the prediction of QC properties solely from a CG model resolution, without recourse to AA representations via backmapping protocols. While this approach discards the ability to make any inference on the AA fluctuations around individual CG configurations, reliance on AA backmapping carries its own costs. Placement of individual atoms must be carefully considered, and new target systems generally require new changes to the backmapping algorithm. Ill-formed AA configurations propagate these errors to calculations of the desired observable, which can manifest as extreme observable predictions that present themselves without context. Moreover, modeling is limited by the observable's computational cost at the AA resolution, which must be performed, ad nauseum, for every subsequently backmapped AA configuration; in the case of QC-derived observables, this cost can be severe. In contrast, the width of the DKL-ECG predictive distribution is informed by both the local CG configuration and the entire training dataset, and contributions to the predictive variance can be separately interpreted. Predictions can be made on new observables or CG representations by training a new instance of DKL-ECG, with suitably replaced labels or input data. Once trained, DKL-ECG can make predictions of the observable on out-of-sample configurations at a significant speedup compared to AA calculations.

We demonstrate that the prediction error of DKL-ECG converges to the intrinsic physical noise limits induced by application of the CG mapping operator across a wide range of CG model resolutions and multiple chemistries. Moreover, DKL-ECG reproduces physically intuitive noise behavior as a function of temperature and CG model resolution. We demonstrate the ability of DKL-ECG to produce the full thermodynamic distribution of the EDOS of states using only CG models, obtaining high accuracy even in the tails of the EDOS. This work focused on a single choice of system featurization, which currently presents obstacles in analyzing systems much larger than we have considered, or more complex condensed phase effects. There remain open questions in the size of the required training set, choosing certain model hyperparameters, and scaling to more realistic case studies, but there also exists a large design space in the field of machine learning to improve the proposed framework. Given the results shown for these initial cases, the DKL-ECG methodology should find use not only in the reproduction of QC-derived electronic properties from a CG model, but also for the preservation of any AA information traditionally "lost" during CG model development.

VI. SUPPLEMENTARY MATERIAL

Coarse-grained mapping operators, details of AA MD simulations, details of DKL-ECG training, resolution-dependent RMSRE of test predictions, details of DKL-ECG training convergence, sampling uncertainty in evaluating DKL-ECG predictions, weighted averages of GMM-clustered data, details of IBI convergence, DKL-projected IBI trajectories.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation Chemical Theory, Models, and Computation division under award CHE-2154916. We acknowledge support from the Dreyfus Program for Machine Learning in the Chemical Sciences and Engineering during this project. We thank Dr. Phil Rauscher for a critical reading of the manuscript.

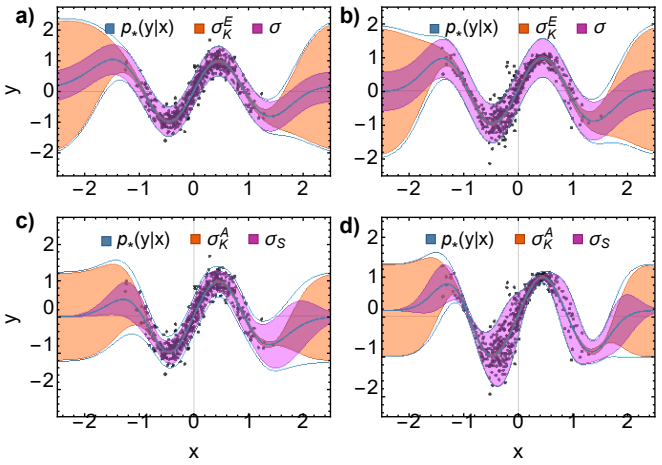

FIG. 8: Comparison of predictive distributions for exact and approximate GPR when applied to toy homoscedastic and heteroscedastic data sets. The predictive distribution of approximate GPR matches the noise present in both data sets, while the predictive variance of exact GPR in data-rich regions is always the mean of the observed variance. In data-rich regions, the predictive variance is dominated by the terms s and s_S. Outside of this region, both GPR methods return to the GPR prior. Uncertainty bandwidths are two standard deviations of the given distributions.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Appendix A: Appendixes

1. Variance Decomposition in Exact and Approximate Gaussian Process Regression

Exact GPR generates a predictive distribution for test points by marginalizing the GP prior distribution over the latent function at the training points:

$$p(f|y) = \frac{1}{p(y)} \int df\, p(y|f)\, p(f;f) \tag{A1}$$

$$\Rightarrow p(f|y) = N\left(m_K^E; s_K^E\right) \tag{A2}$$

$$p(y|y) = N\left(m_K^E; s_K^E + s^2\right) \tag{A3}$$

where specific expressions have been previously defined in Equation 5. Evaluating s_K^E is a $O(N_{tr}^3)$ in the number of training points, limiting the scalability of GPR when inferring on large data sets. The GP prior distribution $p(f;f)$ could be exactly expressed in terms of the conditional distribution on a further set of inducing variables, u. Approximate GPR modifies this expression such that f and f are conditionally independent given u.

$$p(f;f) = \int du\, p(f;f|u)\, p(u) \tag{A4}$$

$$p(f;f|u) \approx p(f|u)\, p(f|u) \tag{A5}$$

This approximation implies that predictions on a test set are made by inference on the set of inducing points, which can remain fixed while the size of the training set increases.

$$p(f|u) = N\left(K_u K_{uu}^{-1} u; K - K_u K_{uu}^{-1} K_u^{\top}\right) \tag{A6}$$

The definition of the inducing point distribution is needed to generate the final predictive distribution. Some methods express this distribution as a subset of the training observations. The variational inducing point method treats the inducing distribution as a further set of hyperparameters of the model,

$$p(u) = N(m; S) \tag{A7}$$

The locations of the inducing points, as well as the parameters m and S, are learned based on the training observations, and result in an approximation of the full training distribution, from only the fixed-size inducing set.

$$p(f|y) = \int du\, p(f|u)\, p(u|y) \tag{A8}$$

$$\int du\, p(f|u)\, p(u) \tag{A9}$$

The last step is made because the inducing variational distribution is independent of y, but parameterized through training based on y. The approximate predictive distribution can now be explicitly calculated, using known properties of normal distributions.

$$\Rightarrow p(y|y) = N\left(K_u K_{uu}^{-1} m; \right. \tag{A10}$$

$$\left. K_u K_{uu}^{-1} K_u \right. \tag{A11}$$

$$\left. + K_u K_{uu}^{-1} S K_{uu}^{-1} K_u + s^2\right) \tag{A12}$$

$$= N\left(m_K^A; s_K^A + s_S^A + s^2\right) \tag{A13}$$

Figure 8 gives a practical visualization of the differences between the predictive distributions of exact and approximate GPR when applied to one-dimensional toy data sets. Both data sets have the same underlying sinusoidal mean function, but differ in their local noise. One data set has a uniform noise at all values of x, while the other has strongly heteroscedastic noise. While both methods are always able to accurately predict the local mean $p(f|x)$, clear differences appear in the predictive variance. Outside the domain sampled by the training data, both models predict the GP prior $p(f)$. Inside the training domain, exact GPR is only able to predict the mean variance of the full data set, which cannot match local noise fluctuations in the heteroscedastic set. The approximate GPR, however, is able to accurately predict the local variance of both sets. Furthermore, the predictive variance inside the training region is dominated by the s_S and s^2 terms, while the predictive variance outside this region is controlled by the s_K term. Decomposing the variance calculation for a

new point whose location relative to the training domain is unknown could then give more information on the model's fit uncertainty at that point.

[1] J. J. de Pablo, "Coarse-grained simulations of macromolecules: from dna to nanocomposites," Annu. Rev. Phys. Chem 62, 555–574 (2011).

[2] G. A. Voth, Coarse-Graining of Condensed Phase and Biomolecular Systems (Taylor Francis Ltd, Hoboken, NJ, 2008).

[3] N. E. Jackson, "Coarse-Graining Organic Semiconductors: The Path to Multiscale Design," J. Phys. Chem. B 125, 485–496 (2021).

[4] W. G. Noid, "Perspective: Coarse-grained models for biomolecular systems," J. Chem. Phys. 139, 090901 (2013).

[5] S. Dhamankar and M. A. Webb, "Chemically specific coarse-graining of polymers: Methods and prospects," J. Polym. Sci. 59, 2613–2643 (2021).

[6] V. Rühle, C. Junghans, A. Lukyanov, K. Kremer, and D. Andrienko, "Versatile object-oriented toolkit for coarse-graining applications," J. Chem. Theory Comput. 5, 3211–3223 (2009).

[7] R. Alessandri, J. J. Uusitalo, A. H. de Vries, R. W. A. Havenith, and S. J. Marrink, "Bulk heterojunction morphologies with atomistic resolution from coarse-grain solvent evaporation simulations," J. Am. Chem. Soc. 139, 3697–3705 (2017).

[8] S. Izvekov and G. A. Voth, "A Multiscale Coarse-Graining Method for Biomolecular Systems," J. Phys. Chem. B 109, 2469–2473 (2005).

[9] A. Chaimovich and M. S. Shell, "Coarse-graining errors and numerical optimization using a relative entropy framework," J. Chem. Phys. 134, 094112 (2011).

[10] D. Reith, M. Pütz, and F. Müller-Plathe, "Deriving effective mesoscale potentials from atomistic simulations," J. Comput. Chem. 24, 1624–1636 (2003).

[11] E. Pretti and M. S. Shell, "A microcanonical approach to temperature-transferable coarse-grained models using the relative entropy," J. Chem. Phys. 155, 094102 (2021).

[12] J. Jin, A. Yu, and G. A. Voth, "Temperature and phase transferable bottom-up coarse-grained models," J. Chem. Theory Comput. 16, 6823–6842 (2020).

[13] Y. Han, J. Jin, and G. A. Voth, "Constructing many-body dissipative particle dynamics models of fluids from bottom-up coarse-graining," J. Chem. Phys. 154, 084122 (2021).

[14] J. F. Rudzinski and T. Bereau, "Coarse-grained conformational surface hopping: Methodology and transferability," J. Chem. Phys. 153, 214110.

[15] F. Noé, A. Tkatchenko, K.-R. Müller, and C. Clementi, "Machine Learning for Molecular Simulation," Ann. Rev. Phys. Chem. 71, 361–390 (2020).

[16] J. Wang, S. Olsson, C. Wehmeyer, A. Pérez, N. E. Charron, G. de Fabritiis, F. Noé, and C. Clementi, "Machine Learning of Coarse-Grained Molecular Dynamics Force Fields," ACS Cent. Sci. 5, 755–767 (2019).

[17] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, "Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons," Phys. Rev. Lett. 104, 136403 (2010).

[18] S. T. John and G. Csányi, "Many-body coarse-grained interactions using gaussian approximation potentials," J. Phys. Chem. B 121, 10934–10949.

[19] Y. B. Varolgüneş, T. Bereau, and J. F. Rudzinski, "Interpretable embeddings from molecular simulations using gaussian mixture variational autoencoders," Mach. Learn.: Sci. Technol. 1, 015012 (2020).

[20] T. A. Wassenaar, K. Pluhackova, R. A. Böckmann, S. J. Marrink, and D. P. Tieleman, "Going Backward: A Flexible Geometric Approach to Reverse Transformation from Coarse Grained to Atomistic Models," J. Chem. Theory Comput. 10, 676–690 (2014).

[21] J. Peng, C. Yuan, R. Ma, and Z. Zhang, "Backmapping from Multiresolution Coarse-Grained Models to Atomic Structures of Large Biomolecules by Restrained Molecular Dynamics Simulations Using Bayesian Inference," J. Chem. Theory Comput. 15, 3344–3353 (2019).

[22] A. J. Rzepiela, L. V. Schäfer, N. Goga, H. J. Risselada, A. H. De Vries, and S. J. Marrink, "Reconstruction of atomistic details from coarse-grained structures," J. Comput. Chem 31, 1333–1343 (2010).

[23] F. Müller-Plathe, "Coarse-graining in polymer simulation: From the atomistic to the mesoscopic scale and back," Chem. Phys. Chem. 3, 754–769.

[24] W. Wang and R. Gómez-Bombarelli, "Coarse-graining auto-encoders for molecular dynamics," npj Comput. Mater. 5, 1–9 (2019).

[25] M. Schöberl, N. Zabaras, and P.-S. Koutsourelakis, "Predictive coarse-graining," J. Comput. Phys. 333, 49–77 (2017).

[26] W. Li, C. Burkhart, P. Polińska, V. Harmandaris, and M. Doxastakis, "Backmapping coarse-grained macromolecules: An efficient and versatile machine learning approach," J. Chem. Phys. 153, 041101 (2020).

[27] Y. An and S. A. Deshmukh, "Machine learning approach for accurate backmapping of coarse-grained models to all-atom models," Chem. Commun. 56, 9312–9315 (2020).

[28] K. A. Louison, I. L. Dryden, and C. A. Laughton, "GLIMPS: A machine learning approach to resolution transformation for multiscale modeling," J. Chem. Theory Comput. 17, 7930–7937 (2021).

[29] M. Stieffenhofer, T. Bereau, and M. Wand, "Adversarial reverse mapping of condensed-phase molecular structures: Chemical transferability," APL Materials 9, 031107 (2021).

[30] W. Pezeshkian, M. König, T. A. Wassenaar, and S. J. Marrink, "Backmapping triangulated surfaces to coarse-grained membrane models," Nat. Commun. 11, 2296 (2020).

[31] V. L. Deringer, A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti, and G. Csányi, "Gaussian Process Regression for Materials and Molecules," Chem. Rev. 121, 10073–10141 (2021).

[32] L. Zhang, J. Han, H. Wang, R. Car, and W. E, "Deepcg: Constructing coarse-grained models via deep neural networks," The Journal of Chemical Physics 149, 034101 (2018), https://doi.org/10.1063/1.5027645.

[33] Z. Li, G. P. Wellawatte, M. Chakraborty, H. A. Gandhi, C. Xu, and A. D. White, "Graph neural network based coarse-grained mapping prediction," Chem. Sci. 11, 9524–9531 (2020), publisher: The Royal Society of Chemistry.

[34] P. Gkeka, G. Stoltz, A. Barati Farimani, Z. Belkacemi, M. Ceriotti, J. D. Chodera, A. R. Dinner, A. L. Ferguson, J.-B. Maillet, H. Minoux, C. Peter, F. Pietrucci, A. Silveira, A. Tkatchenko, Z. Trstanova, R. Wiewiora, and T. Lelièvre, "Machine learning force fields and coarse-grained variables in molecular dynamics: Application to materials and biological systems," Journal of Chemical Theory and Computation 16, 4757–4775 (2020), pMID: 32559068, https://doi.org/10.1021/acs.jctc.0c00355.

[35] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, and A. Aspuru-Guzik, "Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules," ACS Cent. Sci. 4, 268–276 (2018).

[36] F. Noé, S. Olsson, J. Köhler, and H. Wu, "Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning," Science 365, eaaw1147 (2019).

[37] N. E. Jackson, A. S. Bowen, L. W. Antony, M. A. Webb, V. Vishwanath, and J. J. de Pablo, "Electronic structure at coarse-grained resolutions from supervised machine learning," Sci. Adv. 5, eaav1190 (2019).

[38] N. E. Jackson, A. S. Bowen, and J. J. de Pablo, "Efficient Multiscale Optoelectronic Prediction for Conjugated Polymers," Macromolecules 53, 482–490 (2020).

[39] G. Sivaraman and N. E. Jackson, "Coarse-grained density functional theory predictions via deep kernel learning," J. Chem. Theory Comput. 18, 1129–1141 (2022).

[40] C. E. Rasmussen and C. K. I. Williams, Gaussian processes for machine learning, Adaptive computation and machine learning (MIT Press, Cambridge, Mass, 2006) oCLC: ocm61285753.

[41] E. Snelson and Z. Ghahramani, "Sparse Gaussian Processes using Pseudo-inputs," in Advances in Neural Information Processing Systems, Vol. 18 (MIT Press, 2005).

[42] J. Hensman, N. Fusi, and N. D. Lawrence, "Gaussian processes for big data," in Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence, UAI'13 (AUAI Press, Arlington, Virginia, USA, 2013) p. 282–290.

[43] M. Titsias, "Variational Learning of Inducing Variables in Sparse Gaussian Processes," in Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics (PMLR, 2009) pp. 567–574, iSSN: 1938-7228.

[44] M. Jankowiak, G. Pleiss, and J. Gardner, "Parametric Gaussian Process Regressors," in Proceedings of the 37th International Conference on Machine Learning (PMLR, 2020) pp. 4702–4712, iSSN: 2640-3498.

[45] S. Plimpton, "Fast parallel algorithms for short-range molecular dynamics," Journal of Computational Physics 117, 1–19 (1995).

[46] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives, "Development and testing of the OPLS all-atom force field on conformational energetics

and properties of organic liquids," J. Am. Chem. Soc. 118, 11225–11236 (1996).

[47] F. Neese, F. Wennmohs, U. Becker, and C. Riplinger, "The ORCA quantum chemistry program package," J. Chem. Phys. 152, 224108 (2020).

[48] A. K. Soper, "Empirical potential monte carlo simulation of fluid structure," Chem. Phys. 202, 295–306 (1996).

[49] D. Burt, C. E. Rasmussen, and M. V. D. Wilk, "Rates of convergence for sparse variational gaussian process regression," in Proceedings of the 36th International Conference on Machine Learning (PMLR, 2019) pp. 862–871, ISSN: 2640-3498.

[50] A. G. Wilson, Z. Hu, R. Salakhutdinov, and E. P. Xing, "Deep kernel learning," in Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (PMLR, 2016) pp. 370–378, ISSN: 1938-7228.

[51] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," arXiv:1312.6114 [cs, stat] (2014), 1312.6114.

[52] J. R. Gardner, G. Pleiss, D. Bindel, K. Q. Weinberger, and A. G. Wilson, "GPyTorch: Blackbox Matrix-Matrix Gaussian Process Inference with GPU Acceleration," arXiv:1809.11165 [cs, stat] (2021), arXiv: 1809.11165.

[53] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in Advances in Neural Information Processing Systems 32, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (2019) pp. 8024–8035.

[54] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res. 12, 2825–2830 (2011).

[55] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," Biometrika 52, 591–611 (1965).

[56] Y. Tan, N. C. Casetti, B. W. Boudouris, and B. M. Savoie, "Molecular design features for charge transport in nonconjugated radical polymers," Journal of the American Chemical Society 143, 11994–12002 (2021), pMID: 34279095, https://doi.org/10.1021/jacs.1c02571.

[57] H. Bässler, D. Kroh, F. Schauer, V. Nádaždy, and A. Köhler, "Mapping the density of states distribution of organic semiconductors by employing energy resolved–electrochemical impedance spectroscopy," Advanced Functional Materials 31, 2007738 (2021), https://onlinelibrary.wiley.com/doi/pdf/10.1002/adfm.202007738.

[58] A. Davtyan, G. A. Voth, and H. C. Andersen, "Dynamic force matching: Construction of dynamic coarse-grained models with realistic short time dynamics and accurate long time dynamics," The Journal of Chemical Physics 145, 224107 (2016), https://doi.org/10.1063/1.4971430.

[59] S. W. Ober, C. E. Rasmussen, and M. v. d. Wilk, "The promises and pitfalls of deep kernel learning," in Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence (PMLR, 2021) pp. 1206–1216, ISSN: 2640-3498.

[60] J. Quiñonero-Candela and C. E. Rasmussen, "A Unifying View of Sparse Approximate Gaussian Process Regression," J. Mach. Learn. Res. 6, 1939–1959 (2005).

[61] A. Khot, S. B. Shiring, and B. M. Savoie, "Evidence of information limitations in coarse-grained models," J. Chem. Phys. 151, 244105 (2019).

[62] B. Anderson, T. S. Hy, and R. Kondor, "Cormorant: Covariant molecular neural networks," in Advances in Neural Information Processing Systems, Vol. 32, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Inc., 2019).

[63] K. Tran, W. Neiswanger, J. Yoon, Q. Zhang, E. Xing, and Z. W. Ulissi, "Methods for comparing uncertainty quantifications for material property predictions," Mach. Learn.: Sci. Technol. 1, 025006 (2020).

[64] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, "SchNet – A deep learning architecture for molecules and materials," J. Chem. Phys. 148, 241722 (2018).

[65] W. G. Noid, J.-W. Chu, G. S. Ayton, V. Krishna, S. Izvekov, G. A. Voth, A. Das, and H. C. Andersen, "The multiscale coarse-graining method. I. A rigorous bridge between atomistic and coarse-grained models," J Chem Phys 128, 244114 (2008).

[66] A. P. Bartók, S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csányi, and M. Ceriotti, "Machine learning unifies the modeling of materials and molecules," Sci. Adv. 3, e1701816 (2017).