ACCELERATED BAYESIAN INFERENCE FOR MOLECULAR SIMULATIONS USING LOCAL GAUSSIAN PROCESS SURROGATE MODELS

A PREPRINT

B. L. Shanks, H. W. Sullivan, . R. Shazed, M. P. Hoepfner
Department of Chemical Engineering
University of Utah
Salt Lake City, UT
. L. Shanks brennon.shanks@chemeng.utah.edu
M. P. Hoepfner michael.hoepfner@utah.edu

April 2, 2024

BSTR CT

While Bayesian inference is the gold standard for uncertainty quantification and propagation, its use within physical chemistry encounters formidable computational barriers. These bottlenecks are magnified for modeling data with many independent variables, such as X-ray/neutron scattering patterns and electromagnetic spectra. To address this challenge, we employ local Gaussian process (LGP) surrogate models to accelerate Bayesian optimization over these complex thermophysical properties. The time-complexity of the LGPs scales linearly in the number of independent variables, in stark contrast to the computationally expensive cubic scaling of conventional Gaussian processes. To illustrate the method, we trained a LGP surrogate model on the radial distribution function of liquid neon and observed a 1,760,000-fold speed-up compared to molecular dynamics simulation, beating a conventional GP by three orders-of-magnitude. We conclude that LGPs are robust and efficient surrogate models, poised to expand the application of Bayesian inference in molecular simulations to a broad spectrum of experimental data.

1 Introduction

Molecular simulations are able to estimate a broad array of complex experimental observables, including scattering patterns from neutron and X-ray sources and spectra from near-infrared [1], terahertz [2], sum frequency generation [3, 4], and nuclear magnetic resonance [5]. Recent interest in these experiments to study hydrogen bonding networks of water at interfaces [6, 7], electrolyte solutions [8], and biological systems [9] has motivated the continued advancement of simulations to calculate these properties from first-principles [10–12]. However, the ability to estimate these complex properties comes with a high computational cost. This barrier greatly limits our ability to quantify how experimental, model, and parametric uncertainty impact molecular simulation predictions, making it difficult to know whether a model is an appropriate representation of nature or if it is simply over-fitting to a given training set. Therefore, what is needed is a computationally efficient and rigorous uncertainty quantification/propagation (UQ/P) method to link molecular models to large and complex experimental datasets.

Bayesian methods are the gold standard for these aims [13], with examples spanning from neutrino and dark matter detection [14], materials discovery and characterization [15–18], quantum dynamics [19, 20], to molecular simulation [21–31]. The Bayesian probabilistic framework is a rigorous, systematic approach to quantify probability distribution functions on model parameters and credibility intervals on model predictions, enabling robust and reliable parameter optimization and model selection [32, 33]. Interest in Bayesian methods and uncertainty quantification for molecular simulation has surged [34–39] due to its flexible and reliable estimation of uncertainty, ability to identify weaknesses or missing physics in molecular models, and systematically quantify the credibility of simulation predictions. Additionally,

standard inverse methods including relative entropy minimization, iterative Boltzmann inversion, and force matching have been shown to be approximations to a more general Bayesian field theory [40].

The biggest problem plaguing Bayesian inference is its massive computational cost. The two major pinch points are (1) sampling in high-dimensional spaces, commonly known as the "curse of dimensionality", and (2) the large number of model evaluations required to get accurate uncertainty estimates. In computational chemistry, these bottlenecks are magnified since these models are typically expensive. Therefore, rigorous and accurate uncertainty estimation is challenging, or even impossible, without accelerating the simulation prediction time. One way to achieve this speed-up is by approximating simulation outputs with an inexpensive machine learning model. These so-called surrogate models have been developed from neural networks [29, 41], polynomial chaos expansions [42, 43], configuration-sampling-based methods [44] and Gaussian processes [45–47].

Gaussian processes (GPs) are a compelling choice as surrogate models thanks to several distinct advantages. GPs are non-parametric, kernel-based function approximators that can interpolate function values in high-dimensional input spaces. GPs with an appropriately selected kernel also have analytical derivatives and Fourier transforms, making them well-suited for physical quantities such as potential energy surfaces [48, 49]. Additionally, kernels can encode physics-informed prior knowledge, alleviating the "black box" nature inherent to many machine learning algorithms. In fact, a comparison of various nonlinear regressors for molecular representations of ground-state electronic properties in organic molecules demonstrated that kernel regressors drastically outperformed other techniques, including convolutional graph neural networks [50].

Perhaps the most widely adopted application of GP surrogate models in computational chemistry is for model optimization. In the last decade, GP surrogates of simple thermophysical properties including density, heat of vaporization, enthalpy, diffusivity and pressure have been used for force field design [51–56]. However, to our knowledge there are no Bayesian optimization studies that apply GP surrogate models to thermophysical properties with many independent variables, such as structural correlation functions or electromagnetic spectra. In this work, independent variables (IVs) are defined as the fixed quantities over which a measurement is made (*e.g.* frequencies along a spectrum or radial positions along a radial distribution function) and the outcomes of those measurements are referred to as quantities-of-interest (QoIs).

Measurements of complex QoIs with many IVs are often available or easily obtained, yet are rarely included as observations in Bayesian optimization of molecular models. One reason why this may be the case is that previous literature has not outlined accurate and robust approaches to design Gaussian process surrogates for such data. For example, Angelikopoulous and coworkers did not use GP surrogate models for their Bayesian analysis on the radial distribution function (RDF) of liquid Ar [51], despite the fact that doing so would significantly reduce computation time. It is likely that GPs have not been previously used for complex QoIs due to high training and evaluation costs. Specifically, GPs have a cubic time-complexity in the number of IVs, which quickly becomes prohibitively expensive as experimental measurements obtain higher ranges and resolutions.

Local Gaussian processes (LGPs) are an emerging class of accelerated GP methods that are well-equipped to handle large sets of experimental data. These so-called "greedy" Gaussian process approximations are constructed by separating a GP into a subset of GPs trained at distinct locations in the input space [46, 57–59]. Computation on the LGP subset scales linearly with the number of IVs, is trivially parallelizable, and easily implemented in high-performance computing (HPC) architectures [60, 61]. State-of-the-art LGP models have been used to design Gaussian approximation potentials (GAPs) [62], a type of machine learning potential used to study atomic [63–65] and electron structures [62, 66], as well as nuclear magnetic resonance chemical shifts [67] with uncertainty quantification [34]. However, to our knowledge LGPs have not been applied as surrogate models for UQ/P on complex experimental data in computational chemistry.

In this study, we detail a simple and effective surrogate modeling approach for complex experimental observables common in physical chemistry. LGPs unlock the capability for existing Bayesian optimization schemes to incorporate complex data efficiently and accurately at a previously inaccessible computational scale. The key feature of the LGP surrogate model is the reduction in time-complexity with respect to the number of QoIs from cubic to linear, resulting in orders-of-magnitude speed-ups to evaluate complex observable surrogate models and perform posterior estimation. The computational speed-up results from reducing the dimensionality of matrix operations and therefore enables Bayesian UQ/P on experimental data with many IVs. For illustration, consider that a typical Fourier transformed infrared spectroscopy (FT-IR) measurement may contain data between 4000-400 cm^{-1} at a resolution of 2 cm^{-1}, giving a total number of QoIs around $\eta = 1800$. According to the time-complexity scaling in η, a LGP is estimated to accelerate this computation compared to a standard GP by approximately 3,240,000x. Source code and a tutorial on building LGP surrogate models is provided on GitHub.

To demonstrate the method, we trained a LGP surrogate model on the RDF of the (λ-6) Mie fluid and performed Bayesian optimization to fit the parameters of the Mie fluid model to a neutron scattering derived RDF for liquid neon

(Ne). The LGP was found to accelerate the $\eta = 73$ independent variable surrogate model calculation approximately 1,760,000x faster than molecular dynamics (MD) and 2100x faster than a conventional GP with accuracy comparable to the uncertainty in the reported experimental data. Bayesian posterior distributions were then calculated with Markov chain Monte Carlo (MCMC) and used to draw conclusions on model behavior, uncertainty, and adequacy. Surprisingly, we find evidence that Bayesian inference conditioned on the radial distribution function significantly constrains the (λ-6) Mie parameter space, highlighting opportunities to improve force field optimization and design based on neutron scattering experiments.

2 Computational Methods

In the following sections, an outline of standard approaches for Bayesian inference and surrogate modeling with Gaussian processes is presented. Then, we describe the local Gaussian process approximation and highlight key differences in their implementation and computational scaling.

2.1 Bayesian Inference

Bayes' law, derived from the definition of conditional probability, is a formal statement of revising one's prior beliefs based on new observations. Bayes' theorem for a given model, set of model input parameters, θ, and set of experimental QoIs, $\mathbf{y}$, is expressed as,

$$p\ \theta|\mathbf{y}) \propto p\ \mathbf{y}|\theta)p\ \theta) \tag{1}$$

where $p\ \theta)$ is the 'prior' probability distribution over the model parameters, $p\ \mathbf{y}|\theta)$ is the 'likelihood' of observing $\mathbf{y}$ given parameters θ, and $p\ \theta|\mathbf{y})$ is the 'posterior' probability that the underlying parameter θ models or explains the observation $\mathbf{y}$. Equality holds in eq (1) if the right-hand-side is normalized by the 'marginal likelihood', $p\ \mathbf{y})$, but including this term explicitly is unnecessary since the posterior probability distribution can be normalized *post hoc*. In molecular simulations, θ is the set of unknown parameters in the selected model, usually the force field parameters in the Hamiltonian, to the experimental QoI that the simulation estimates. The observations, $\mathbf{y}$, can be any QoI or combination of QoIs (*e.g.* RDFs, spectra, densities, diffusivities, etc). This construction, known as the standard Bayesian scheme, is generalizable to any physical model and its corresponding parameters including density functional theory (DFT), *ab initio* molecular dynamics (AIMD), and path integral molecular dynamics (PIMD).

Calculating the posterior distribution then just requires prescription of prior distributions on the model input parameters and evaluation of the likelihood function. In this work, Gaussian distributions are used for both the prior and likelihood functions, which is a standard choice according to the central limit theorem. The Gaussian likelihood has the form,

$$p\ \mathbf{y}|\theta) = \quad \frac{1}{\sqrt{2\pi}\sigma_n}\Big)^{\eta} \exp\Big[\quad \frac{1}{2\sigma_n^2}\sum_{i=1}^{\eta}\left[\mathbf{y}_\theta \quad \mathbf{y}\right]^2\Big] \tag{2}$$

where η is the number of observables in $\mathbf{y}$, $\mathbf{y}_\theta$ is the model predicted observables at model input θ, and σ_n is a nuisance parameter describing the unknown variance of the Gaussian likelihood. Cailliez and coworkers choose the nuisance parameter as the sum of simulation and experiment variances ($\sigma_n^2 \approx \sigma_{sim}^2 + \sigma_{exp}^2$) [52]; however, if these variances are unknown or one wishes to explore the distribution of variances, the nuisance parameter can be inferred via the Bayesian inference. Hence, the resulting posterior distribution on the nuisance parameter includes the unknown uncertainty arising due to the sum of the model and the experimental variances. In this work, the nuisance parameter is treated as an unknown to be inferred along with the explicit model parameters. Note that in some cases a different likelihood function may be more appropriate based on physics-informed prior knowledge of the distribution of the observable of interest (*e.g.* the multinomial likelihood in relative entropy minimization between canonical ensembles [68]).

The computationally expensive part of calculating eq 2 is determining $\mathbf{y}_\theta$ at a sufficient number of points in the parameter space. Generally, this can be achieved by calculating $\mathbf{y}_\theta$ at dense, equally spaced points in the parameter space of interest (grid method), sampling the parameter space with Markov chain Monte Carlo (MCMC) to estimate the posterior with a histogram (approximate sampling method), or assuming that the posterior distribution has a specific functional form (*i.e.* Laplace approximation). Regardless of the selected method, each of these posterior distribution characterization techniques require a prohibitive number of molecular simulations to adequately sample the parameter space (often on the order of 10^5 10^6), which is infeasible for even modest sized molecular systems.

2.2 Gaussian Process Surrogate Models

Gaussian processes accelerate the Bayesian likelihood evaluation by approximating $\mathbf{y}_\theta$ with an inexpensive matrix calculation. A Gaussian process is a stochastic process such that every finite set of random variables (position, time, etc) has a multivariate normal distribution [45]. The joint distribution over all random variables in the system therefore defines a functional probability distribution. The expectation of this distribution maps a set of model parameters, θ^*, and IVs, $\mathbf{r}$, to the most probable QoI given the model parameters, $S\ \mathbf{r}|\theta^*)$, such that,

$$[GP] : \theta^* \times \mathbf{r} \mapsto S\ \mathbf{r}|\theta^*) \tag{3}$$

where the expectation operator is written in terms of a kernel matrix, $\mathbf{K}$, training set parameter matrix, $\hat{\mathbf{X}}$, and training set output matrix, $\hat{\mathbf{Y}}$, according to the equation,

$$[GP\ \theta^*, \mathbf{r})] = \mathbf{K}_{\ \theta^*,\mathbf{r}),\hat{\mathbf{X}}}[\mathbf{K}_{\hat{\mathbf{X}},\hat{\mathbf{X}}} + \sigma^2_{noise}\mathbf{I}]^{\ 1}\hat{\mathbf{Y}} \tag{4}$$

where σ^2_{noise} is the variance due to noise and $\mathbf{I}$ is the identity matrix. Note that in general the IVs, $\mathbf{r}$, can be multi-dimensional. As an example, consider the case a GP maps a set of force field parameters to the angular RDF of a liquid. We now have a 2-dimensional space of IVs since the angular RDF gives the atomic density along the radial and angular dimensions. In the following mathematical development, it is assumed that the QoI is 1-dimensional for sake of convenience and note that extending the method to higher-dimensional observables just requires redefining the IVs in accordance with eq (4).

The kernel matrix, $\mathbf{K}$, quantifies the relatedness between input parameters and can be selected based on prior knowledge of the physical system. A standard kernel for physics-based applications is the squared-exponential (or radial basis function) since the resulting GP is infinitely differentiable, smooth, continuous, and has an analytical Fourier transform [69]. The squared-exponential kernel function between input points $\theta_m, r_m)$ and $\theta_n, r_n)$ is given by,

$$K_{mn} = \ ^2 \exp\left(\frac{r_m \quad r_n)^2}{2\ell_r^2} \quad \sum_{o=1}^{\dim\ \theta)} \frac{\theta_{o,m} \quad \theta_{o,n})^2}{2\ell_{\theta_o}^2} \right) \tag{5}$$

where o indexes over $\dim(\theta)$ and the hyperparameters ^{2} and ℓ are the kernel variance and correlation length scale of parameter , respectively. Hyperparameter optimization can be performed by log marginal likelihood maximization, k-fold cross validation [45] or marginalization with an integrated acquisition function [70], but can be computationally expensive and is usually avoided if accurate estimates of the hyperparameters can be made from prior knowledge of the chemical system.

To train a standard GP surrogate model, N training samples are generated in the input parameter space and a molecular simulation is performed for each training set sample to calculate N predictions over the number of target QoIs, η. The training set, $\hat{\mathbf{X}}$, is then a $(N\eta \times \dim(\theta) + 1)$ matrix of the following form,

$$\hat{\mathbf{X}} = \begin{bmatrix} \theta_{1,1} & \theta_{2,1} & & r_1 \\ \theta_{1,1} & \theta_{2,1} & & r_2 \\ \vdots & \vdots & \vdots & \vdots \\ \theta_{1,1} & \theta_{2,1} & & r_\eta \\ \theta_{1,2} & \theta_{2,2} & & r_1 \\ \vdots & \vdots & \vdots & \vdots \\ \theta_{1,N} & \theta_{2,N} & & r_\eta \end{bmatrix} \tag{6}$$

where the $\theta_{i,j}$ are the i^{th} model parameter for sample index j and r_k are the IVs of the target QoI. Note that the training sample index, $j = 1, \quad ,N$, is updated in the model parameters only after η rows spanning the domain of the observable, giving $N\eta$ total rows. Therefore, the training set matrix represents all possible combinations of the training parameters in the θ parameter input space. The training set observations, $\hat{\mathbf{Y}}$, are a $(N\eta \times 1)$ column vector of the observable outputs from the training set,

$$\hat{\mathbf{Y}} = [S\ \theta_1, r_1), \quad ,S\ \theta_1, r_\eta), S\ \theta_2, r_1), \quad ,S\ \theta_N, r_\eta)]^T \tag{7}$$

where $S\ \theta_j, r_k) = y\ \theta_j, r_k) \quad \mu_{GP}^{prior}\ \theta_j, r_k)$ is the difference between the training set observation of model parameters θ_j at IV r_k and a GP prior mean function. Of course, the GP prior mean, μ_{GP}^{prior}, is the same shape as the training set observations matrix,

$$\mu_{GP}^{prior} := [\mu\ \theta_1, r_1), \quad ,\mu\ \theta_1, r_\eta), \mu\ \theta_2, r_1), \quad ,\mu\ \theta_N, r_\eta)]^T \tag{8}$$

where $\mu\ \theta_j, r_k)$ is the GP prior mean for parameter set θ_j at r_k. Note that the selection of a prior mean can impact the quality of fit of the GP surrogate model and should reflect physically justified prior knowledge of the physical system.

Conceptually, since a Gaussian process is a Bayesian model, the prior serves as a current state of knowledge that can encode an initial guess for the QoI before the GP sees any training data. The subtraction of the GP prior mean from the model output effectively shifts the QoI by this pre-specified mean function. Hence, the GP is trained on these mean shifted observations rather than the observations themselves. Although shifting the data by another function seems like it shouldn't change the ability of the GP to estimate the QoI, it actually can have an important impact on the stochastic properties of the data as a function of the IVs. By construction, GPs are stationary, meaning that the means, variances, and covariances are assumed to be equal along all QoI. But for complex data, this is often not the case. For example, it is known that the RDF is zero for small r values and has asymptotic tailing behavior to unity at long-range. The GP prior mean effectively shifts this non-stationary data and makes it behave as if it were stationary by removing any r dependencies.

The expectation of the GP for a new set of parameters, $S^*\ \mathbf{r}|\theta^*)$, is then a ($\eta$ x 1) column vector calculated with eq (4),

$$S^*\ \mathbf{r}|\theta^*) = [S^*\ r_1|\theta^*), \quad ,S^*\ r_\eta|\theta^*)]^T \tag{9}$$

where $S^*\ \mathbf{r}|\theta^*)$ is the most probable difference function between the model and GP prior mean. Hence, to obtain a comparison to the experimental QoI you simply add the GP prior mean at θ^*, $\mu_{GP}^{*,prior}\ \theta^*, \mathbf{r})$, back to $S^*\ \mathbf{r}|\theta^*)$.

The GP expectation calculation is burdened by the inversion of the training-training kernel matrix with $\mathscr{O}\ N^3\eta^3)$ time complexity and the $(\eta \times N\eta) \times (N\eta \times N\eta) \times (N\eta \times 1)$ matrix product with $\mathscr{O}\ N^2\eta^3)$ time complexity. Note that these estimates are for naive matrix multiplication. Regardless, the cubic scaling in η dominates the time-complexity for observables with many QoIs. For example, to build a GP surrogate model for the density of a noble gas ($\eta = 1$) with Lennard-Jones interactions (dim(θ) = 2) would give a training set matrix of $(2N \times 3)$. Similarly, a surrogate model for an infrared spectrum of water from 600-4000 cm ^{1} at a resolution of 4 cm ^{1} ($\eta = 850$) estimated with a 3 point water model of Lennard-Jones type interactions (dim(θ) = 6) would generate a training set matrix of size ($850N \times 7$). Clearly, the complexity of the output QoI causes a significant increase in the computational cost of the matrix operations.

2.3 The Local Gaussian Process Surrogate Model

The time-complexity of the training-kernel matrix inversion and the matrix product can be substantially reduced by fragmenting the full Gaussian process of eq (4) into η Gaussian processes. This method is also referred to as the subset of regressors approximation [71] and is considered a "greedy" approximation [45]. Under this construction, an individual GP_k is trained to map a set of model parameters to an individual QoI,

$$[GP_k] : \theta \mapsto S\ r_k) \tag{10}$$

where $\mathbf{r}$ is no longer an input parameter. The training set matrix, $\hat{\mathbf{X}}'$, is now a $(N \times \dim(\theta))$ matrix,

$$\hat{\mathbf{X}}' = \begin{bmatrix} \theta_{1,1} & \theta_{2,1} & \\ \theta_{1,2} & \theta_{2,2} & \\ \vdots & \vdots & \vdots \\ \theta_{1,N} & \theta_{2,N} & \end{bmatrix} \tag{11}$$

while the training set observations, $\hat{\mathbf{Y}}'_k$, is a $(N \times 1)$ column vector of the QoIs from the training set at r_k,

$$\hat{\mathbf{Y}}'_k = [S\ \theta_1, r_k), \quad ,S\ \theta_N, r_k)]^T \tag{12}$$

where $S\ \theta_j, r_k) = y\ \theta_j, r_k) \quad \mu_{LGP,k}^{prior}\ r_k)$ and k indexes over IVs. The LGP prior mean $\mu_{LGP,k}^{prior}\ r_k)$ is now,

$$\mu_{LGP,k}^{prior} := [\mu\ \theta_1, r_k), \quad ,\mu\ \theta_N, r_k)]^T \tag{13}$$

such that $\mu\ \theta_j, r_k)$ is the GP prior mean for parameter θ_j at r_k. The squared-exponential kernel function is now,

$$K_{mn} = \ ^2 \exp \quad \sum_{o=1}^{\dim\ \theta)} \frac{\theta_{o,m} \quad \theta_{o,n})^2}{2\ell_{\theta_o}^2} \Bigg) \tag{14}$$

The LGP surrogate model expectation for the observable at r_k, at a new set of parameters, θ^*, is just the expectation of the k^{th} Gaussian process given the training set data,

$$S_{loc}^*\ r_k|\theta^*) = \quad [GP_k\ \theta^*)] = \mathbf{K}_{\theta^*,\hat{\mathbf{X}}'}[\mathbf{K}_{\hat{\mathbf{X}}',\hat{\mathbf{X}}'} + \sigma_{noise}^2\mathbf{I}] \ \ ^1\hat{\mathbf{Y}}'_k \tag{15}$$

We then just combine the local results from the subset of η GPs to obtain a prediction for the difference between the model and LGP prior mean,

$$S_{loc}^*\ \mathbf{r}|\theta^*) = [S_{loc}^*\ r_1|\theta^*), \quad ,S_{loc}^*\ r_\eta|\theta^*)]^T \tag{16}$$

and subsequently add back the LGP prior mean to obtain the estimated QoI, $y_{loc}^*\ \mathbf{r}|\theta^*) = S_{loc}^*\ \mathbf{r}|\theta^*) + \mu_{LGP,k}^{prior}\ \theta^*, \mathbf{r})$.

By reducing the dimensionality of the relevant matrices, the time complexity of the matrix calculations are drastically reduced compared to a standard GP. The single step inversion of the training-training kernel matrix is now of $\mathcal{O}\ N^3)$ time complexity while the η step $(1 \times N) \times (N \times N) \times (N \times 1)$ matrix products are reduced to $\mathcal{O}\ N^2\eta)$ time complexity. If the number of training samples, N, the number of IVs, η, and the number of model evaluations, G, are equal between the full and LGP algorithms, then a LGP approximation reduces the evaluation time complexity in a standard GP from cubic-scaling, η^3, to embarrassingly parallelizable linear-scaling, η.

In summary, a local Gaussian process is an approximation in which the QoIs are modeled as independent random variables, each described by their own Gaussian process. This amounts to assuming that the random variables are stochastically independent. For time-independent data including scattering measurements and spectroscopy, this approximation is appropriate since each observation is an independent measurement at each independent variable. Finally, it is well-established that low rank approximations of Gaussian processes can compromise the accuracy of the estimated uncertainty, so the use of LGP regressors should be carefully scrutinized based on the risk/consequences of misrepresenting the resulting functional distributions.

Complex experimental observables can be reconstructed by this set of LGPs through a series of relatively straightforward matrix operations with linear time-complexity in the number of IVs. Furthermore, the LGP has all of the primary advantages of Bayesian methods, including built-in UQ and analytical derivatives and Fourier transforms. In the following section, we demonstrate the computational enhancement and accuracy of the LGP approach by modeling the RDF of neon at $42K$. The LGP surrogate model is then implemented within a Bayesian framework to exemplify the power of UQ/P for molecular modeling.

3 Local Gaussian Process Surrogate for the RDF of Liquid Ne

To explore the computational advantages of LGP surrogate models for Bayesian inference, we studied the experimental RDF of liquid Ne [72] under a (λ-6) Mie fluid model. The (λ-6) Mie force field is a flexible Lennard-Jones type potential with variable repulsive exponent,

$$v_2^{Mie}\ r) = \frac{\lambda}{\lambda \quad 6} \quad \frac{\lambda}{6}\Bigg)^{\frac{6}{\lambda\ 6}} \varepsilon \Bigg[\quad \frac{\sigma}{r}\Bigg)^{\lambda} \quad \frac{\sigma}{r}\Bigg)^{6}\Bigg] \tag{17}$$

where λ is the short-range repulsion exponent, σ is the collision diameter ( ), and ε is the dispersion energy (kcal/mol) [73].

MD simulations were performed from a Sobol sampled set spanning a prior range based on existing force field models [74–76] ($\lambda = [6\ 1, 18]$, $\sigma = [0\ 88, 3\ 32]$, and $\varepsilon = [0, 0\ 136]$) to generate a RDF training set matrix of the form in eq 11. Prior parameter ranges were selected so that training samples were restricted to the liquid regime of the (λ-6) Mie phase

Table 1: Average relative time and speed-up to QoI evaluation and training set matrix inversion for a standard and local Gaussian process for 960 training samples and a RDF with $\eta = 73$ points.

Model	QoI Eval. Time (s)	Speed Up (t/t_{sim})	Inv. Time (s)
Simulation	1,251	1	-
GP	1.52	822	355
LGP	0.0007	1,760,267	0.01

diagram [77, 78]. A sequential sampling approach was used in which we Sobol sample the prior range of parameters, calculate the training sample with the best-fit to the experimental data (lowest root mean squared error), center the new space on this training sample, and then narrow the sample range around this center point by a user selected ratio γ. This procedure was repeated three times with 320 samples per round (960 total training simulations) with $\gamma = 0\ 8$. This ratio was selected so that the final range would span >3 standard deviations of the posterior distributions estimated in prior literature [51, 75]. Subsequently, 320 test simulations were randomly sampled from the final range and used to determine whether or not the surrogate model provides accurate model predictions. A visualization of this procedure is provided in the Supporting Information.

The number of observed points η in the radial distribution function was calculated by dividing the reported $r_{max} \quad r_{min} \approx 15\ 3$ by the effective r-space resolution given by, $\quad r = \pi/Q_{max}$, where $\quad r = 0\ 21 \quad$ for reported $Q_{max} = 15 \quad ^{1}$. This relation indicates that the appropriate number of observed independent r-values in the RDF is $\eta = 73$.

The training set matrix and training observation matrix were then constructed from the 960 training samples according to eqs (11) and (12), respectively. As a prior mean, we selected the RDF determined analytically from the dilute limit potential of mean force (PMF),

$$\mu^{prior}_{PMF,k}\ \theta_j, r_k) := g\ \theta_j, r_k) = \exp[\ \ \beta V\ \theta_j, r_k)] \tag{18}$$

where $g\ \theta_j, r_k)$ and $V\ \theta_j, r_k)$ are the analytical dilute limit RDF and (λ-6) Mie potential for parameters θ_j at r_k, respectively. A PMF prior mean yields physically realistic short-range ($g\ r) = 0$) and long-range behavior ($g\ r) \to 1$). The PMF prior had improved RMSE compared to an ideal gas prior ($\forall r \in \mathbb{R}^{+}_{0},\ g\ r) = 1$), but this difference did not significantly impact the Bayesian posterior estimate (see Supporting Information). Finally, LGP hyperparameter optimization was performed using brute force to maximize the log-marginal likelihood [79] over the training set.

Quantitative analysis of model sensitivity can be performed with probabilistic derivatives of the QoI with respect to model parameters (see Supporting Information) and subsequently related to temperature derivatives of radial distribution functions [80].

3.1 Computational Efficiency and ccuracy

Now that we have constructed the training set matrix, we simply evaluate the expectation at each r_k according to eq (15) and combine the results into a single array as in eq (16). The average computational time to invert the training set matrix and evaluate the surrogate model for both a standard GP and LGP are shown below in Table 1. The LGP surrogate accelerates the RDF evaluation time compared to molecular dynamics by a factor of 1,700,000 for the $\eta = 73$ independent variable QoI with 960 training simulations. This 6 orders-of-magnitude speed-up beats a standard GP by 3 orders-of-magnitude (2141x). With respect to the training-training kernel matrix inversion, the LGP wins out on the standard GP by a factor of 31,565.

In summary, the LGP significantly accelerates both computational bottlenecks for Gaussian process surrogate modeling; namely, the training set matrix inversion and surrogate model evaluation time. Of course, the exact speed-ups depend on numerous factors including the number of IVs η, the number of training samples used to construct the training set matrix N, the level of code parallelization, and hyperparameter optimization procedure. Which step is rate limiting depends on the surrogate modeling application. For instance, if the surrogate model doesn't need to be evaluated a large number of times, the training set generation, matrix inversion and hyperparameter optimization will be the rate limiting steps. On the other hand, applications that require a large number of model evaluations, such as uncertainty quantification and propagation, result in the surrogate model evaluation time being rate limiting. Typically, designing a surrogate model is only necessary in the latter case.

Clearly the LGP is fast, but is it accurate? In other words, does the LGP provide QoI predictions that are within a reasonable level of accuracy to serve as a true surrogate model for the molecular dynamics predictions? To evaluate the accuracy of the local predictions, a test set of 320 (λ-6) Mie parameters was randomly sampled from the final range of

the sequential sampling method (see Supporting Information) and the RMSE computed between simulated and LGP predicted radial distribution functions along all radial positions, r. The results are summarized below in Figure 1.

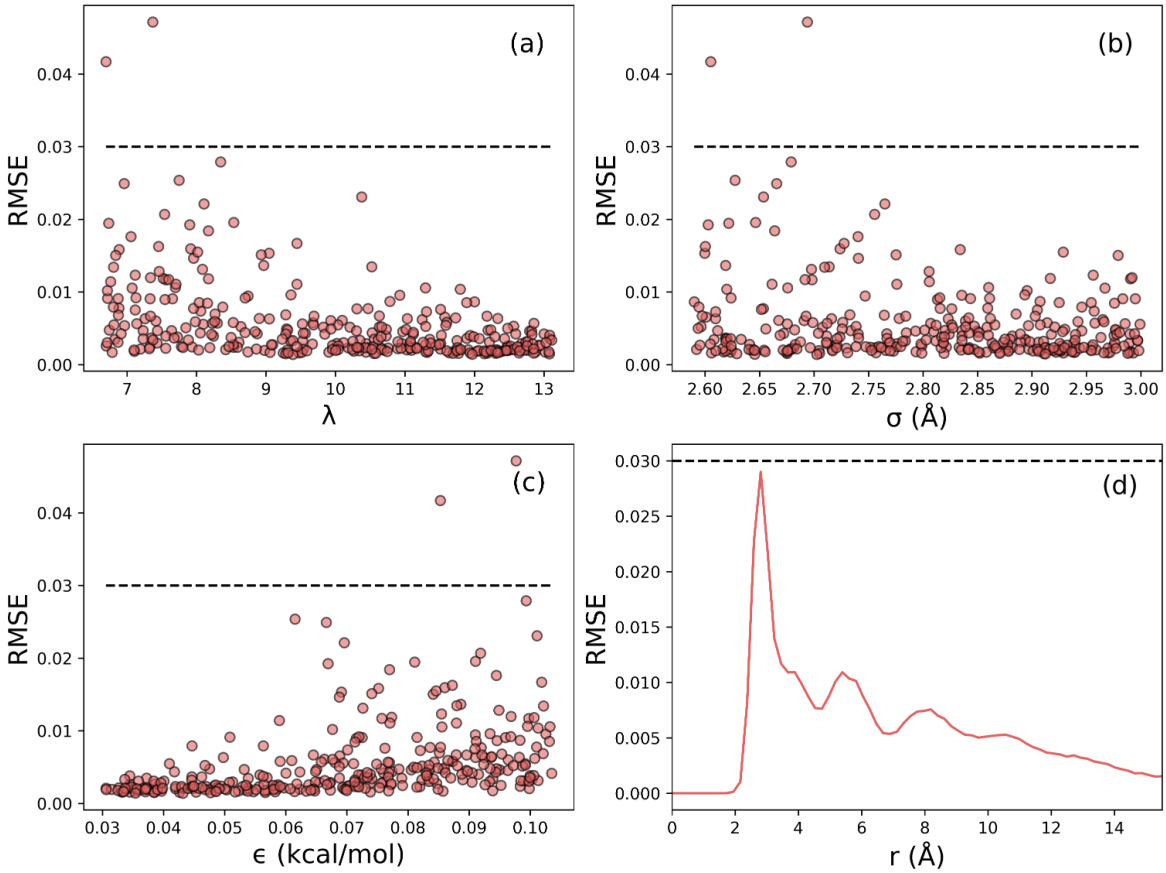

Figure 1: (a)-(c) Test set samples over each parameter plotted against the RMSE between simulated and LGP data. (d) Average RMSE over the 320 test set samples as a function of r. The dashed line represents the reported error from the experiment [72].

The RMSE for all radial positions is less than 0.03, which is excellent considering that this error is smaller than the reported experimental uncertainty ($\sim$0.03). Of course, the acceptable RMSE over the QoI is user-defined and largely subjective based on the surrogate model application, but can be improved with additional training and hyperparameter optimization if necessary (an example is included in the Supporting Information).

3.2 Learning from the Ne RDF Surrogate Model with Bayesian nalysis

Our fast and accurate LGP surrogate model now allows us to explore the underlying probability distributions on the (λ-6) Mie parameter space. This example is provided to show how one can use Bayesian analysis to learn about correlations and relationships between model parameters as well as model adequacy. This analysis can provide robust insight into the nature of the model and provide quantifiable evidence for whether or not the model is appropriate for a target application. Bayesian inference yields a probability distribution function over the model parameters called the joint posterior probability distribution. The maximum of the joint posterior, referred to as the *maximum a posteriori* (M P), represents the set of parameters with the highest probability of explaining the given experimental data. In force field design, the M P would be an appropriate choice for an optimal set of model parameters. However, the power of the Bayesian approach lies in the fact that, not only can we identify the optimal parameters, but we can also examine the probability distribution of the parameters around these optima. For instance, the width of the distribution provides evidence for how important a parameter in the model is for representing the target data. For a given parameter, a wide distribution indicates that the parameter has little influence on the model prediction. On the other hand, a narrow distribution indicates that the parameter is critical to the model prediction. Additionally, the joint posterior may exhibit