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.

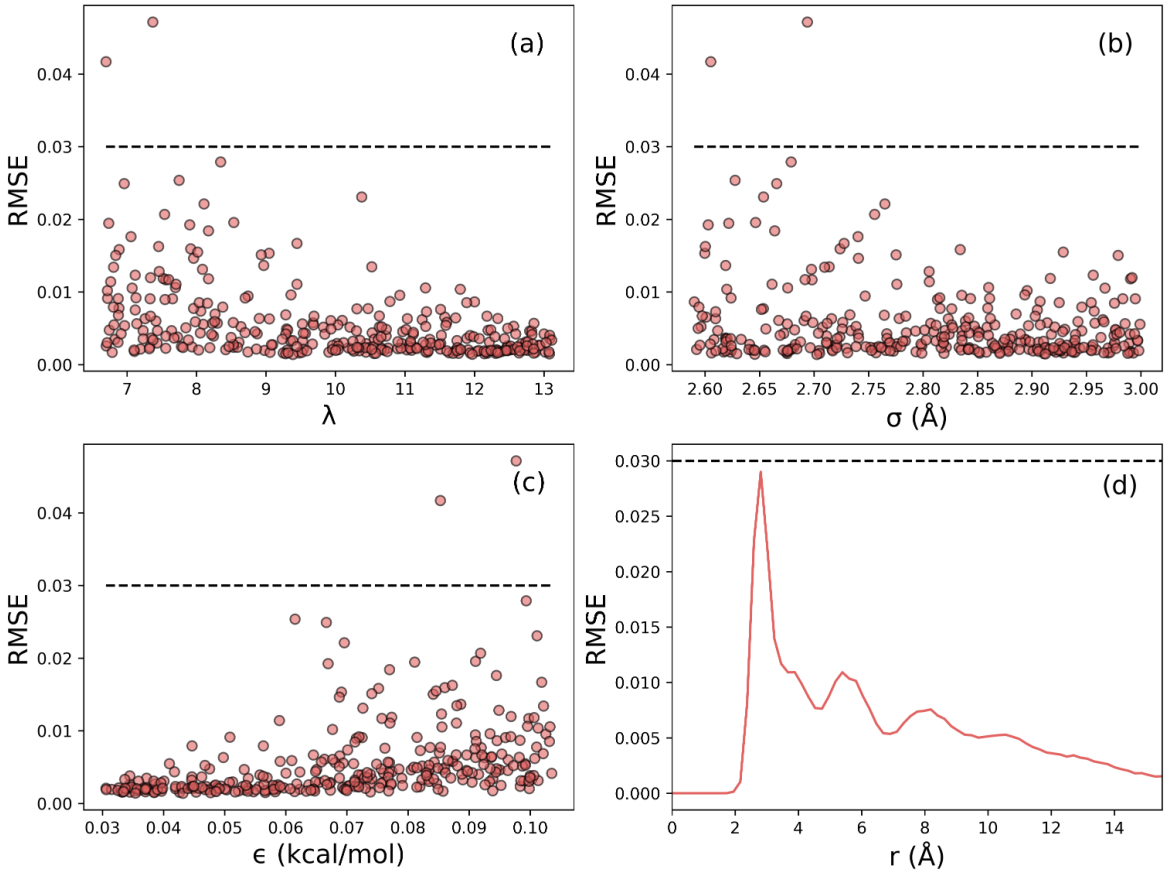

Figure 1: (a)-(c) Test set samples over each parameter plotted against the RMSE between simulated and LGP data. (d) Average RMSE over the 320 test set samples as a function of r. The dashed line represents the reported error from the experiment [72].

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

multiple peaks, or modes. A multimodal joint posterior suggests that there are multiple sets of model parameters that reproduce the target data, which may be a symptom of model inadequacy. Finally, the symmetry of the distribution provides information on relationships and correlations between parameters, providing a framework to diagnose subtle relationships that may otherwise go unnoticed.

Usually, the joint posterior distribution is a high-dimensional quantity that cannot be visualized directly. However, we can visualize the joint posterior along one dimension by integrating out the contributions over all other parameters. The resulting distributions are called marginal distributions. Marginal distributions computed over the (λ-6) Mie potential parameters optimized to the RDF of liquid Ne are shown in Figure 2.

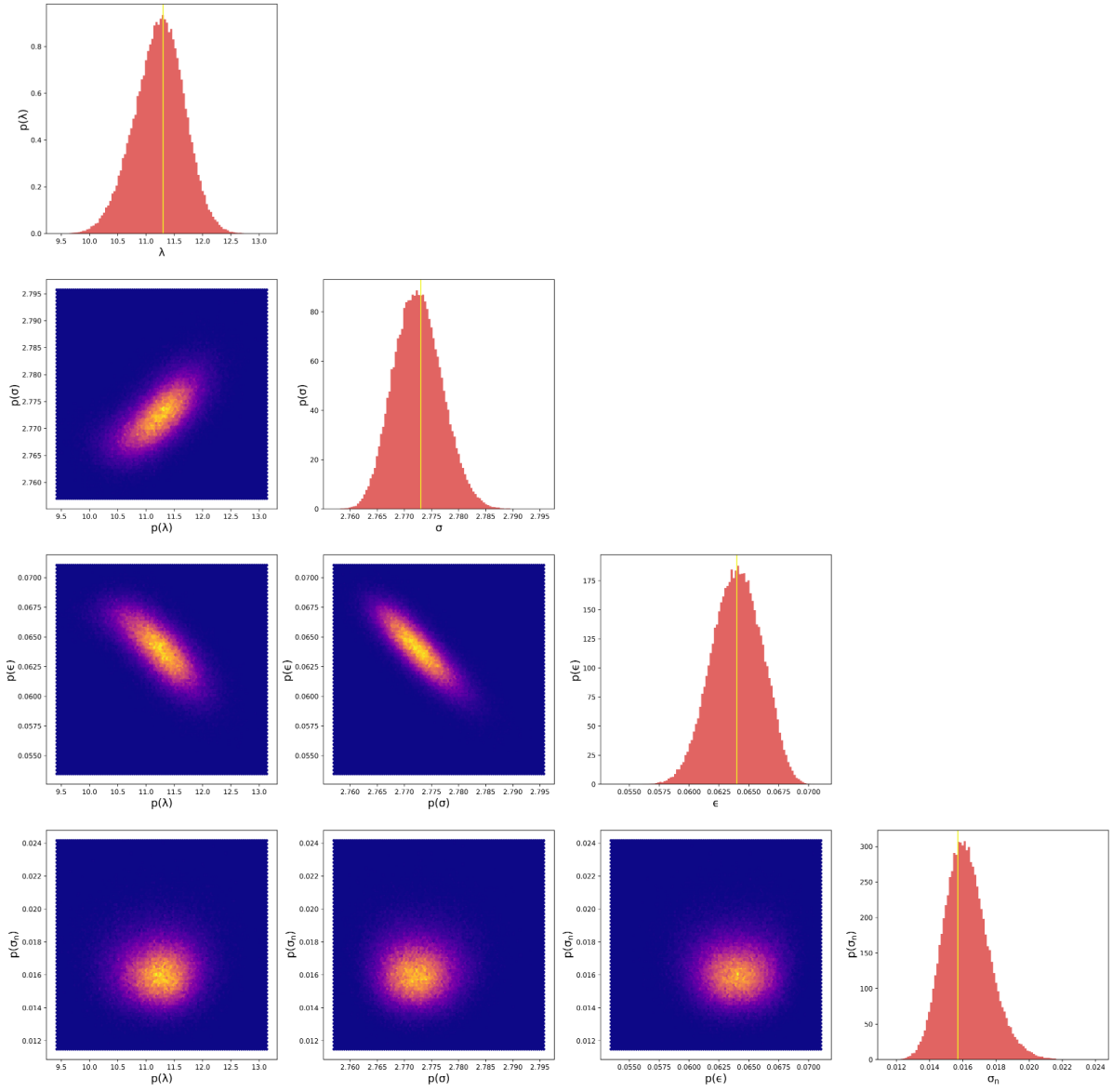

Figure 2: (Diagonal) 1D marginal distributions for the (λ-6) Mie fluid parameters. Prior distributions are not depicted since they are flat lines near 0 probability. Yellow vertical lines represent the *maximum a posterior* ($M\ \ P$) estimate. (Off-Diagonal) 2D marginal histograms showing parameter correlations.

For each parameter, the resulting marginal posterior distributions are unimodal and symmetric. This result is not surprising in the context of recent results that show iterative Boltzmann inversion, which is a maximum likelihood approach to the structural inverse problem, is convex for Lennard-Jones type fluids [81]. Observing the 2D marginal

distributions in Figure 2, we can also see that each of the parameters are correlated to one other. For example, the negative correlation between σ and ε suggests that increasing the size of the particle should be accompanied by a decrease in the effective particle attraction. Conceptually this makes sense, if the particles are larger, then they would need to have a weaker attractive force to give the same atomic structure. This result is consistent with Bayesian analysis on liquid Ar [51]. The nuisance parameter distribution shows that the unknown standard deviation between the LGP surrogate model and the experimental data is around 0.016.

One surprising characteristic of the posterior distribution is that it is extremely narrow. Recall that narrow distributions indicate that the parameters are important, or have tight control, over the model quality-of-fit to the experimental data. From our Bayesian analysis, we can therefore confidently conclude that detailed interatomic force information is contained within the experimental RDF. This observation is in stark contrast to over 60 years of prior literature which has unanimously asserted that only the excluded volume or collision diameter can be ascertained from experimental scattering data [82–84]. In fact, the Bayesian analysis shows that it is possible to determine values for λ, σ, and ε within ± 2, ± 0 02 , and ± 0 0075 kcal/mol with 95% certainty. This result leads to two important conclusions: (1) Scattering data can effectively constrain the force field model parameter space and (2) the data must be sufficiently accurate to do so. These results provide evidence that scattering data could be invaluable to inform accurate force fields, particularly for structure and self-assembly applications.

The joint posterior can also be used for model parameter selection given the experimental observation. Specifically, the optimal parameters are given by the $M \;\; P$, corresponding to the maximum of the joint posterior distribution. The $M \;\; P$ is presented in Table 2 along with two other existing force fields for liquid Ne.

Table 2: Summary of ($\lambda \quad 6$) Mie potential parameters optimized for Ne. Values for the repulsive exponent parameter are rounded to the nearest integer.

Force Field	QoI	λ	σ ( )	ε (kcal/mol)
Mick (2015)	VLE	11	2.794	0.064
SOPR (2022)	RDF	11	2.778	0.063
This Work	RDF	11	2.773	0.064

The estimated Mie parameters are in agreement with the Mick [75] and structure optimized potential refinement (SOPR) [76] models. This result confirms that the radial distribution function contains sufficient information to determine transferable force field parameters in simple liquids.

Some interesting questions arise considering that both the Mie fluid model and SOPR, which is a probabilistic iterative Boltzmann method for experimental scattering data, give similar predictions for the structure-optimized potentials. The key difference between the Bayesian optimization performed in this work and SOPR is that the former is parametric while the latter is non-parametric, both of which have strengths and weaknesses. Specifically, parametric models are less complex but may not be flexible enough to describe subtle details of the experimental observation. On the other hand, non-parametric models can describe nuanced experiments but may over-fit to non-physical features of the data. It is then natural to wonder: Is a (λ-6) Mie model adequate to describe the experimental scattering data? Or does the scattering data complexity necessitate the use of non-parametric iterative potential refinement techniques like SOPR?

We can investigate the first question of model adequacy by propagating parameter uncertainty through the LGP to construct a distribution of RDF predictions - referred to as the posterior predictive. The posterior predictive can be estimated by evaluating the LGP for all MCMC samples and computing the mean,

$$[S^*_{loc} \;\; \mathbf{r}_k)] \approx \frac{1}{N} \sum_{i=1}^{N} S^*_{loc} \;\; \mathbf{r}_k|\theta_i) \tag{19}$$

and variance,

$$\mathbb{V}[S^*_{loc} \;\; \mathbf{r}_k)] \approx \frac{1}{N} \sum_{i=1}^{N} \; S^*_{loc} \;\; \mathbf{r}_k|\theta_i) \qquad [S^*_{loc} \;\; \mathbf{r}_k)])^2 \tag{20}$$

of the resulting QoI predictions. Recall that the nuisance parameter distribution is also sampled to account for unknown uncertainties in the LGP surrogate model and experimental data. The posterior predictive therefore quantifies of how accurately we know the QoI given experimental, model, and parametric uncertainty estimated with Bayesian

inference. If the model is adequate, the Bayesian credibility interval ($\mu \pm 2\sigma$) should contain approximately 95% of the experimental data. The posterior predictive and residuals ($g_{exp}\ r)\quad \mu\ r)$) estimated for the liquid Ne RDF are shown in Figure 3.

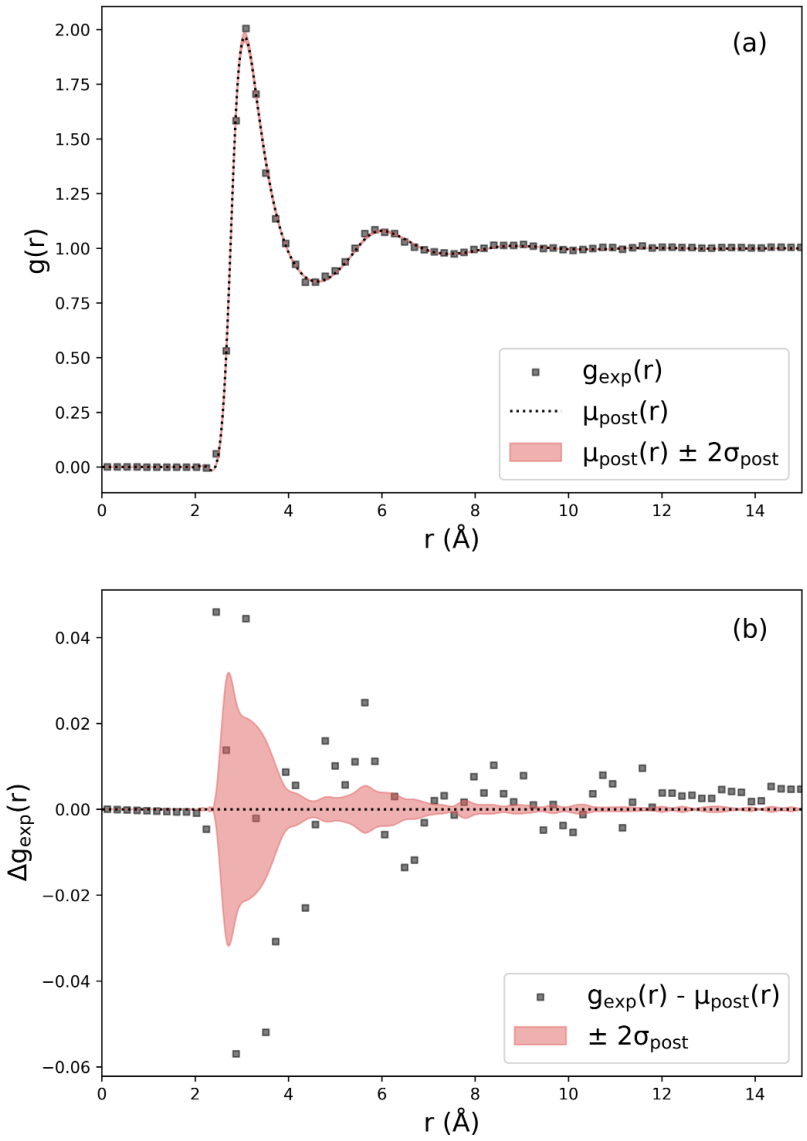

Figure 3: (a) RDF mean and credibility interval propagated from the parameter uncertainty quantified with Bayesian inference. (b) Residual analysis comparing the experimental data with the posterior predictive distribution.

Clearly, the agreement between the posterior predictive mean and the experimental data is excellent. However, the residuals often lie outside of the $2\sigma_{post}$ credibility interval. These differences between the experiment and model could be explained by a number of different factors, including errors arising from Fourier transform truncation, background scattering corrections or model inadequacy, among others. However, without rigorous uncertainty quantification on the experimental scattering data, it is currently not possible to determine which factor or combination of factors results in the model disagreement. We argue that this knowledge gap necessitates rigorous UQ/P studies on scattering data as well as iterative potential refinement methods. Combining these approaches with Bayesian inference on molecular dynamics models could then shed light on what physical interactions can be learned from scattering experiments.

In summary, we have shown that a LGP surrogate model enables rapid and accurate uncertainty quantification and propagation with Bayesian inference. We then showed how the posterior distribution is an indispensable tool to learn

subtle relationships between model parameters, identify how important each model parameter is to describe the outcome of experiments, and quantify our degree of belief that our model adequately describes our observations. The power of Bayesian inference is evident.

4 Conclusions

We have shown that local Gaussian process surrogate models trained on an experimental RDF of liquid neon improves the computational speed of QoI prediction 1,760,000-fold with exceptional accuracy from only 960 training simulations. The 3 orders-of-magnitude evaluation time speed-up for a local versus standard Gaussian process was shown to accelerate Bayesian inference without the need for advanced sampling techniques such as on-the-fly learning. Furthermore, since the LGP linearly scales with the number of output QoIs, significantly higher speed-ups are expected for more complex data, such as infrared spectra or high resolution scattering experiments, or for multiple data sources simultaneously (*e.g.* scattering, spectra, density, diffusivity, etc). We conclude that local Gaussian processes are an accurate and reliable surrogate modeling approach that can accelerate Bayesian analysis of molecular models over a broad array of complex experimental data.

5 cknowledgements

We would like to thank Dr. Jürgen Dölz for helpful discussions on the implementation of local Gaussian processes, hyperparameter optimization and uncertainty quantification. Dr. Valeria Molinero for providing feedback on writing and scientific communication in preparation of the manuscript. Finally, we thank Sean T. Smith and Philip J. Smith for their guidance on Bayesian methodology. This study is supported by the National Science Foundation Award No. CBET-1847340. The support and resources from the Center for High Performance Computing at the University of Utah are gratefully acknowledged.

6 uthor Contributions

BL Shanks - conceptualization, formal analysis, code development, manuscript writing and preparation. HW Sullivan - conceptualization, formal analysis, code development, manuscript preparation. AR Shazed - molecular simulations. MP Hoepfner - conceptualization, funding acquisition, manuscript preparation.

References

(1) Czarnecki, M. A.; Morisawa, Y.; Futami, Y.; Ozaki, Y. *Chem. Rev.* **2015**, *115*, 9707–9744.
(2) Schmuttenmaer, C. A. *Chem. Rev.* **2004**, *104*, 1759–1780.
(3) Nihonyanagi, S.; Yamaguchi, S.; Tahara, T. *Chem. Rev.* **2017**, *117*, 10665–10693.
(4) Hosseinpour, S.; Roeters, S. J.; Bonn, M.; Peukert, W.; Woutersen, S.; Weidner, T. *Chem. Rev.* **2020**, *120*, 3420–3465.
(5) Mishkovsky, M.; Frydman, L. *nnu. Rev. Phys. Chem.* **2009**, *60*, 429–448.
(6) Roget, S. A.; Carter-Fenk, K. A.; Fayer, M. D. . *m. Chem. Soc.* **2022**, *144*, 4233–4243.
(7) Li, P.; Jiang, Y.; Hu, Y.; Men, Y.; Liu, Y.; Cai, W.; Chen, S. *Nat. Catal* **2022**, *5*, 900–911.
(8) Wang, T.; Tian, Z.; You, Z.; Li, Z.; Cheng, H.; Li, W.; Yang, Y.; Zhou, Y.; Zhong, Q.; Lai, Y. *Energy Storage Mater.* **2022**, *45*, 24–32.
(9) Meng, W.; Peng, H.-C.; Liu, Y.; Stelling, A.; Wang, L. *J. Phys. Chem. B* **2023**, *127*, 2351–2361.
(10) Bally, T.; Rablen, P. R. *J. Org. Chem.* **2011**, *76*, 4818–4830.
(11) Thomas, M.; Brehm, M.; Fligg, R.; Vöhringer, P.; Kirchner, B. *Phys. Chem. Chem. Phys.* **2013**, *15*, 6608–6622.
(12) Gastegger, M.; Behler, J.; Marquetand, P. *Chem. Sci.* **2017**, *8*, 6924–6935.
(13) Psaros, A. F.; Meng, X.; Zou, Z.; Guo, L.; Karniadakis, G. E. *J. Comput. Phys.* **2023**, *477*, 111902.
(14) Eller, P.; Fienberg, A. T.; Weldert, J.; Wendel, G.; Böser, S.; Cowen, D. F. *Nucl. Instrum. Methods Phys. Res. : ccel. Spectrom. Detect. ssoc. Equip.* **2023**, *1048*, 168011.
(15) Todorovic, M.; Gutmann, M. U.; Corander, J.; Rinke, P. *Npj Comput. Mater.* **2019**, *5*.
(16) Zuo, Y.; Qin, M.; Chen, C.; Ye, W.; Li, X.; Luo, J.; Ong, S. P. *Mater. Today.* **2021**, *51*, 126–135.
(17) Fang, L.; Guo, X.; Todorović, M.; Rinke, P.; Chen, X. *J. Chem. Inf. Model.* **2023**, *63*, 745–752.

(18) Sharma Priyadarshini, M.; Romiluyi, O.; Wang, Y.; Miskin, K.; Ganley, C.; Clancy, P. *Mater. Horiz.* **2024**, *11*, 781–791.

(19) V. Krems, R. *Phys. Chem. Chem. Phys.* **2019**, *21*, 13392–13410.

(20) Deng, Z.; Tutunnikov, I.; Averbukh, I. S.; Thachuk, M.; Krems, R. V. *J. Chem. Phys.* **2020**, *153*, 164111.

(21) Frederiksen, S. L.; Jacobsen, K. W.; Brown, K. S.; Sethna, J. P. *Phys. Rev. Lett.* **2004**, *93*, 165501.

(22) Cooke, B.; Schmidler, S. C. *Biophys. J.* **2008**, *95*, 4497–4511.

(23) Cailliez, F.; Pernot, P. *J. Chem. Phys.* **2011**, *134*, 054124.

(24) Farrell, K.; Oden, J. T.; Faghihi, D. *J. Comput. Phys.* **2015**, *295*, 189–208.

(25) Wu, S.; Angelikopoulos, P.; Papadimitriou, C.; Moser, R.; Koumoutsakos, P. *Philos. Trans. R. Soc.* **2016**, *374*, 20150032.

(26) Patrone, P. N.; Dienstfrey, A.; Browning, A. R.; Tucker, S.; Christensen, S. *Polymer* **2016**, *87*, 246–259.

(27) Messerly, R. A.; Knotts, T. A.; Wilding, W. V. *J. Chem. Phys.* **2017**, *146*, 194110.

(28) Dutta, R.; Brotzakis, Z. F.; Mira, A. *J. Chem. Phys.* **2018**, *149*, 154110.

(29) Wen, M.; Tadmor, E. B. *Npj Comput. Mater.* **2020**, *6*, 1–10.

(30) Bisbo, M. K.; Hammer, B. *Phys. Rev. Lett.* **2020**, *124*, 086102.

(31) Xie, Y.; Vandermause, J.; Ramakers, S.; Protik, N. H.; Johansson, A.; Kozinsky, B. *Npj Comput. Mater.* **2023**, *9*, 1–8.

(32) Gelman, A.; Carlin, J. B.; Stern, H. S.; Rubin, D. B., *Bayesian Data nalysis*; Chapman and Hall/CRC: New York, 1995.

(33) Shahriari, B.; Swersky, K.; Wang, Z.; Adams, R. P.; de Freitas, N. *Proc. IEEE* **2016**, *104*, 148–175.

(34) Musil, F.; Willatt, M. J.; Langovoy, M. A.; Ceriotti, M. *J. Chem. Theory Comput.* **2019**, *15*, Publisher: American Chemical Society, 906–915.

(35) Cailliez, F.; Pernot, P.; Rizzi, F.; Jones, R.; Knio, O.; Arampatzis, G.; Koumoutsakos, P. In *Uncertainty Quantification in Multiscale Materials Modeling*; Elsevier: 2020, pp 169–227.

(36) Vandermause, J.; Torrisi, S. B.; Batzner, S.; Xie, Y.; Sun, L.; Kolpak, A. M.; Kozinsky, B. *Npj Comput. Mater.* **2020**, *6*, 1–11.

(37) Köfinger, J.; Hummer, G. *Eur. Phys. J. B* **2021**, *94*, 245.

(38) Vandermause, J.; Xie, Y.; Lim, J. S.; Owen, C. J.; Kozinsky, B. *Nat. Commun.* **2022**, *13*, 5183.

(39) Blasius, J.; Zaby, P.; Dölz, J.; Kirchner, B. *J. Chem. Phys.* **2022**, *157*, 014505.

(40) Lemm, J. C., *Bayesian Field Theory*; JHU Press: 2003.

(41) Li, C.; Gilbert, B.; Farrell, S.; Zarzycki, P. *J. Chem. Inf. Model.* **2023**.

(42) Ghanem, R. G.; Spanos, P. D., *Stochastic Finite Elements: Spectral pproach*; Courier Corporation: 2003.

(43) Jacobson, L. C.; Kirby, R. M.; Molinero, V. *J. Phys. Chem. B* **2014**, *118*, 8190–8202.

(44) Messerly, R. A.; Razavi, S. M.; Shirts, M. R. *J. Chem. Theory Comput.* **2018**, *14*, 3144–3162.

(45) Rasmussen, C. E.; Williams, C. K. I., *Gaussian processes for machine learning*; MIT Press: Cambridge, Mass, 2006.

(46) Nguyen-Tuong, D.; Seeger, M.; Peters, J. *dv Robot.* **2009**, *23*, 2015–2034.

(47) Burn, M. J.; Popelier, P. L. A. *J. Chem. Phys.* **2020**, *153*, 054111.

(48) Dai, J.; Krems, R. V. *J. Chem. Theory Comput.* **2020**, *16*, 1386–1395.

(49) Yang, N.; Hill, S.; Manzhos, S.; Carrington, T. *J. Mol. Spectrosc.* **2023**, *393*, 111774.

(50) Faber, F. A.; Hutchison, L.; Huang, B.; Gilmer, J.; Schoenholz, S. S.; Dahl, G. E.; Vinyals, O.; Kearnes, S.; Riley, P. F.; von Lilienfeld, O. A. *J. Chem. Theory Comput.* **2017**, *13*, 5255–5264.

(51) Angelikopoulos, P.; Papadimitriou, C.; Koumoutsakos, P. *J. Chem. Phys.* **2012**, *137*, 144103.

(52) Cailliez, F.; Bourasseau, A.; Pernot, P. *J. Comput. Chem.* **2014**, *35*, 130–149.

(53) Kulakova, L.; Arampatzis, G.; Angelikopoulos, P.; Hadjidoukas, P.; Papadimitriou, C.; Koumoutsakos, P. *Sci. Rep.* **2017**, *7*, 16576.

(54) Befort, B. J.; DeFever, R. S.; Tow, G. M.; Dowling, A. W.; Maginn, E. J. *J. Chem. Inf. Model.* **2021**, *61*, 4400–4414.

(55) C. Madin, O.; R. Shirts, M. *Digit. Discov.* **2023**, *2*, 828–847.

(56) Wang, N.; Carlozo, M. N.; Marin-Rimoldi, E.; Befort, B. J.; Dowling, A. W.; Maginn, E. J. *J. Chem. Theory Comput.* **2023**, *19*, 4546–4558.

(57) Das, K.; Srivastava, A. N. In *2010 IEEE International Conference on Data Mining*, 2010, pp 791–796.

(58) Park, C.; Apley, D. Patchwork Kriging for Large-scale Gaussian Process Regression, en, 2018.
(59) Terry, N.; Choe, Y. *PLoS One* **2021**, *16*, DOI: 10.1371/journal.pone.0256470.
(60) Gramacy, R. B.; Apley, D. W. *J. Comput. Graph. Stat.* **2015**, *24*, 561–578.
(61) Broad, J.; Wheatley, R. J.; Graham, R. S. *J. Chem. Theory Comput.* **2023**, *19*, 4322–4333.
(62) Deringer, V. L.; Bartók, A. P.; Bernstein, N.; Wilkins, D. M.; Ceriotti, M.; Csányi, G. *Chem. Rev.* **2021**, *121*, 10073–10141.
(63) Caro, M. A.; Deringer, V. L.; Koskinen, J.; Laurila, T.; Csányi, G. *Phys. Rev. Lett.* **2018**, *120*, 166101.
(64) Deringer, V. L.; Bernstein, N.; Bartók, A. P.; Cliffe, M. J.; Kerber, R. N.; Marbella, L. E.; Grey, C. P.; Elliott, S. R.; Csányi, G. *J. Phys. Chem. Lett.* **2018**, *9*, 2879–2885.
(65) L. Deringer, V.; Merlet, C.; Hu, Y.; Hoon Lee, T.; A. Kattirtzi, J.; Pecher, O.; Csányi, G.; R. Elliott, S.; P. Grey, C. *Chem. Comm.* **2018**, *54*, 5988–5991.
(66) Cheng, B.; Mazzola, G.; Pickard, C. J.; Ceriotti, M. *Nature* **2020**, *585*, 217–220.
(67) Paruzzo, F. M.; Hofstetter, A.; Musil, F.; De, S.; Ceriotti, M.; Emsley, L. *Nat. Commun.* **2018**, *9*, 4501.
(68) Shell, M. S. *J. Chem. Phys.* **2008**, *129*, 144108.
(69) Ambrogioni, L.; Maris, E. In *IST TS*, PMLR: 2018, pp 217–225.
(70) Snoek, J.; Larochelle, H.; Adams, R. P. In *dv. Neural Inf. Process.* Curran Associates, Inc.: 2012; Vol. 25.
(71) Silverman, B. W. *J. R. Stat. Soc. Ser. B Methodol.* **1985**, *47*, 1–21.
(72) Bellissent-Funel, M. C.; Buontempo, U.; Filabozzi, A.; Petrillo, C.; Ricci, F. P. *Phys. Rev. B* **1992**, *45*, 4605–4613.
(73) Mie, G. *nn. Phys.* **1903**, *316*, 657–697.
(74) Vrabec, J.; Stoll, J.; Hasse, H. *J. Phys. Chem. B* **2001**, *105*, 12126–12133.
(75) Mick, J. R.; Soroush Barhaghi, M.; Jackman, B.; Rushaidat, K.; Schwiebert, L.; Potoff, J. J. *J. Chem. Phys.* **2015**, *143*, 114504.
(76) Shanks, B. L.; Potoff, J. J.; Hoepfner, M. P. *J. Phys. Chem. Lett.* **2022**, *13*, 11512–11520.
(77) Widom, B.; Rowlinson, J. S. *J. Chem. Phys.* **1970**, *52*, 1670–1684.
(78) Ramrattan, N.; Avendaño, C.; Müller, E.; Galindo, A. *Mol. Phys.* **2015**, *113*, 932–947.
(79) Sundararajan, S.; Keerthi, S. S. *Neural. Comput.* **2001**, *13*, 1103–1118.
(80) Piskulich, Z. A.; Thompson, W. H. *J. Chem. Phys.* **2020**, *152*, 011102.
(81) Hanke, M. *J. Stat. Phys.* **2018**, *170*, 536–553.
(82) Clayton, G. T.; Heaton, L. *Phys. Rev.* **1961**, *121*, 649–653.
(83) Jovari, P. *Mol. Phys.* **1999**, *97*, 1149–1156.
(84) Hansen, J.-P.; McDonald, I. R., *Theory of Simple Liquids: with pplications to Soft Matter*; Academic Press: San Diego, 2013.
(85) Anderson, J. A.; Glaser, J.; Glotzer, S. C. *Comput. Mater. Sci.* **2020**, *173*, 109363.
(86) Ramasubramani, V.; Dice, B. D.; Harper, E. S.; Spellings, M. P.; Anderson, J. A.; Glotzer, S. C. *Comput. Phys. Commun.* **2020**, *254*, 107275.
(87) Heinonen, M.; Mannerström, H.; Rousu, J.; Kaski, S.; Lähdesmäki, H. In *Proceedings of the 19th International Conference on rtificial Intelligence and Statistics*, PMLR: 2016, pp 732–740.
(88) Angelikopoulos, P.; Papadimitriou, C.; Koumoutsakos, P. *Comput. Methods. ppl. Mech. Eng.* **2015**, *289*, 409–428.
(89) Foreman-Mackey, D.; Hogg, D. W.; Lang, D.; Goodman, J. *Publ. stron. Soc. Pac.* **2013**, *125*, 306.

7 ppendix

Molecular Dynamics Simulation of Mie Fluids

Computer generated radial distribution functions (RDFs) were calculated using molecular dynamics (MD) simulations in the HOOMD-Blue package [85]. Simulations were initiated with a lattice configuration of 864 particles and compressed to a reduced density of $\rho = 0\ 02477$ atom/ ^{3} and thermal energy $T = 42\ 2$ K. The HOOMD NVT integrator was used for a 0.25 nanosecond equilibration step and a 0.25 nanosecond production step (dt = 0.5 femtosecond). Potentials were truncated at 3σ with an analytical tail correction, and RDFs were calculated using the Freud package [86].

Table 3: Estimated boundaries for physics-constrained prior space based on the (λ - 6) Mie fluid phase diagram. $m = 6$ is the attractive tail exponent of the (λ - 6) Mie potential. $*$) The maximum λ was selected to be substantially larger than previously reported values [74–76].

Param.	Min.	Min. Criteria	Max.	Max. Criteria
λ	6.1	$m = 6 \implies \lambda > 6$	18	Literature*
σ	2.55	Vapor-Liquid Equil.	3.32	Solid-Liquid Equil.
ε	0.00	$\varepsilon < 0$ undefined	0.136	Vapor-Solid Equil.

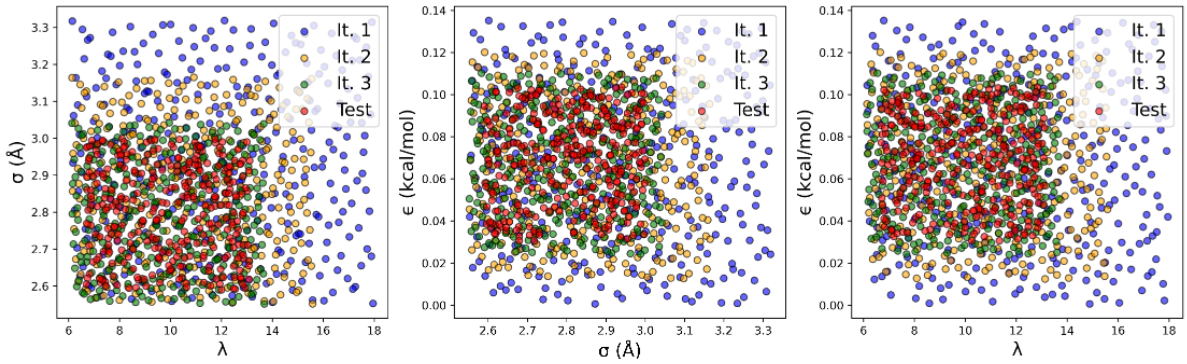

Figure 4: 2D training and sample parameter set used to train and test the LGP surrogate model.

B Training and Test Set Generation

The first step to design a LGP surrogate model is to generate a training set of model input parameter input and QoI outputs. To generate the training set, we need dense samples of model parameters in the region of the parameter space that well-represents the target experimental data. In general, it is not known *a priori* where this region is, particularly if there is no prior knowledge of what model parameters are best with respect to an experimental observation, $\mathbf{y}$. However, there are parameter regions that we can exclude *a priori* based on the physics of the (λ-6) Mie fluid. For instance, Ne is a liquid at the experimental thermodynamic conditions, so we can use well-established (λ-6) Mie fluid phase diagrams and vapor-liquid transitions [77] to restrict the parameter ranges to the liquid phase only. Specifically, given a fixed temperature (T = 42.2K) and density (ρ = 0.024 ^{3}), it is trivial to determine the σ and ε parameter ranges reported in the manuscript via relations for the scaled temperature ($T^* = k_b T/\varepsilon$) and scaled density ($\rho^* = \rho\sigma^3$). The parameter ranges determined using the Mie fluid phase diagram are presented in Table 3. Restricting the parameter to physically justified ranges is important to avoid a "garbage in, garbage out" scenario for an LGP surrogate model. Given this prior range, we then performed the sequential sampling approach outlined in the manuscript. A visualization of this procedure is shown below in Figure 4.

B.1 The Gaussian Process Prior Mean

In this manuscript, an analytical solution for the RDF based on the dilute limit potential of mean force (PMF) was used as the GP prior mean. This choice is appropriate as an RDF prior since it will have the same features that we expect a liquid RDF to have, *i.e.* RDF values of zero at low r and a long-range tail that asymptotically approaches unity. However, note that even a prior guess that doesn't encode this information can still produce accurate LGP surrogate models for RDFs. For example, in Figure 5 we can see that an ideal gas RDF prior, which amounts to approximating that the RDF is unity everywhere ($g^{IG}\ r) = 1$), can still be learned by the local Gaussian processes with RMSE values close to the more physically justified PMF.

Clearly, the RMSE along r is a similar magnitude for the ideal gas and PMF prior, but the r-dependent behavior is noticeably different. For the ideal gas prior, we see that there is high RMSE at low r, which is inconsistent with our intuition for a liquid RDF due to the excluded volume of atoms. What is occurring here is that the GP estimate is being "pulled" towards the prior at low r. On the other hand, the PMF prior exhibits behavior in line with our physical intuition; namely, near zero error at r values smaller than the relative diameter of the atom. Perhaps surprisingly, we see in Figure 6 that the choice of prior mean doesn't have a large impact on the posterior distribution or MAP estimates. We attribute this to the fact that the RMSE is sufficiently small for both the ideal gas and PMF priors that the posterior distribution isn't significantly modified. However, it does influence the posterior predictive distribution as evidenced by

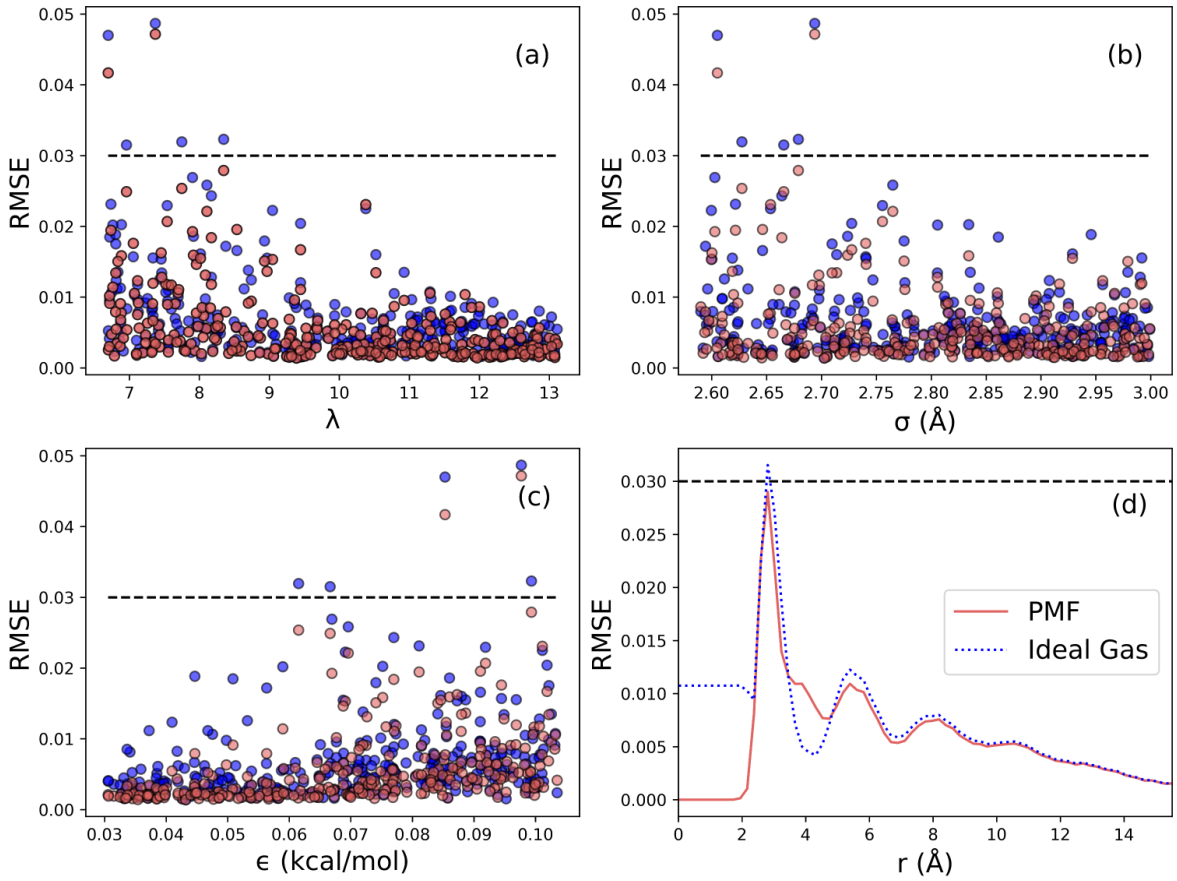

Figure 5: RMSE as a function of r for an ideal gas (blue) and potential of mean force (red) prior.

Table 4: Optimum hyperparameter values under the ideal gas and PMF prior computed from 25000 random samples over the reported test range.

Name	Test Range	Ideal Gas	PMF
ℓ_λ	0.5-4	3.31	3.58
ℓ_σ	0.01-0.05	0.046	0.048
ℓ_ε	0.001-0.01	0.0098	0.0093
	1E-4-0.1	0.094	0.095
σ_{noise}	1E-4-0.01	7.2E-4	8.3E-4

Figure 7. Specifically, note that there is uncertainty at low r for the ideal gas prior, whereas this uncertainty vanishes for the PMF prior.

B.2 Hyperparameter Selection

The final step is to learn a set of LGP hyperparameters that provide accurate estimates of the target QoI. A standard approach to selecting hyperparameters is to maximize the model evidence [45] or apply an expected improvement criterion based on an integrated acquisition function [70]. Here we applied a brute force search based on minimizing the leave-one-out (LOO) error for 25,000 hyperparameter options randomly sampled over a prior range using the method of Sundararajan and coworkers [79] (Table 4). This method gives relatively similar hyperparameter estimations for both an ideal gas and PMF GP prior. The prior range was selected based on the (λ-6) Mie parameter sensitivity analysis of Mick and coworkers [75].

A limitation of the brute force approach to hyperparameter selection is that we don't account for potential hyperparameter uncertainty in the LGP prediction. However, the self-consistency of our predictions with existing literature on liquid neon

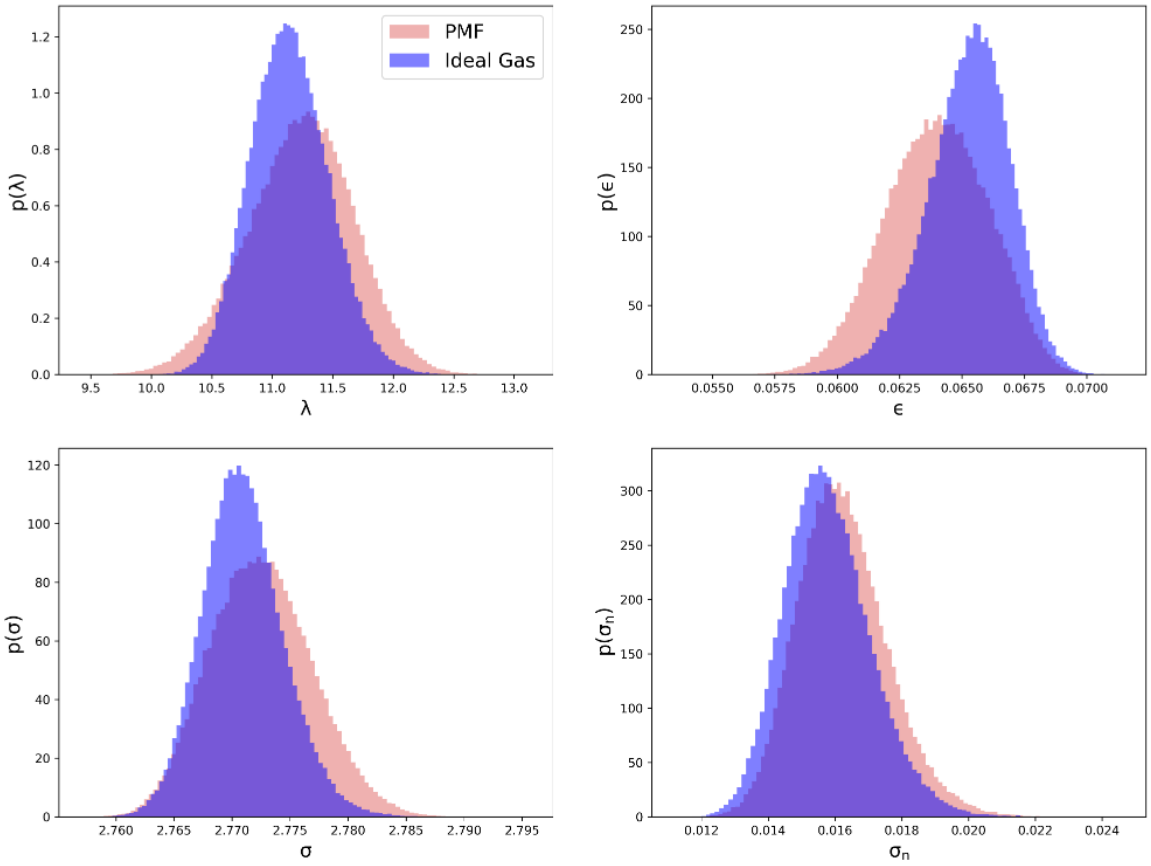

Figure 6: Marginal posteriors for the ideal gas PMF (red) and ideal gas (blue) priors.

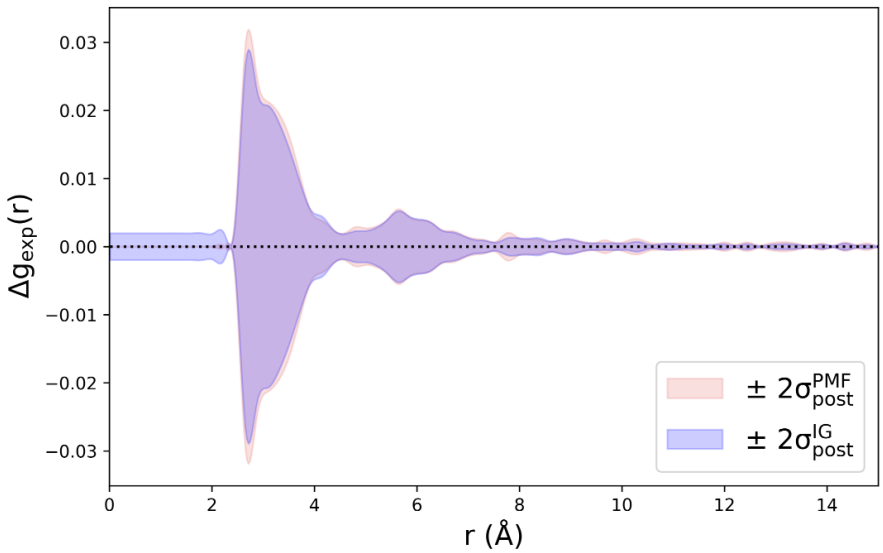

Figure 7: Posterior predictives for the ideal gas PMF (red) priors ideal gas (blue) priors.

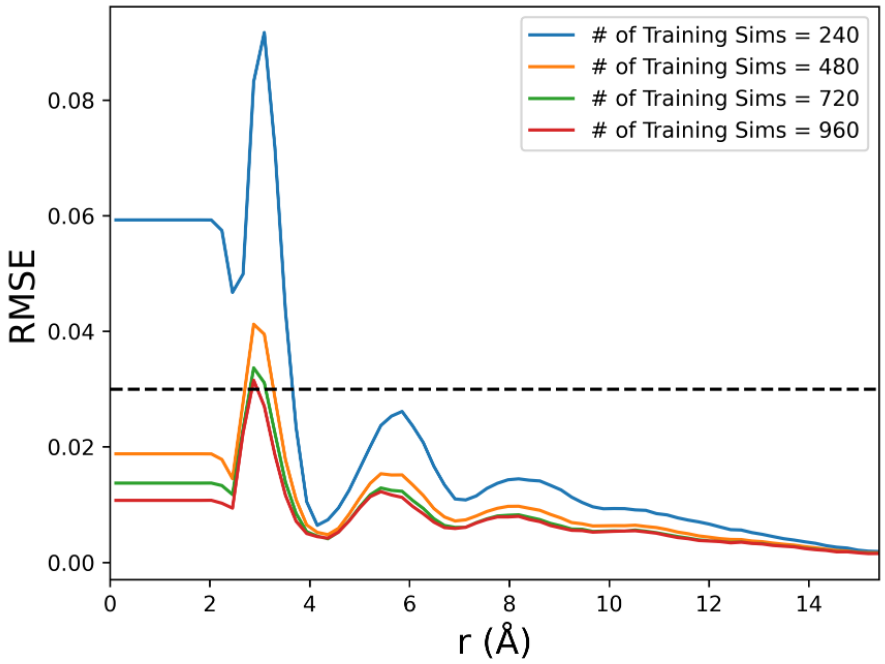

Figure 8: RMSE along r for varying numbers of training simulations under an ideal gas prior.

[75, 76] suggests that this uncertainty is likely insignificant. Note that one could propagate hyperparameter uncertainty by performing Bayesian optimization over the hyperparameters, sampling the resultant hyperparameter posterior distribution, and propagating the samples through the posterior predictive estimation step. Finally, hyperparameter optimization for the LGP model is non-trivial since the LGP is an approximation to a non-stationary stochastic process [87].

What if the previously described method fails to yield an accurate surrogate model? In this case, one can repeat the sequential sampling by adding more training simulations at each range to retrain the LGP until the RMSE is sufficiently small. As an example, Figure 8 demonstrates that surrogate model accuracy improves as more samples are added at each range. Note that the accuracy of the surrogate will not improve beyond the statistical uncertainty of the underlying model.

A more rigorous, but non-trivial method for surrogate model training, is to use adaptive or on-the-fly learning, in which the uncertainty in the LGP prediction is used to decide whether or not a new simulation is needed in the training set. This approach has been used in prior work[51, 88] but was found to be unnecessary for our purposes due to the efficiency and accuracy of the LGP with relatively few training samples.

B.3 Using the LGP Surrogate Model for Parameter Sensitivity nalysis

Sensitivity of a QOI to a model parameter, θ_i, can be quantified using the analytical derivative of the local GP surrogate model according to the following equation,

$$\frac{\partial \ [GP_k \ \theta^*)]}{\partial \theta_i} = \ \frac{\theta_i \quad \theta_i^*}{\ell^2_{\theta_i}}\Big)\mathbf{K}_{\theta^*,\hat{\mathbf{X}}'}[\mathbf{K}_{\hat{\mathbf{X}}',\hat{\mathbf{X}}'} + \sigma^2_{noise}\mathbf{I}] \ \ ^1\hat{\mathbf{Y}}'_k \tag{21}$$

where k is the QoI index. The Gaussian process derivative is a quantitative measure of the influence of a perturbation in θ_i to the expectation of the observable $\hat{\mathbf{Y}}'_k$. Therefore, eq (21) can approximate the impact of changing a molecular simulation parameter on the QOI (*e.g.* how much does changing the effective particle size impact the RDF at any r). Figure 9 shows probabilistic local GP derivatives calculated for the (λ-6) Mie parameters.

The repulsive exponent derivative exhibits a small magnitude and has a minimum at the RDF half maximum. This behavior suggests that increasing the repulsive exponent, which determines the "hardness" of the particles, steepens the slope of the first peak in the RDF. This result is intuitive considering that in a hard-particle model there is a

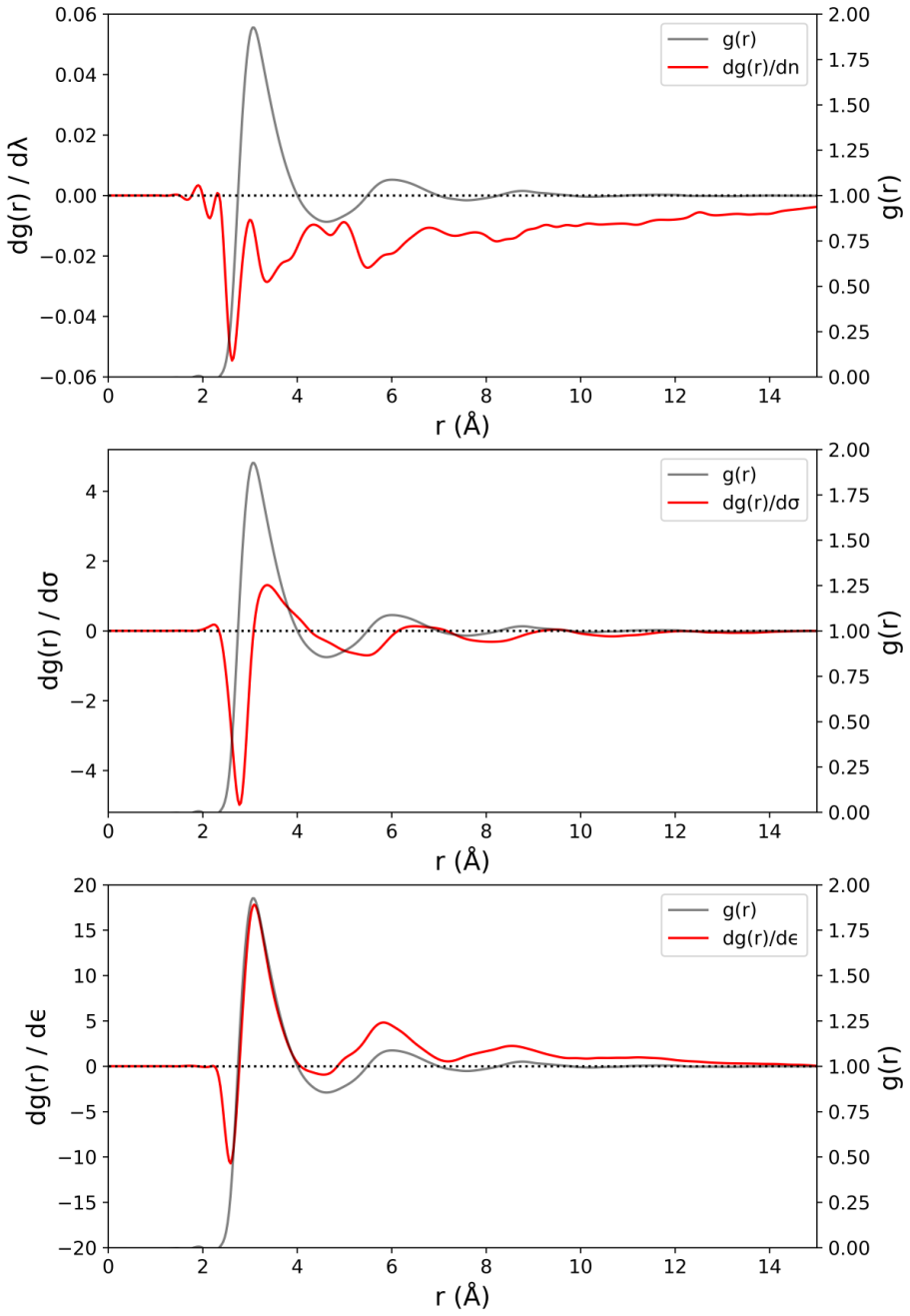

Figure 9: Derivatives of the local GP along the RDF calculated from eq (21).

Table 5: Prior parameters on the (λ-6) Mie model parameters.

Parameter	Distribution	μ	s
λ	Normal	12.0	9
σ		2.7	1.8
ε		0.112	0.225
σ_n	Log-Normal	1	1

discontinuous jump at the hard-particle radius (infinite slope) that progressively softens with the introduction of an exponential repulsive decay function. In the case of the collision diameter, zeros of the derivative occur at RDF peaks and troughs, while local extrema align with the half-maximum positions. Consequently, increasing the effective particle size shifts the RDF to the right while maintaining relatively constant peak heights. Regarding the dispersion energy, its derivative displays zeros at the half-maximum positions of the RDF and local extrema at peaks and troughs. This behavior indicates that an increase in the dispersion energy leads to an increased magnitude of the RDF peaks and greater liquid structuring.

Derivatives of structure with respect to thermodynamic state variables (T, P, μ, etc) can be computed with fluctuation theory. Let's now take as an example the ε-derivative of the RDF in Ne. We find that an increase in the dispersion energy deepens the interatomic potential well, resulting in greater attraction and a more structured liquid. Noting that the reduced temperature, T^*, is inversely related to ε by,

$$T^* = \frac{k_B T}{\varepsilon} \tag{22}$$

then the $g\ r)$ derivative with respect to ε at constant T, is equal to the $g\ r)$ derivative with respect to the reduced thermodynamic beta,

$$\frac{\partial g\ r)}{\partial \varepsilon} = \frac{\partial g\ r)}{\partial \beta^*} \tag{23}$$

where $\beta^* = T^*/k_B T$. In summary, an increase in ε is equivalent to a decrease in temperature. It is therefore expected that the ε derivative and temperature derivative behave the same; specifically, a decrease in temperature should increase result in greater fluid structuring without significantly impacting peak positions. Unsurprisingly, this behavior is exactly what was observed in recent work that computed temperature derivatives of the O-O pair RDF in water using a fluctuation theory approach [80].

C The Standard Bayesian Framework

For simplicity of notation, let $\theta = \{\lambda, \sigma, \varepsilon, \sigma_n\}$ represent the model parameters and $\mathscr{Y} = S_d\ Q)$ be the RDF observations. The nuisance parameter, σ_n, represents the width of the Gaussian likelihood and is considered a model parameter since nothing is known about this parameter *a priori*. Calculating the posterior probability distribution with Bayesian inference then requires two components: (1) prescription of prior distributions on the model parameters, $p\ \theta)$, and (2) evaluation of the RDF likelihood, $p\ \mathscr{Y}|\theta)$. The prior distribution over the $\lambda\quad 6)$ Mie parameters is assumed to be a multivariate normal distribution,

$$\theta \sim \qquad \mu_\theta, \sigma_\theta^2) \tag{24}$$

where μ_θ and σ_θ^2 are the prior mean and variance of each $\lambda\quad 6)$ Mie parameter in θ, respectively. A wide, multivariate normal distribution was selected because it is non-informative and conjugate to the Gaussian likelihood equation. The prior on the nuisance parameter is assumed to be log-normal,

$$\log \sigma_n \sim \qquad \mu_{\sigma_n}, \sigma_{\sigma_n}^2) \tag{25}$$

where μ_{σ_n} and σ_{σ_n} are the prior mean and variance of the nuisance parameter. The log-normal prior imposes the constraint that the nuisance parameter is non-negative, which is obviously true because a negative variance in the observed data is undefined. For reference, the prior parameters used in this study are summarized in Table 5.

The likelihood function is assumed to be Gaussian according to the central limit theorem,

$$p\ \mathscr{Y}|\theta) \propto \frac{1}{\sigma_n^{n_{samples}}} \exp\left[\quad \frac{1}{2\sigma_n^2} \sum_i [S_{\theta_i}\ Q_j) \quad S_d\ Q_j)]^2 \right] \tag{26}$$

where $S_\theta\ Q_i)$ is the molecular simulation predicted RDF and j indexes over discrete points along the momentum vector. Bayes' theorem is then expressed as,

$$p\ \theta|\mathscr{Y}) \propto p\ \mathscr{Y}|\theta) p\ \theta) \tag{27}$$

where equivalence holds up to proportionality. This construction is acceptable since the resulting posterior distribution can be normalized *post hoc* to find a valid probability distribution.

C.1 Markov Chain Monte Carlo

To populate the Bayesian likelihood distribution, Markov Chain Monte Carlo (MCMC) samples over the model parameters $\theta = \{\lambda, \sigma, \varepsilon, \sigma_n\}$ are passed to the surrogate model, evaluated, and compared to the experimental RDF. 960,000 MCMC samples were calculated using the emcee package [89] from 160 walkers (5000 samples/walker) with a 1000 sample burn-in per walker. The MCMC moves applied were differential evolution (DE) at a 0.8 ratio and DE Snooker at a 0.2 ratio, which is known to give good results for multimodal distributions. The acceptance ratio obtained from this sampling procedure was $\sim$0.27 and the autocorrelation between steps was 16 moves.