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

A PREPRINT

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

October 31, 2023

ABSTRACT

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 apply a Bayesian framework accelerated via local Gaussian process (LGP) surrogate models. We show that the time-complexity of 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 experimental radial distribution function of liquid neon, and observed a remarkable 288,000-fold speed-up compared to molecular dynamics with insignificant loss in predictive accuracy. 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 ever-advancing experimental data.

1 Introduction

Molecular simulations are becoming increasingly able to estimate complex experimental observables, including scattering patterns from neutron and X-ray sources and spectra from near-infrared [1], terrahertz [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 significant computational expense of molecular simulation greatly limits our understanding of how experimental, model, and parametric uncertainty impact these predictions. Without this understanding, it is 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 complex experimental data.

Bayesian methods are the gold standard for these aims [13], with examples spanning from neutrino and dark matter detection [14], materials discovery and characterization [15–17], quantum dynamics [18, 19], to molecular simulation [20–30]. 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 [31, 32]. Interest in Bayesian methods for molecular simulation has surged [33–37] 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 [38].

The biggest problem plaguing Bayesian inference is its massive computational cost. The two major computational 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. For molecular simulation, these bottlenecks are magnified because molecular simulations are 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 [28, 39], polynomial chaos expansions [40, 41], configuration-sampling-based methods [42] and Gaussian processes [43–45].

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 [46] and structural correlations [47]. 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 [48].

Despite the numerous advantages of GPs, their adoption for Bayesian inference in physical chemistry remains limited. Bayesian inference on the self-diffusion coefficient and radial distribution function (RDF) of gaseous and liquid argon has been demonstrated using a highly-parallelizable transitional Markov chain Monte Carlo method with an adaptive GP surrogate model [49, 50]. These studies clearly show that using advanced sampling methods can drastically reduce computational cost. However, prior work in physical chemistry applications has not emphasized improving the computational efficiency of the GP surrogate model directly. This is significant since GPs have a cubic time-complexity in the number of independent variables (*e.g.* frequencies measured along a spectrum or radii measured along a RDF), which becomes more expensive as we include more data types into the Bayesian inference and as experimental measurements obtain higher ranges and resolution.

There is an emerging class of accelerated GP methods that are well-equipped to handle large sets of experimental data. Local Gaussian processes (LGPs) are constructed by separating a GP into a set of GPs trained at distinct locations in the input space [44, 51–53]. Computation on the LGP subset is trivially parallelizable and easily implemented in high-performance computing (HPC) architectures [54, 55]. State-of-the-art LGP models have been used to design Gaussian approximation potentials (GAPs) [56], a type of machine learning potential used to study atomic [57–59] and electron structures [56, 60], as well as nuclear magnetic resonance chemical shifts [61]. However, to our knowledge LGPs have not been applied as molecular simulation surrogate models for complex experimental observables.

In this study, we detail a simple and effective LGP surrogate model approach for complex experimental measurements common in physical chemistry. We show that the LGP approximation reduces the GP surrogate model evaluation time-complexity with respect to the number of QOIs from cubic to linear. The computational speed-up results from reducing the dimensionality of matrix operations and therefore enables Bayesian UQ/P on large scale experimental datasets. For example, 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, we estimate that a LGP would accelerate this computation compared to a standard GP by approximately 3,240,000x.

As a direct application of the method, we trained a LGP surrogate model on a neutron scattering derived RDF for liquid neon (Ne). The LGP surrogate is found to reproduce the $\eta = 73$ independent variable output space approximately 288,000x faster than molecular dynamics (MD) and 3500x faster than a conventional GP with accuracy comparable to the uncertainty in the reported experimental data (RMSE = 0.02). Accelerated Bayesian inference of the RDF under a (λ-6) Mie model is then used to draw conclusions on model behavior, uncertainty, and adequacy. Surprisingly, we find evidence that the (λ-6) Mie model is too inflexible to match fine structural details of liquid neon, highlighting opportunities for improving force field optimization and design.

2 Computational Methods

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 parameters that we want to optimize in the selected model, usually the force field parameters in the Hamiltonian, to the experimental QOI that we expect the simulation to reproduce. 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, *ab initio* molecular dynamics, and path integral molecular dynamics.

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, we employ Gaussian distributions for both the prior and likelihood functions, which is a standard choice according to the central limit theorem. The Gaussian likelihood has the form,

$$p(\mathbf{y}|\theta) = \left(\frac{1}{\sqrt{2\pi}\sigma_n}\right)^{\eta} \exp\left[-\frac{1}{2\sigma_n^2}\sum_i [\mathbf{y}_\theta - \mathbf{y}]^2\right] \tag{2}$$

where η is the number of observables in $\mathbf{y}$, σ_n is a nuisance parameter describing the unknown variance of the Gaussian likelihood, and $\mathbf{y}_\theta$ is the model predicted observables at model input θ. 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 [62]).

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 (often on the order of $10^5 - 10^6$) to adequately sample the parameter space, which is completely infeasible for even modest sized molecular systems.

2.1 Gaussian Process Surrogate Models

GPs accelerate the Bayesian likelihood evaluation by approximating $\mathbf{y}_\theta$ with an inexpensive matrix calculation. Therefore, GP surrogate models offer a cost-effective substitute for computationally expensive molecular simulations. By leveraging matrix-based approximations, this ML-accelerated approach enables the efficient calculation of extensive sets of thermodynamic data at a scale that is infeasible for molecular simulations.

A Gaussian process is a stochastic process such that every finite set of random variables (position, time, etc) has a multivariate normal distribution [43]. The joint distribution over all random variables in the system therefore defines a probability distribution over functions. The expectation of this distribution maps a set of model parameters, θ^*, and independent variables, $\mathbf{r}$, to the most probable QOI function, $S(\mathbf{r}|\theta^*)$, such that,

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

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,

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

where σ_{noise}^2 is the variance due to noise and $\mathbf{I}$ is the identity matrix. Note that in general the independent variables, $\mathbf{r}$, can be multidimensional. As an example, consider the case where we want a GP to map a set of force field parameters to the angular RDF of a liquid. We now have a 2-dimensional space of independent variables since the angular RDF gives the atomic density along the radial and angular dimensions. Another example is if one wanted to build a GP for an RDF and FT-IR spectra simultaneously. In the following mathematical development, we assume that the QOI

is 1-dimensional for the sake of notational convenience and note that extending the argument to higher-dimensional observables is trivial.

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 [63]. The squared-exponential kernel function between input points (θ_m, r_m) and (θ_n, r_n) is given by,

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

where o indexes over $\dim(\theta)$ and the hyperparameters α^2 and ℓ_A are the kernel variance and correlation length scale of parameter A, respectively. Hyperparameter optimization can be performed by log marginal likelihood maximization, k-fold cross validation [43] or marginalization with an integrated acquisition function [64], 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, we need to generate N training samples in the input parameter space and run a molecular simulation 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,

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

where the $\theta_{i,j}$ are the i^{th} model parameter for sample index j and r_k are the independent variables of the target QOI. Note that the training sample index, $j = 1,...,N$, 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, $\hat{\mathbf{Y}}$, is then a ($N\eta \times 1$) column vector of the observable outputs from the training set,

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

where $S(\theta_j, r_k)$ is the training set observation of model parameters θ_j at independent variable r_k. The expectation of the GP for a new set of parameters, $S(\mathbf{r}|\theta^*)$, is then a (η x 1) column vector calculated with eq (4) to give the most probable QOI output,

$$S^*(\mathbf{r}|\theta^*) = [S^*(r_1|\theta^*), ..., S^*(r_\eta|\theta^*)]^T. \tag{8}$$

The GP expectation calculation is burdened by the inversion of the training-training kernel matrix with $\mathcal{O}(N^3\eta^3)$ time complexity and the $(\eta \times N\eta) \times (N\eta \times N\eta) \times (N\eta \times 1)$ matrix product with $\mathcal{O}(N^2\eta^3)$ 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 ($\dim(\theta) = 2$) 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 ($\dim(\theta) = 6$) 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 matrix operations.

2.2 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 along the independent variables of the QOI. Under this construction, an individual GP_k is trained to map a set of model parameters to an individual QOI,

$$\mathbb{E}[GP_k] : \theta \mapsto S(r_k) \tag{9}$$

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

$$\hat{\mathbf{X}}' = \begin{bmatrix} \theta_{1,1} & \theta_{2,1} & \dots \\ \theta_{1,2} & \theta_{2,2} & \dots \\ \vdots & \vdots & \vdots \\ \theta_{1,N} & \theta_{2,N} & \dots \end{bmatrix} \tag{10}$$

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,

$$\hat{\mathbf{Y}}'_k = [S(\theta_1, r_k), ..., S(\theta_N, r_k)]^T \tag{11}$$

where the k indexes over independent variables. The LGP surrogate model prediction for the observable at r_k, $S^*_{loc}(r_k)$, at a new set of parameters, θ^*, is just the expectation of the k^{th} Gaussian process given the training set data,

$$S^*_{loc}(r_k) = \mathbb{E}[GP_k(\theta^*)] = \mathbf{K}_{\theta^*,\hat{\mathbf{X}}'}[\mathbf{K}_{\hat{\mathbf{X}}',\hat{\mathbf{X}}'} + \sigma^2_{noise}\mathbf{I}]^{-1}\hat{\mathbf{Y}}'_k \tag{12}$$

We can then just combine the local results from the subset of η GPs to obtain a prediction for the QOIs,

$$S^*_{loc}(\mathbf{r}) = [S^*_{loc}(r_1), ..., S^*_{loc}(r_\eta)]^T. \tag{13}$$

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 independent variables, 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 parallel linear-scaling, η.

In summary, a local Gaussian process is an approximation in which all of the QOIs are modeled as independent random variables, each described by their own Gaussian process. This amounts to assuming that the random variables are independent and identically distributed (i.i.d.), which may or may not be true for a given experimental observation. For time-independent data including scattering measurements and spectroscopy, this approximation is appropriate since each observation is an independent measurement at each independent variable. In contrast, this approximation would not hold for time-dependent data in which a measurement at time t depends on previous measurements $t_i < t$.

We show that complex experimental observables can be reconstructed by this set of LGPs through a series of relatively straightforward matrix operations with linear time-complexity with respect to the number of independent variables. We also note that 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 show the computational enhancement and accuracy of the LGP approach by modeling the RDF of neon at $42K$. We then use the surrogate model within a Bayesian framework to exemplify the power of UQ/P for molecular simulations.

3 A 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 [65] under a (λ-6) Mie fluid model. The (λ-6) Mie force field is flexible Lennard-Jones type potential with variable repulsive exponent,

$$v_2^{Mie}(r) = \frac{\lambda}{\lambda - 6}\left(\frac{\lambda}{6}\right)^{\frac{6}{\lambda-6}} \varepsilon \left[\left(\frac{\sigma}{r}\right)^{\lambda} - \left(\frac{\sigma}{r}\right)^{6}\right] \tag{14}$$

where λ is the short-range repulsion exponent, σ is the collision diameter (Å), and ε is the dispersion energy (kcal/mol) [66].

MD simulations were performed from a sobol sampled set spanning a prior range based on existing force field models [67–69] ($\lambda = [5,20]$, $\sigma = [1.5,4.0]$, and $\varepsilon = [0,0.5]$) to generate a RDF training set matrix of the form in eq 10. The number of training simulations performed was $N = 480$, the number of r points in the experimental RDF was $\eta = 73$, and the number of surrogate model calls is $G = 8 \times 10^5$. The number of points η in the radial distribution was calculated by dividing the reported $r_{max} - r_{min} \approx 15.3$ by the effective r-space resolution given by, $\Delta r = \pi / Q_{max}$, where $\Delta r = 0.21$ Å for reported $Q_{max} = 15$ Å^{-1}. The total number of training simulations was selected based on a sequential sampling approach in which 160 simulations were performed and progressively narrowed based on the RMSE between the simulation and experimental RDF. Three iterations (480 total) was found to provide an RMSE similar to the reported experimental uncertainty. Details on the MD simulations, training set, and Bayesian framework are provided in the Appendix.

3.1 Computational Efficiency and Accuracy

Now that we have constructed the training set matrix, we simply evaluate the expectation at each r according to eq (12) and combine the results into a single array as in eq (13). The average computational time to evaluate the RDF and invert the training set matrix are shown below in Table 1. The LGP surrogate model is found to accelerate the model QOI prediction time compared to MD by a factor of 287,984. Furthermore, this 5 orders-of-magnitude speed-up is shown to outpace a standard GP by 3 orders-of-magnitude (3533x). The total time to compute the Bayesian posterior distribution with MCMC was thereby reduced from an estimated ∼4.9 years for MD or ∼22 days for a standard GP surrogate model, to less than 9 minutes on our local cluster. In fact, the LGP is so fast that we can calculate the Bayesian likelihood without introducing non-trivial sampling techniques such as active learning and/or translational Markov chain Monte Carlo.

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

Model	Time to QOI (s)	Speed Up (t/t_{sim})	Inv. Time (s)	Inv. Speed Up (t_{GP}/t_{LGP})
Simulation	86.25	1	-	-
GP	1.06	81.19	87	1
LGP	0.0003	287,984	0.02	4183

We've established that 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 MD? To evaluate the accuracy of the local predictions, we ran a 160 sample test set around the posterior distribution and computed the root mean squared error (RMSE) between simulated and LGP predicted radial distributions. The results are summarized in Figure 1.

The total RMSE over the test set gives a value of 0.02, which is excellent considering that this is less than the uncertainty reported in the experimental measurement (∼0.03). A plot of the RMSE as a function r-coordinate reveals period and asymptotically decaying accuracy in r corresponding approximately to the peak positions, which is expected because the magnitude of the RDF exhibits similar behavior. Overall, the LGP RMSE is considered excellent for a RDF, but could be improved by including additional training set samples.

One notable limitation of the LGP surrogate model is that it may have numerical instabilities that reduce the accuracy of the GP posterior distribution [70, 71]. For this work, we argue that this limitation is minor since we do not use the GP posterior distribution directly; rather, we use only its expectation. However, if an application requires a high accuracy estimation of GP posterior distribution (*e.g.* active learning or forecasting), more advanced greedy approximation methods may be required [70–73].

3.2 Learning from Surrogate Models with Bayesian Analysis

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 between model parameters, relationships between model parameters and the QOI, and model adequacy. This knowledge can provide deep 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* (*MAP*), represents

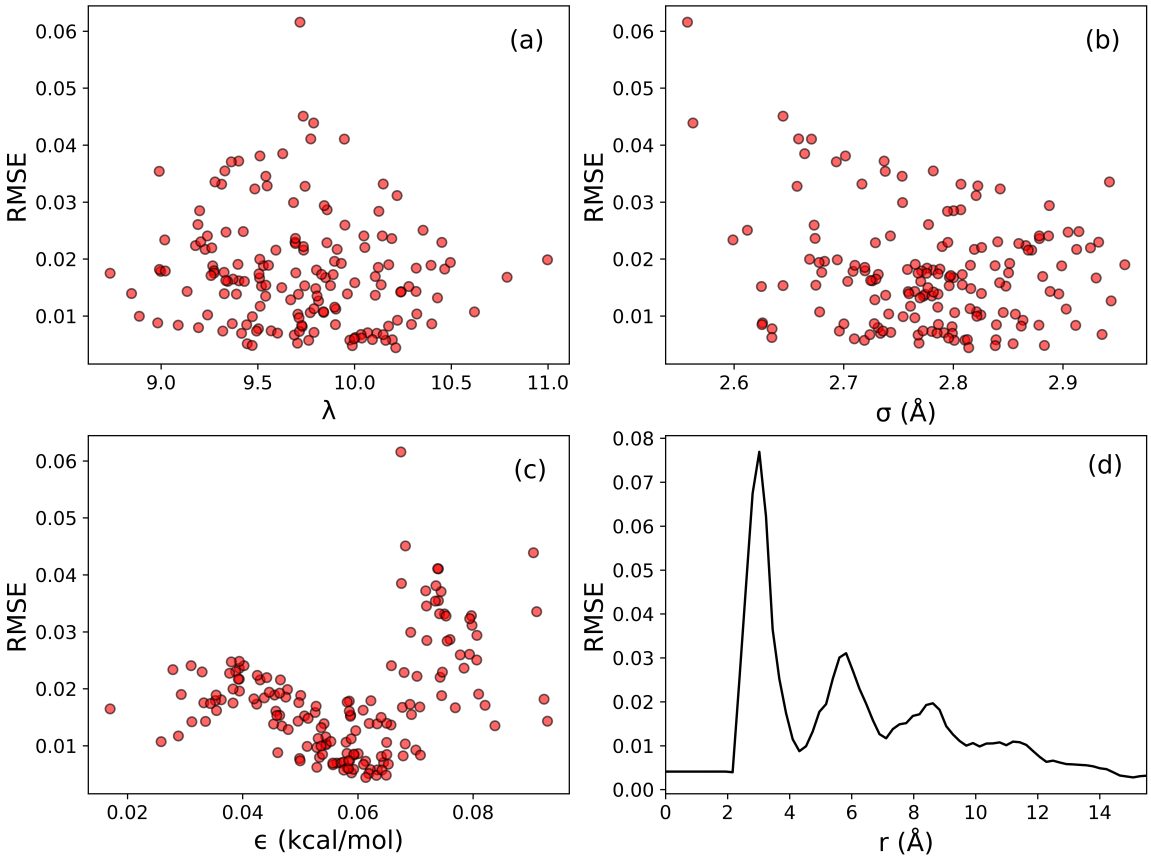

Figure 1: (a)-(c) Test set samples over each parameter plotted against the RMSE between simulated and LGP data. (d) Average RMSE over the 160 test set samples as a function of r,

the set of parameters with the highest probability of explaining the given experimental data. In force field design, the *MAP* 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.

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 [74]. Observing the 2D marginal distributions in the second row of 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

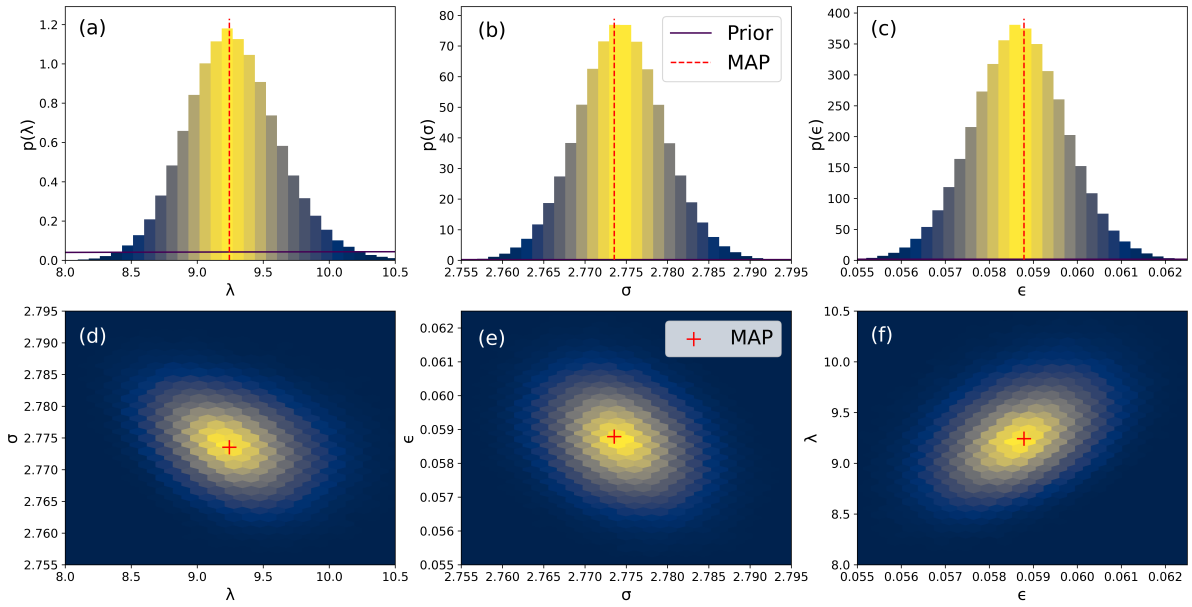

Figure 2: (a)-(c) 1D marginal distributions for the (λ-6) Mie fluid parameters. Blue lines are the prior distributions and red dotted lines represent the *maximum a posterior* (*MAP*). (d)-(f) 2D marginal histograms showing parameter correlations. Red marker indicates the 2D *MAP*.

Bayesian analysis on liquid Ar [49]. Quantitative analysis of model sensitivity can also be performed with probabilistic derivatives of the QOI with respect to model parameters (see Appendix).

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 [75–77]. In fact, the Bayesian analysis shows that we can actually determine values for λ, σ, and ε within ± 1.5, ± 0.03Å, and ± 0.005 kcal/mol with 95% certainty. This result leads to two important conclusions. First, scattering data effectively constrains the force field model parameter space. Second, scattering data must be accurate to design structure-optimized force fields.

The joint posterior can also be used for model parameter optimization. Specifically, the optimal parameters are given by the *MAP*, corresponding to the maximum of the joint posterior distribution. The *MAP* is presented in Table 2 along with two other existing force fields for liquid Ne.

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

Force Field	QOI	λ	σ (Å)	ε (kcal/mol)
Mick (2015)	VLE	11	2.794	0.064
SOPR (2022)	RDF	11	2.778	0.063
This Work	RDF	9.27	2.774	0.059

We can clearly see that the optimal repulsive exponent and dispersion energy prediction are lower than the Mick [68] and structure optimized potential refinement (SOPR) [69] models, while the optimal collision diameter is the same. Interestingly, the optimal ε parameter derived from SOPR falls just outside of the 95% credibility interval, despite being trained on the same experimental RDF. The difference between the Mie fluid and SOPR model 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 the Mie model missing important physics? Or is the SOPR model over-fitting to imperfect data?

We can investigate these questions further by propagating parameter uncertainty through the Mie model to construct a distribution of RDF predictions - referred to as the posterior predictive. The posterior predictive quantifies of how accurately we know the QOI given experimental, model, and parametric uncertainty. 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 ($g_{exp}(r) - \mu(r)$) for the liquid Ne RDF are shown in Figure 3.

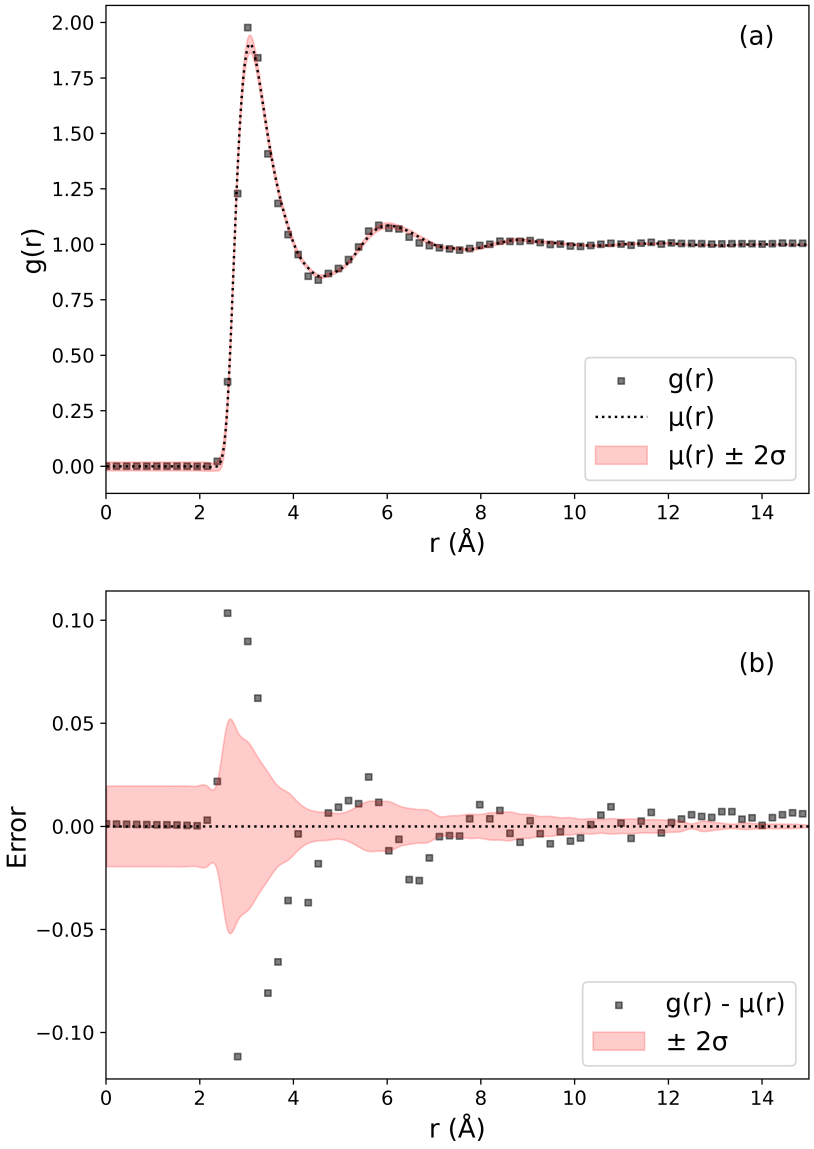

Figure 3: (a) RDF mean and credibility interval propagated from the parameter uncertainty quantified with Bayesian inference. (b) Residual analysis comparing the experimental data with the Bayesian posterior distribution.

We notice that the residuals between the experiment and posterior predictive mean are evenly distributed but often lie outside of the 95% credibility interval. Therefore, the Mie model captures the general structural behavior of Ne but fails to describe nuanced features present in the RDF data. This result suggests that there are potentially missing physics that could be incorporated to better present the structural behavior of the fluid. We therefore suspect that the non-parametric SOPR model is capturing these more realistic interactions, which is supported by its excellent agreement with structure

and thermodynamic properties. Of course, this remains an open hypothesis and should be scrutinized through empirical testing and uncertainty quantification of the SOPR method.

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 nearly 300,000-fold with exceptional accuracy from only 480 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, we expect significantly higher speed-up 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 be extremely useful for Bayesian inference of molecular models over a broad array of complex experimental data.

5 Acknowledgements

This study is supported by the National Science Foundation Award No. CBET-1847340. We would like to thank Sean T. Smith and Philip J. Smith for their invaluable advice on the implementation of Gaussian processes.

6 Author 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. *Annu. Rev. Phys. Chem.* **2009**, *60*, 429–448.
(6) Roget, S. A.; Carter-Fenk, K. A.; Fayer, M. D. . *Am. 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. A: Accel. Spectrom. Detect. Assoc. 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) V. Krems, R. *Phys. Chem. Chem. Phys.* **2019**, *21*, 13392–13410.
(19) Deng, Z.; Tutunnikov, I.; Averbukh, I. S.; Thachuk, M.; Krems, R. V. *J. Chem. Phys.* **2020**, *153*, 164111.
(20) Frederiksen, S. L.; Jacobsen, K. W.; Brown, K. S.; Sethna, J. P. *Phys. Rev. Lett.* **2004**, *93*, 165501.
(21) Cooke, B.; Schmidler, S. C. *Biophys. J.* **2008**, *95*, 4497–4511.
(22) Cailliez, F.; Pernot, P. *J. Chem. Phys.* **2011**, *134*, 054124.
(23) Farrell, K.; Oden, J. T.; Faghihi, D. *J. Comput. Phys.* **2015**, *295*, 189–208.
(24) Wu, S.; Angelikopoulos, P.; Papadimitriou, C.; Moser, R.; Koumoutsakos, P. *Philos. Trans. R. Soc. A* **2016**, *374*, 20150032.
(25) Patrone, P. N.; Dienstfrey, A.; Browning, A. R.; Tucker, S.; Christensen, S. *Polymer* **2016**, *87*, 246–259.
(26) Messerly, R. A.; Knotts, T. A.; Wilding, W. V. *J. Chem. Phys.* **2017**, *146*, 194110.
(27) Dutta, R.; Brotzakis, Z. F.; Mira, A. *J. Chem. Phys.* **2018**, *149*, 154110.
(28) Wen, M.; Tadmor, E. B. *Npj Comput. Mater.* **2020**, *6*, 1–10.
(29) Bisbo, M. K.; Hammer, B. *Phys. Rev. Lett.* **2020**, *124*, 086102.
(30) Xie, Y.; Vandermause, J.; Ramakers, S.; Protik, N. H.; Johansson, A.; Kozinsky, B. *Npj Comput. Mater.* **2023**, *9*, 1–8.
(31) Gelman, A.; Carlin, J. B.; Stern, H. S.; Rubin, D. B., *Bayesian Data Analysis*; Chapman and Hall/CRC: New York, 1995.
(32) Shahriari, B.; Swersky, K.; Wang, Z.; Adams, R. P.; de Freitas, N. *Proc. IEEE* **2016**, *104*, 148–175.
(33) Musil, F.; Willatt, M. J.; Langovoy, M. A.; Ceriotti, M. *J. Chem. Theory Comput.* **2019**, *15*, 906–915.
(34) 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.
(35) Vandermause, J.; Torrisi, S. B.; Batzner, S.; Xie, Y.; Sun, L.; Kolpak, A. M.; Kozinsky, B. *Npj Comput. Mater.* **2020**, *6*, 1–11.
(36) Köfinger, J.; Hummer, G. *Eur. Phys. J. B* **2021**, *94*, 245.
(37) Vandermause, J.; Xie, Y.; Lim, J. S.; Owen, C. J.; Kozinsky, B. *Nat. Commun.* **2022**, *13*, 5183.
(38) Lemm, J. C., *Bayesian Field Theory*; JHU Press: 2003.
(39) Li, C.; Gilbert, B.; Farrell, S.; Zarzycki, P. *J. Chem. Inf. Model.* **2023**.
(40) Ghanem, R. G.; Spanos, P. D., *Stochastic Finite Elements: A Spectral Approach*; Courier Corporation: 2003.
(41) Jacobson, L. C.; Kirby, R. M.; Molinero, V. *J. Phys. Chem. B* **2014**, *118*, 8190–8202.
(42) Messerly, R. A.; Razavi, S. M.; Shirts, M. R. *J. Chem. Theory Comput.* **2018**, *14*, 3144–3162.
(43) Rasmussen, C. E.; Williams, C. K. I., *Gaussian processes for machine learning*; MIT Press: Cambridge, Mass, 2006.
(44) Nguyen-Tuong, D.; Seeger, M.; Peters, J. *Adv Robot.* **2009**, *23*, 2015–2034.
(45) Burn, M. J.; Popelier, P. L. A. *J. Chem. Phys.* **2020**, *153*, 054111.
(46) Dai, J.; Krems, R. V. *J. Chem. Theory Comput.* **2020**, *16*, 1386–1395.
(47) Yang, N.; Hill, S.; Manzhos, S.; Carrington, T. *J. Mol. Spectrosc.* **2023**, *393*, 111774.
(48) 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.
(49) Angelikopoulos, P.; Papadimitriou, C.; Koumoutsakos, P. *J. Chem. Phys.* **2012**, *137*, 144103.
(50) Kulakova, L.; Arampatzis, G.; Angelikopoulos, P.; Hadjidoukas, P.; Papadimitriou, C.; Koumoutsakos, P. *Sci. Rep.* **2017**, *7*, 16576.
(51) Das, K.; Srivastava, A. N. In *2010 IEEE International Conference on Data Mining*, 2010, pp 791–796.
(52) Park, C.; Apley, D.
(53) Terry, N.; Choe, Y. *PLoS One* **2021**, *16*, e0256470.
(54) Gramacy, R. B.; Apley, D. W. *J. Comput. Graph. Stat.* **2015**, *24*, 561–578.
(55) Broad, J.; Wheatley, R. J.; Graham, R. S. *J. Chem. Theory Comput.* **2023**, *19*, 4322–4333.
(56) Deringer, V. L.; Bartók, A. P.; Bernstein, N.; Wilkins, D. M.; Ceriotti, M.; Csányi, G. *Chem. Rev.* **2021**, *121*, 10073–10141.
(57) Caro, M. A.; Deringer, V. L.; Koskinen, J.; Laurila, T.; Csányi, G. *Phys. Rev. Lett.* **2018**, *120*, 166101.
(58) 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.

(59) 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.
(60) Cheng, B.; Mazzola, G.; Pickard, C. J.; Ceriotti, M. *Nature* **2020**, *585*, 217–220.
(61) Paruzzo, F. M.; Hofstetter, A.; Musil, F.; De, S.; Ceriotti, M.; Emsley, L. *Nat. Commun.* **2018**, *9*, 4501.
(62) Shell, M. S. *J. Chem. Phys.* **2008**, *129*, 144108.
(63) Ambrogioni, L.; Maris, E. In *AISTATS*, PMLR: 2018, pp 217–225.
(64) Snoek, J.; Larochelle, H.; Adams, R. P. In *Adv. Neural Inf. Process.* Curran Associates, Inc.: 2012; Vol. 25.
(65) Bellissent-Funel, M. C.; Buontempo, U.; Filabozzi, A.; Petrillo, C.; Ricci, F. P. *Phys. Rev. B* **1992**, *45*, 4605–4613.
(66) Mie, G. *Ann. Phys.* **1903**, *316*, 657–697.
(67) Vrabec, J.; Stoll, J.; Hasse, H. *J. Phys. Chem. B* **2001**, *105*, 12126–12133.
(68) Mick, J. R.; Soroush Barhaghi, M.; Jackman, B.; Rushaidat, K.; Schwiebert, L.; Potoff, J. J. *J. Chem. Phys.* **2015**, *143*, 114504.
(69) Shanks, B. L.; Potoff, J. J.; Hoepfner, M. P. *J. Phys. Chem. Lett.* **2022**, 11512–11520.
(70) Foster, L.; Waagen, A.; Aijaz, N.; Hurley, M.; Luis, A.; Rinsky, J.; Satyavolu, C.; Com, M.
(71) Rasmussen, C. E.; Quiñonero-Candela, J. In *Proceedings of the 22nd international conference on Machine learning*, Association for Computing Machinery: New York, NY, USA, 2005, pp 689–696.
(72) Wacker, J.; Filippone, M. *Procedia Computer Science* **2022**, *207*, 987–996.
(73) Lu, Y.; Ma, J.; Fang, L.; Tian, X.; Jiang, J. In 2023, pp 21950–21959.
(74) Hanke, M. *J. Stat. Phys.* **2018**, *170*, 536–553.
(75) Clayton, G. T.; Heaton, L. *Phys. Rev.* **1961**, *121*, 649–653.
(76) Jovari, P. *Mol. Phys.* **1999**, *97*, 1149–1156.
(77) Hansen, J.-P.; McDonald, I. R., *Theory of Simple Liquids: with Applications to Soft Matter*; Academic Press: San Diego, 2013.
(78) Anderson, J. A.; Glaser, J.; Glotzer, S. C. *Comput. Mater. Sci.* **2020**, *173*, 109363.
(79) Ramasubramani, V.; Dice, B. D.; Harper, E. S.; Spellings, M. P.; Anderson, J. A.; Glotzer, S. C. *Comput. Phys. Commun.* **2020**, *254*, 107275.
(80) Angelikopoulos, P.; Papadimitriou, C.; Koumoutsakos, P. *Comput. Methods. Appl. Mech. Eng.* **2015**, *289*, 409–428.
(81) Foreman-Mackey, D.; Hogg, D. W.; Lang, D.; Goodman, J. *Publ. Astron. Soc. Pac.* **2013**, *125*, 306.
(82) Piskulich, Z. A.; Thompson, W. H. *J. Chem. Phys.* **2020**, *152*, 011102.

7 Appendix

A Notes on Gaussian Process Hyperparameter Selection

The kernel of the Gaussian process depends on a set of hyperparameters that can influence the surrogate model prediction. A standard approach to selecting hyperparameters is to maximize the model evidence [43] or apply an expected improvement criterion based on an integrated acquisition function [64]. Here we have applied a brute force search based on minimizing the leave-one-out cross validation (LOOCV) error, which creates a limitation in the uncertainty quantification since we are not accounting for hyperparameter uncertainty. However, the self-consistency of our approach with other methods, including vapor-liquid equilibria optimized [68] and structure optimized potential refinement [69], suggests that this choice is appropriate for structural modeling in these systems. However, implementation of a fully Bayesian treatment of the Gaussian process hyperparameters may be necessary in more complex structural problems.

B Molecular Dynamics Simulation of Mie Fluids

Computer generated radial distribution functions were calculated using molecular dynamics (MD) simulations in the HOOMD-Blue package [78]. Simulations were initiated with a lattice configuration of 864 particles and compressed to a reduced density of $\rho = 0.02477$ atom/Å^3 and temperature $T = 42.2$ 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 radial distribution functions were calculated using the Freud package [79].

B.1 Constructing the Surrogate Model Training Set

To generate a useful training set, we need dense samples of model parameters in the region of the parameter space that can well-represent the target quantity-of-interest (QOI). However, it is not known *a priori* where this region is, particularly if there is no prior knowledge of what model parameters best represent the target observable. In this work, we use a sequential sampling approach in which we sample a wide range of parameters, determine the best-fit to the target observable, and then narrow the parameter range around this point for the next set of training samples. We repeated this procedure three times with 160 samples per iteration. A visualization of this procedure is shown below in Figure 4.

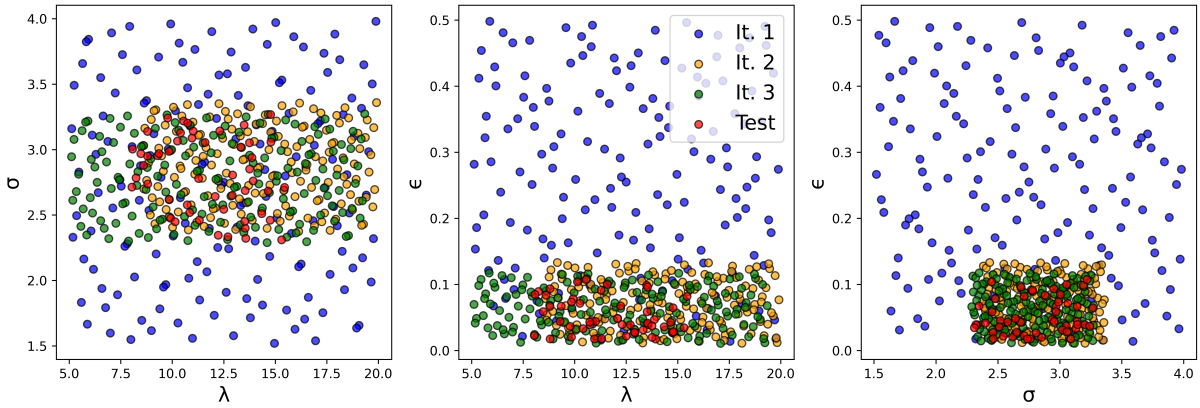

Figure 4: 2D training and sample parameter set used to train and evaluate the local Gaussian process surrogate model.

The updated range on the training set samples are centered on the parameters that provided the lowest RMSE compared to the experiment and the range was narrowed by a fixed interval based on inspection of the RMSE between the training set and the experimental data. One could also use an adaptive or on-the-fly learning approach in which the uncertainty in the local GP prediction is used to decide whether or not a new simulation is needed. This approach was used in prior work [49, 80] but was unnecessary in this study due to the efficiency and accuracy of the local GP.

C The Standard Bayesian Framework

Bayesian inference was implemented to calculate parameter probability distributions as a function of structure factor uncertainty. For simplicity of notation, let $\theta = \{\lambda, \sigma, \varepsilon, s_n\}$ represent the model parameters and $\mathscr{Y} = S_d(Q)$ be the structure factor observations. The nuisance parameter, s_n, represents the width of the Gaussian likelihood and is considered an unknown 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 likelihood, $p(\mathscr{Y}|\theta)$. The prior distribution over the $(\lambda - 6)$ Mie parameters is assumed to be a multivariate normal distribution,

$$\theta \sim \mathscr{N}(\mu_\theta, s_\theta^2) \tag{15}$$

where μ_θ and s_θ^2 are the prior mean and variance of each $(\lambda - 6)$ 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,

$$\log s_n \sim \mathscr{N}(\mu_{s_n}, s_{s_n}^2) \tag{16}$$

where μ_{s_n} and s_{s_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 3.

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

Table 3: 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
s_n	Log-Normal	1	1

$$p(\mathscr{Y}|\theta) \propto \frac{1}{s_n^{n_{samples}}} \exp\left[-\frac{1}{2s_n^2}\sum_i \left[S_{\theta_i}(Q_j) - S_d(Q_j)\right]^2\right] \tag{17}$$

where $S_\theta(Q_i)$ is the molecular simulation predicted QOI and j indexes over discrete independent variables of the QOI. Bayes' theorem is then expressed as,

$$p(\theta|\mathscr{Y}) \propto p(\mathscr{Y}|\theta)p(\theta) \tag{18}$$

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.

To populate the Bayesian likelihood distribution, Markov Chain Monte Carlo (MCMC) samples are passed to the surrogate model, evaluated, and compared to the experimental RDF. 800,000 MCMC samples were calculated using the emcee package [81] using 160 walkers and a 1000 sample burn-in. The acceptance ratio obtained from this sampling procedure was $\sim$0.3 and the autocorrelation between steps was 15 moves.

D Sensitivity Analysis with Gaussian Process Derivatives

One of the most useful applications of Bayesian inference is parameter sensitivity analysis. As was shown in the main text, studying the correlations and shape of the joint posterior distribution can reveal insight into how parameters influence the prediction of a QOI. However, the joint posterior is not the only method to study parameter sensitivity. Specifically, analytical derivatives of the local GP surrogate model can directly quantify the differential rate of change of the QOI to a model parameter. As a demonstration, figure 5 plots local GP derivatives calculated for the (λ-6) Mie parameters.

We can now precisely study how each parameter changes the radial distribution function. For instance, the repulsive exponent derivative exhibits a small magnitude and has a minimum at the radial distribution function half maximum. This behavior suggests that increasing the repulsive exponent, which determines the "hardness" of the particles, steepens the slope of the first peak. This makes sense, just consider that in a hard-particle model there is a 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 structure factor peaks and troughs, while local extrema align with the half-maximum positions. Consequently, increasing the effective particle size shifts the radial distribution function to the right while maintaining relatively constant peak heights. Regarding the dispersion energy, its derivative displays zeros at the half-maximum positions of the structure factor and local extrema at peaks and troughs. This behavior indicates that an increase in the dispersion energy leads to an increased magnitude of the radial distribution function peaks and greater liquid structuring.

It is also possible to relate GP derivatives with thermodynamic variable derivatives. Let's take as an example the ε-derivative of the radial distribution function. 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 for a Mie fluid by,

$$T^* = \frac{k_B T}{\varepsilon} \tag{19}$$

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,

$$\frac{\partial g(r)}{\partial \varepsilon} = \frac{\partial g(r)}{\partial \beta^*} \tag{20}$$

where $\beta^* = T^*/k_B T$. In summary, an increase in ε is equivalent to a decrease in temperature. We should therefore expect that the ε derivative and temperature derivative to 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 radial distribution function in water using a fluctuation theory approach [82]. This qualitative agreement suggests that Gaussian process surrogate models can also offer additional insight into thermodynamic derivatives of microscopic quantities.

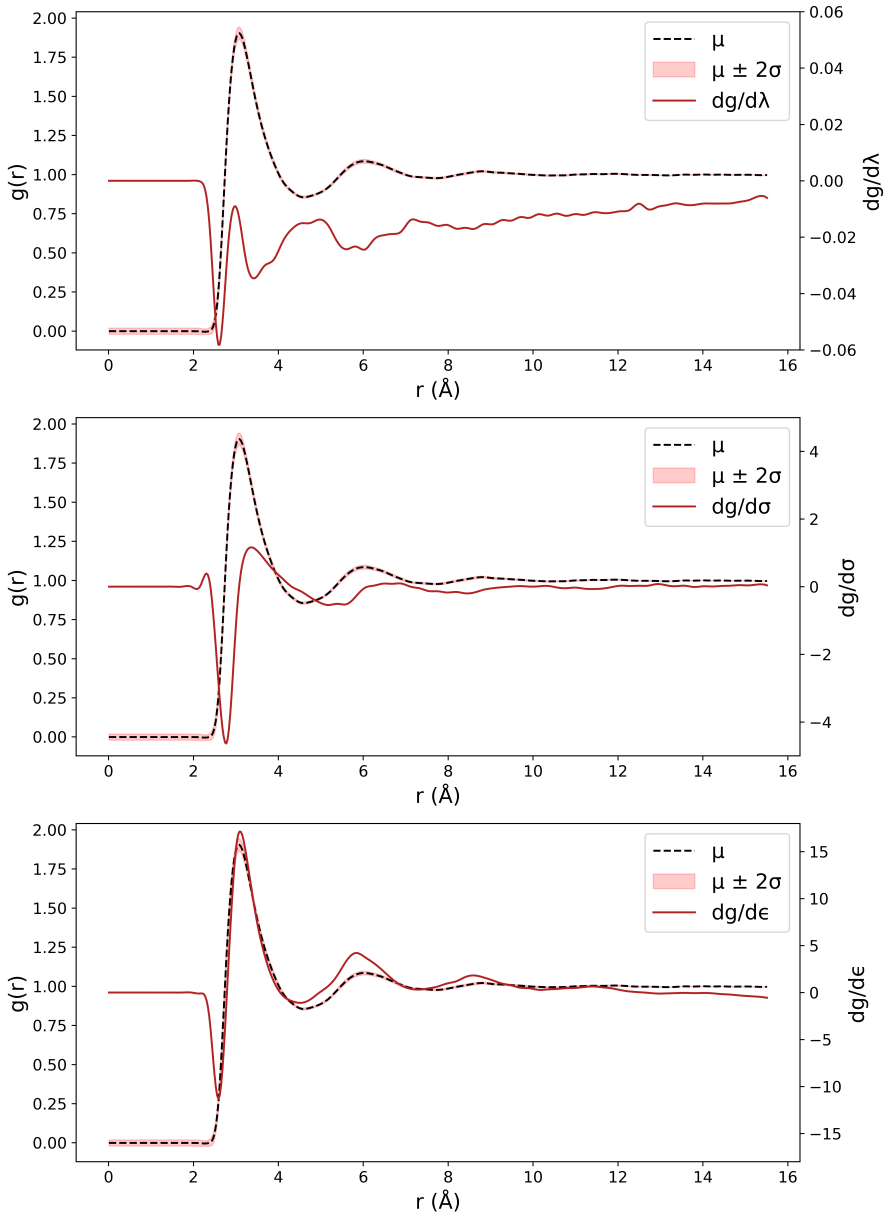

Figure 5: Derivatives of the local GP along the radial distribution function calculated from eq (5).