THE JOURNAL OF PHYSICAL CHEMISTRY B
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

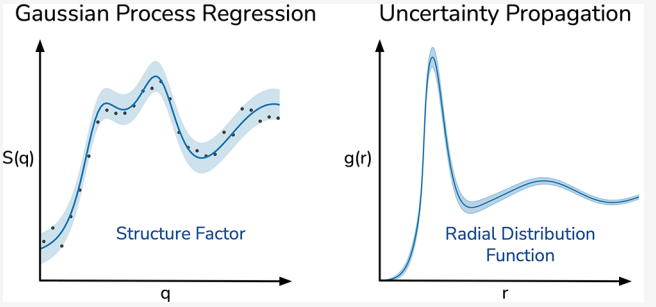
pubs.acs.org/JPCB
Article

Physics-Informed Gaussian Process Inference of Liquid Structure from Scattering Data

Harry Winston Sullivan, Matej Cervenka, Brennon L. Shanks,* and Michael P. Hoepfner*

Cite This: J. Phys. Chem. B 2025, 129, 11802−11815
Read Online

ABSTRACT: We present a nonparametric Bayesian framework to infer radial distribution functions from experimental scattering measurements with uncertainty quantification using nonstationary Gaussian processes. The Gaussian process prior mean and kernel functions are designed to mitigate well-known numerical challenges with the Fourier transform, including discrete measurement binning and detector windowing, while encoding fundamental yet minimal physical knowledge of the liquid structure. We demonstrate uncertainty propagation of the Gaussian process posterior to unmeasured quantities of interest. Experimental radial distribution functions of liquid argon and water with uncertainty quantification are provided as both a proof of principle for the method and a benchmark for molecular models.



INTRODUCTION

The radial distribution function (RDF), which characterizes the spatial arrangement of atoms, is a cornerstone in liquid state theory and serves as a vital benchmark for molecular simulations. Our understanding of the liquid state relies heavily on established theoretical relationships that link the RDF to thermodynamic properties and interatomic forces. These include the Ornstein−Zernike relation,[1] Henderson's inverse theorem,[2] the Born−Bogilubov−Green−Kirkwood−Yvon hierarchy,[3] and Kirkwood-Buff integrals,[4] among others. Despite these profound and intricate connections, structure is often relegated to a validation step in molecular modeling, with preference typically given to training force field parameters using macroscopic thermodynamic data[5] or to interatomic potentials computed from quantum mechanical methods.[6] While both approaches can yield models that accurately reproduce the thermophysical properties of fluids, they often struggle to fully capture structural features observed in experiments.[7−9] We argue that to more closely align molecular models with the principles of statistical mechanics, greater emphasis should be placed on experimentally derived RDFs in force field optimization and design.

In X-ray and neutron scattering experiments, the observed quantity is the momentum-space static structure factor, from which RDFs are subsequently inferred.[10] For single-atom systems, the static structure factor is related to the RDF, or $g(r)$, via a Fourier transform with radial symmetry,

$$S(q) - 1 = 4\pi\rho \int_0^\infty (g(r) - 1)\frac{\sin(qr)}{qr} r^2 \mathrm{d}r \tag{1}$$

where q is the momentum transfer and ρ is the atomic number density. In mixtures or molecular liquids, the total structure factor, $F(q)$, can be expressed as a combination of site−site partial structure factors, S_{ij}, between atoms i, j, such that,

$$F(q) = \sum_{i \geq j} [2 - \delta_{ij}] w_{ij} S_{ij}(q) \tag{2}$$

where w_{ij} is a (possibly q-dependent in the case of X-rays) weighting factor depending on the scattering length density and atomic concentration of the i, j pair, and δ_{ij} is the Kronecker delta. This linear system, known as the Faber−Ziman decomposition,[11] is ill-posed when the number of measured total structure factors is fewer than the number of unique site−site partial structure factors, which for a system with N distinct atom types has $N(N + 1)/2$ unique S_{ij} terms. To constrain this underdetermined linear system, scattering measurements can be performed on isotopologues (systems differing only by isotopic substitutions), which alter the scattering length densities without changing the underlying structure. However, in practice, obtaining a sufficient number of isotopologue measurements is often prohibitively expensive in terms of both experimental time and the cost of purified isotopes. As a result, solutions to the Faber−Ziman decomposition have historically been approximated using

Received: July 19, 2025
Revised: October 17, 2025
Accepted: October 21, 2025
Published: October 31, 2025

https://doi.org/10.1021/acs.jpcb.5c05024
J. Phys. Chem. B 2025, 129, 11802−11815

iterative molecular simulation methods to close the linear system with simulated structure data, such as reverse Monte Carlo (RMC)[12] and empirical potential structure refinement (EPSR).[13]

Assuming that the partial structure factors are known, they can then be Fourier transformed with eq 1 to obtain real-space site−site pair distribution functions, $g_{i,j}(r)$, which quantify the atomic density of type i within a spherical shell around any atom of type j. The $g_{i,j}(r)$ describes the relative likelihood of finding a neighboring atom at distance r; in liquids, it goes to zero at small r due to atom−atom impenetrability, exhibits oscillations that reflect local structure, and approaches unity at large r where correlations vanish.Eqs 1 and 2 are conceptually appealing, but their practical implementation faces several challenges. First, the finite size of individual neutron detectors constrains structure factor measurements to discrete momentum transfer values, $\Delta q = q_i - q_{i-1}$, which, according to the Peterson-Middleton sampling theorem,[14] can result in aliasing if the sampling efficiency is <1. Second, finite detector coverage windows the measurement to a range between some $q_{\min}$ and $q_{\max}$, preventing the evaluation of the full integral specified in eq 1. Windowing can introduce truncation artifacts (ripples), reduce the real-space resolution, and, when a smooth windowing correction is applied, artificially broaden the RDF peaks. Finally, measurement uncertainty of neutron counts and momentum transfer positions (i.e., time-of-flight uncertainty) introduces noise that can corrupt the underlying signal.[15,16]

In practice, a discrete radial Fourier transform (rFT) must be computed over N uncertain observations,

$$g(r) \approx 1 + \frac{1}{2\pi^2\rho}\sum_{i=1}^{N}\frac{1}{2}\left(S(q_{i-1})\frac{\sin(q_{i-1}r)}{q_{i-1}r}q_{i-1}^2 - S(q_i)\frac{\sin(q_i r)}{q_i r}q_i^2\right)\Delta q \tag{3}$$

where the sum is from some nonzero $q_{\min}$ to some finite $q_{\max} = N\Delta q$. The key problem is that, depending on the degree of undersampling and the choice of window function, the discrete Fourier transform can systematically distort the predicted RDF relative to the ground truth. These distortions, in turn, increase the uncertainty in the inferred fluid structure. This uncertainty may partially explain why scattering data are not more widely used as an optimization target in force field design.

The most well-studied problem in prior literature is addressing the $q_{\max}$ cutoff using so-called modification functions.[17] The essential idea here is to smoothly transition the structure factor from a data-dominated section (as measured by the neutron/X-ray detector) to a model-driven section (dictated by prior physical knowledge of the structure factor). Modification functions are designed to force the contribution of the experimental data to 0 near $q_{\max}$, effectively nullifying any features in the data and strictly relying on the physical model alone. Usually, the data are transitioned into a Poisson point process ideal gas model (i.e., $S_{\text{Ideal}}(q) = 1$).[18] Mathematically, this modifies the integral of eq 1 into,

$$g(r) = 1 + \frac{1}{2\pi^2\rho}\int_0^\infty (S(q) - 1)M(q)\frac{\sin(qr)}{qr}q^2\mathrm{d}q \tag{4}$$

where $M(q)$ is the q-dependent modification function. Common choices for the modification function are the first Bessel function,[19] second Bessel function,[20,21] cosine cutoff,[22] and dynamic functions.[23] However, as pointed out by Proctor et al.,[17] eq 4 is an approximate Bayesian predictive model where the modification function transitions into a prior $S(q)$ model. To see this, one can rewrite eq 4 in the following way,

$$=1 + \frac{1}{2\pi^2\rho}\int_0^\infty \left(\underbrace{(S(q) - 1)M(q)}_{\text{Data Driven Predictive}} + \underbrace{(S_{\text{Ideal}}(q) - 1)(1 - M(q))}_{\text{Model Driven Predictive}}\right)\frac{\sin(qr)}{qr}q^2\mathrm{d}q \tag{5}$$

where we have split the two contributions of the integrand into "data-driven" and "model-driven" parts, regulated by the modification function. Here, the $M(q)$ is viewed as a discrete posterior probability mass, meaning that all we have done is expressed the structure factor as a weighted mixture of two outcomes, either data or model, at each q value. An extensive analysis of commonly used modification functions can be found in ref 17.

While using prior information to constrain the space of possible RDFs is a valuable idea, the formulation above does not naturally support uncertainty quantification in the RDF predictions within a probabilistic framework. Prior studies have attempted to estimate uncertainty in scattering data by averaging RDF predictions across multiple experiments and computing standard deviations,[24] propagating experimental structure factor errors through the Fourier transform,[25] or combining both methods.[23] The primary limitation of simply comparing different data sets or analysis methods is that such estimates become unreliable if all sources share a common systematic error. Similarly, the Fourier transform error propagation method produces the largest uncertainties at small r, precisely where theories of interatomic forces dictate that the RDF must vanish. This results in unrealistic uncertainty estimates. A more rigorous approach has been introduced in which Bayesian uncertainty quantification is applied to the interatomic potential using experimentally derived RDFs as observations.[26] However, this parametric approach relies on a predefined functional form for the potential (for example, a Lennard−Jones or Mie potential), inherently constraining the model and introducing unnecessary bias. What is needed, therefore, is a probabilistic framework that can incorporate known physical features of the RDF, such as short-range exclusion and long-range decay, while remaining flexible enough to avoid biases imposed by assumed potential forms or molecular simulation models.

We propose that a mathematically rigorous version of the RDF posterior satisfying these requirements can be computed through the use of Bayesian inference on the experimental structure factor directly. Specifically, by placing a Gaussian process (GP) prior over the experimental structure factor, multiplying it with an appropriate likelihood function, and finally computing the rFT over the resulting q-space posterior distribution, we obtain a Bayesian posterior distribution on the RDF. The use of a prior regularizes the infinite set of possible functions that could fit the finite observed data set, while the likelihood serves as a data fit penalty. The resultant posterior distribution on the RDF represents a direct uncertainty quantification over the real-space structure, given the momentum space scattering observations.

GPs have been used extensively in solving ill-posed inverse problems in computational chemistry,[6] including Fourier

https://doi.org/10.1021/acs.jpcb.5c05024
J. Phys. Chem. B 2025, 129, 11802−11815

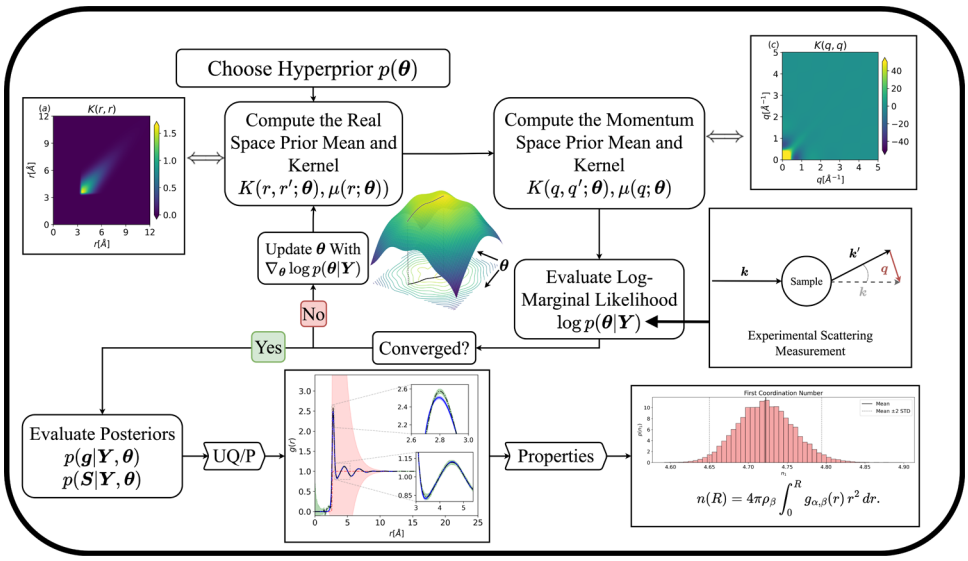


Figure 1. A flowchart corresponding to the GPFT algorithm applied to scattering data.

analysis of noisy and truncated signals, which plagues the scattering problem.[27] A GP-based approach was recently developed to analyze small-angle neutron scattering data, demonstrating that GP predictions can optimize neutron beamtime usage and increase experimental throughput without compromising data quality.[28] Furthermore, GPs naturally resolve many of the current challenges of scattering analysis cited earlier. For example, they can infer the structure factor on a continuum of momentum values with a domain consistent with the rFT (from $q = 0$ to $\lim_{q\to\infty} S(q)$). As long as the GP mean and kernel selection are physics-informed and flexible enough to represent the data, Bayesian inference will be robust up to available experiments and our theoretical understanding of the structure. Such a framework is more elegant and satisfactory than, say, neural networks or other black-box machine learning tools that often ignore expert knowledge and do not have uncertainty quantification built into their mathematical formalism. The GP framework, therefore, supports inference while maintaining transparency and expert interpretability.

In this study, we present a probabilistic machine learning framework to estimate total or partial RDFs with uncertainty quantification from background and inelastic corrected scattering data using nonstationary GP regression. We show how nonstationary GPs with a physics-informed mean and kernel conditioned on experimental scattering data enable the complete reconstruction of the atomic structure from both simulation and experimentally derived total structure factors. The mean and kernel selection reflect simple and indisputable properties of the RDF, including the correct limiting behaviors for realistic bulk fluids ($\lim_{r\to 0} g(r) = 0$, $\lim_{r\to\infty} g(r) = 1$), continuity and differentiability, and the presence of bonded and nonbonded contributions.

As test cases for the nonstationary GP model, we performed RDF inference for a simple liquid (argon) and a complex liquid (water) from structure factors derived from both simulation and experiment. All inferred RDF distributions are free from spurious Fourier artifacts and preserve tailing behaviors as dictated by the nonstationary kernel. For liquid argon, we find that the nonstationary GP prediction shows near-perfect agreement with a gold-standard neutron scattering analysis from Yarnell.[29] For water structure obtained from a classical water model with flexible bonds, the nonstationary GP regression reconstructs the ground truth data even with significant noise introduced to the structure factor signal. Once the nonstationary GP was validated, we investigated an X-ray scattering data set of liquid water[23] to obtain a novel interpretation of the oxygen-oxygen RDF distribution that can be compared with molecular models of water.

At first glance, it might seem relatively uninteresting to perform Bayesian inference over the RDF prediction. However, this type of computation can serve as a key result for validation of molecular dynamics simulation[23,30] and help unlock emerging methods in computational chemical physics. For instance, access to an RDF distribution as a GP could serve as a link between the increasingly popular Gaussian approximation potential (GAP) framework[31] and experimental scattering data. Additionally, force field optimization algorithms such as structure optimized potential refinement (SOPR),[32,33] which models the interatomic potential as a GP, can now propagate uncertainty directly from experimental observation into the estimation of interatomic potentials. The same is true of parametric Bayesian force field inference, which can be employed with the methods presented here to estimate how well a given molecular model represents complex experimental data. Finally, as Bayesian interpretations become more frequently integrated into chemical physics, these approaches will be necessary to understand model uncertainties with respect to experiments, a critical step of the scientific method that has been largely under-reported in the existing literature due to a lack of rigorous approaches to estimate uncertainty in complex experimental observables. This framework, therefore, allows us to leverage all available data without throwing away established physical knowledge accumulated through generations of scientific discovery (Figure 1).

THEORY AND METHODS

Consider a structure factor, which is unknown to us, that lives within a *distribution* of possible functions that are known to obey specific physical characteristics. We can write this mathematically,[1] in the context of a GP by stating that any

https://doi.org/10.1021/acs.jpcb.5c05024
J. Phys. Chem. B 2025, 129, 11802−11815

evaluation (or set of evaluations) of the unknown function, $S(q)$, is distributed as a multivariate Gaussian,

$$S(q) \sim \mathcal{GP}(\mu(q), K(q, q')) \tag{6}$$

where the mean, $\mu(q)$, represents the a priori expected value of the function at each point in the input space, and the covariance (or kernel) function, $K(q, q')$, represents the relatedness of the output quantities with respect to the process inputs. The mean and kernel constrain the set of possible functions that could represent the experimental observation to those consistent with physical intuition. The nonstationary behavior of such a GP refers to the fact that the Fourier transform of the true structure factor has limiting behavior with certainty (i.e., $\lim_{r\to 0} g(r) = 0$, $\lim_{r\to\infty} g(r) = 1$), meaning that the functional distribution has covariances that change with respect to its inputs.[34]

A physics-informed GP model enables Bayesian inference of a structure factor posterior distribution conditioned on the experimental scattering data. This posterior reflects uncertainties from both known physical principles and experimental observations, offering a reliable and robust estimation of the uncertainty in the liquid structure. Finally, due to the linearity of the rFT, this uncertainty can be propagated into the prediction of the RDF and subsequently compared to molecular simulation predictions. This rigorous representation of our current knowledge of the liquid structure serves as a powerful validation tool for molecular models and helps to pinpoint key measurements necessary to resolve gaps in our understanding of the organization of molecules in liquids.

The GP Model. Bayes' theorem provides a natural route to compute the structure factor posterior, $p(\boldsymbol{S}|\boldsymbol{Y}, \boldsymbol{\theta})$, according to Bayes theorem,

$$\begin{aligned} p(\boldsymbol{S}|\boldsymbol{Y}, \boldsymbol{\theta}) &= \frac{p(\boldsymbol{Y}|\boldsymbol{S}, \boldsymbol{\theta})p(\boldsymbol{S}|\boldsymbol{\theta})}{p(\boldsymbol{Y}|\boldsymbol{\theta})} \\ &= p(\boldsymbol{Y}|\boldsymbol{S}, \boldsymbol{\theta})p(\boldsymbol{S}|\boldsymbol{\theta})\left[\int p(\boldsymbol{Y}|\boldsymbol{S}, \boldsymbol{\theta})p(\boldsymbol{S}|\boldsymbol{\theta})\mathrm{d}\boldsymbol{S}\right]^{-1} \end{aligned} \tag{7}$$

where $p(\boldsymbol{S}|\boldsymbol{\theta})$, $p(\boldsymbol{Y}|\boldsymbol{S}, \boldsymbol{\theta})$, and $p(\boldsymbol{Y}|\boldsymbol{\theta})$ are the prior, likelihood, and model evidence, respectively. Here, $\boldsymbol{Y}$ is the set of experimental observations, $\boldsymbol{S}$ is the value of the structure factor due to some inducing vector, and $\boldsymbol{\theta}$ is the set of GP hyperparameters.

The *GP prior* over the inducing index vector of momentum transfer values $\boldsymbol{q}$ is defined by a mean, $\mu(\boldsymbol{q}) = \boldsymbol{\mu}_{\boldsymbol{q}}$, and a kernel, $K(\boldsymbol{q}, \boldsymbol{q}) = \widehat{\boldsymbol{K}}_{\boldsymbol{q},\boldsymbol{q}}$ function, which, when evaluated on the index set, produces a vector and a matrix, respectively,

$$\begin{aligned} p(\boldsymbol{S}|\boldsymbol{\theta}) = (2\pi)^{-d/2}\det|\widehat{\boldsymbol{K}}_{\boldsymbol{q},\boldsymbol{q}}|^{-1/2}\exp((\boldsymbol{S}-\boldsymbol{\mu}_{\boldsymbol{q}})^T \\ \widehat{\boldsymbol{K}}_{\boldsymbol{q},\boldsymbol{q}}^{-1}(\boldsymbol{S}-\boldsymbol{\mu}_{\boldsymbol{q}})) \end{aligned} \tag{8}$$

where the determinant is required due to the nondiagonal covariance between the latent function values. Note that this quantity is just a GP representation of the structure factor distribution before seeing any data.

The *likelihood* is the probability that the observed data are generated by a particular instance of $\boldsymbol{S}$. Assuming that the structure factor has approximately spatially uncorrelated and constant Gaussian noise (which is the case for reactor source neutron scattering[10]), an appropriate likelihood is a homoscedastic normal distribution,

$$p(\boldsymbol{Y}|\boldsymbol{S}, \boldsymbol{\theta}) = (2\pi\omega^2)^{-d/2}\exp((\boldsymbol{S}-\boldsymbol{Y})^T(\omega^2\hat{\boldsymbol{I}})^{-1}(\boldsymbol{S}-\boldsymbol{Y})) \tag{9}$$

where ω is a noise parameter, d is the number of observed data points, and $\hat{\boldsymbol{I}}$ is the identity matrix. Modified versions of this expression would be required for systems with significant heteroscedastic noise, such as spallation source neutron instruments. In this case, the likelihood can be generalized by replacing the scalar variance ω^2 with a function $\omega^2(\boldsymbol{q})$ varying along the diagonal of the covariance matrix while remaining jointly Gaussian.[35] In general, as long as the observations are independent, the likelihood can represent any noise distribution, underscoring the flexibility of the Bayesian framework.[36] Note that the dependence of these expressions on hyperparameters $\boldsymbol{\theta}$ stems from the underlying kernel and mean functions used to evaluate $\boldsymbol{S}$.

Finally, the conjugacy of Gaussian distributions for both the prior and likelihood enables analytical integration of the *model evidence* (also known as the marginal likelihood),[37,38] and upon taking the log to improve numerical stability, gives,

$$\begin{aligned} \log p(\boldsymbol{Y}|\boldsymbol{\theta}) = -\frac{1}{2}(\boldsymbol{Y}-\boldsymbol{\mu}_{\boldsymbol{q}})^T(\widehat{\boldsymbol{K}}_{\boldsymbol{q},\boldsymbol{q}}+\omega^2\hat{\boldsymbol{I}})^{-1}(\boldsymbol{Y}-\boldsymbol{\mu}_{\boldsymbol{q}}) \\ -\frac{1}{2}\log\det|\widehat{\boldsymbol{K}}_{\boldsymbol{q},\boldsymbol{q}}+\omega^2\hat{\boldsymbol{I}}| - \frac{d}{2}\log 2\pi \end{aligned} \tag{10}$$

Combining these expressions into eq 7, the posterior distribution over the latent function $\boldsymbol{S}$ evaluated at some m-sized index vector $\boldsymbol{q}^*$ is then,

$$\begin{aligned} p(\boldsymbol{S}|\boldsymbol{Y}, \boldsymbol{\theta}) = (2\pi)^{-m/2}\det|\widehat{\boldsymbol{\Sigma}}_{\mathrm{Post}}|^{-1/2}\exp((\boldsymbol{S}-\boldsymbol{\mu}_{\mathrm{Post}})^T \\ \widehat{\boldsymbol{\Sigma}}_{\mathrm{Post}}^{-1}(\boldsymbol{S}-\boldsymbol{\mu}_{\mathrm{Post}})) \end{aligned} \tag{11}$$

where the posterior mean and variance are given by,

$$\boldsymbol{\mu}_{\mathrm{Post}} = \boldsymbol{\mu}_{\boldsymbol{q}*} + \widehat{\boldsymbol{K}}_{\boldsymbol{q}*,\boldsymbol{q}}(\widehat{\boldsymbol{K}}_{\boldsymbol{q},\boldsymbol{q}}+\omega^2\hat{\boldsymbol{I}})^{-1}(\boldsymbol{Y}-\boldsymbol{\mu}_{\boldsymbol{q}}) \tag{12}$$

$$\widehat{\boldsymbol{\Sigma}}_{\mathrm{Post}} = \widehat{\boldsymbol{K}}_{\boldsymbol{q}*,\boldsymbol{q}*} - \widehat{\boldsymbol{K}}_{\boldsymbol{q}*,\boldsymbol{q}}(\widehat{\boldsymbol{K}}_{\boldsymbol{q},\boldsymbol{q}}+\omega^2\hat{\boldsymbol{I}})^{-1}\widehat{\boldsymbol{K}}_{\boldsymbol{q},\boldsymbol{q}*} \tag{13}$$

Posterior Estimation of the RDF. The next step is to propagate uncertainty from the structure factor distribution into real space. By the fluctuation−dissipation theorem, the RDF is related to the structure factor through a 3D Fourier transform, which, assuming spherical symmetry, can be written as the well-known rFT, which we denote $\widetilde{\mathcal{H}}$,

$$\begin{aligned} \widetilde{\mathcal{H}}_q[f(q)] &= \frac{1}{2\pi^2\rho}\int_0^\infty f(q)\frac{\sin(qr)}{qr}q^2\mathrm{d}q,\ \widetilde{\mathcal{H}}_r^{-1}[f(r)] \\ &= 4\pi\rho\int_0^\infty f(r)\frac{\sin(qr)}{qr}r^2\mathrm{d}r \end{aligned} \tag{14}$$

which maps a function of r to a function of q. Notably, the inverse of the rFT proportional to the operator is itself up to a proportionality constant ($\widetilde{\mathcal{H}} = 8\pi^3\rho^2\widetilde{\mathcal{H}}^{-1}$, see Supporting Information Section S2 for details) and is linear with respect to the input function. This operator is related to the Hankel transform.[39] The RDF structure factor relationship is then,

$$S(q) = 1 + \widetilde{\mathcal{H}}_r^{-1}[g(r)-1],\ g(r) = 1 + \widetilde{\mathcal{H}}_q[S(q)-1] \tag{15}$$

At first glance, it may seem unclear how to apply eq 15 to a distribution of structure factors; however, the normality of the GP, in tandem with the linearity of the operator, can alleviate

https://doi.org/10.1021/acs.jpcb.5c05024
J. Phys. Chem. B 2025, 129, 11802−11815

nearly all of the difficulty since the resulting distribution is trivially Gaussian. This nice property is due to the well-known fact that the linear transformation of a finite-dimensional Gaussian distribution is again Gaussian,

$$\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{\mu}, \widehat{\boldsymbol{\Sigma}}) \tag{16}$$

$$\Rightarrow \widehat{\boldsymbol{A}}\boldsymbol{z} \sim \mathcal{N}(\widehat{\boldsymbol{A}}\boldsymbol{\mu}, \widehat{\boldsymbol{A}}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{A}}^{T}) \tag{17}$$

where $\hat{\boldsymbol{A}}$ is a linear operator acting on a finite-dimensional vector $\boldsymbol{z}$. Assuming that the linear operator is bounded and densely defined, the same property holds for GPs.[40] This approach is often leveraged in the analysis of partial differential equations[41] and can be applied to the rFT integral operator defined in eq 15.

Eq 17 has important implications for relating kernels between the Fourier duals of momentum and real space. For example, we can now construct new kernels in the Fourier dual space by applying the linear rFT operator,

$$K(r, r') = \mathrm{cov}(g(r), g(r')) = \widetilde{\mathcal{H}}_q[\widetilde{\mathcal{H}}_{q'}[K(q, q')]] \tag{18}$$

$$K(r, q') = \mathrm{cov}(g(r), S(q')) = \widetilde{\mathcal{H}}_q[K(q, q')] \tag{19}$$

$$K(q, r') = \mathrm{cov}(S(q), g(r')) = \widetilde{\mathcal{H}}_r^{-1}[K(r, r')] \tag{20}$$

$$K(q, q') = \mathrm{cov}(S(q), S(q')) = \widetilde{\mathcal{H}}_r^{-1}[\widetilde{\mathcal{H}}_{r'}^{-1}[K(r, r')]] \tag{21}$$

In essence, the RDF posterior distribution reflects correlations between observed data in q-space projected into r-space, giving the overall probability of the RDF $\boldsymbol{g}$ evaluated on an n-sized index vector $\boldsymbol{r}$ as,

$$p(\boldsymbol{g}|\boldsymbol{Y}, \boldsymbol{\theta}) = (2\pi)^{n/2}\det|\widehat{\boldsymbol{\Sigma}}_{\mathrm{Post,RDF}}|^{-1/2}\exp((\boldsymbol{g} - \boldsymbol{\mu}_{\mathrm{Post,RDF}})^{T} \widehat{\boldsymbol{\Sigma}}_{\mathrm{Post}}^{-1}(\boldsymbol{g} - \boldsymbol{\mu}_{\mathrm{Post,RDF}})) \tag{22}$$

with posterior mean and variance,

$$\boldsymbol{\mu}_{\mathrm{Post,RDF}} = \boldsymbol{\mu}_{\boldsymbol{r}} + \widehat{\boldsymbol{K}}_{\boldsymbol{r},\boldsymbol{q}}(\widehat{\boldsymbol{K}}_{\boldsymbol{q},\boldsymbol{q}} + \omega^2\hat{\boldsymbol{I}})^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}_{\boldsymbol{q}}) \tag{23}$$

$$\widehat{\boldsymbol{\Sigma}}_{\mathrm{Post,RDF}} = \widehat{\boldsymbol{K}}_{\boldsymbol{r},\boldsymbol{r}} - \widehat{\boldsymbol{K}}_{\boldsymbol{r},\boldsymbol{q}}(\widehat{\boldsymbol{K}}_{\boldsymbol{q},\boldsymbol{q}} + \omega^2\hat{\boldsymbol{I}})^{-1}\widehat{\boldsymbol{K}}_{\boldsymbol{q},\boldsymbol{r}} \tag{24}$$

where $\boldsymbol{\mu}_{\boldsymbol{r}}$ is just $\widetilde{\mathcal{H}}_q[\mu(q) - 1]$ evaluated at $\boldsymbol{r}$. Formally, these expressions may also be obtained by application of the rFT $\widetilde{\mathcal{H}}_q$ operator to the $S(q)$ posterior evaluated at a single inducing point q.

While the above method works in theory, not all of the integrals are analytically tractable and conducive to pen and paper computation. Indeed, it is more practical to use an approximate operator, $\widetilde{\mathbb{H}}$, computed with simple numerical quadrature,

$$\widetilde{\mathcal{H}}_q[f(q)] \approx \sum_{i=1}^{N} \frac{1}{4\pi^2\rho} \left(f(q_{i-1})\frac{\sin(q_{i-1}r)}{q_{i-1}r}q_{i-1}^2 - f(q_i)\frac{\sin(q_i r)}{q_i r}q_i^2\right)\Delta q = \widetilde{\mathbb{H}}_q[f(\boldsymbol{q})] \tag{25}$$

where the grid of q values is over the range of the integral. Note that the approximate operator $\widetilde{\mathbb{H}}$ acts on a discretized grid of function values $f(\boldsymbol{q})$ and produces a single number which corresponds to the implicit radial argument r. The choice of grid spacing can affect the resulting RDFs and uncertainty predictions; therefore, care must be taken to ensure the grid resolves all relevant features of the kernel and mean functions used. Details on the exact grid spacing and limits used are provided in the Supporting Information. This approximate rFT operator retains linearity while generalizing to custom GP prior means and kernels that do not have analytical rFTs. This strategy holds connections with typical Bayesian quadrature techniques.[42] The inverse is discretized similarly to an alternate prefactor.

Designing a GP for Liquid Structure Factors. The crux of designing any GP model is choosing an appropriate prior, which for a GP is fully specified by its mean and kernel functions and their corresponding hyperparameters. This step is also the most critical for enforcing physical behaviors and constraints in the GP regression.

Because physical correlations are less transparent in momentum space, it is more intuitive to impose constraints directly on the real-space RDF, where structural features are more easily interpreted and well-understood. Given a real-space mean $\mu(r)$ and kernel $K(r, r')$, we can then perform an rFT using one of the techniques from the previous section to obtain $K(q, q')$ as well as the log marginal likelihood in eq 10. Past work has shown that capturing the limiting behaviors correctly can greatly improve the transform procedure.[17] Therefore, it is crucial to ensure proper boundary behaviors in the GP prior. We know there must be an excluded volume, as well as a trend toward 1 at the limit, to preserve the overall density of the fluid. Mathematically, these boundary conditions are expressed as,

$$\lim_{r\to 0} g(r) = 0 \quad \lim_{r\to\infty} g(r) = 1 \tag{26}$$

which can be incorporated into the GP model directly using a nonstationary kernel.

Nonstationary Kernel Selection. Unlike stationary kernels, which assume that the covariance only depends on the distance between inducing input locations ($k(r, r') = k(|r - r'|)$), nonstationary kernels allow the covariance to change across different regions of the input space. A cleverly designed nonstationary kernel can improve the model's predictive power by ensuring that the GP adheres to known physical constraints. To see why this is the case, consider that when the kernel evaluation approaches zero, the GP distribution tends toward the mean (c.f. eqs 12 and 13). This provides a natural strategy for enforcing the boundary conditions: force the mean to a known limiting behavior while forcing the covariance to vanish. Specifically, the limits we are after are,

$$\lim_{r \text{ or } r'\to 0 \text{ or } \infty} K(r, r') = 0, \quad \lim_{r\to 0}\mu(r) = 0, \quad \lim_{r\to\infty}\mu(r) = 1 \tag{27}$$

ensuring that the kernel captures localized variations away from the boundaries ($r = 0$ and $r \to\infty$), while the mean function encodes the global boundary behaviors.

Although the RDF is technically a map from $\mathbb{R}^+$ to $\mathbb{R}^+$, the GP itself is not restricted to this domain. To account for this, we impose symmetry with respect to $r = 0$,

$$K_{\mathrm{Sym.}}(r, r') = K(r, r') + K(-r, r') \tag{28}$$

https://doi.org/10.1021/acs.jpcb.5c05024
J. Phys. Chem. B 2025, 129, 11802−11815

Symmetrizing the kernel prevents artificial asymmetries in the model and ensures that the process behaves consistently across the full input space (for further details on improving numerical stability of kernel calculations, see Supporting Information Section S3). Additionally, we know that $g(r)$ must be continuous and differentiable, as it is required to belong to the radially symmetric Schwartz space to be Fourier transformable.[2] Therefore, we based our kernel on the widely used squared exponential kernel, but with nonstationary behavior introduced through r-dependent length scale $l(r)$ and a width scale $\sigma(r)$ functions. The kernel of this type is known as the Gibbs kernel,[43] which is highly flexible and allows for spatially varying properties,

$$K(r, r') = \sigma(r)\sigma(r')\sqrt{\frac{2l(r)l(r')}{l(r) + l(r')}}\exp\left(\frac{-(r - r')^2}{l(r) + l(r')}\right) \tag{29}$$

The flexibility of the Gibbs kernel makes it particularly well-suited for systems where the properties of the process change over space in a known way. By parametrizing $\sigma(r)$ and $l(r)$ using a chosen functional form, we can further incorporate known behaviors and enhance generalizability. Following the strategy outlined above, we aim to embed as much physically relevant behavior as possible into $\mu(r)$, while selecting $\sigma(r)$ and $l(r)$ to account for deviations from the mean.

In the fluid structures of concern to this work, it is atypical to see large length scale changes as a function of r (with the exception of the bonded vs nonbonded structure handled in the mean). This allows us to choose $l(r)$ to be a constant. The limiting behavior of the kernel at large or small inputs is then encoded within $\sigma(r)$. By ensuring that $\sigma(r)$ tends to zero as r tends to 0, ∞ the kernel will satisfy eq 27. Simple functional forms that satisfy these constraints are a constant length scale function and a decaying sigmoid for the width function

$$l(r) = l, \quad \sigma(r) = \frac{\text{Max} \times \exp(\text{Decay} \times \text{Loc})}{1 + \exp(-\text{Slope} \times (r - \text{Loc}))}\exp(-r \times \text{Decay})) \tag{30}$$

where the hyperparameters Max, Decay, Loc, and Slope control the height, decay rate, peak location, and sharpness of the peak in the sigmoid, respectively.

The presented phenomenological kernel represents an Occam's razor strategy to kernel design. However, although this kernel satisfies the relevant physical constraints, alternative functional forms with comparable properties and varying levels of rigor are certainly possible. In principle, both $l(r)$ and $\sigma(r)$ could be derived from first principles, modeled as latent functions (e.g., GPs with their own mean and kernel structures), or selected phenomenologically, as we do here. Conveniently, the Bayesian GP framework provides a principled way to compare kernels through the computation of Bayes factors, which are ratios of their respective marginal likelihoods, $\text{BF} = \frac{p(y \mid K_1)}{p(y \mid K_2)}$. The higher the value of the Bayes factor, the more the data support K_1. To compute the marginal likelihood for a candidate kernel, hierarchical inference over the kernel hyperparameters must be performed to fully account for uncertainty. Although this approach is rigorous, its high computational cost places it beyond the scope of the present study. Nonetheless, it remains a promising direction for future work in physics-informed kernel design and selection.

Mean Selection. The simplest information to include in the mean μ is the hard-particle-like repulsive shell and bond information. In simulations, bonds are often modeled using a harmonic oscillator, resulting in a sharp, approximately Gaussian peak in $g(r)$. This leads to the bonded portion of the mean being represented as a sum of Normal distributions,

$$\mu_{\text{Bonded}}(r) = \sum_{b=1}^{B} h_b \mathcal{N}(r|r_b, s_b) \tag{31}$$

Here, the sum is taken over each unique structural peak in the particular molecule. For instance, when studying the oxygen−hydrogen correlation of water, we would expect at least one peak to correspond to the oxygen−hydrogen bond. However, for larger molecules, the situation can become more complex. Consider the hydrogen−hydrogen correlations in benzene. Although each hydrogen atom is exclusively bonded to carbon, we still observe bond-like peaks in $g(r)$ due to the intramolecular hydrogen atoms still being in proximity with one another. These manifest as approximately normal peaks as if they were directly bonded. In principle, one could then relate the parameter r_b to the equilibrium bond lengths, s_b to the strength of the bonds, and h_b to the typical number of atoms at the distance r_b. The excluded volume part is then represented as a simple sigmoid. This choice aligns with the limit behavior of the mean outlined above,

$$\mu_{\text{Non-Bonded}}(r) = \frac{1}{1 + \exp(-s_0(r - r_0))} \tag{32}$$

Overall, the mean function for the GP model is then,

$$\mu(r) = \mu_{\text{Bonded}}(r) + \mu_{\text{Non-Bonded}}(r) \tag{33}$$

$$= \sum_{b=1}^{B} h_b \mathcal{N}(r|r_b, s_b) + \frac{1}{1 + \exp(-s_0(r - r_0))} \tag{34}$$

Consequences of selecting this particular real-space GP mean on the structure factor are further discussed in Supporting Information Section S4.

Hyperparameter Optimization. The final step of the method involves inferring the GP hyperparameters, given the experimental structure factor. For this task, we learn a Bayesian hyperposterior using a hierarchical inference scheme (type II maximum likelihood) with appropriately defined priors as described in Supporting Information Section S3. The computational cost of the nonstationary GP method is dominated by this hyperparameter optimization step, which entails repeated cubic-scaling GP evaluations in the number of experimental q-space points. The difficulty of optimization grows with the number of hyperparameters, as this typically goes hand-in-hand with a rougher objective function. The numerical Fourier transform unique to this work scales linearly with the number of r-space grid points and is negligible in comparison to the matrix inverse.

Despite the theoretical complexity, in our experiments, we found the optimization to run end-to-end within a couple of hours on a laptop given modest-sized data sets (200−500 observations). However, for a given set of hyperparameters, a single GP inference is of relatively trivial computational cost given standard memory and processor capabilities of modern personal computers up to approximately 10^3 observations. The

https://doi.org/10.1021/acs.jpcb.5c05024
J. Phys. Chem. B 2025, 129, 11802−11815

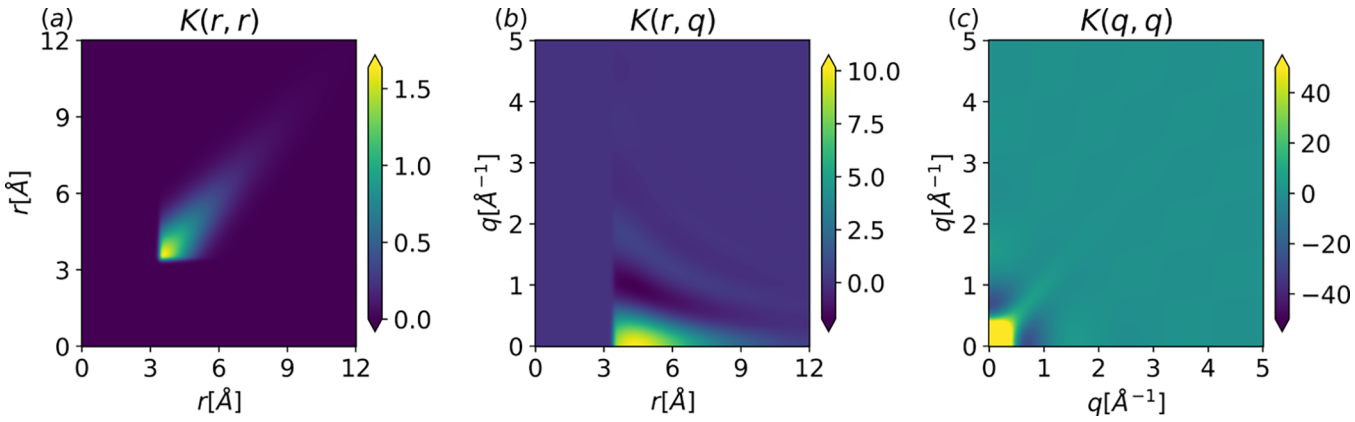


Figure 2. Gaussian process kernels after hyperparameter fitting of argon at a temperature T = 85 [K] and density ρ = 0.02125 [atom/Å^3]. Left corresponds to eq 18, middle corresponds to eq 19, and right corresponds to eq 21. The color bar represents the range of values indicated by the colormap. Any values outside the specified range are clipped and displayed by using the colors corresponding to the nearest boundary.

establishment of standardized ranges or additional physical constraints on the hyperparameters would support rapid characterization of numerous samples with the added benefit of uncertainty quantification. Our implementation of the optimization and inference procedure is available on GitHub at https://github.com/hoepfnergroup/LiquidStructureGP-Sullivan.

RESULTS

Having established the theoretical framework, we now demonstrate its utility on both simple and complex liquids through synthetic and experimental scattering data. (1) In a liquid argon scattering experiment, we demonstrate excellent agreement between GP-derived structure factors and results from a gold-standard neutron scattering analysis. Beyond numerical accuracy, the nonstationary GP provides enhanced physical interpretability through kernel heat maps and posterior covariance matrices, which visualize the relationship between momentum- and real-space features. (2) To validate the nonstationary GP framework in a molecular system, we applied the method to simulated liquid water with a known ground truth. We find that the GP reconstructs the real-space RDF, even under moderate noise. (3) Finally, we apply the framework to an experimental X-ray scattering experiment of liquid water. Here, the GP yields a novel prediction for the oxygen−oxygen RDF with uncertainty quantification. The posterior is then propagated to estimate posterior predictive statistics for the first and second peak positions and heights, as well as the coordination number. Both the visualization of error bars using a noise-free posterior and the GP-based inference of coordination numbers are detailed in Supporting Information Sections S5 and S6, respectively.

Liquid Argon. We begin by examining the quintessential neutron-weighted argon structure factor measured by Yarnell.[29] Although the data set incorporates post hoc modifications to address multiple scattering, background scattering, finite sample volumes, and noise, it remains widely regarded as a benchmark data set in the field. However, a key issue is that denoising alters the uncertainty estimation in the nonstationary GP method. To approximate the original, predenoised data and preserve realistic uncertainty estimates, we reintroduced constant Gaussian noise (σ^2_{noise} = 0.04) to the input estimated from a figure in Yarnell's manuscript.

Now, for the nonstationary GP construction of liquid argon, which has no bonded contributions, the hyperparameter vector is reduced to $\boldsymbol{\theta} = [r_0, s_0, l, \text{Max}, \text{Slope}, \text{Loc}, \text{Decay}, \omega]^T$. Figure 2 visualizes eqs 18−21 before hyperparameter optimization. Notably, the kernel matrices exhibit the distinct structural characteristics enforced through the prior. Specifically, the lack of structure at low radius values in eqs 18 and 19 corresponds to the excluded volume of the argon atoms. The presence of this feature in the prior distribution suggests that the atomic size is learned during the LMLH optimization of the prior mean hyperparameters rather than conditioning on the observed data. At medium to large values of r, there is a clear periodic structure in q, indicating both positive and negative correlations. This is an expected feature due to the underlying integration factor in eq 1 being a decaying sinusoid. Lastly, notice the magnitudes involved in each correlation. While the maximum value in the $K(r, r')$ correlation is typical of Ar, the magnitudes in $K(q, q')$ are greater than what is observed in experiments. This results in increased flexibility in the low q region that is inconsistent with known limiting behaviors of $S(q)$.

For example, the high variance at low q ($\sigma^2_q > 40$, Figure 2c) arises from model misspecification. This discrepancy appears to result from the absence of constraints on the total density and isothermal compressibility. To understand this, consider the case q = 0, where the sinc term in eq 1 tends to 1 in the $q \to 0$ limit so that $S(0) = 1 + 4\pi\rho\int_0^\infty (g(r) - 1)r^2 dr$. The behavior at q = 0 is then determined by the well-known compressibility equation, which relates $S(q)$ to the isothermal compressibility, $\lim_{q\to 0} S(q) = \rho k_B T \chi_T$. Hence, the large prior variance at q = 0 indicates that the function distribution does not have a fixed isothermal compressibility. Thermodynamic quantities of this type could be incorporated into the model directly as constraints, ensuring that all realizations of the experimental data are consistent with known thermodynamic quantities. Several strategies exist to enforce such consistency, including the use of warping functions, the design of kernel and mean functions that inherently satisfy the constraints, or the incorporation of auxiliary data to implicitly embed them.[41] Although such constraints are not utilized here, they remain a promising direction for future exploration of GPs as mathematical models in liquid state theory.

With these kernels, we perform hyperparameter optimization by minimizing eq 10 (see Supporting Information Section

https://doi.org/10.1021/acs.jpcb.5c05024
J. Phys. Chem. B 2025, 129, 11802−11815

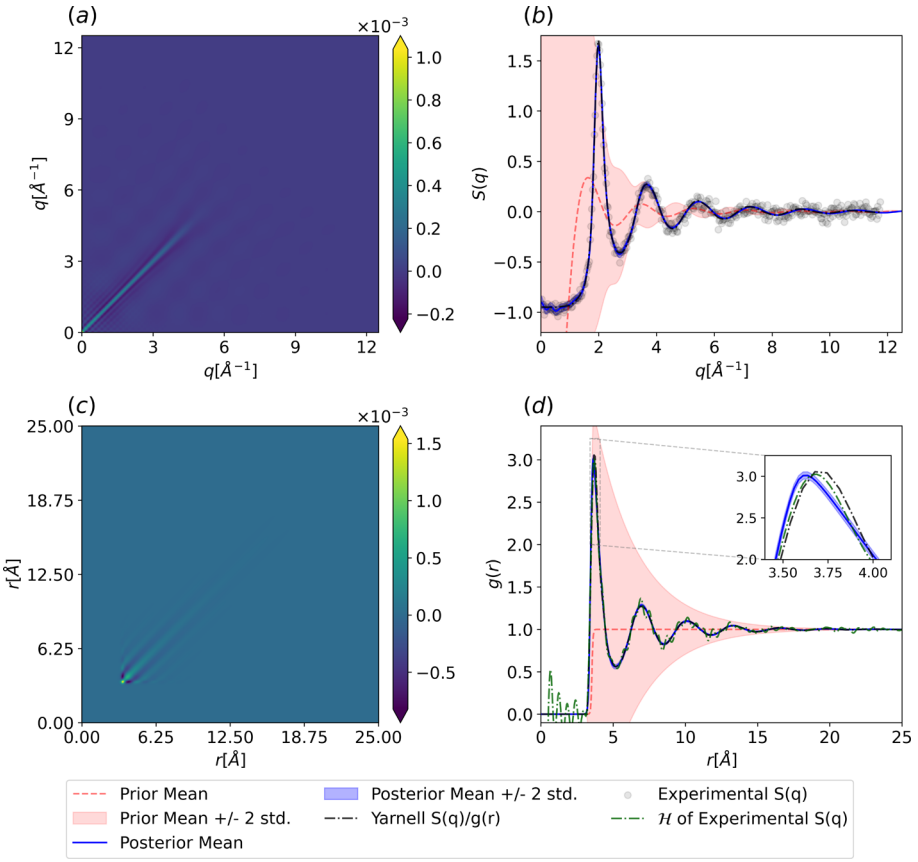


Figure 3. Posterior of the Gaussian process fit to argon at a temperature of T = 85 [K] and a density of ρ = 0.02125 [atom/Å^3]. (a) Posterior covariance in q-space from eq 13. (b) Prior and posterior distributions for the argon structure factor from eq 12. (c) Posterior covariance in r-space from eq 24. (d) Prior and posterior argon RDF from eq 23. The color bars are clipped similarly to Figure 2.

S7 for a table of initial and trained hyperparameters), after which the prior distribution over structure factors is conditioned on the data to yield a posterior distribution, along with the associated distribution over the RDF. This procedure is described by eqs 11−13 and 22−24. The posterior mean and covariance in $\boldsymbol{q}$ and $\boldsymbol{r}$ space are presented in Figure 3.

The direct rFT of the data in Figure 3d clearly exhibits $q_{\max}$ cutoff errors, manifested as high-frequency oscillations. As commented by Lorch in 1969, these high-frequency oscillations are often "erroneously identified as truncation ripples by other workers."[19] In the Yarnell interpretation, truncation ripples were removed through an iterative procedure. First, the structure factor $S(q)$ was artificially extended to the compressibility limit (q = 0), then directly Fourier-transformed to produce an initial estimate of $g(r)$. Next, $g(r)$ was set to zero in the low-r region ($0 \leq r \leq 0.8d$, where d is an estimated atomic diameter) and inverse Fourier-transformed back to yield an updated $S(q)$. This process was repeated iteratively, until it was no longer necessary to set $g(r)$ to zero at low-r. This iterative procedure is well-established in neutron scattering analysis, and the nonstationary GP framework naturally preserves the spirit of this procedure. The process of optimizing eq 10 formally uses the same scheme. In practice, both Yarnell's method and the nonstationary GP yield nearly identical predictions of the real-space structure, with deviations likely arising from the denoising procedure or imperfect hyperparameter optimization. To account for this type of uncertainty in the GP formalism, one would increase the hierarchy of the optimization and propagate $p(\boldsymbol{\theta}|Y)$ into the $g(r)$ distribution. Due to the associated computational cost as well as the negligible difference to Yarnell's results, we did not explore this avenue; however, it provides a clear direction for future work.

The nonstationary GP methodology also provides us with direct access to the posterior covariance matrix in real space. We can see in Figure 3c that the posterior covariance in real-space exhibits a highly nonstationary structure that is fully consistent with the physical constraints imposed in the kernel design stage. Namely, the zero covariance at short-range exactly mimics a certain low-r limit constraint, while the decaying covariance at high-r reflects the decay of the RDF oscillations to unity. In momentum space, the decreasing posterior covariance of the structure factor as a function of q demonstrates that the physics-informed prior on the RDF naturally leads to a physically consistent estimate of $S(q)$ through the rFT (Figure 3a).

Water. We now turn our attention to liquid water, a considerably more complex system due to the presence of chemical bonds and three partial structure factors. When inferring real-space structures in bonded systems, the nonstationary GP framework requires an additional prior mean term, given by eq 31, to handle the bonded part of the structure for the oxygen−hydrogen and hydrogen−hydrogen partial structure factors. Other than the additional hyperparameters introduced by the prior mean, the nonstationary GP regression proceeds in the exact same manner as in the liquid argon case.

https://doi.org/10.1021/acs.jpcb.5c05024
J. Phys. Chem. B 2025, 129, 11802−11815

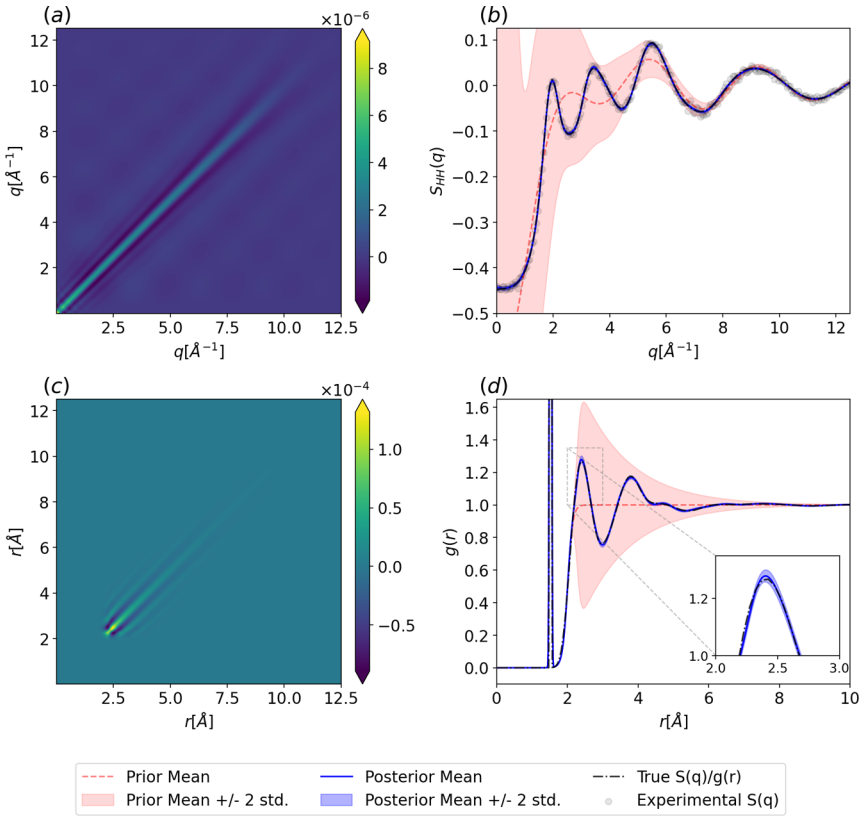


Figure 4. Posterior of the Gaussian process fit to structure factors derived from NVT simulations at a temperature of 298.15 K and density of 1 g cm^{-3} with flexible TIP4P/2005f water. (a) Posterior covariance in q-space from eq 13. (b) Prior and posterior distributions for the hydrogen−hydrogen structure factor from eq 12. (c) Posterior covariance in r-space from eq 24. (d) Prior and posterior hydrogen−hydrogen RDF from eq 23. The sharp feature observed at ∼1.6 Å represents the distance between the hydrogen atoms in a water molecule.

Simulated Liquid Water. First, we analyzed an artificial noisy unweighted structure of simulated water obtained from the flexible TIP4P/2005f water model[44] (for simulation and GP training details, see Supporting Information Section S8). The goal of this test was to verify that the nonstationary GP accurately recovers the ground truth real-space structure from noisy momentum-space data. This validation step is essential, as it builds confidence in the methodology before applying it to experimental data where the ground truth real-space structure is unknown.

In Figure 4, we show the posterior covariance and hydrogen−hydrogen partial structure factor and RDF including the GP prior (red), posterior (blue), perturbed structure factor (black dots), and ground truth (dashed black line). The hydrogen−hydrogen partial structure factor is presented here because it includes both bonded and nonbonded contributions, resulting in a more complex correlation structure with more hyperparameters than the oxygen−oxygen case, and thus poses a greater challenge for the method. The data in Figure 4b clearly show that the hyperparameter learning and regression result in an excellent posterior representation of the hydrogen−hydrogen partial structure factor as well as its rFT to real space in Figure 4d. The bonded and nonbonded regions of the hydrogen−hydrogen partial RDF are well captured by the model, with the ground truth lying within the estimated RDF posterior distribution. While we only show the hydrogen−hydrogen partial posterior distributions here, we note that the oxygen−hydrogen and oxygen−oxygen posterior distributions exhibit strong qualitative agreement and are provided in Supporting Information Section S8.

Experimental X-ray Scattering of Liquid Water. We now turn to the analysis of a broadened X-ray scattering data set for liquid water reported by Skinner and co-workers.[23] In X-ray scattering experiments on water, the signal is dominated by oxygen−oxygen correlations due to the weak scattering cross section of hydrogen arising from its single electron. As a result, assuming that background scattering corrections in the original data set were appropriately handled, the nonstationary GP model in this case needs only to infer the oxygen−oxygen correlation. We then compared our predictions to those of Skinner's interpretation. In their analysis, a variable Lorch modification function developed by Soper and Barney[21] was applied with fixed parameters (a = 2.8 and b = 0.5 Å), and error propagation was performed using the method of Weitkamp.[25]

The nonstationary GP applied to the structure factor provides an excellent representation of the underlying structure given the noisy experimental X-ray scattering data (for GP training details, see Supporting Information Section S9). More insightful, however, is the comparison between the nonstationary GP model and Skinner's interpretation shown in Figure 5d,e. While the mean predictions from both methods closely align at and beyond the first peak, significant differences emerge at lower distances. Specifically, Skinner's interpretation exhibits nonphysical oscillations below the collision diameter of the oxygen atom, an artifact of Fourier truncation, that leads to nonphysical negative values of $g(r)$. Even more striking are the differences in uncertainty estimates between the two methods. Skinner's uncertainty rapidly increases at low-r, primarily reflecting known limitations of the applied error-

https://doi.org/10.1021/acs.jpcb.5c05024
J. Phys. Chem. B 2025, 129, 11802−11815

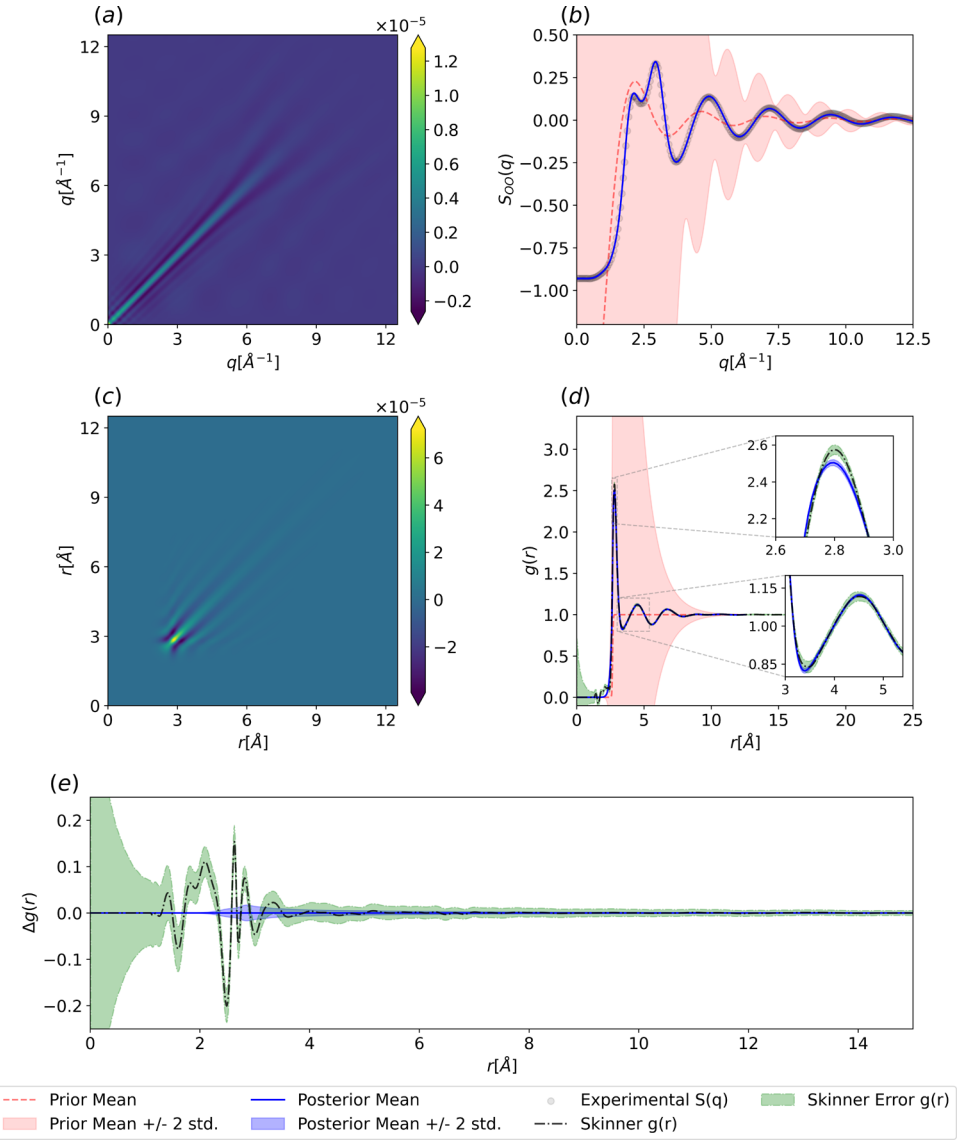


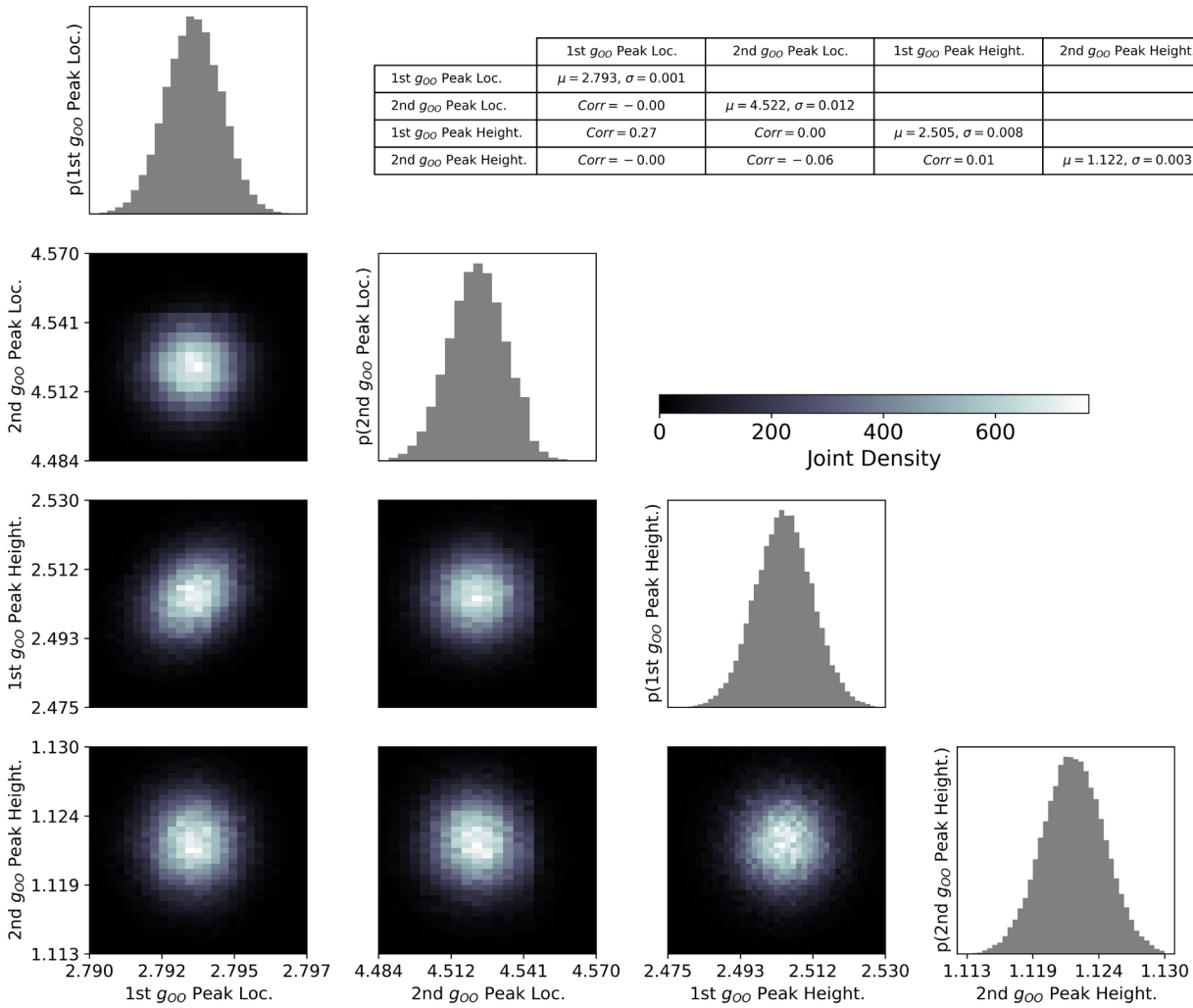
Figure 5. Posterior of the Gaussian process fit to the X-ray scattering data. (a) Posterior covariance in q-space from eq 13. (b) Prior and posterior distributions for the oxygen−oxygen structure factor from eq 12. (c) Posterior covariance in r-space from eq 24. (d) Prior and posterior oxygen−oxygen RDF from eq 23. (e) GP Mean subtracted comparison between the uncertainty estimates from the nonstationary GP approach and Skinner's interpretation.[23]

estimation procedure,[25] making it unclear whether these uncertainties genuinely represent data-informed variability or methodological artifacts. Conversely, the nonstationary GP uncertainty profile matches physical expectations: negligible uncertainty at low-r, a pronounced increase reaching a maximum around the first peak, followed by a gradual decay to negligible uncertainty at large distances. Notably, our interpretation predicts a slightly larger uncertainty in the local structure of water in the first solvation shell.

A physically justified posterior distribution on the RDF can subsequently be used to estimate statistics on other observables, such as peak heights and peak positions (Figure 6) as well as coordination number (Figure 7). Since the distributions are nearly Gaussian, we present the mean plus or minus two standard deviations ($\mu \pm 2\sigma$) as a summary statistic. The first peak location is estimated to be 2.793 ± 0.002, and the first peak height is 2.505 ± 0.016. The joint 2D marginal distributions over peak location and peak height show near-zero correlation for every two parameter sets, aside from a slight positive correlation (0.27) between the first peak height and first peak location. Finally, the estimated first coordination number is 4.722 ± 0.07, which is in good agreement with the generally accepted value determined from X-ray scattering data of 4.7.[45]

Given that water is of central importance for a wide variety of fields and is the subject of substantial investigation over its local structure, it is worth reflecting on which structural interpretations should serve as benchmarks for molecular modeling. The nonstationary GP framework presented here offers not only a physics-informed reconstruction of the RDF, but also a posterior predictive distribution that quantifies uncertainty in a principled way. This makes it particularly well-suited for benchmarking simulations, where one expects model predictions -- in this case RDFs, coordination numbers, peak heights, and peak positions -- to fall within credible intervals that reflect both experimental noise and structural ambiguity. While no interpretation is free from assumptions or limitations, the GP-based approach provides a transparent and reprodu-

https://doi.org/10.1021/acs.jpcb.5c05024
J. Phys. Chem. B 2025, 129, 11802−11815



	1st g_{OO} Peak Loc.	2nd g_{OO} Peak Loc.	1st g_{OO} Peak Height.	2nd g_{OO} Peak Height.
1st g_{OO} Peak Loc.	$\mu = 2.793$, $\sigma = 0.001$			
2nd g_{OO} Peak Loc.	$Corr = -0.00$	$\mu = 4.522$, $\sigma = 0.012$		
1st g_{OO} Peak Height.	$Corr = 0.27$	$Corr = 0.00$	$\mu = 2.505$, $\sigma = 0.008$	
2nd g_{OO} Peak Height.	$Corr = -0.00$	$Corr = -0.06$	$Corr = 0.01$	$\mu = 1.122$, $\sigma = 0.003$

Figure 6. This corner plot shows the joint distribution of the first and second RDF peak locations and heights. The table shows the mean and standard deviation values of the marginals along its diagonal. The off-diagonal terms are the Pearson correlation coefficients in the joint marginal distributions.

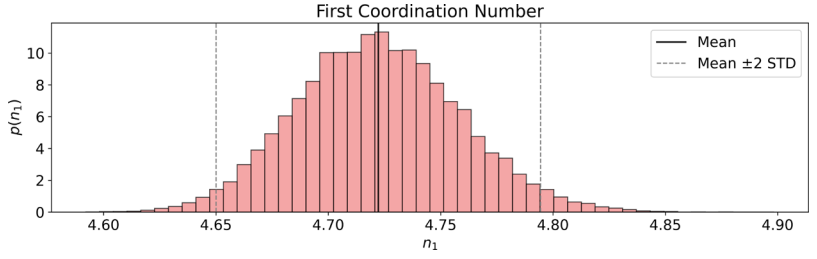


Figure 7. Histogram estimation of the coordination number probability density derived from samples of the nonstationary GP posterior.

cible framework that balances physics with statistical rigor and is likely to be a more robust choice for comparison against molecular models.

DISCUSSION

The nonstationary GP framework offers a principled, data-driven approach to structural inference by combining Bayesian inference with physics-informed priors. We demonstrate accurate fits for test systems ranging from liquid Ar and TIP4P/2005f water to experimental X-ray scattering of liquid water with only modest physical assumptions. The method operates on a continuous domain in both momentum and real space to mitigate problems with binning artifacts or q_{max} truncation errors. The model effectively filters normal random noise present in the experimental observations, which ultimately stabilizes any subsequent analysis of downstream properties, such as peak heights, locations, and coordination numbers. By avoiding simulation-specific biases (e.g., from force field parameters, thermostats, or numerical integrators), the GP posterior provides a reproducible and physically grounded benchmark. Moreover, its flexibility allows for the inclusion of additional physics-based constraints (e.g., isothermal compressibility limits, virial equations, Kirkwood−

https://doi.org/10.1021/acs.jpcb.5c05024
J. Phys. Chem. B 2025, 129, 11802−11815

Buff integrals[4]), making it a powerful tool for bridging scattering data with macroscopic thermodynamic properties.

The nonstationary GP method also hints at a deeper connection between structure and interatomic potentials when there is noise present in the data. The Henderson inverse theorem, which shows that the RDF for pairwise additive and homogeneous systems has a pair potential that is unique up to an additive constant, is derived in the noiseless limit.[2] However, as shown empirically by Soper, when noise is present, there is an ensemble of potential energy functions corresponding to the scattering data target.[46] We hypothesize that the nonstationary GP method can be considered a Laplace approximation on the true structural posterior determined by such a potential energy ensemble. Adopting this philosophical perspective could significantly enhance our understanding of structure−thermodynamics relationships and improve the accuracy of thermodynamic predictions.

Despite these advantages, several limitations remain. The nature of scattering experiments themselves poses the problem that species with low scattering length densities may be effectively invisible in the total signal. Additionally, all structure-analysis methods confront the underdetermined nature of the Faber−Ziman decomposition for multicomponent systems and mixtures, which permits multiple RDF solutions consistent with the same total scattering data, leading to nonuniqueness.[47] The GP framework could be extended to these cases by representing each partial structure factor as a linear combination of nonstationary GPs; however, this approach quickly increases the number of hyperparameters to be inferred, leading to higher computational cost. It may also be possible to develop methods analogous to EPSR that consistently integrate experimental data with molecular models of the interatomic potential within a Bayesian framework (c.f. ref 32), though, to our knowledge, no such algorithm has been attempted.

Furthermore, absent *perfect* experimental data processing procedures (e.g., background, multiple scattering, inelasticity, etc.) and scattering statistics resulting from an infinite radiation flux, no fluid structure interpretation can be entirely unbiased. While our approach assumes that these corrections are accurate, extending the GP framework to model them explicitly would be a valuable step toward a more complete and uncertainty-aware analysis. In principle, the GP framework can incorporate such systematic corrections hierarchically, for example, by using non-Gaussian likelihoods for scattering corrections, performing Bayesian inference over parametric models of systematic errors, and estimating time-of-flight uncertainty via error-in-variables approaches.

Finally, the GP prior mean and kernel used here represent just one of many choices that can satisfy the physical constraints. Future work could explore alternative priors that more tightly enforce thermodynamic behavior or integrate knowledge of interatomic potentials obtained through other Bayesian schemes.[26,32] As the methodology evolves, refining and exploring alternative priors are likely to improve interpretability, computational efficiency, and predictive accuracy.

CONCLUSIONS

We introduce a method for rigorous uncertainty quantification and propagation in experimentally derived radial distribution functions using physics-informed, nonstationary GP regression. This approach constructs a minimal yet physically expressive kernel that preserves the Fourier duality between the structure factor and the RDF. By addressing pervasive challenges in the Fourier transformation of momentum-space scattering data and incorporating Bayesian inference, our approach offers a robust and interpretable alternative to traditional structural analysis methods.

Applied to both simple and complex liquids, the model yields physically reasonable posterior distributions for radial distribution functions that capture both mean behavior and structural uncertainty. Crucially, the nonstationary GP framework achieves this without relying on computationally intensive molecular simulations that may be affected by systematic model bias imposed by force field assumptions. Its flexibility allows for the principled incorporation of physical knowledge and integration of data preprocessing steps through hierarchical modeling. Taken together, we conclude that the Bayesian framework established in this study may represent the best path toward an unbiased assessment of fluid structure.

ASSOCIATED CONTENT

Data Availability Statement

Structure factors, radial distribution functions with credibility intervals, posterior means and covariances, and nonstationary GP hyperparameters are available in the Supporting Information, from the corresponding authors upon reasonable request, and also provided on Github at https://github.com/hoepfnergroup/LiquidStructureGP-Sullivan.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpcb.5c05024.

Python notebook for importing.txt files of the posterior mean and covariances for liquid argon, TIP4P−F water, and liquid water at atmospheric pressure and 295 K (ZIP)

Section S1: notation; Section S2: deriving the radial Fourier transform operator; Section S3: numerical considerations and hyperparameter optimization; Section S4: discussion on prior mean selection; Section S5: understanding error bar visualization for GPs; Section S6: GP coordination number analysis; Section S7: neutron weighted argon structure analysis (Table S1: initial and final GP hyperparameters; Figure S1: optimization analysis, posteriors, and residuals); Section S8: simulated unweighted water data set (Figure S2: simulated oxygen−oxygen predictions; Figure S3: simulated oxygen−hydrogen predictions; Figures S4−S6: optimization progression, posteriors and residuals for the oxygen−oxygen, oxygen−hydrogen, and hydrogen−hydrogen partials, respectively); Section S9 debroadened X-ray water analysis (Table S2: initial and final GP hyperparameters; Figure S7: optimization analysis, posteriors, and residuals) (PDF)

AUTHOR INFORMATION

Corresponding Authors

Brennon L. Shanks − *Institute of Organic Chemistry and Biochemistry of the Czech Academy of Sciences, 166 10 Prague 6, Czech Republic;* orcid.org/0000-0002-3453-7258; Email: shanks.brennon@uochb.cas.cz

Michael P. Hoepfner − *Department of Chemical Engineering, University of Utah, Salt Lake City, Utah 84112-9203,*

https://doi.org/10.1021/acs.jpcb.5c05024
J. Phys. Chem. B 2025, 129, 11802−11815

United States; orcid.org/0000-0001-9648-6911; Email: michael.hoepfner@utah.edu

Authors

Harry Winston Sullivan − *Department of Chemical Engineering and Material Science, University of Minnesota - Twin Cities, Minneapolis, Minnesota 55455, United States;* orcid.org/0009-0009-4062-6132

Matej Cervenka − *Institute of Organic Chemistry and Biochemistry of the Czech Academy of Sciences, 166 10 Prague 6, Czech Republic*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpcb.5c05024

Author Contributions

H.W.S.: conceptualization (equal), algorithm development (lead), code implementation (lead), writing - original draft (equal) B.L.S.: conceptualization (equal), algorithm development (supporting), writing - original draft (equal) M.C.: validation (supporting) M.P.H.: algorithm development (supporting), writing - review and editing (lead), funding acquisition (lead).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This study is supported by the EFRC-MUSE, an Energy Frontier Research Center funded by the U.S. Department of Energy, Office of Science, Basic Energy Sciences under Award No. DE-SC0019285. The support and resources from the Center for High Performance Computing at the University of Utah are gratefully acknowledged. We would like to thank Aryan Deshwal for reading a preliminary version of the manuscript and providing helpful comments on Gaussian processes and kernel design.

■ ADDITIONAL NOTES

[1]For a concise list of notation used in this manuscript, see Supporting Information Section S1.

[2]Notably one can extend the domain of the FT to the tempered distributions, we do not consider such cases in this work.

■ REFERENCES

(1) Ornstein, L. S.; Zernike, F. Accidental deviations of density and opalescence at the critical point of a single substance. *Proc. Acad. Sci. Amsterdam* **1914**, *17*, 793−806.

(2) Henderson, R. L. A uniqueness theorem for fluid pair correlation functions. *Phys. Lett. A* **1974**, *49*, 197−198.

(3) Hansen, J.; McDonald, I. R. *Theory of Simple Liquids: With Applications to Soft Matter*; Academic Press, 2013.

(4) Kirkwood, J. G.; Buff, F. P. The statistical mechanical theory of solutions. *I. J. Chem. Phys.* **1951**, *19*, 774−777.

(5) Mackerell, A. D., Jr. Empirical force fields for biological macromolecules: Overview and issues. *J. Comput. Chem.* **2004**, *25*, 1584−1604.

(6) Deringer, V. L.; Bartók, A. P.; Bernstein, N.; Wilkins, D. M.; Ceriotti, M.; Csányi, G. Gaussian process regression for materials and molecules. *Chem. Rev.* **2021**, *121*, 10073−10141.

(7) Headen, F. T.; Cullen, P. L.; Patel, R.; Taylor, A.; Skipper, N. T. The structures of liquid pyridine and naphthalene: The effects of heteroatoms and core size on aromatic interactions. *Phys. Chem. Chem. Phys.* **2018**, *20*, 2704−2715.

(8) Cervenka, M.; Shanks, B. L.; Mason, P. E.; Jungwirth, P. Cation-π Interactions in Biomolecular Contexts by Neutron Scattering and Molecular Dynamics: A Case Study of the Tetramethylammonium Cation. *J. Phys. Chem. B* **2025**, *129*, 6911−6918.

(9) Fan, S.; Mason, P. E.; Chamorro, V. C.; Shanks, B. L.; Martinez-Seara, H.; Jungwirth, P. Charge Scaling Force Field for Biologically Relevant Ions Utilizing a Global Optimization Method. *J. Chem. Theory Comput.* **2025**, *21*, 9023−9034.

(10) Willis, B. T. M.; Carlile, C. J. *Experimental Neutron Scattering*; Oxford University Press, 2017.

(11) Faber, T. E.; Ziman, J. M. A theory of the electrical properties of liquid metals: III. the resistivity of binary alloys. *Philosophical Magazine: A Journal of Theoretical Experimental and Applied Physics* **1965**, *11*, 153−173.

(12) McGreevy, R. L.; Pusztai, L. Reverse Monte Carlo Simulation: A New Technique for the Determination of Disordered Structures. *Mol. Simul.* **1988**, *1*, 359−367.

(13) Soper, A. K. Empirical potential Monte Carlo simulation of fluid structure. *Chem. Phys.* **1996**, *202*, 295−306.

(14) Petersen, D. P.; Middleton, D. Sampling and reconstruction of wave-number-limited functions in *N*-dimensional euclidean spaces. *Information and Control* **1962**, *5*, 279−323.

(15) Neuefeind, J.; Feygenson, M.; Carruth, J.; Hoffmann, R.; Chipley, K. K. The nanoscale ordered materials diffractometer NOMAD at the spallation neutron source SNS. *Nucl. Instrum. Methods. Phys. Res. B* **2012**, *287*, 68−75.

(16) Shanks, B. L.; Sullivan, H. W.; Hoepfner, M. P. Bayesian Analysis Reveals the Key to Extracting Pair Potentials from Neutron Scattering Data. *J. Phys. Chem. Lett.* **2024**, *15*, 12608−12618.

(17) Proctor, J. E.; Pruteanu, C. G.; Moss, B.; Kuzovnikov, M. A.; Ackland, G. J.; Monk, C. W.; Anzellini, S. A comparison of different Fourier transform procedures for analysis of diffraction data from noble gas fluids. *J. Appl. Phys.* **2023**, *134*, 114701.

(18) Torquato, S. Hyperuniform states of matter. *Phys. Rep.* **2018**, *745*, 1−95.

(19) Lorch, E. Neutron diffraction by germania, silica and radiation-damaged silica glasses. *J. Phys. C: Solid State Phys.* **1969**, *2*, 229.

(20) Soper, A. K.; Barney, E. R. On the use of modification functions when Fourier transforming total scattering data. *J. Appl. Crystallogr.* **2012**, *45*, 1314−1317.

(21) Soper, A. K.; Barney, E. R. Extracting the pair distribution function from white-beam X-ray total scattering data. *J. Appl. Crystallogr.* **2011**, *44*, 714−726.

(22) Bellissent-Funel, M. C.; Buontempo, U.; Filabozzi, A.; Petrillo, C.; Ricci, F. P. Neutron diffraction of liquid neon and xenon along the coexistence line. *Phys. Rev. B* **1992**, *45*, 4605−4613.

(23) Skinner, L. B.; Huang, C.; Schlesinger, D.; Pettersson, L. G. M.; Nilsson, A.; Benmore, C. J. Benchmark oxygen-oxygen pair-distribution function of ambient water from x-ray diffraction measurements with a wide Q-range. *J. Chem. Phys.* **2013**, *138*, No. 074506.

(24) Soper, A. K. The radial distribution functions of water as derived from radiation total scattering experiments: Is there anything we can say for sure? *Int. Scholarly Res. Not.* **2013**, *2013*, No. e279463.

(25) Weitkamp, T.; Neuefeind, J.; Fischer, H. E.; Zeidler, M. D. Hydrogen bonding in liquid methanol at ambient conditions and at high pressure. *Mol. Phys.* **2000**, *98*, 125−134.

(26) Shanks, B. L.; Sullivan, H. W.; Shazed, A. R.; Hoepfner, M. P. Accelerated Bayesian inference for molecular simulations using local Gaussian process surrogate models. *J. Chem. Theory Comput.* **2024**, *20*, 3798−3808.

(27) Ambrogioni, L.; Maris, E. Integral transforms from finite data: An application of Gaussian process regression to Fourier analysis. In *AISTATS*, 2018; pp 217−225.

(28) Tung, C.; Yip, S.; Huang, G.; Porcar, L.; Shinohara, Y.; Sumpter, B. G.; Ding, L.; Do, C.; Chen, W. Unlocking hidden information in sparse small-angle neutron scattering measurements. *J. Colloid Interface Sci.* **2025**, *692*, No. 137554.

(29) Yarnell, J. L.; Katz, M. J.; Wenzel, R. G.; Koenig, S. H. Structure factor and radial distribution function for liquid argon at 85K. *Phys. Rev. A* **1973**, *7*, 2130−2144.
(30) Amann-Winkel, K.; Bellissent-Funel, M.; Bove, L. E.; Loerting, T.; Nilsson, A.; Paciaroni, A.; Schlesinger, D.; Skinner, L. X-ray and Neutron Scattering of Water. *Chem. Rev.* **2016**, *116*, 7570−7589.
(31) Deringer, V. L.; Caro, M. A.; Csányi, G. Machine learning interatomic potentials as emerging tools for materials science. *Adv. Mater.* **2019**, *31*, No. 1902765.
(32) Shanks, B. L.; Potoff, J. J.; Hoepfner, M. P. Transferable force fields from experimental scattering data with machine learning assisted structure refinement. *J. Phys. Chem. Lett.* **2022**, *13*, 11512−11520.
(33) Shanks, B. L.; Sullivan, H. W.; Jungwirth, P.; Hoepfner, M. P. Experimental evidence of quantum Drude oscillator behavior in liquids revealed with probabilistic iterative Boltzmann inversion. *J. Chem. Phys.* **2025**, *162*, 164501.
(34) Heinonen, M.; Mannerström, H.; Rousu, J.; Kaski, S.; Lähdesmäki, H. Non-stationary Gaussian process regression with Hamiltonian Monte Carlo. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, 2016; pp 732−740.
(35) Goldberg, P.; Williams, C.; Bishop, C. Regression with Input-dependent Noise: A Gaussian Process Treatment. In *NeurIPS*, **1997**.
(36) Li, R.; John, S. T.; Solin, A. Improving Hyperparameter Learning under Approximate Inference in Gaussian Process Models. In *Proceedings of the 40th International Conference on Machine Learning*, 2023; pp 19595−19615.
(37) Bishop, C. M. *Pattern Recognition and Machine Learning*; Information science and statistics; Springer, 2006.
(38) Rasmussen, C. E.; Williams, C. K. I. *Gaussian Processes for Machine Learning*; MIT Press, 2006.
(39) Baddour, N. Application of the generalized shift operator to the Hankel transform. *SpringerPlus* **2014**, *3*, 246.
(40) Matsumoto, T.; Sullivan, T. J. Images of Gaussian and other stochastic processes under closed, densely-defined, unbounded linear operators. *arXiv* **2024**.
(41) Swiler, L.; Gulian, M.; Frankel, A.; Safta, C.; Jakeman, J. A Survey of Constrained Gaussian Process Regression: Approaches and Implementation Challenges. *arXiv* **2020**.
(42) O'Hagan, A. Bayes−Hermite quadrature. *J. Stat. Plan. Inference* **1991**, *29*, 245−260.
(43) Gibbs, M. N. Bayesian Gaussian Processes for Regression and Classification. Ph.D. thesis, University of Cambridge, 1997.
(44) González, M. A.; Abascal, J. L. F. A flexible model for water based on TIP4P/2005. *J. Chem. Phys.* **2011**, *135*, 224516.
(45) Head-Gordon, T.; Johnson, M. E. Tetrahedral structure or chains for liquid water. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 7973−7977.
(46) Soper, A. K. Tests of the empirical potential structure refinement method and a new method of application to neutron diffraction data on water. *Mol. Phys.* **2001**, *99*, 1503−1516.
(47) Soper, A. K. On the uniqueness of structure extracted from diffraction experiments on liquids and glasses. *J. Phys.: Condens. Matter* **2007**, *19*, 415108.

https://doi.org/10.1021/acs.jpcb.5c05024
J. Phys. Chem. B 2025, 129, 11802−11815