Probabilistic Machine Learning for Statistical Mechanical Inverse Problems

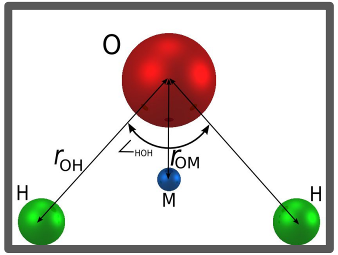
Brennon L. Shanks
Harry W. Sullivan
Abdur R. Shazed
PI: Michael P. Hoepfner

University of Utah, Department of Chemical Engineering

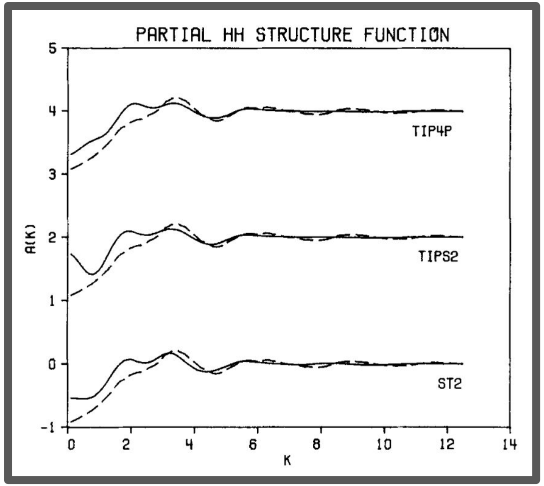
The Forward Problem: The Standard Method for Modeling

Molecular Model
(DFT, MD, AIMD, etc.)

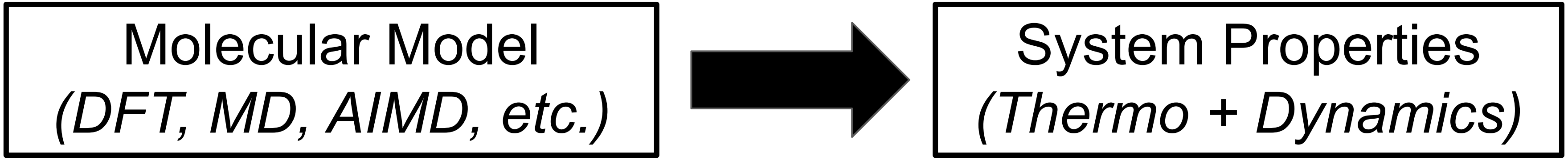
System Properties
(Thermo + Dynamics)



Comparison of simple potential functions for simulating liquid water
W. L. Jorgensen et al. 1983, *J. Chem. Phys.*



The Forward Problem: The Standard Method for Modeling



Molecular Model

(DFT, MD, AIMD, etc.)

System Properties

(Thermo + Dynamics)

The philosophy behind the forward problem is that we create a model for nature and then predict the value of measurements given that model.

Some Challenges with the Forward Problem Approach

What model should I choose?

Quantum?

Classical?

Continuum?

Phenomenological?

Machine Learning?

Some Challenges with the Forward Problem Approach

What model should I choose?

How good are my model parameters?

Quantum?

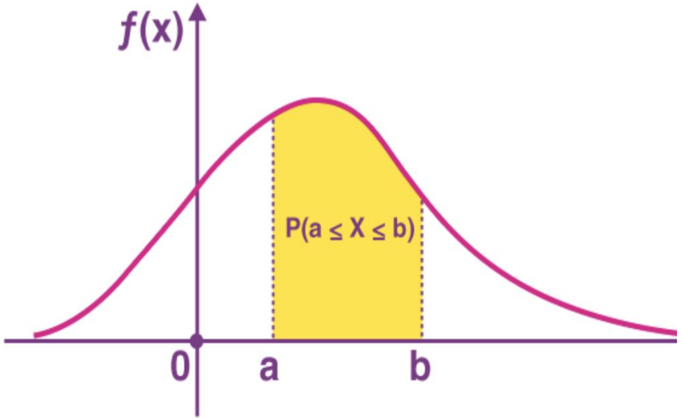
Classical?

Continuum?

Phenomenological?

Machine Learning?

Every model parameter has an associated uncertainty.



Some Challenges with the Forward Problem Approach

What model should I choose?

How good are my model parameters?

Is my model appropriate, or am I overfitting?

Quantum?

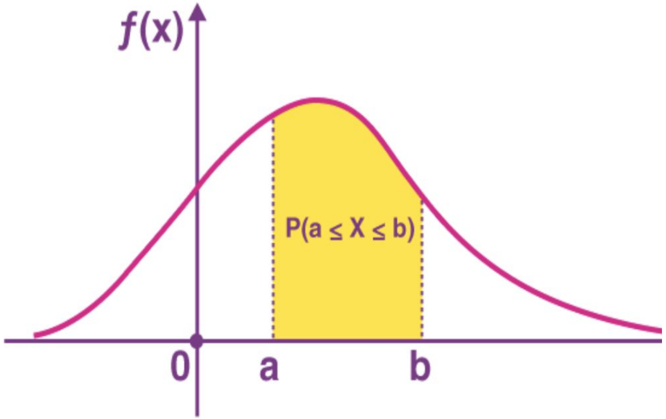
Classical?

Continuum?

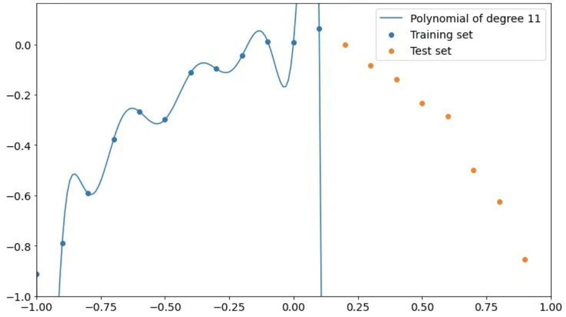
Phenomenological?

Machine Learning?

Every model parameter has an associated uncertainty.



Example: We can always fit an n-degree polynomial to n data points, but does that mean it is a perfect physical model? (NO!)



Some Challenges with the Forward Problem Approach

What model should I choose?

How good are my model parameters?

Is my model appropriate, or am I overfitting?

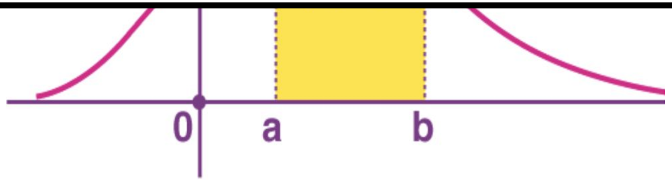
Quantum?

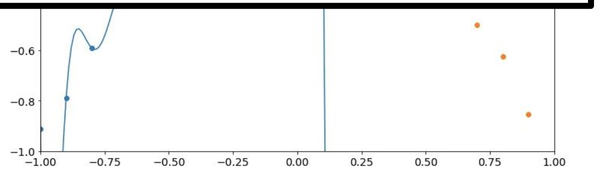
Every model parameter has

Example: We can always fit an

We can address all of these problems with Bayesian uncertainty quantification!

Machine Learning?





Bayesian methods as a framework to quantify uncertainty

The Basic Outline of Bayesian Approaches

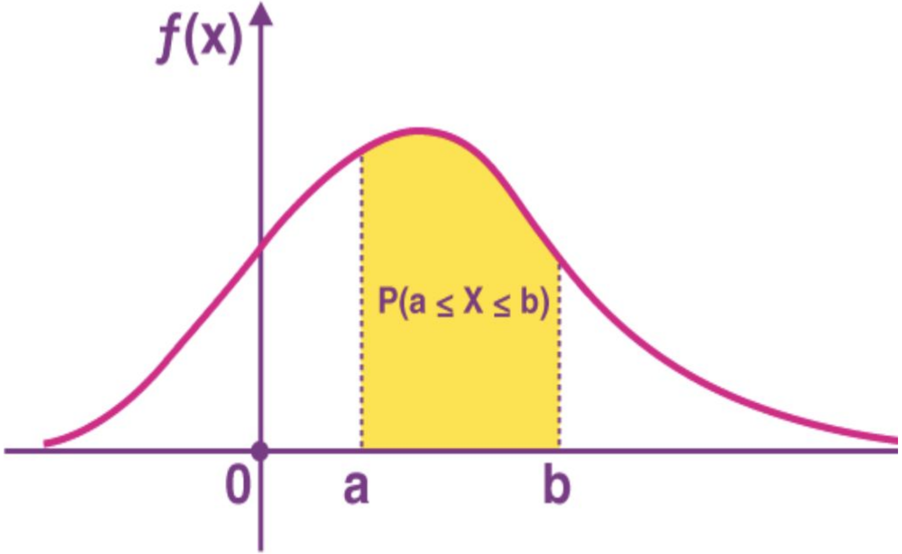
Bayesian methods as a framework to quantify uncertainty

The Basic Outline of Bayesian Approaches

(1) Define **'prior'** probability distributions

$p(\theta)$

The prior is our initial state of knowledge



Should be **wide** and **flat** to allow for all (reasonable) possibilities

Bayesian methods as a framework to quantify uncertainty

The Basic Outline of Bayesian Approaches

(1) Define **‘prior’** probability distributions

(2) Define and evaluate a **‘likelihood’** function

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

$$p(\mathscr{Y}|\theta)p(\theta)$$

The likelihood reflects how accurately our model parameters (θ) fit the experimental data (y)

Bayesian methods as a framework to quantify uncertainty

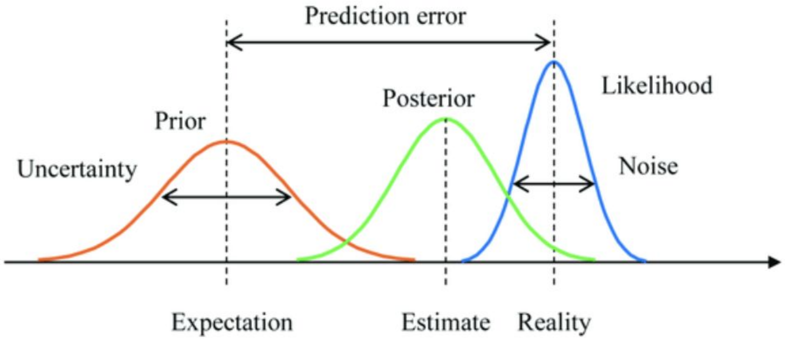
The Basic Outline of Bayesian Approaches

(1) Define **'prior'** probability distributions

(2) Define and evaluate a **'likelihood'** function

(3) Solve for the **'posterior'** distribution

$$p(\theta|\mathscr{Y}) = \frac{p(\mathscr{Y}|\theta)p(\theta)}{p(\mathscr{Y})}$$



The posterior is the new probability of parameters after observations

Bayesian methods as a framework to quantify uncertainty

The Basic Outline of Bayesian Approaches

(1) Define **‘prior’** probability distributions

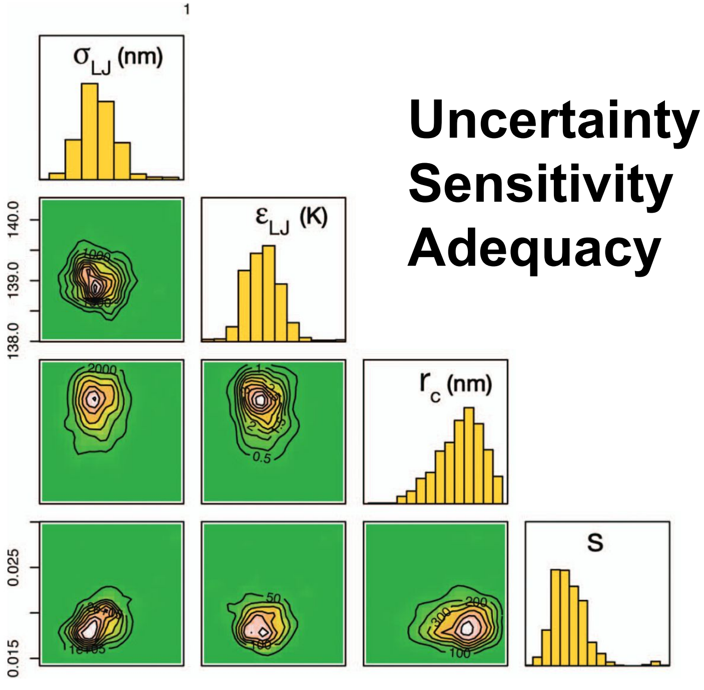
(2) Define and evaluate a **‘likelihood’** function

(3) Solve for the **‘posterior’** distribution

$$p(\theta|\mathscr{Y}) = \frac{p(\mathscr{Y}|\theta)p(\theta)}{p(\mathscr{Y})}$$

The posterior is a direct quantification of parameter uncertainty based on your experimental data, Y.

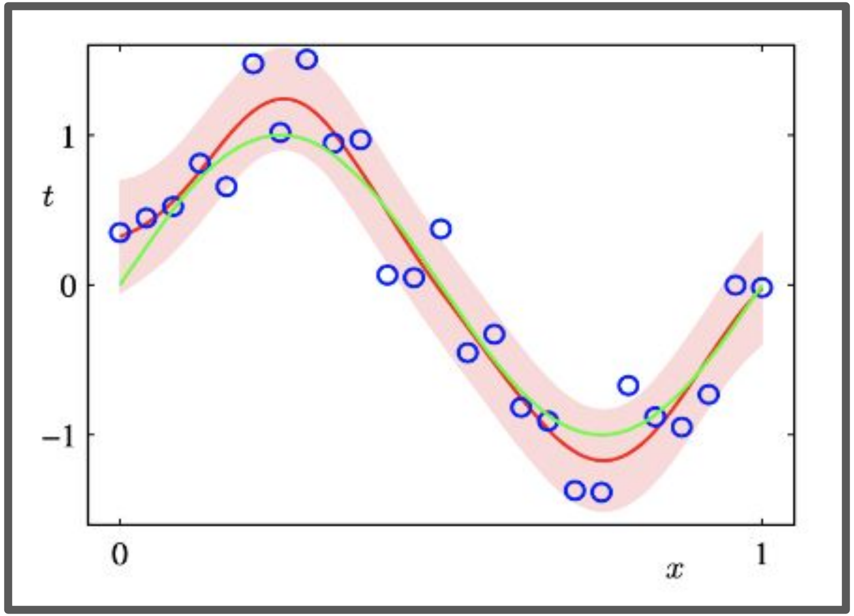
Uncertainty Quantification

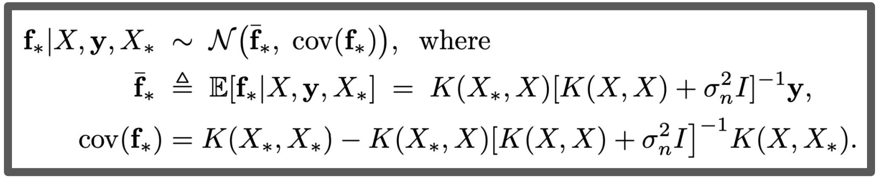


Uncertainty
Sensitivity
Adequacy

Marginal Posteriors on LJ Parameters
Koumoutsakos 2012, *J. Chem. Phys.*

Bayesian Methods on Non-Parametric Functions! - Gaussian Processes



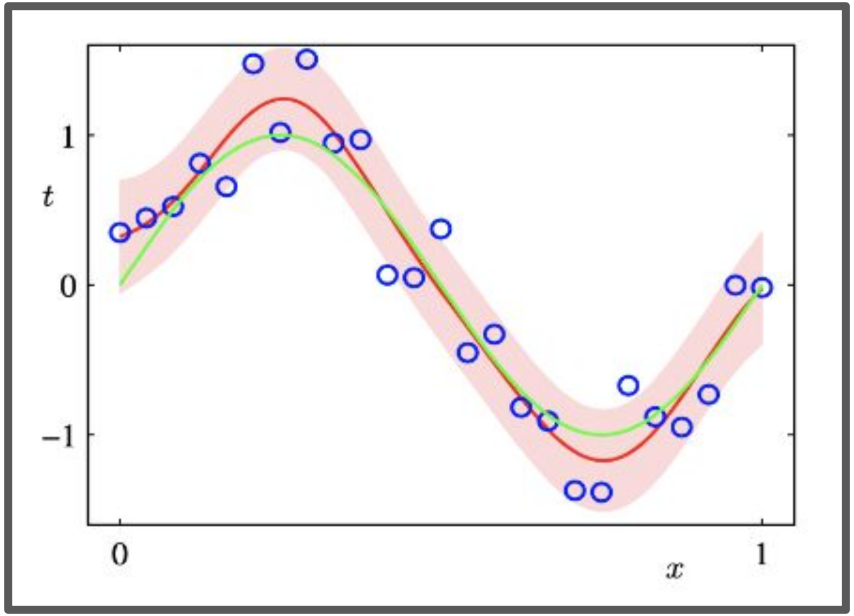


$$\mathbf{f}_*|X,\mathbf{y},X_* \sim \mathcal{N}\big(\bar{\mathbf{f}}_*,\ \mathrm{cov}(\mathbf{f}_*)\big), \text{ where}$$

$$\bar{\mathbf{f}}_* \triangleq \mathbb{E}[\mathbf{f}_*|X,\mathbf{y},X_*] = K(X_*,X)[K(X,X)+\sigma_n^2 I]^{-1}\mathbf{y},$$

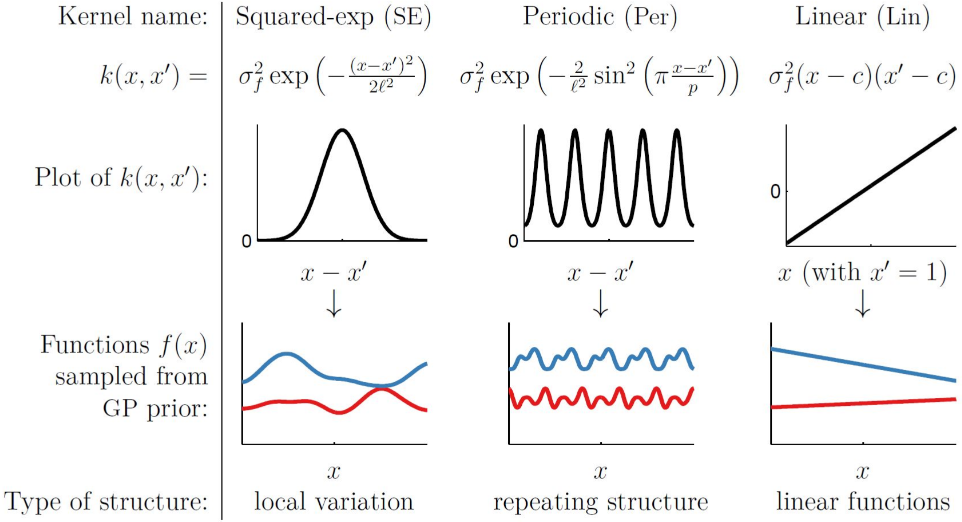
$$\mathrm{cov}(\mathbf{f}_*) = K(X_*,X_*) - K(X_*,X)[K(X,X)+\sigma_n^2 I]^{-1}K(X,X_*).$$

Bayesian Methods on Non-Parametric Functions! - Gaussian Processes



$$\mathbf{f}_*|X,\mathbf{y},X_* \sim \mathcal{N}\big(\bar{\mathbf{f}}_*,\ \operatorname{cov}(\mathbf{f}_*)\big), \text{ where}$$
$$\bar{\mathbf{f}}_* \triangleq \mathbb{E}[\mathbf{f}_*|X,\mathbf{y},X_*] = K(X_*,X)[K(X,X)+\sigma_n^2 I]^{-1}\mathbf{y},$$
$$\operatorname{cov}(\mathbf{f}_*) = K(X_*,X_*) - K(X_*,X)[K(X,X)+\sigma_n^2 I]^{-1}K(X,X_*).$$

Kernels specify the Gaussian process ‘prior’ over functions!

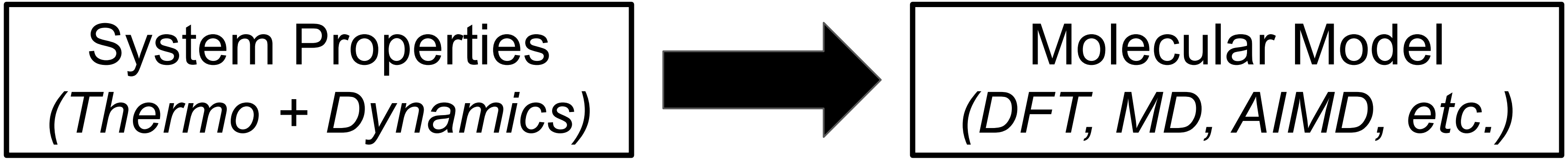


Kernel name:	Squared-exp (SE)	Periodic (Per)	Linear (Lin)
$k(x,x') =$	$\sigma_f^2 \exp\left(-\frac{(x-x')^2}{2\ell^2}\right)$	$\sigma_f^2 \exp\left(-\frac{2}{\ell^2}\sin^2\left(\pi\frac{x-x'}{p}\right)\right)$	$\sigma_f^2(x-c)(x'-c)$
Plot of $k(x,x')$:	$x - x'$	$x - x'$	x (with $x' = 1$)
Functions $f(x)$ sampled from GP prior:	x	x	x
Type of structure:	local variation	repeating structure	linear functions

Make observations and predict function with uncertainty!

Applications of Bayesian Methods in Statistical Mechanical Inverse Problems

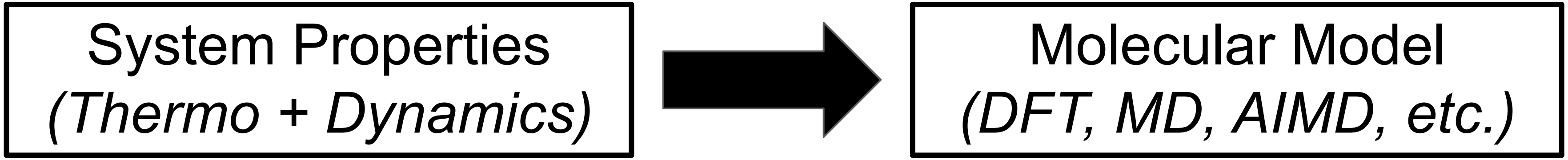
What is an Inverse Problem?



System Properties
(Thermo + Dynamics)

Molecular Model
(DFT, MD, AIMD, etc.)

What is an Inverse Problem?



System Properties
(Thermo + Dynamics)

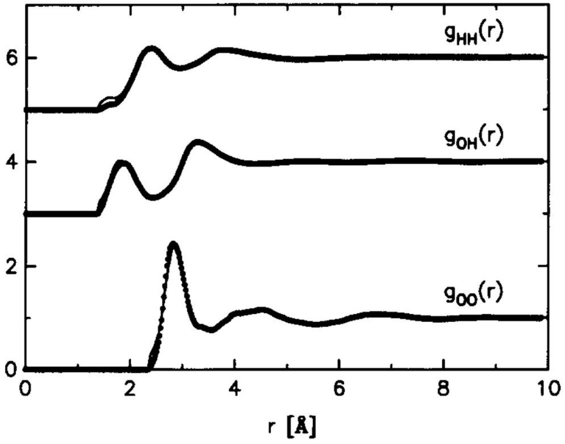
Molecular Model
(DFT, MD, AIMD, etc.)

The philosophy behind inverse problems is that we learn a model for nature based on experimental observation.

This is also the idea behind Bayesian methods!

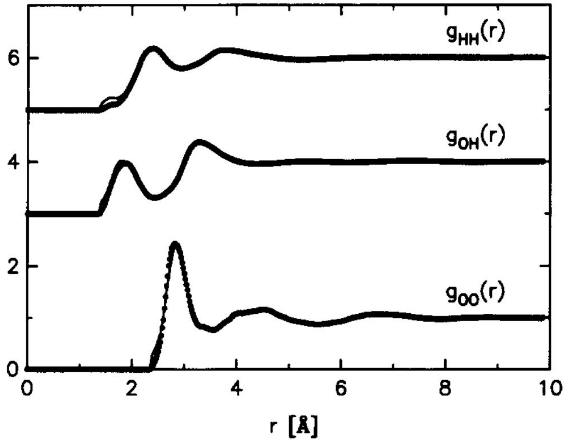
Applications of Inverse Problems for Interesting Chemistry

Scattering Analysis for
Ill-Posed Structure Prediction
A. K. Soper 1996, *Chem.*

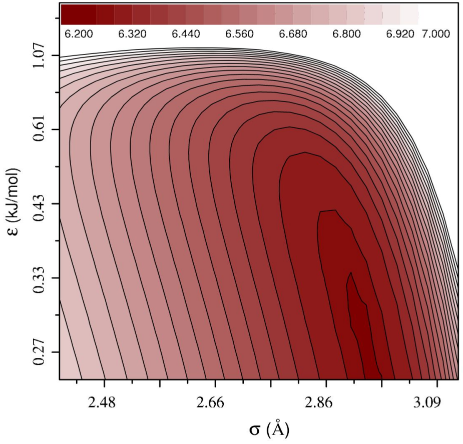


Applications of Inverse Problems for Interesting Chemistry

Scattering Analysis for
Ill-Posed Structure Prediction
A. K. Soper 1996, *Chem.*

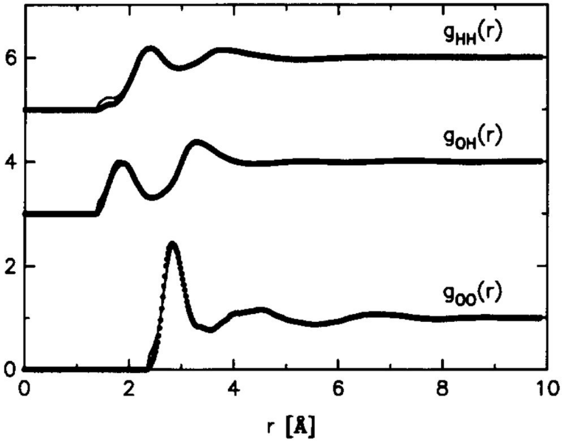


Coarse-Graining
Carmichael et al. 2013,
J. Chem. Phys.

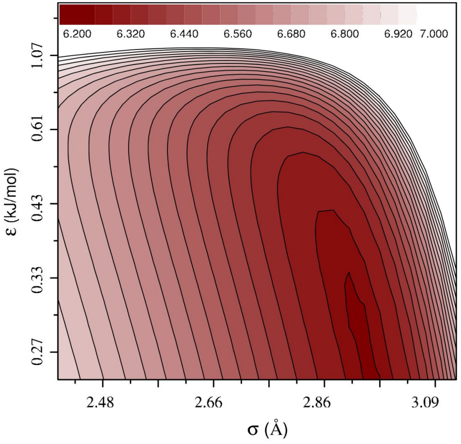


Applications of Inverse Problems for Interesting Chemistry

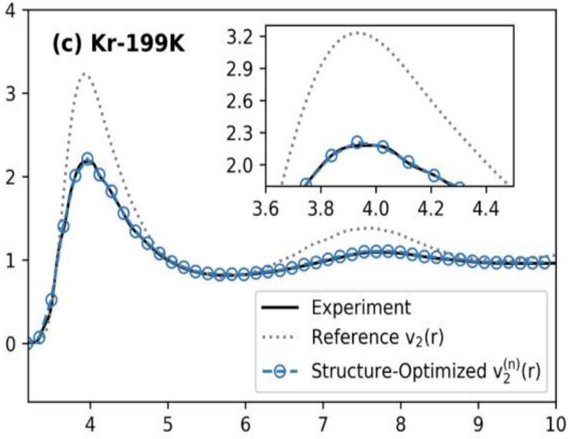
Scattering Analysis for Ill-Posed Structure Prediction
A. K. Soper 1996, *Chem.*



Coarse-Graining
Carmichael et al. 2013,
J. Chem. Phys.



Structure Optimized Potential Refinement
B. L. Shanks 2022, *J. Phys. Chem. Lett.*



I. Structure-Optimized Potential Refinement (SOPR): Learning Interaction Potentials from Scattering Data

Motivation

- From statistical mechanics, we know that we can predict all thermodynamic properties of a system if we know both the structure and potential energy.

- The “inverse problem” involves finding the potential energy given experimental data on the atomic positions (scattering).

- Researchers have been looking for a solution to this inverse problem for over a century, and **no robust and accurate method** has ever been demonstrated.

- We attempted to revisit this problem using the powerful method of Bayesian inference.

Training Force Fields from Scattering Data with SOPR

SOPR Algorithm

(1) Run molecular simulation with potential Vo, calculate simulated RDF

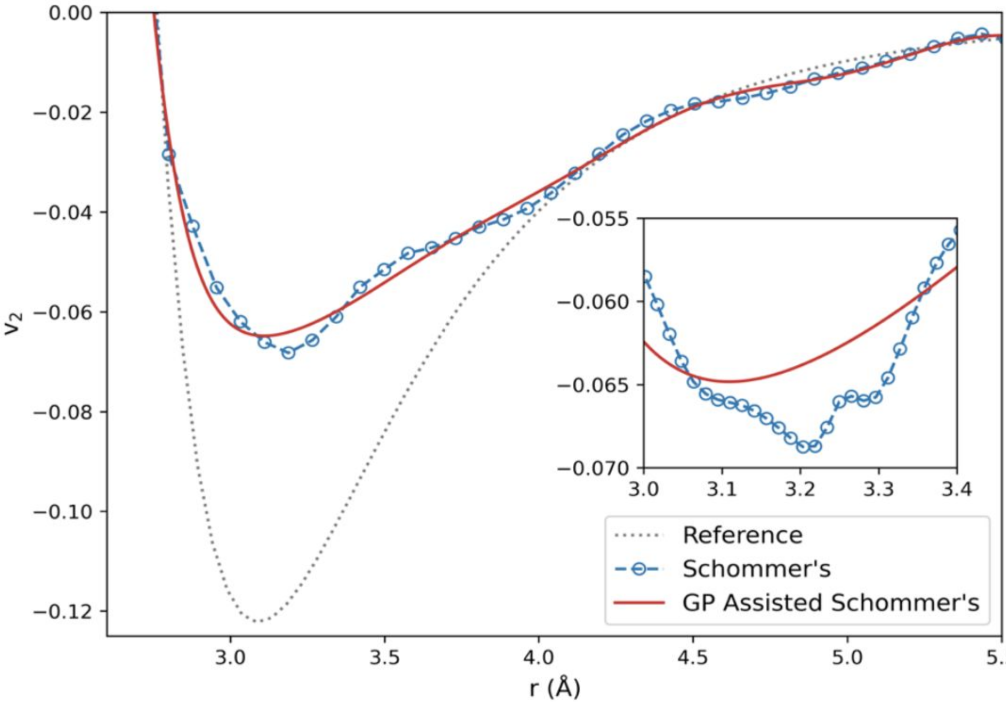
(2) $$v_2^{(n)'}(r_i) = v_2^0(r_i) + \gamma\beta^{-1}\sum_n \Delta g^{(n)'}(r_i)$$

(3) **Gaussian process regression for force stability** (figure to the right).

(4) Run new molecular simulation and check for consistency between exp and sim

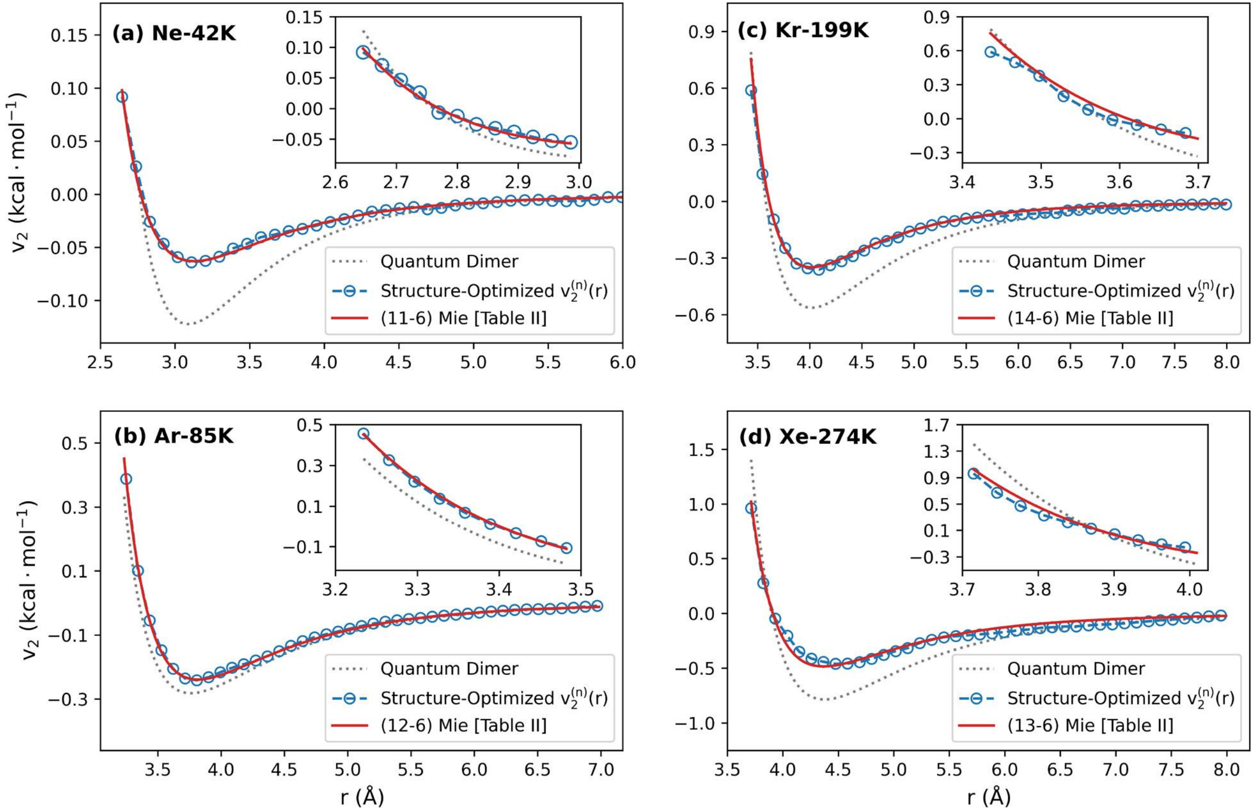
(5) Repeat until converged!

GP for Probabilistic Regression



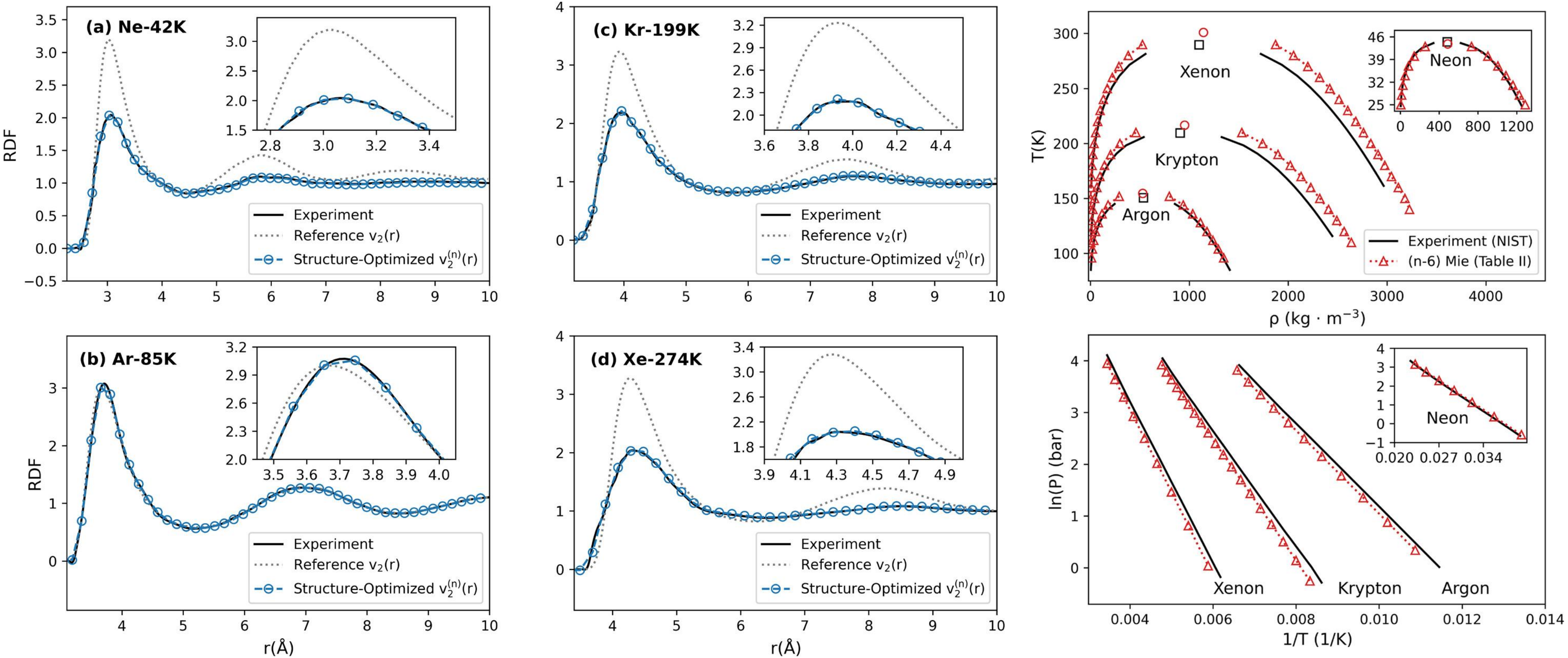
Noble Gas Force Fields from Scattering Data

SOPR Potentials Generated from Neutron Scattering Data



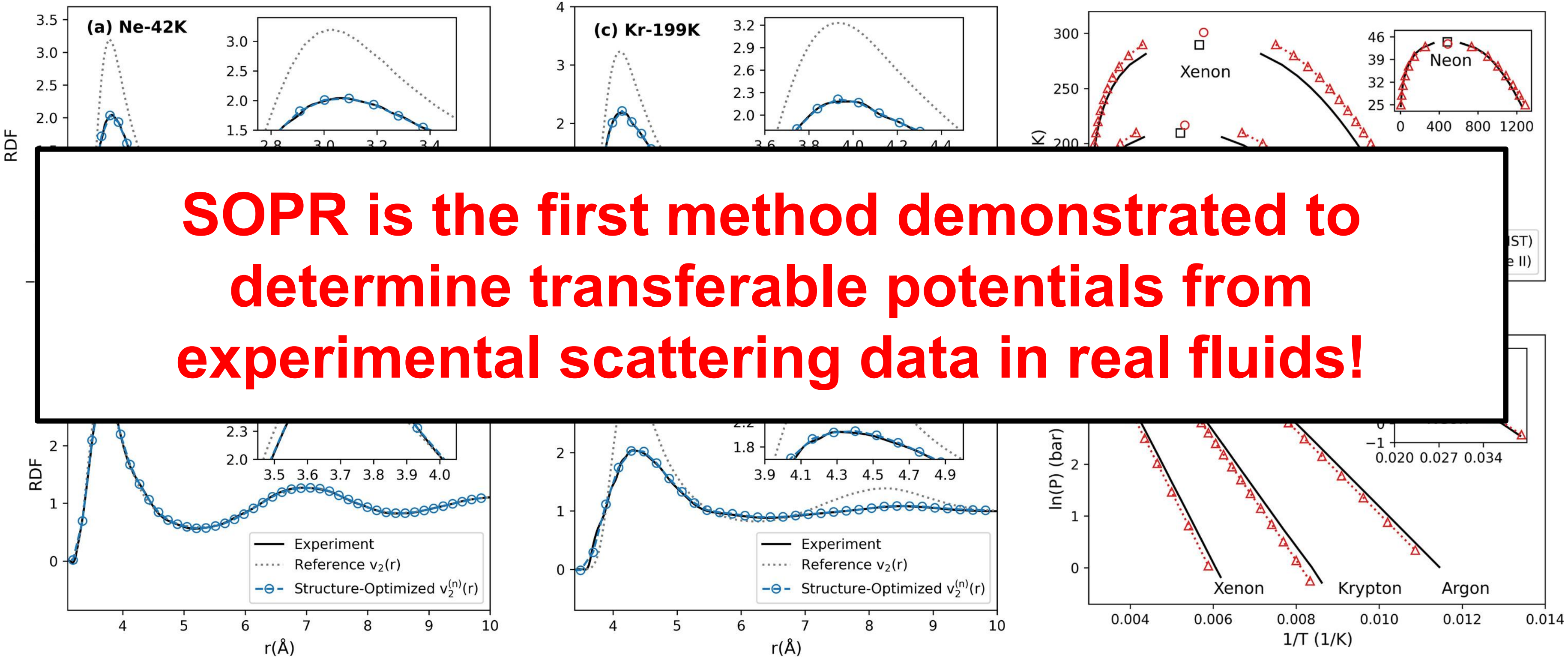
Noble Gas Force Fields from Scattering Data

Radial Distribution Functions and Vapor Liquid Equilibrium Match with Excellent Agreement



Noble Gas Force Fields from Scattering Data

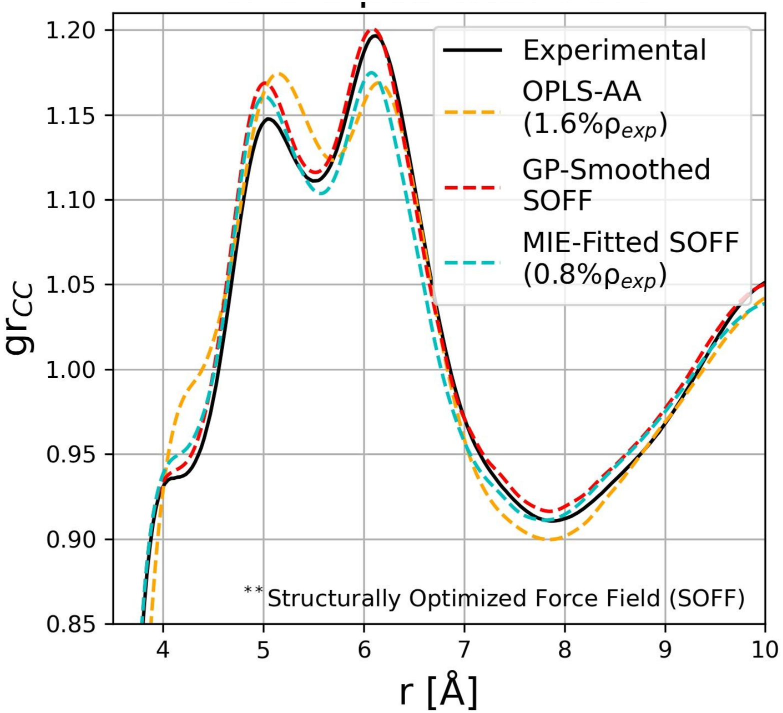
Radial Distribution Functions and Vapor Liquid Equilibrium Match with Excellent Agreement



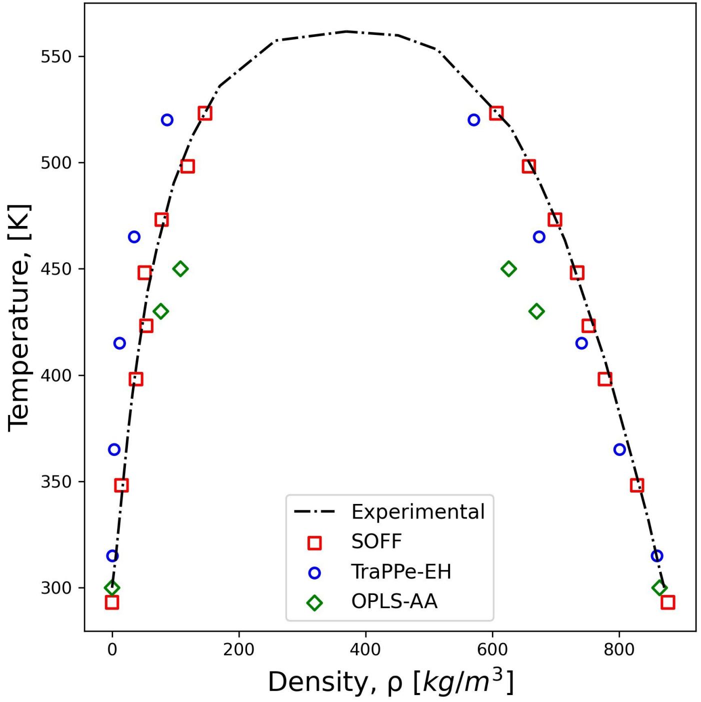
SOPR is the first method demonstrated to determine transferable potentials from experimental scattering data in real fluids!

Extending SOPR Beyond Monatomics - Molecular Liquids

Excellent RDF + VLE Agreement for Water, Benzene and Methane



** Structurally Optimized Force Field (SOFF)



Abdur Shazed

Harry Sullivan

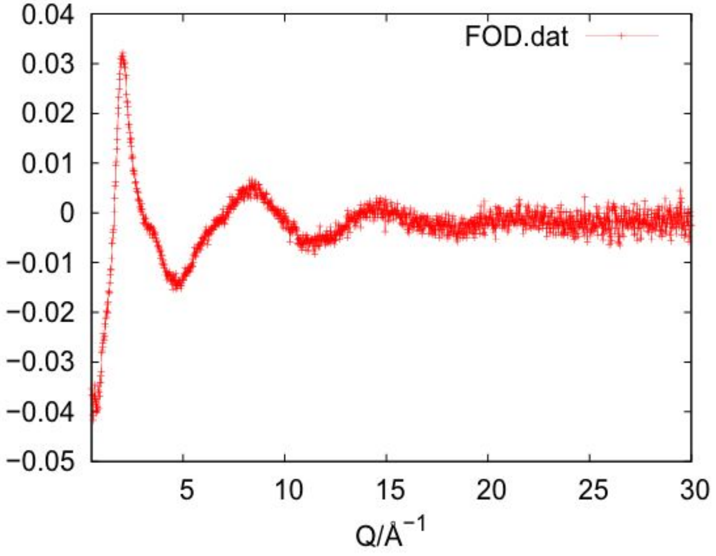
Impact

- SOPR is the first method to find transferable potentials from neutron scattering data!
- Preliminary results show similar efficacy on molecular liquids (water, benzene and methane)
- SOPR offers an efficient way to determine force fields from experiments free from a functional form.

II. How Does Experimental Uncertainty Influence our Potential Predictions?

Understanding Experimental Uncertainty Under a Known Model

Measurement Uncertainty



Noise in the Structure Factor of Water
Neuefiend 2012, *Nuc. Inst. Methods.*

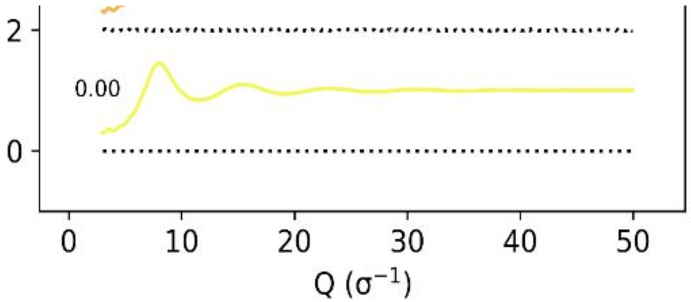
- We argued that SOPR could determine non-parametric potentials that are accurate and flexible
- **However, we don’t know how uncertainty in the experimental data impacts predictions from SOPR**
- Here we use Bayesian inference to quantify this when the model is known (toy problem).

Investigating the impact of measurement uncertainty in Mie fluids

Mie Fluid Interaction Potential

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

Run Model Simulation

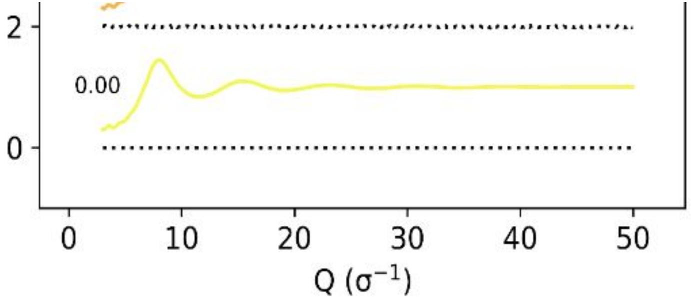


Investigating the impact of measurement uncertainty in Mie fluids

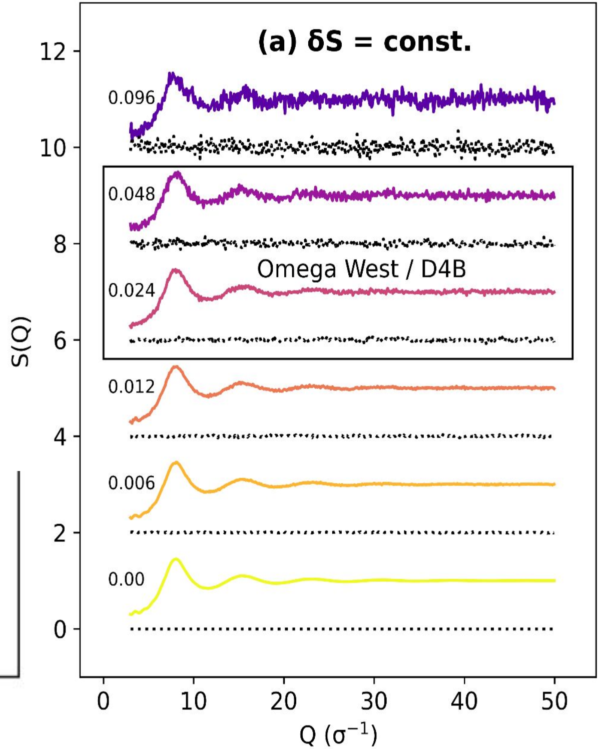
Mie Fluid Interaction Potential

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

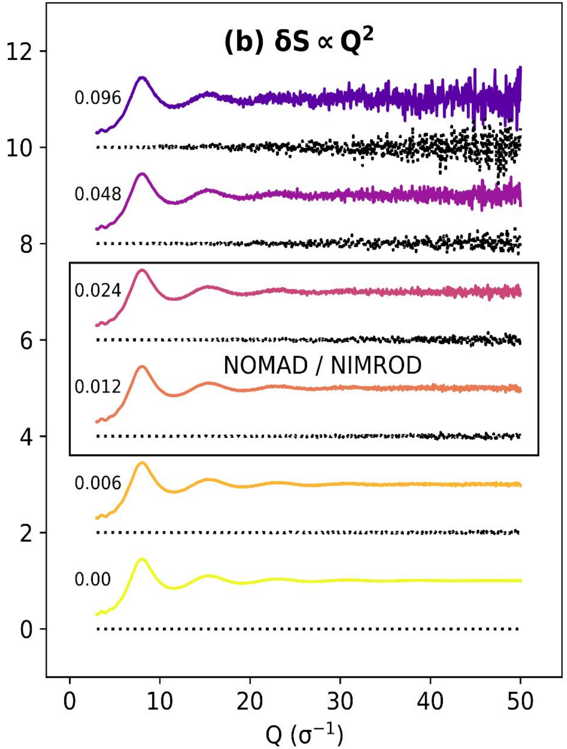
Run Model Simulation



Reactor Source

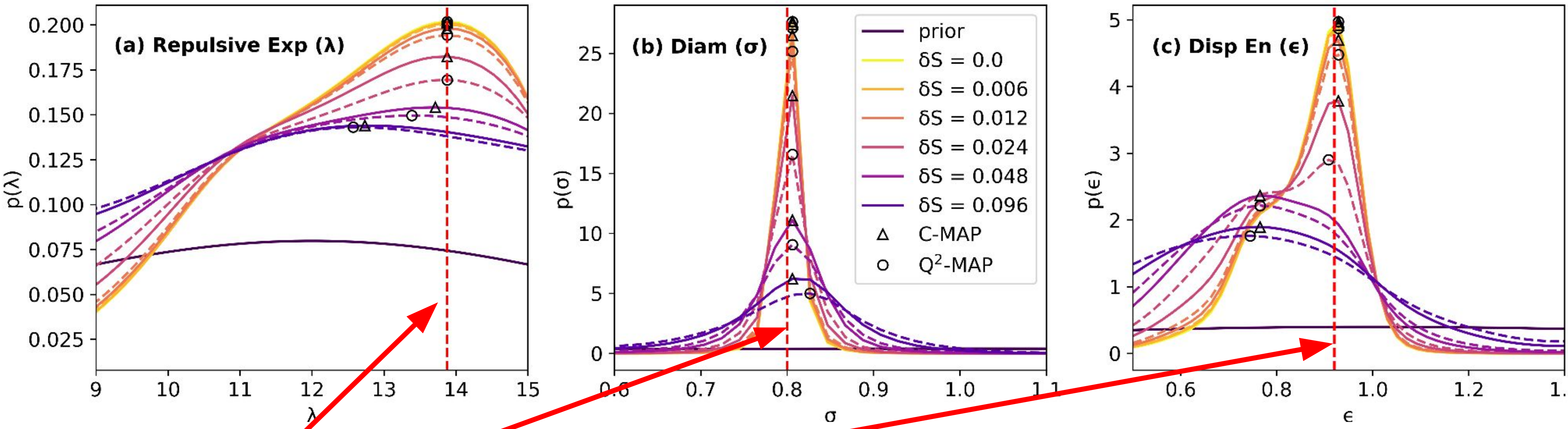


Spallation Source



Can we recover our original model from the structure?

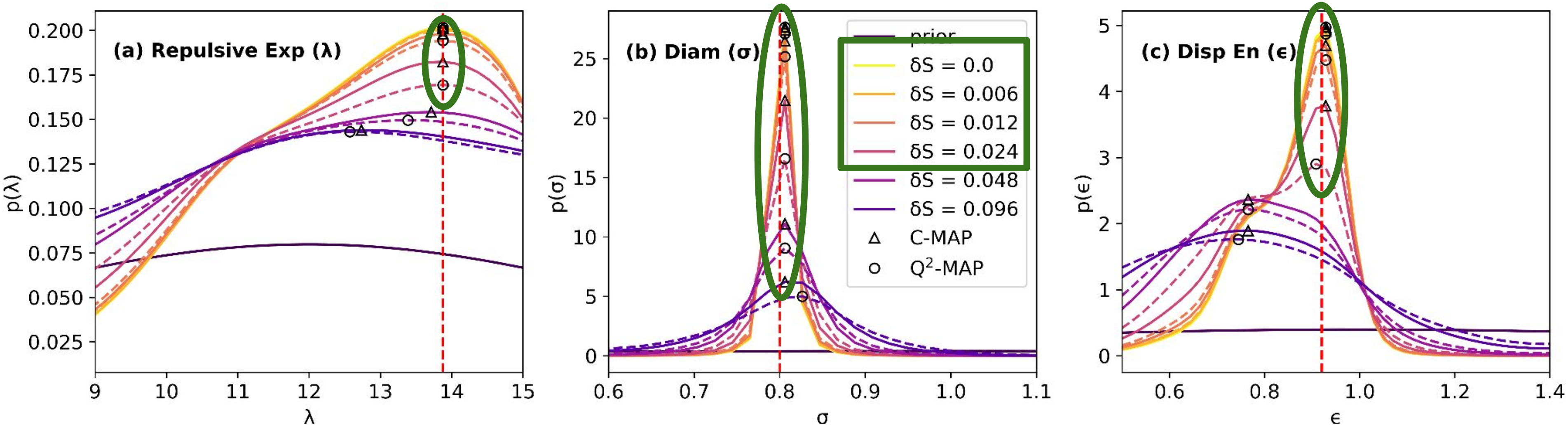
Bayesian Marginal Probability Distribution on Model Parameters



Known Model

Can we recover our original model from the structure?

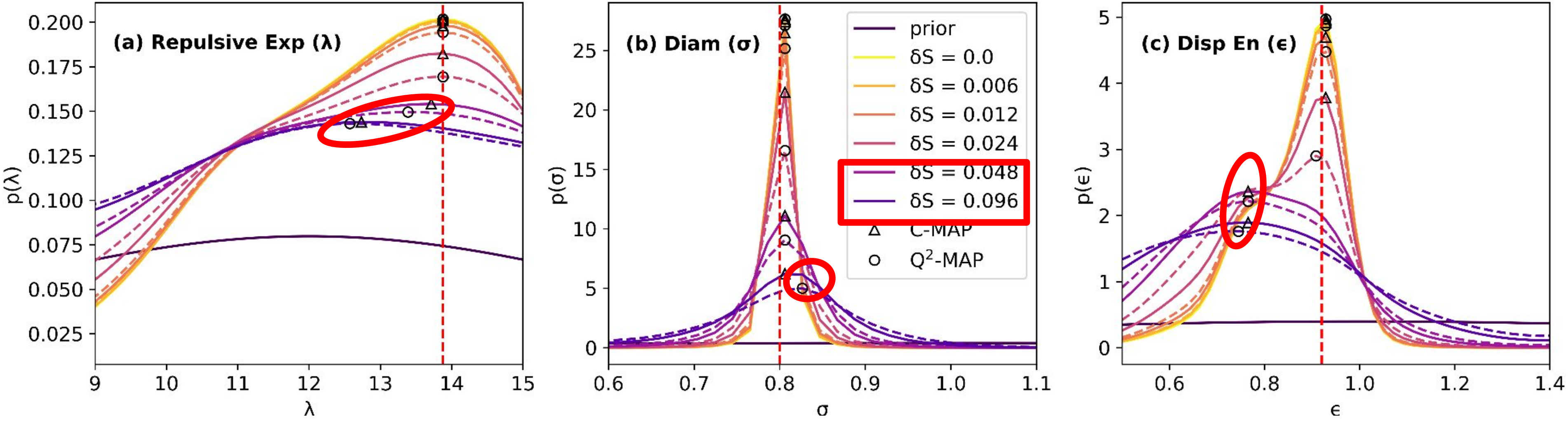
Bayesian Marginal Probability Distribution on Model Parameters



Bayesian optimization recovers force field parameters with high-accuracy for low uncertainty structure factor measurements.

Can we recover our original model from the structure?

Bayesian Marginal Probability Distribution on Model Parameters

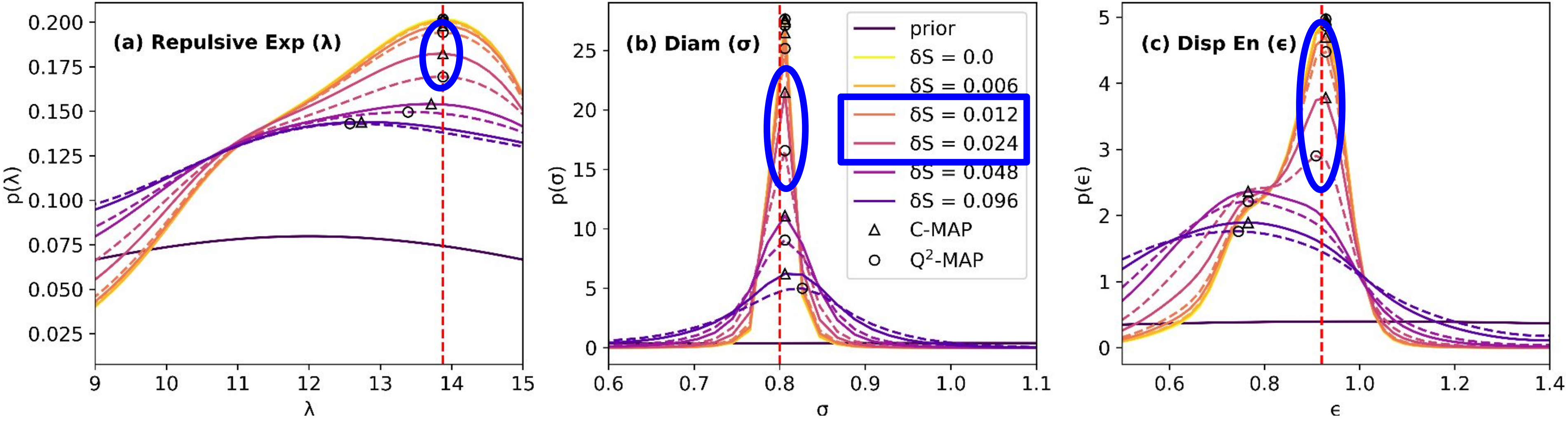


Uncertainty increases and accuracy declines rapidly below a 0.024 variance.

This data quality is representative of the 1960s-1980s neutron sources.

Can we recover our original model from the structure?

Bayesian Marginal Probability Distribution on Model Parameters



Existing instruments (NOMAD/NIMROD) can provide measurements below the precision threshold.

Impact

We have shown that experimental uncertainty can drastically influence the results of inverse methods!

Before this study, many assumed that recovering interaction potential parameters from neutron scattering was not feasible.

We know have evidence that prior work over the last 60 years struggled to find solutions to the inverse problem because **the available data was too low quality!**

III. Designing Surrogate Models for Expensive Calculations

Motivation

- Bayesian methods can answer important questions with respect to uncertainty, but are computationally expensive.

- Each posterior distribution represents results from ~1 million molecular sims!

- **How can we speed up the Bayesian analysis?**

Accelerated Bayesian Inference with Gaussian Process Surrogates

Evaluating the Bayesian likelihood is easy! Just run ~1 million molecular simulations to populate the model parameter space and you’re done!

Accelerated Bayesian Inference with Gaussian Process Surrogates

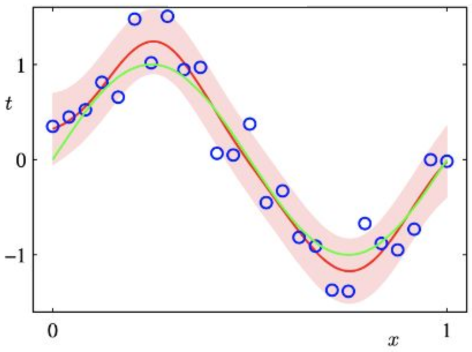
Evaluating the Bayesian likelihood is easy! Just run ~1 million molecular simulations to populate the model parameter space and you're done!

Instead, we train a GP on N ~ 480 simulations
For data containing η independent variables.

~86 fold speed up

$(N\eta \text{ x } \dim(\boldsymbol{\theta}) + 1)$

$GP(\boldsymbol{\theta}^*)$



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

Accelerated Bayesian Inference with Gaussian Process Surrogates

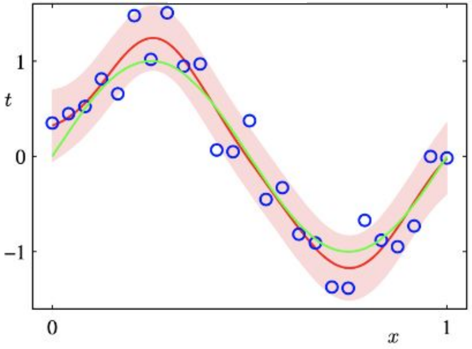
Evaluating the Bayesian likelihood is easy! Just run ~1 million molecular simulations to populate the model parameter space and you’re done!

Instead, we train a GP on N ~ 480 simulations For data containing η independent variables.

Local GPs reduce matrix size and are about **3500 fold faster than full GPs**

~86 fold speed up

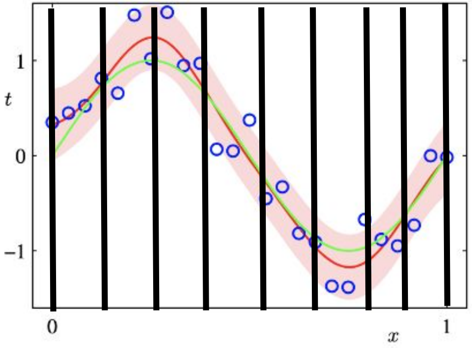
$GP(\boldsymbol{\theta}^*)$



$(N\eta \text{ x } \dim(\boldsymbol{\theta}) + 1)$

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

$GP_k(\boldsymbol{\theta}^*)$



$\eta \quad (N \text{ x } \dim(\boldsymbol{\theta}))$

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

Accelerated Bayesian Inference with Gaussian Process Surrogates

Evaluating the Bayesian likelihood is easy! Just run ~1 million molecular simulations to populate the model parameter space and you're done!

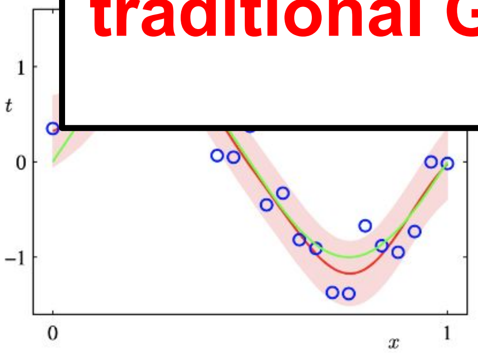
Instead, we train a GP on N ~ 480 simulations

F

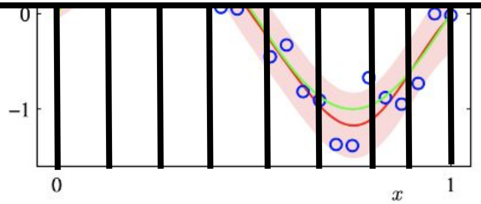
Local GPs reduce matrix size and are about 3500 fold faster than full GPs

Local Gaussian processes are ~3500x faster than traditional GPs by reducing dimension of matrices

θ))

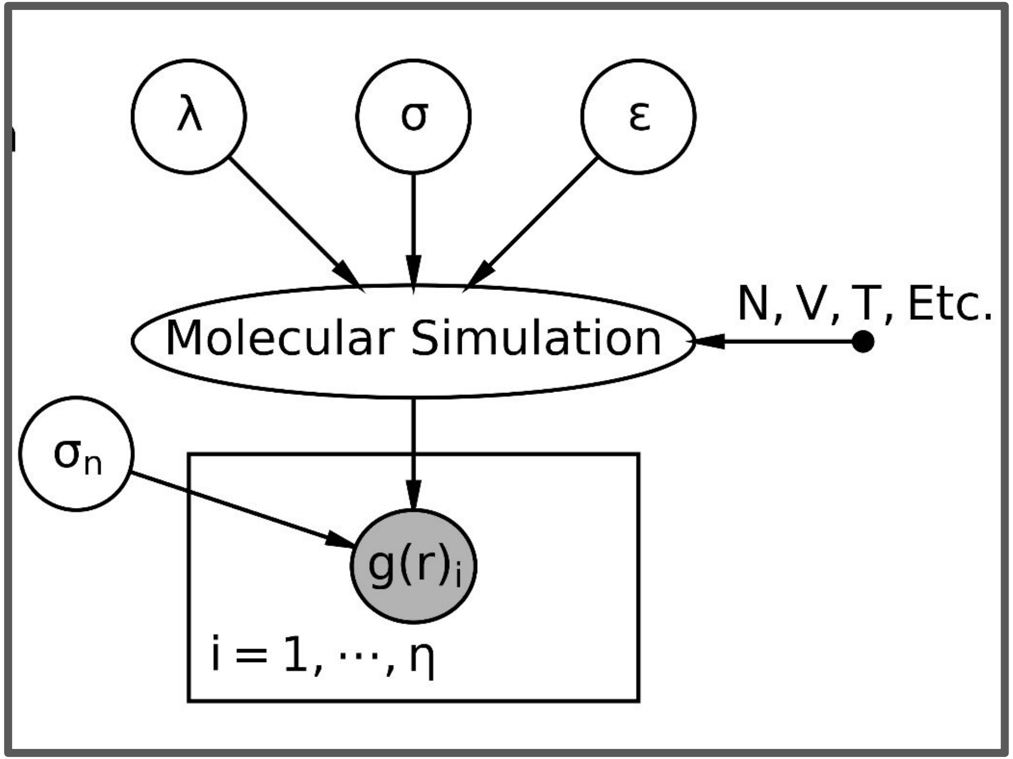


$$\hat{\mathbf{X}} = \begin{bmatrix} \theta_{1,1} & \theta_{2,1} & \dots & r_\eta \\ \theta_{1,2} & \theta_{2,2} & \dots & r_1 \\ \vdots & \vdots & \vdots & \vdots \\ \theta_{1,N} & \theta_{2,N} & \dots & r_\eta \end{bmatrix}$$



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

Example: Building a LGP Surrogate Model for the RDF of Liquid Ne



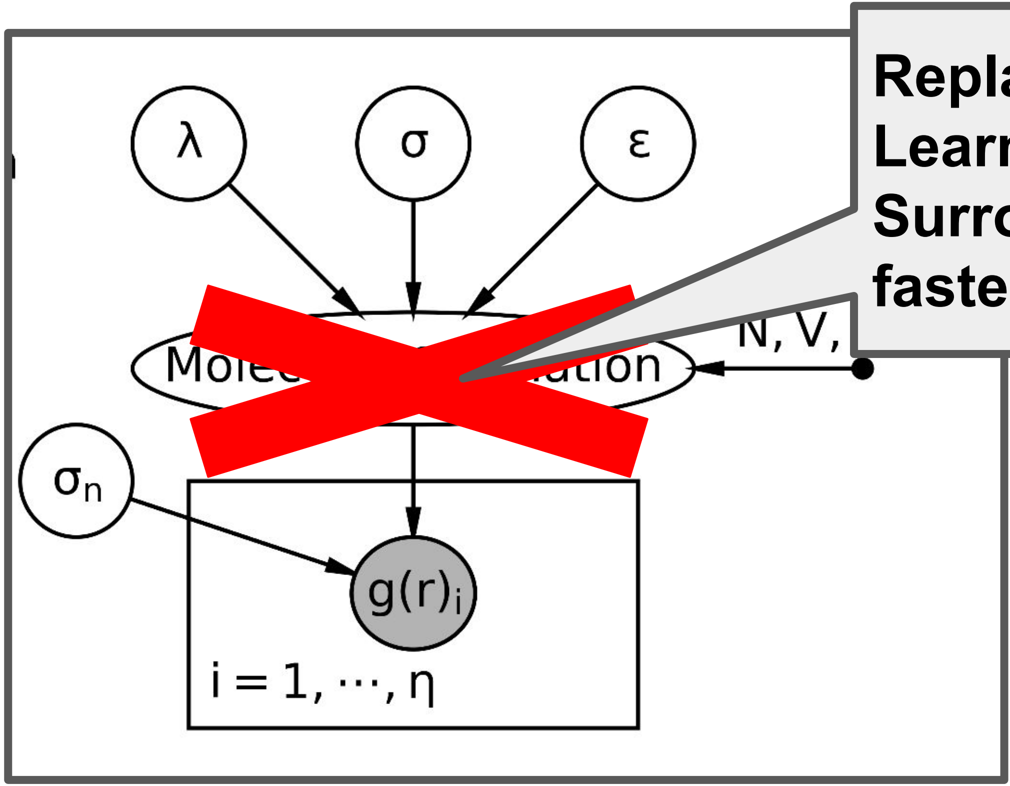
Mie Potential w/ 3 Parameters

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

Bayes Theorem

$$p(\boldsymbol{\theta}|\mathscr{Y}) = \frac{p(\mathscr{Y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathscr{Y})}$$

Example: Building a LGP Surrogate Model for the RDF of Liquid Ne



Replace with Machine Learning (Local GP) Surrogate Model that is faster!

Parameters

$$\left(\frac{\sigma}{r}\right)^{\lambda} - \left(\frac{\sigma}{r}\right)^{6}\right]$$

Bayes Theorem

$$p(\theta|\mathscr{Y}) = \frac{p(\mathscr{Y}|\theta)p(\theta)}{p(\mathscr{Y})}$$

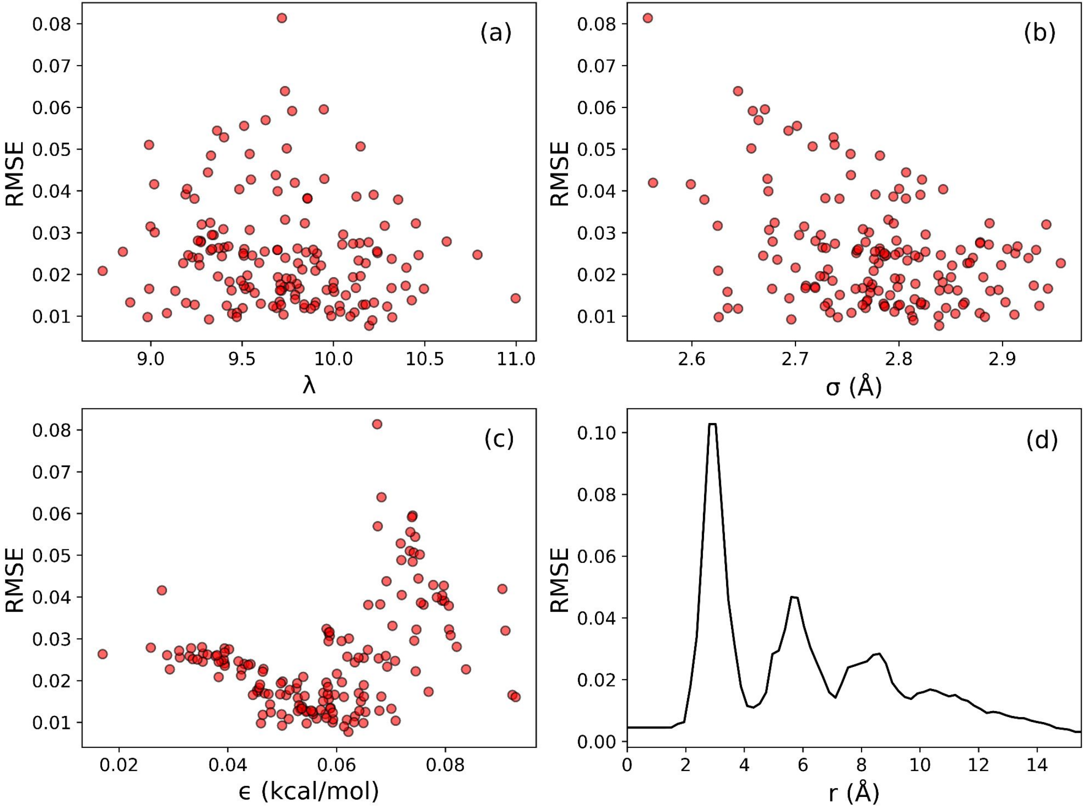
Speed and Accuracy of Local Gaussian Process Surrogate Models

A GP can predict the RDF 86x faster than MD

A local GP can predict the RDF 288,000x faster than MD!!

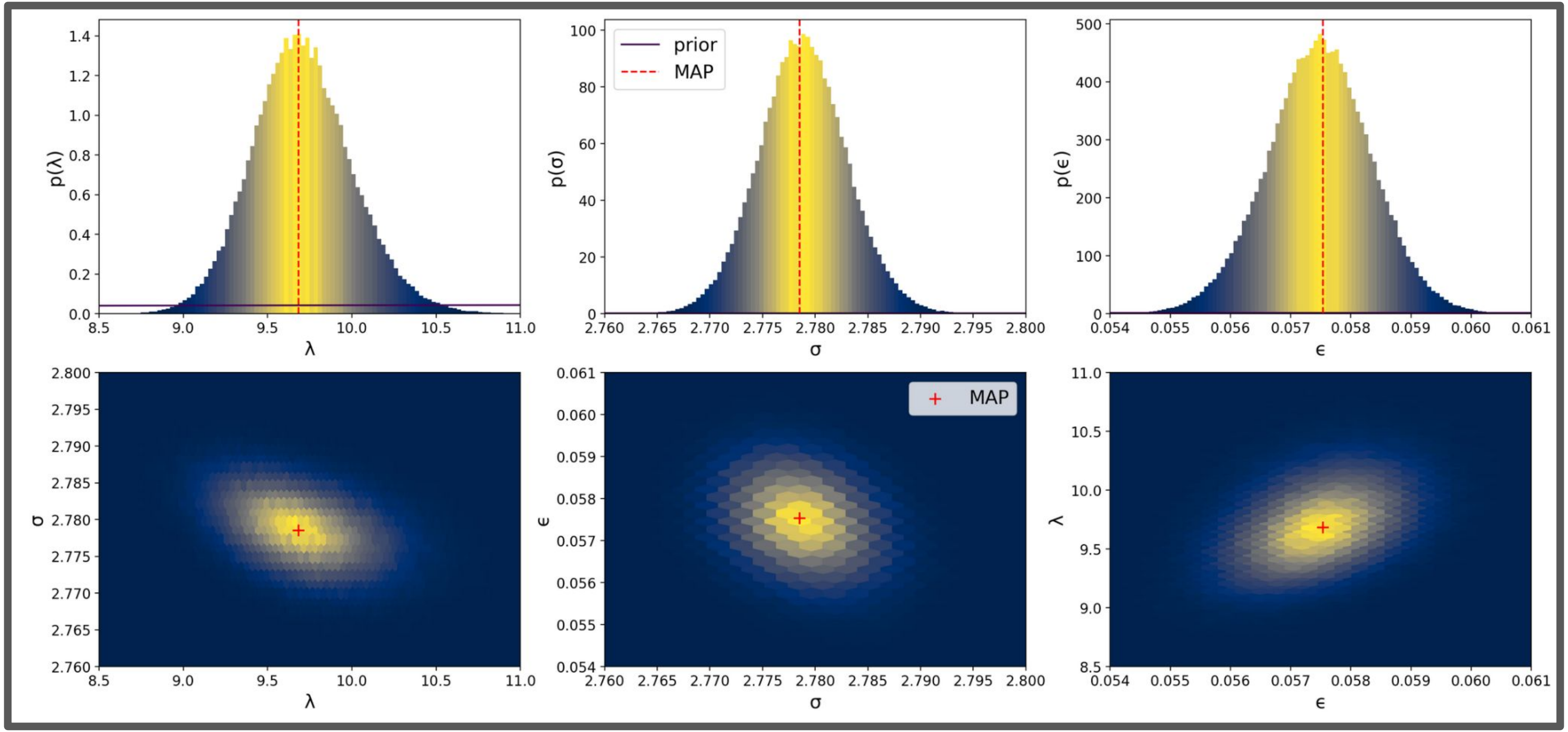
We also find that the RMSE is within the RDF uncertainty

RMSE Over Test Set



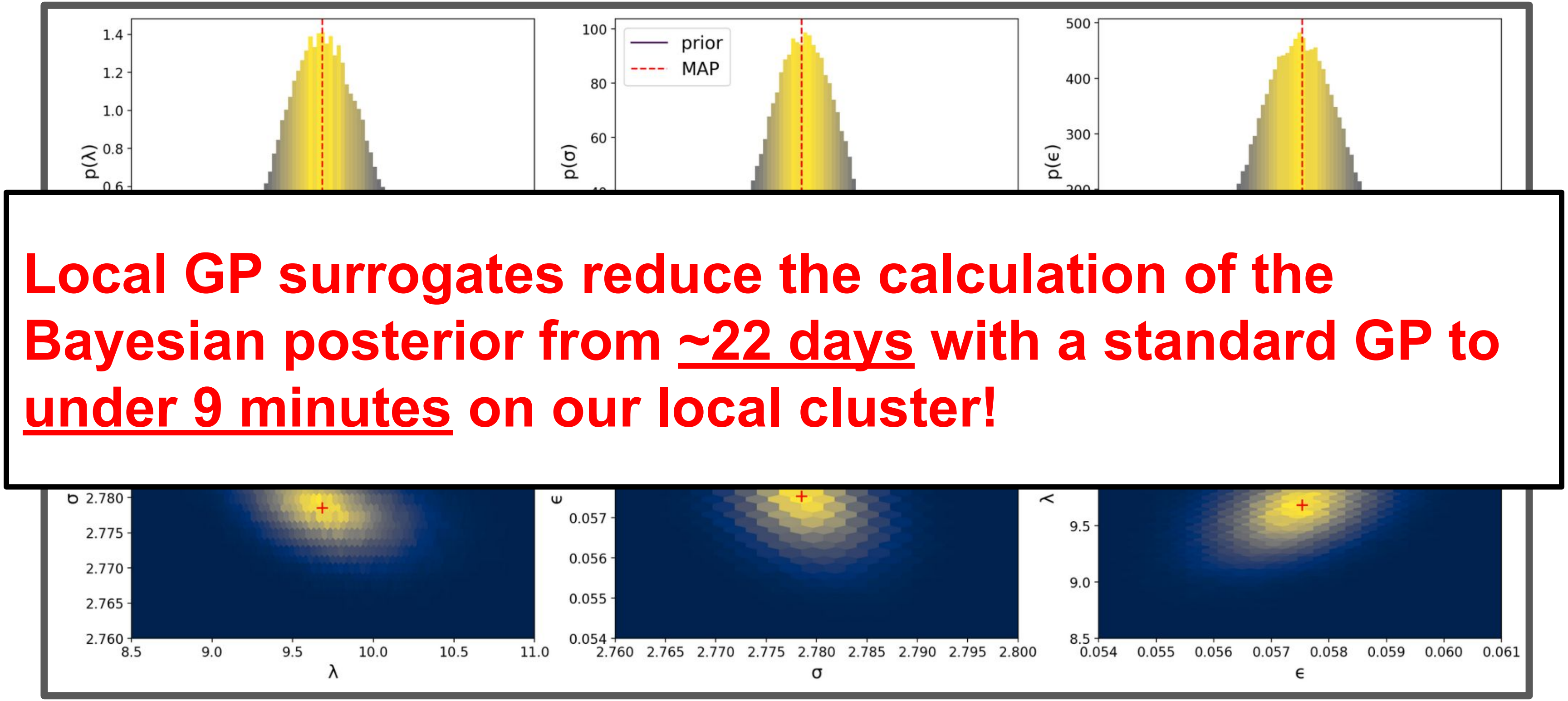
Learning from the Bayesian Posterior Distribution

Posterior marginal distributions are just integrals over the joint posterior



Learning from the Bayesian Posterior Distribution

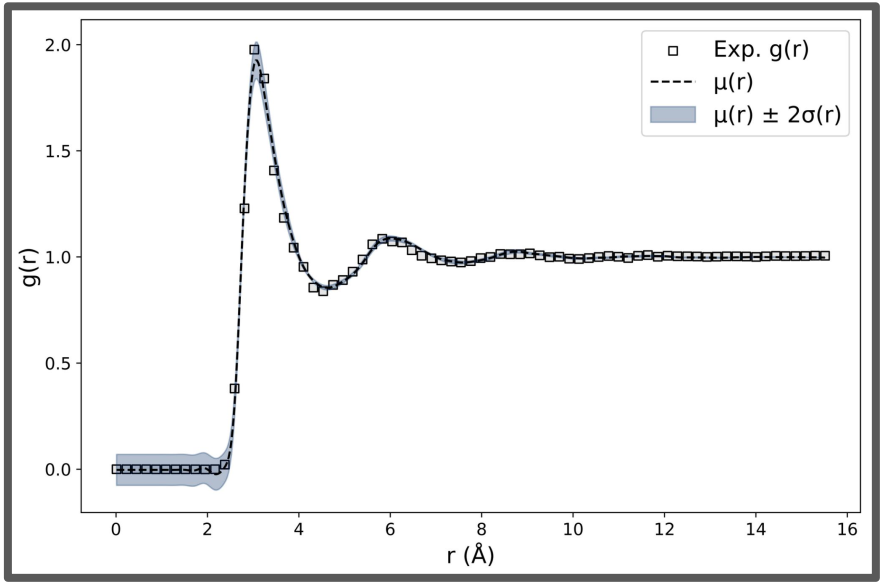
Posterior marginal distributions are just integrals over the joint posterior



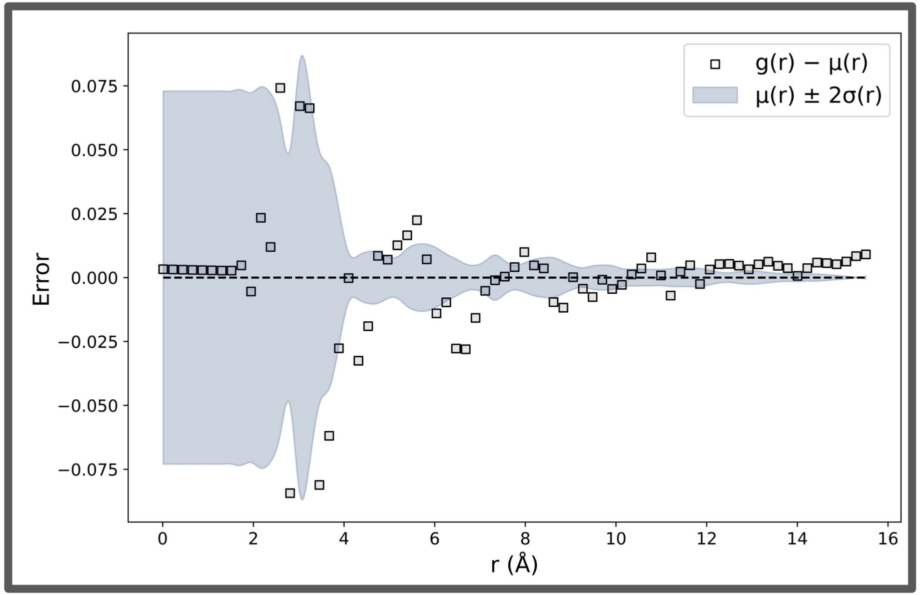
Local GP surrogates reduce the calculation of the Bayesian posterior from ~22 days with a standard GP to under 9 minutes on our local cluster!

Learning from the Posterior Predictive Distribution

Posterior Predictive



Residual Analysis



Experimental data often lies outside of the credibility interval → there is potentially missing physics that we need to incorporate into the model.

Impact

- Local GP surrogate models are reliable and fast!

- They enable Bayesian force field optimization and uncertainty quantification for complex experiments
    - (scattering and spectroscopy data)

- These surrogate models can help with model selection, validation and sensitivity analysis.

Summary and Key Takeaways

- Inverse problems are useful for interesting chemistry, including scattering analysis, coarse-graining, and force field development.

- Bayesian inference is a rigorous framework to quantify uncertainty, which enables detailed study of model sensitivity, uncertainty, and adequacy.

- **Bayesian UQ can answer questions around interatomic forces, enable active learning approaches using decision theory, and rigorously incorporate uncertainty into force field development.**

- Local GPs are effective surrogate models for complex experimental data (scattering, spectra, etc)

Thank you!

Brennon Shanks

Hoepfner Research Group

Email: brennon.shanks@chemeng.utah.edu

Website: https://bshanks.netlify.app/