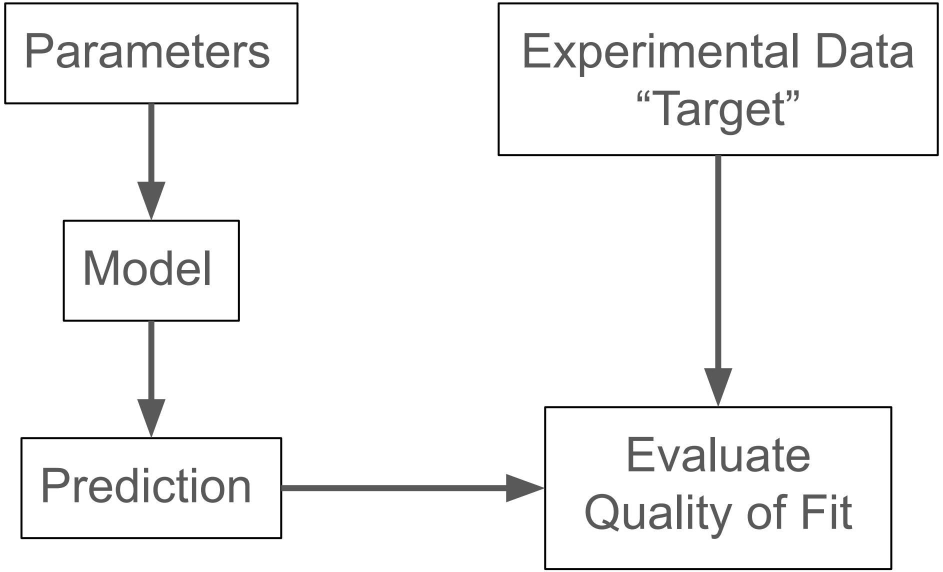
Bayesian Approaches for Uncertainty-Aware Force Field Optimization and Surrogate Modeling

Brennon Shanks

Institute of Organic Chemistry and Biochemistry Prague

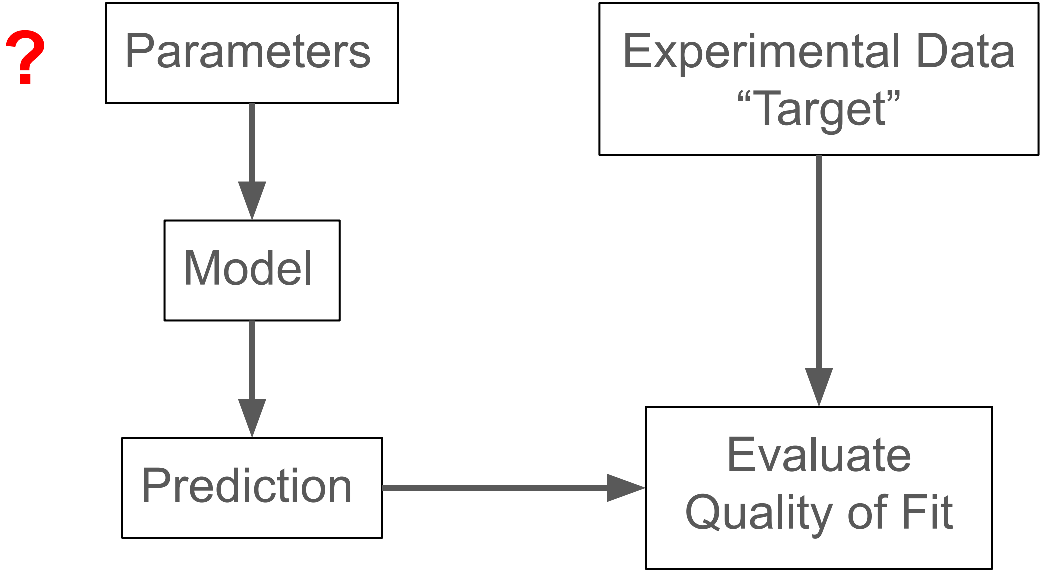
The general question of model optimization

What are the best set of model parameters to represent a data target?



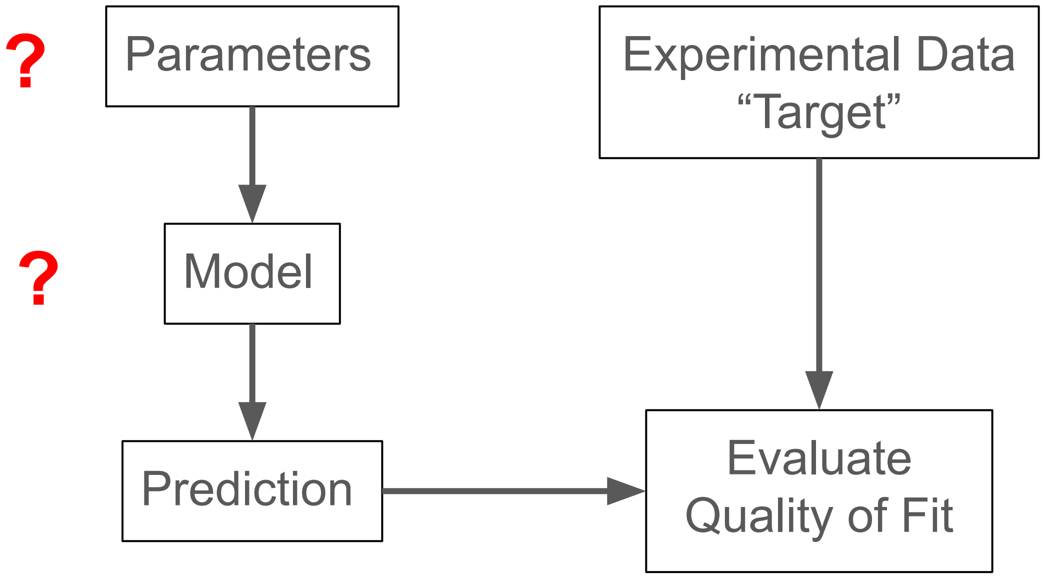
The general question of model optimization

What are the best set of model parameters to represent a data target?



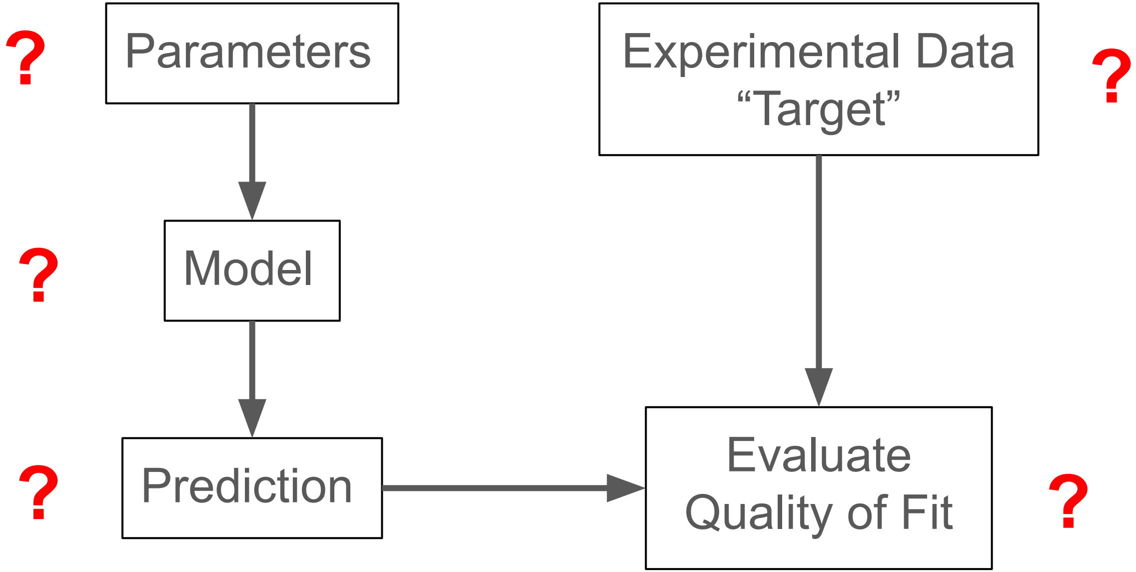
The general question of model optimization

What are the best set of model parameters to represent a data target?



The general question of model optimization

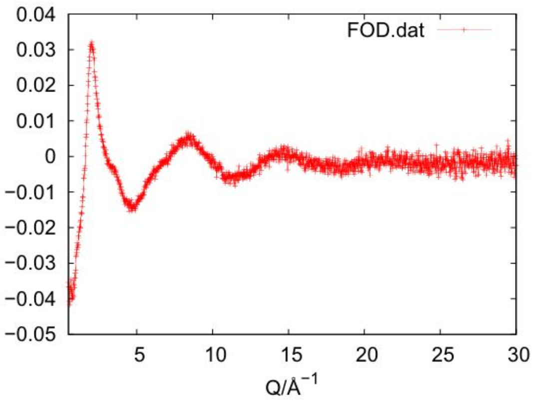
What are the best set of model parameters to represent a data target?



Quantifying uncertainty is a problem from experiment to model

Is my data reliable?

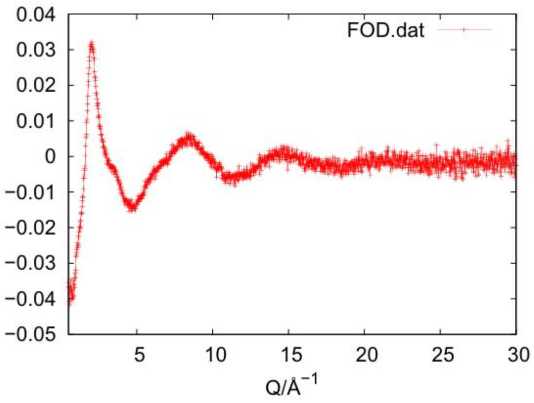
Data…
- Is noisy
- Often require corrections
- Non-trivial to analyze



Quantifying uncertainty is a problem from experiment to model

Is my data reliable?

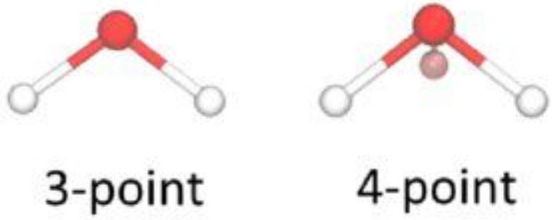
Data…
- Is noisy
- Often require corrections
- Non-trivial to analyze



Did I choose a good model?

We can’t know *a priori* if a model represents nature perfectly.

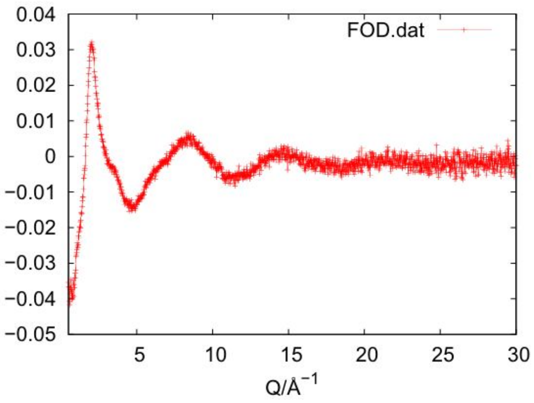
- TIP3P or TIP4P?



Quantifying uncertainty is a problem from experiment to model

Is my data reliable?

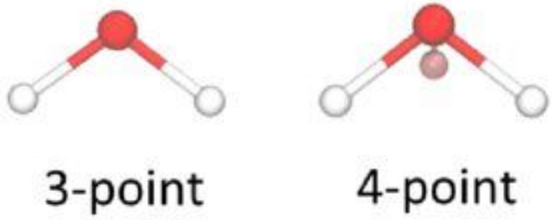
Data…
- Is noisy
- Often require corrections
- Non-trivial to analyze



Did I choose a good model?

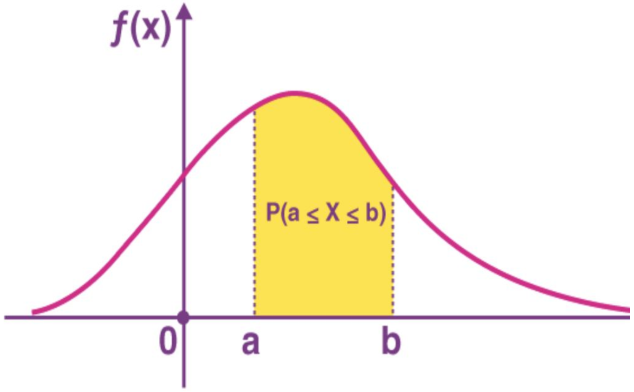
We can't know *a priori* if a model represents nature perfectly.

- TIP3P or TIP4P?



Am I confident in the model parameters?

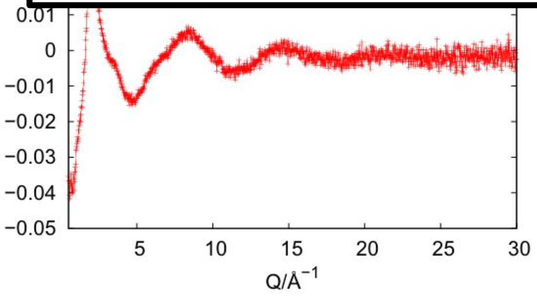
Are the model parameters guaranteed to be correct?

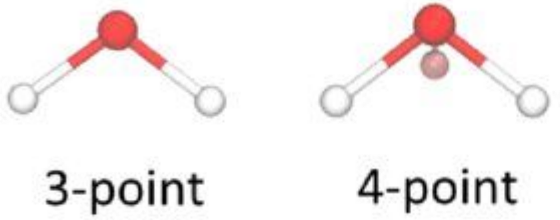


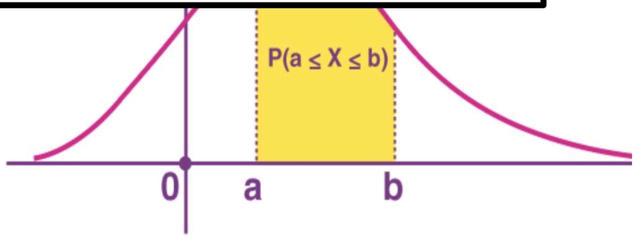
Quantifying uncertainty is a problem from experiment to model

Is my data reliable?	Did I choose a good model?	Am I confident in the model parameters?

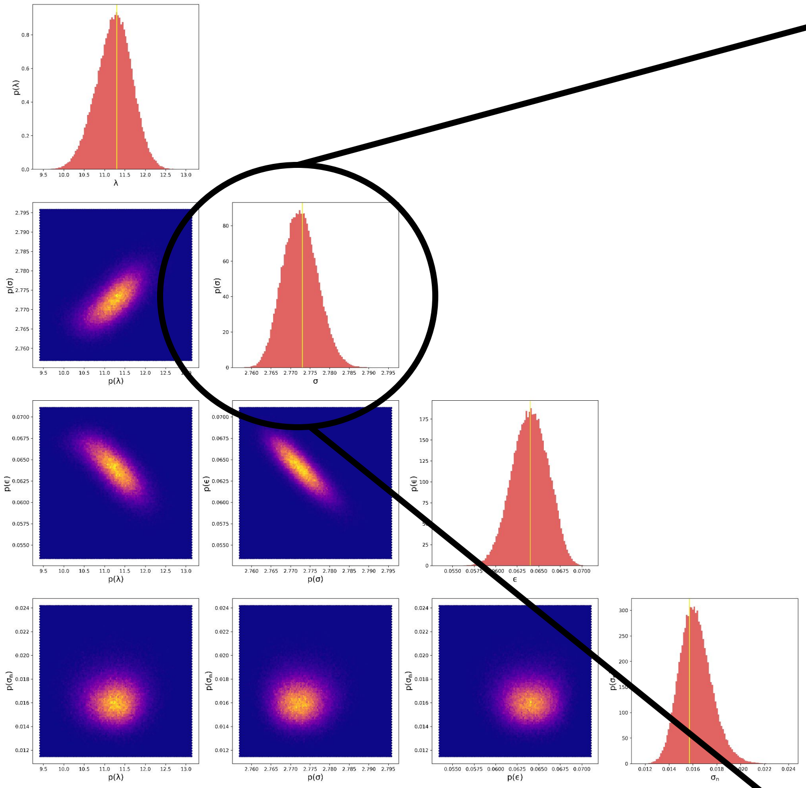
What can we do if we quantify this uncertainty?



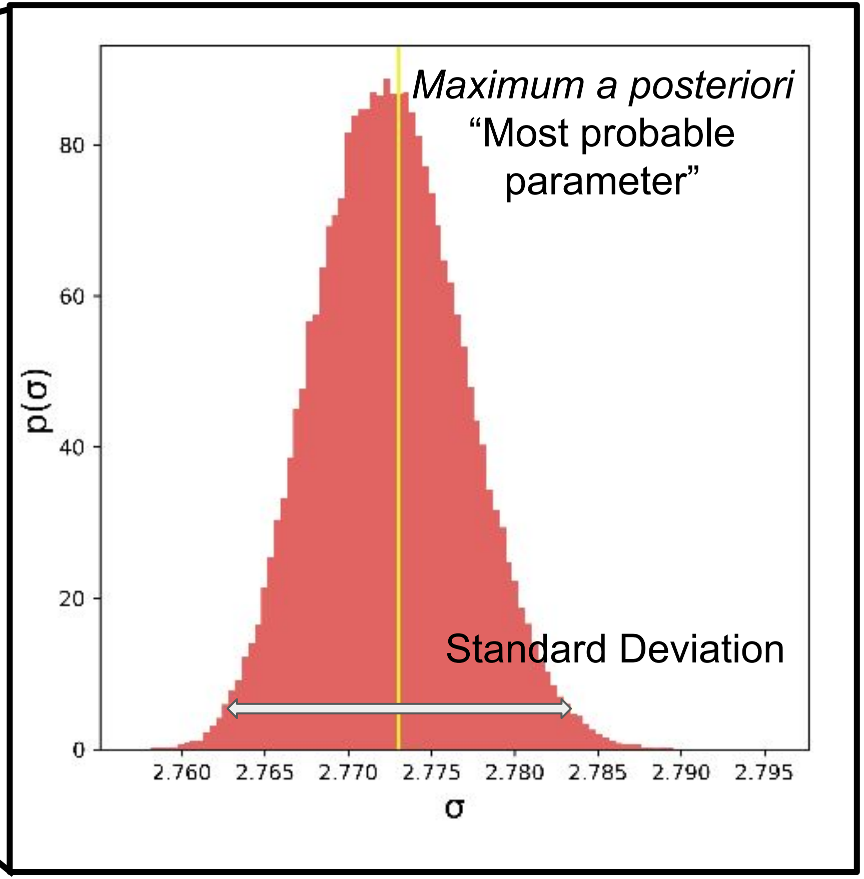




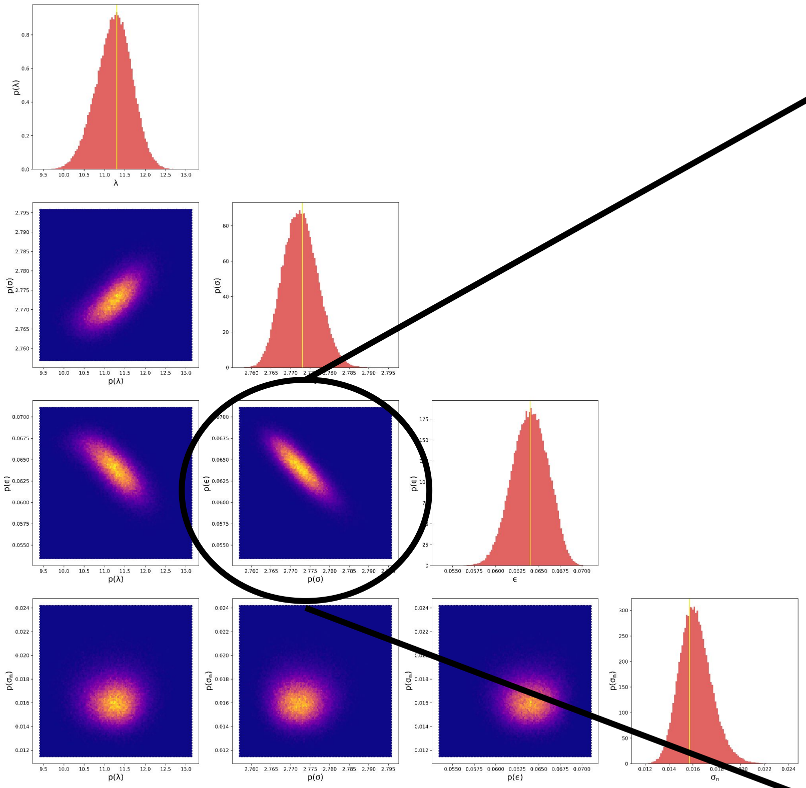
We get a quantification of our biases and model sensitivity



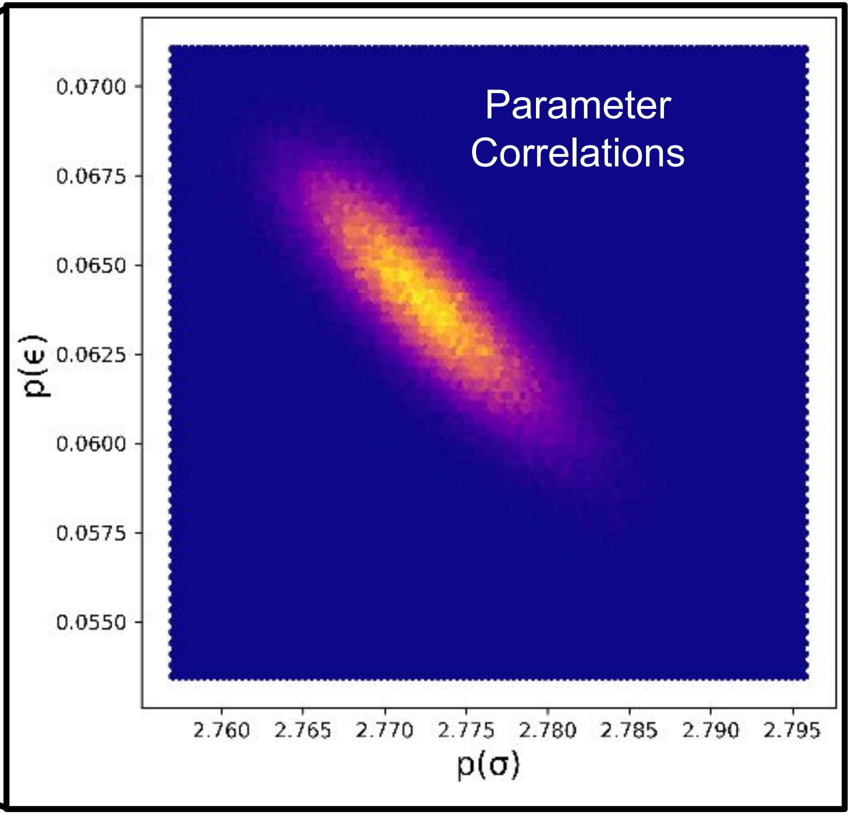
Marginal Posteriors on Mie Parameters
Shanks 2024, *J. Chem. Theory Comput.*



We get a quantification of our biases and model sensitivity



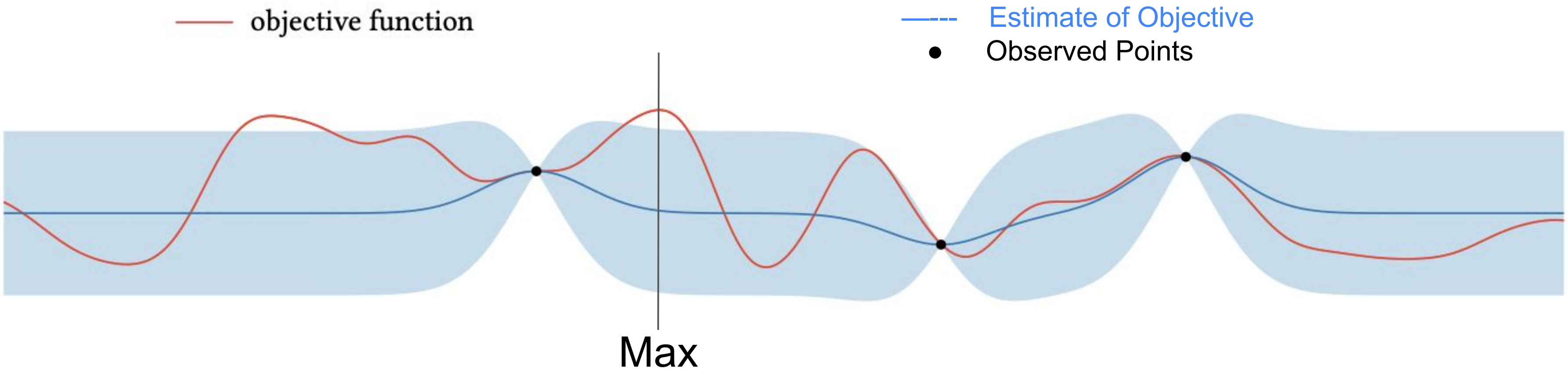
Marginal Posteriors on Mie Parameters
Shanks 2024, *J. Chem. Theory Comput.*



Efficient optimization requires knowledge of uncertainty and risk

Suppose we want to maximize an objective, but obtaining training data is expensive.

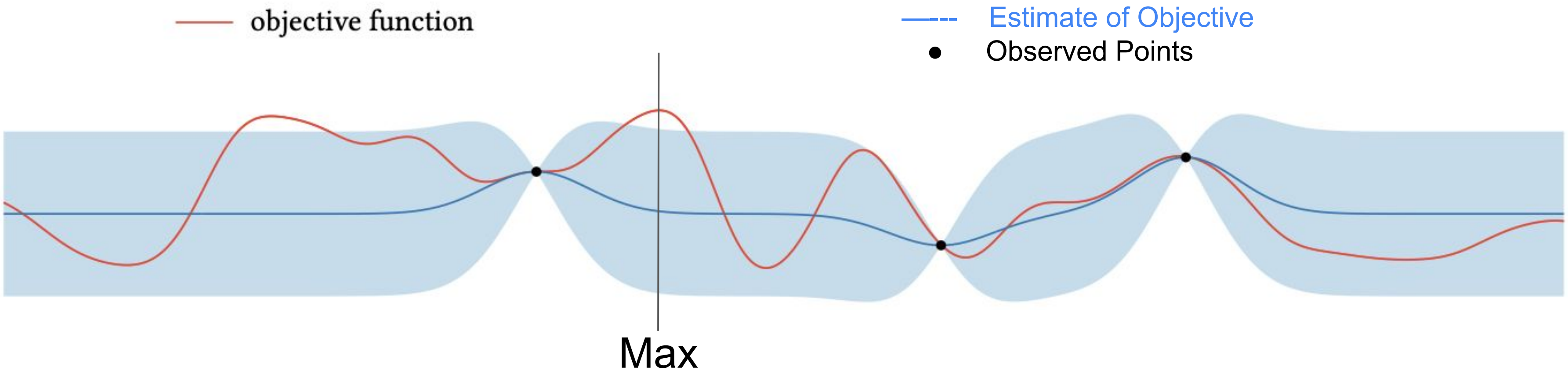
To minimize training time, we want to find the optimal next training point.



Efficient optimization requires knowledge of uncertainty and risk

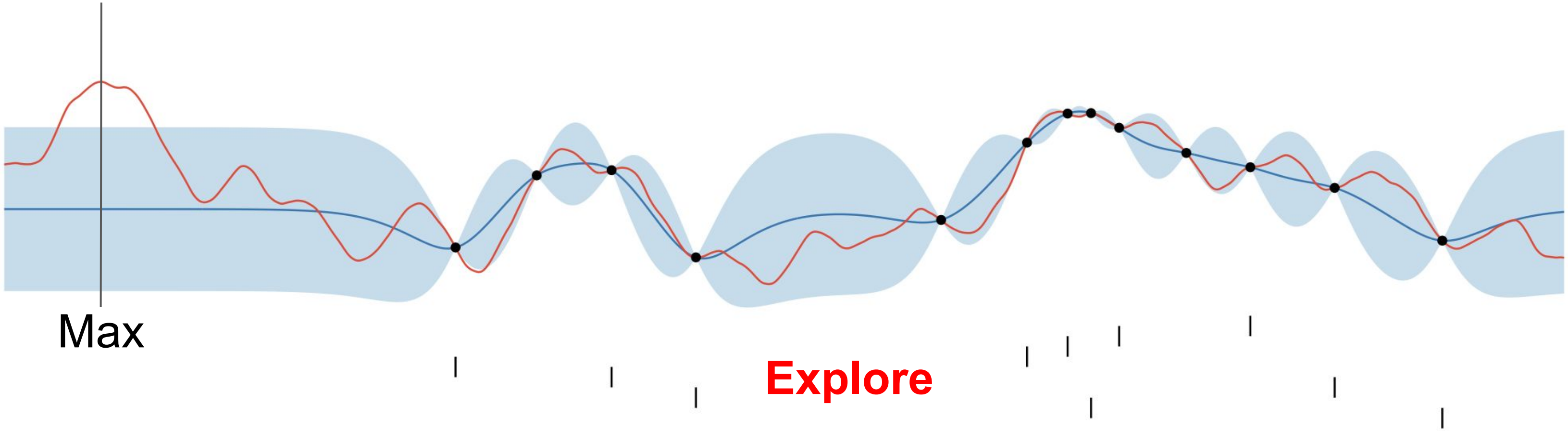
Suppose we want to maximize an objective, but obtaining training data is expensive.

To minimize training time, we want to find the optimal next training point.

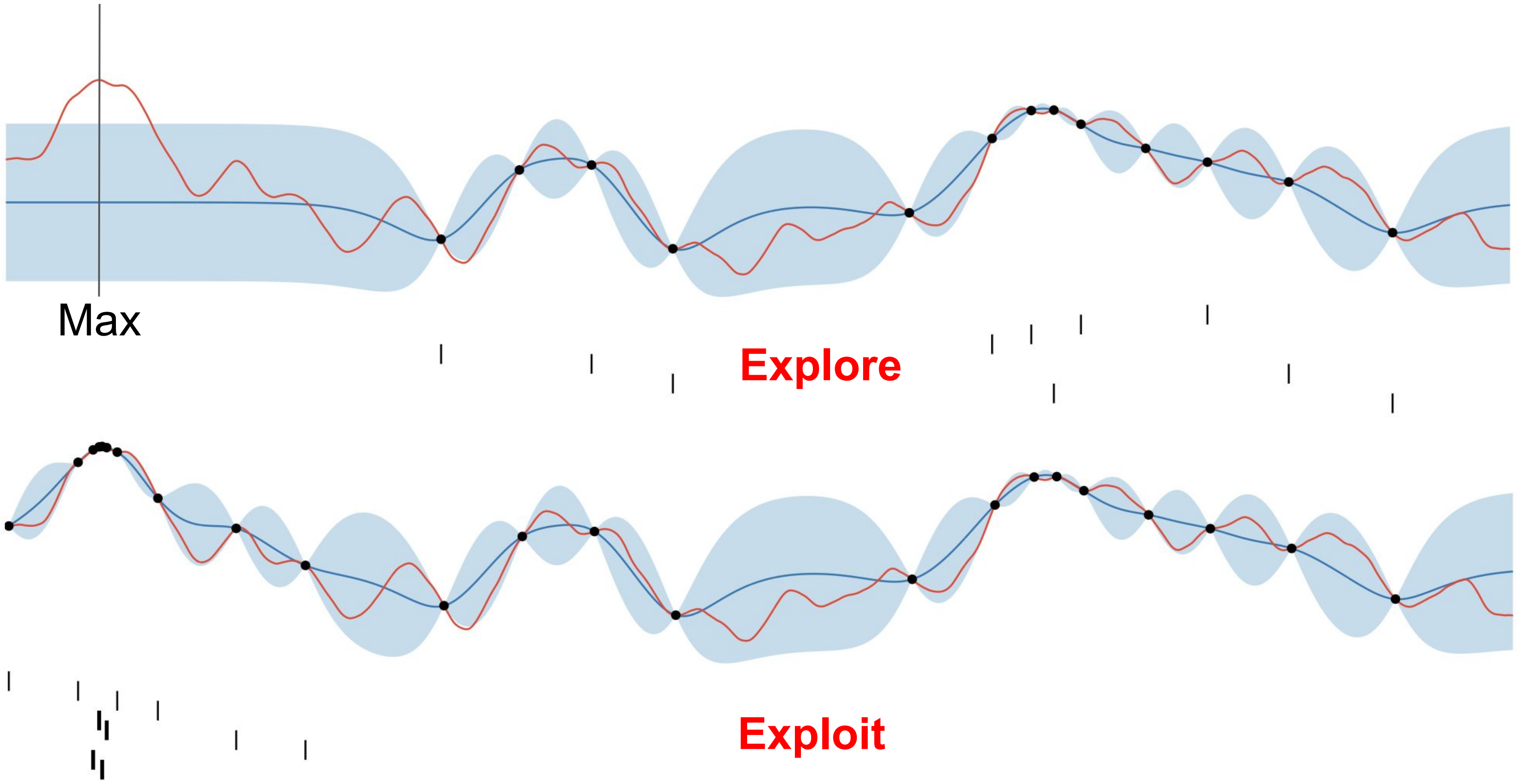


Quantify the change in expected utility and choose the maximum value

Efficient optimization requires knowledge of uncertainty and risk



Efficient optimization requires knowledge of uncertainty and risk



Bayesian optimization as a framework to quantify uncertainty

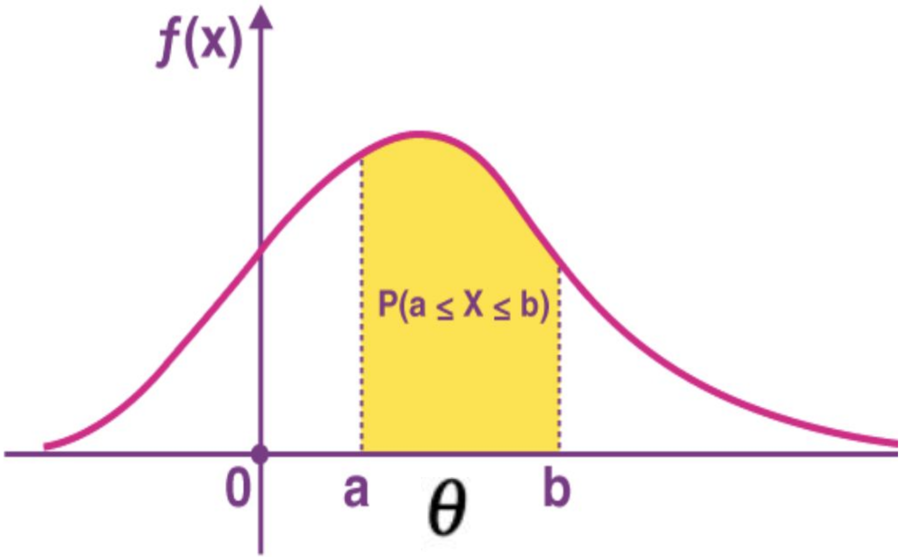
The Basic Outline of Bayesian Approaches

Bayesian optimization 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 optimization as a framework to quantify uncertainty

The Basic Outline of Bayesian Approaches

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

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

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

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

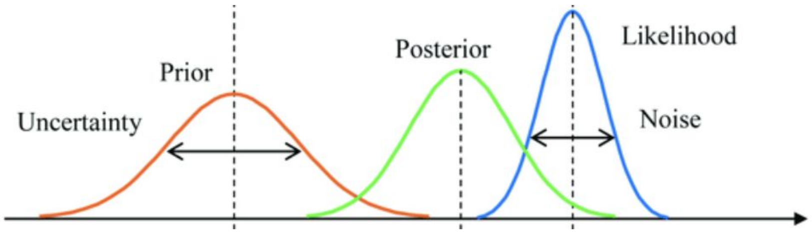
Bayesian optimization as a framework to quantify uncertainty

The Basic Outline of Bayesian Approaches

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

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

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



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

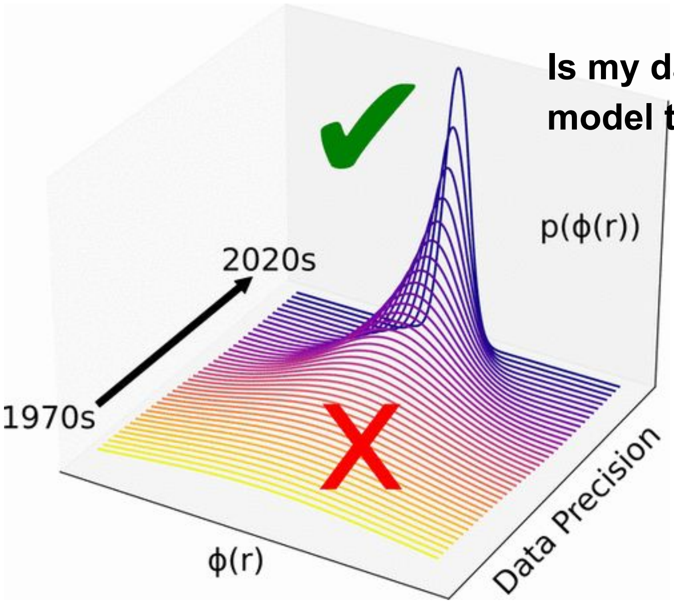
"When the facts change, I change my opinion. What do you do, sir?" - John Maynard Keynes

The posterior is the updated probability of parameters after observations

So what exactly can we use Bayesian inference for?

Bayes can quantify parameter sensitivity and model adequacy

Is Neutron Scattering Data a Good Target for classical force field optimization?



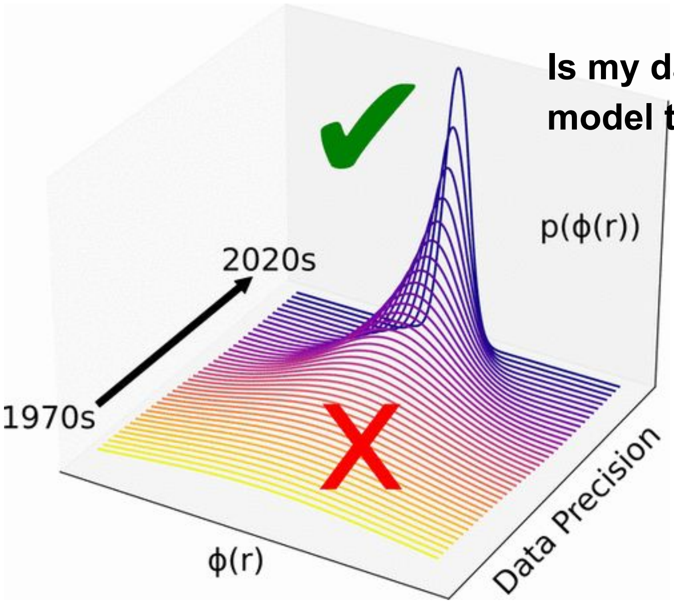
Is my data useful for model training?

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.* 12608–12618 (2024) doi:10.1021/acs.jpclett.4c02941.

Bayes can quantify parameter sensitivity and model adequacy

Is Neutron Scattering Data a Good Target for classical force field optimization?

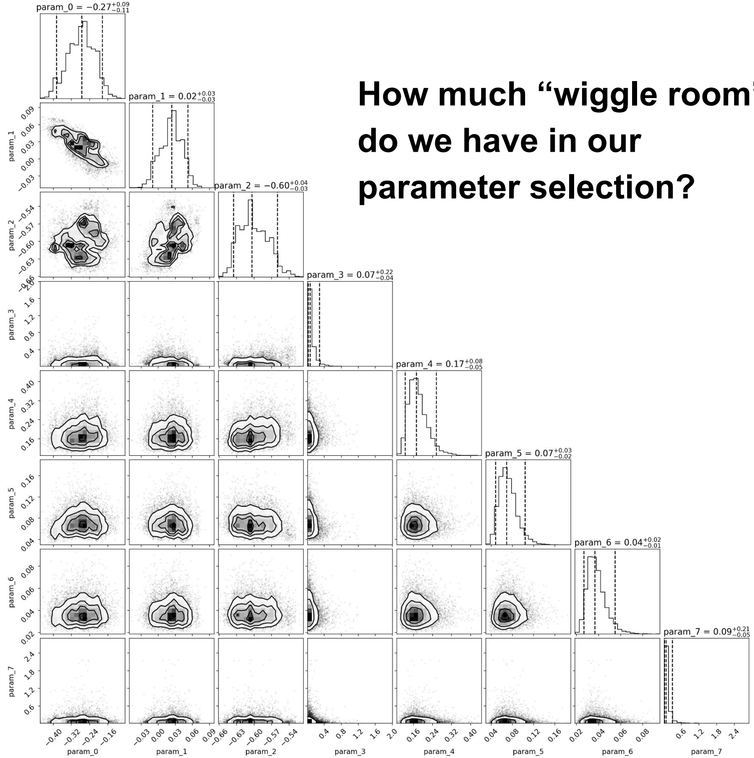
Is my data useful for model training?



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.* 12608–12618 (2024) doi:[10.1021/acs.jpclett.4c02941](https://doi.org/10.1021/acs.jpclett.4c02941).

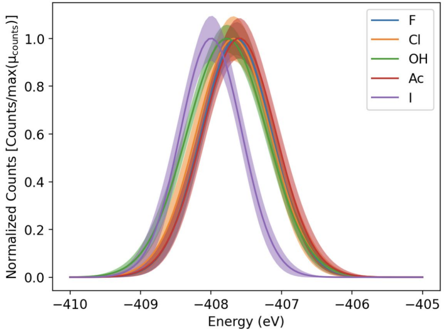
Acetate Partial Charge Optimization

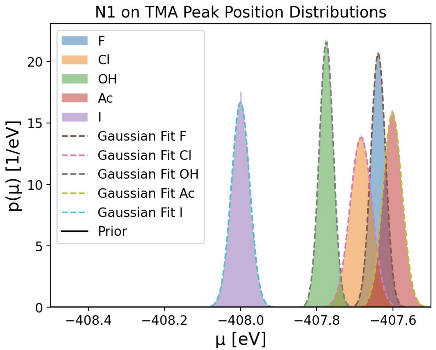
How much “wiggle room” do we have in our parameter selection?

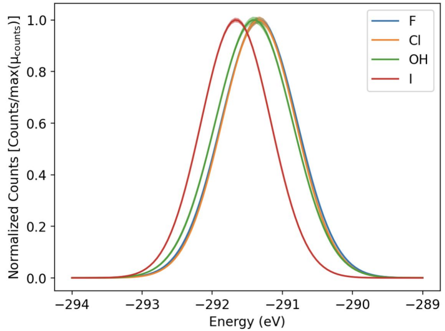


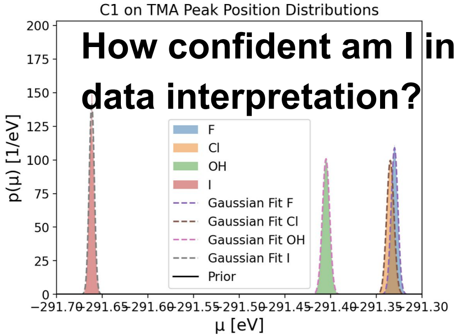
Bayes can help quantify uncertainty in experimental data

Quantifying Uncertainty in Photoelectron Spectra Peak Positions





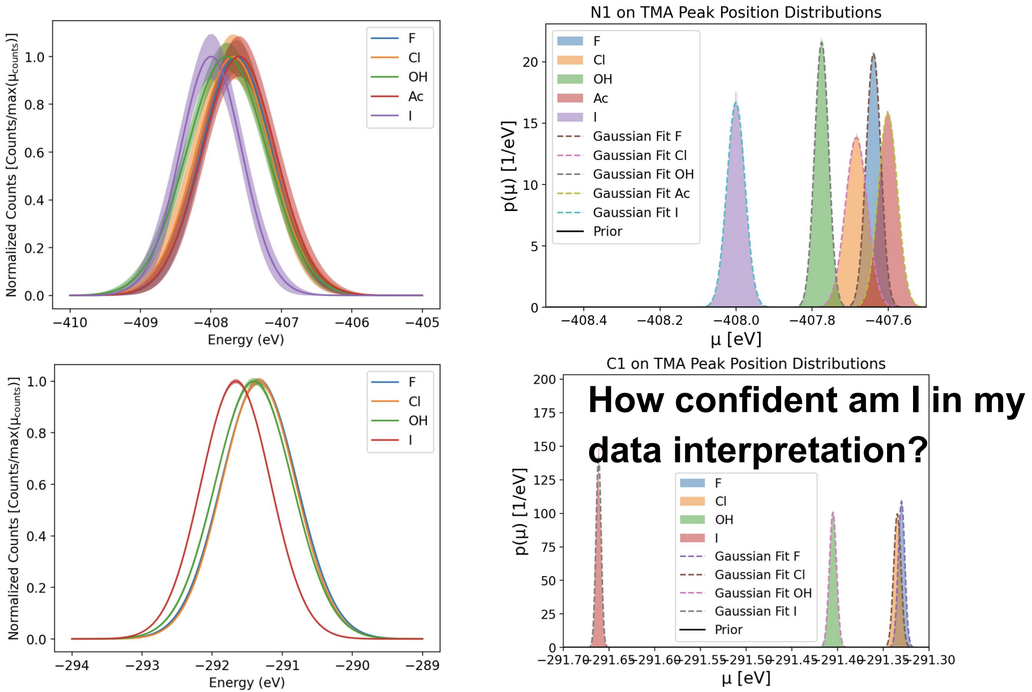




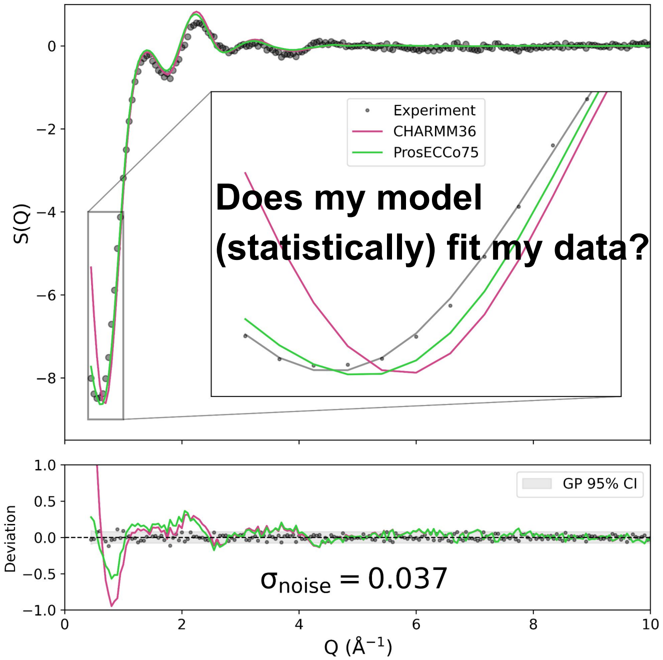
How confident am I in my data interpretation?

Bayes can help quantify uncertainty in experimental data

Quantifying Uncertainty in Photoelectron Spectra Peak Positions

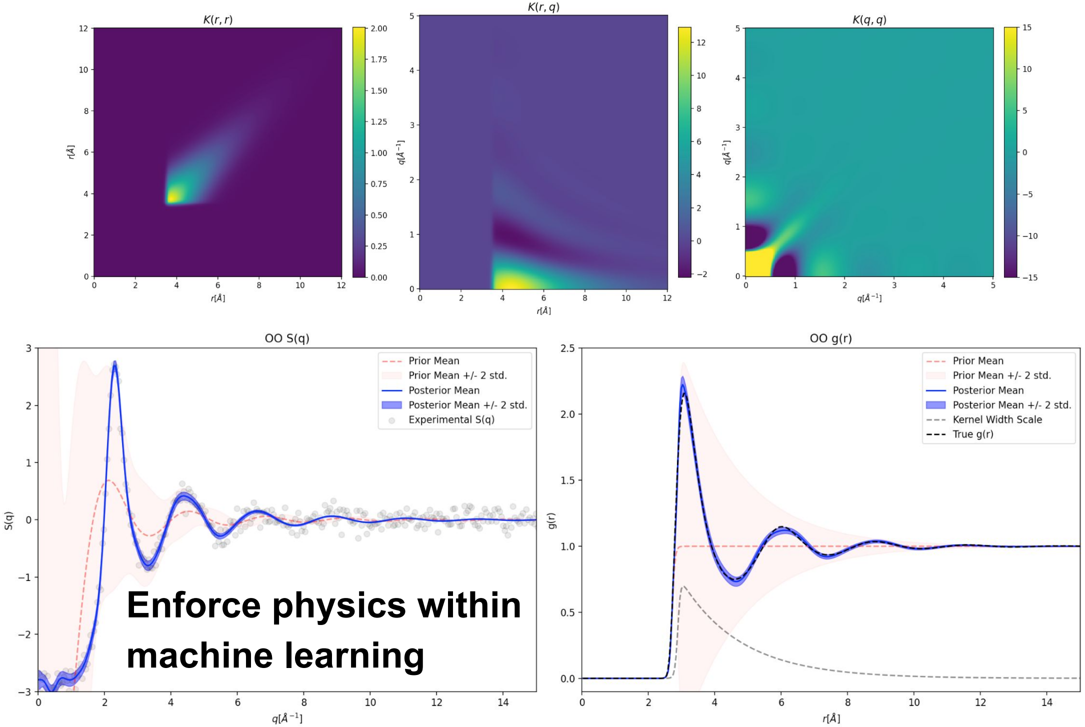


Comparing MD Simulations to Experimental Scattering Data



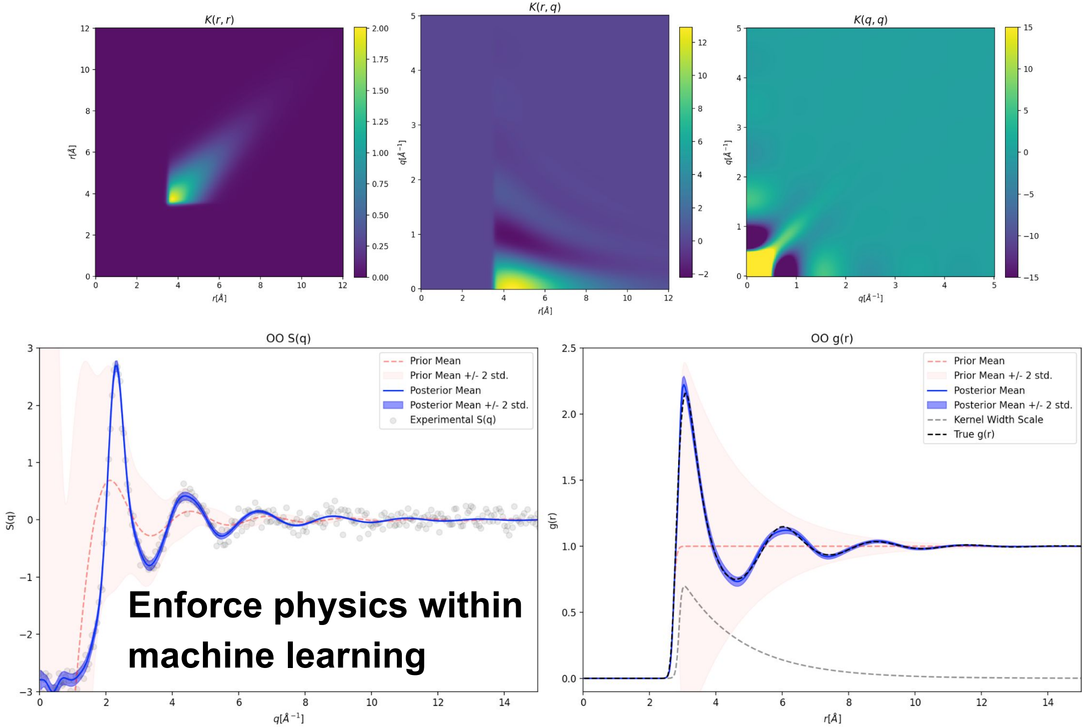
Bayes is a powerful mathematical tool for inverse problems

Neutron and X-ray Scattering Fourier Transforms



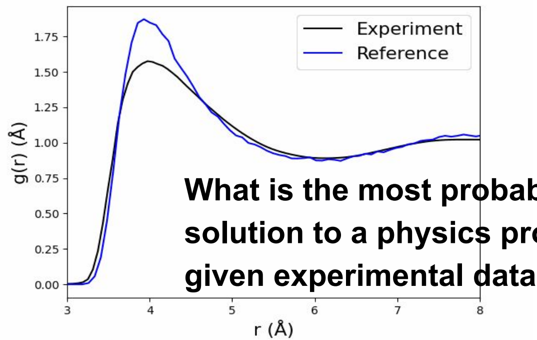
Bayes is a powerful mathematical tool for inverse problems

Neutron and X-ray Scattering Fourier Transforms

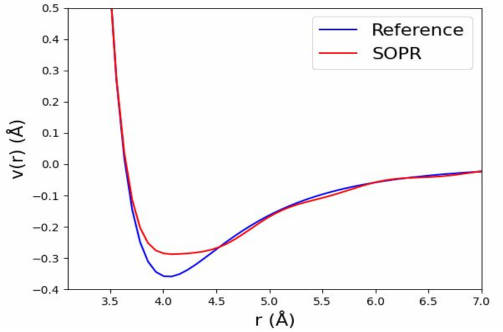


Enforce physics within machine learning

Learning Forces from Structure



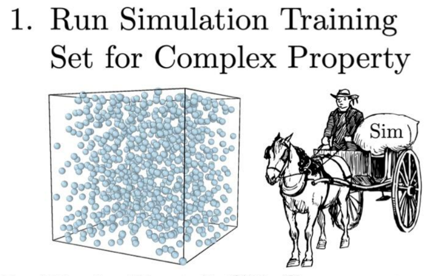
What is the most probable solution to a physics problem given experimental data?



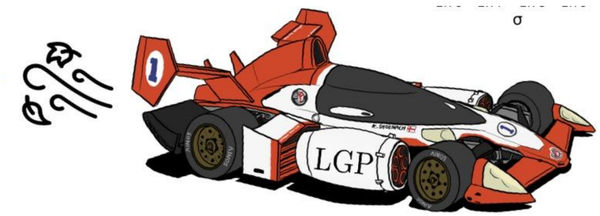
J. Phys. Chem. Lett. 2022, 13, 49, 11512-11520

Surrogate Models – Speeding Up Expensive Calculations with ML

Surrogate models are fast and accurate alternatives to expensive models.

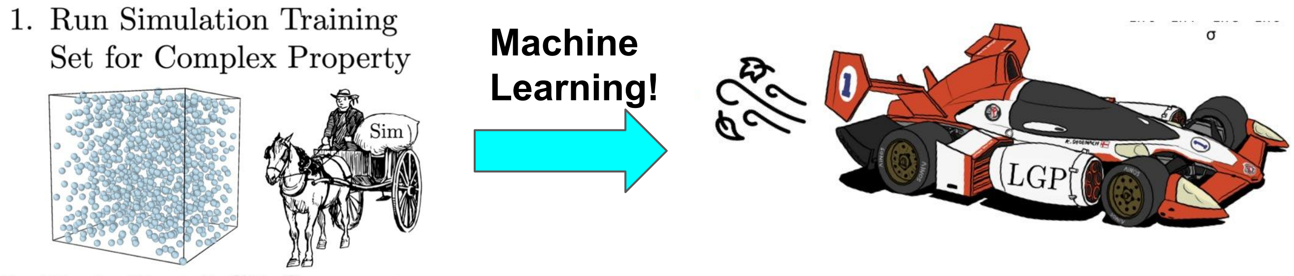


Machine Learning!

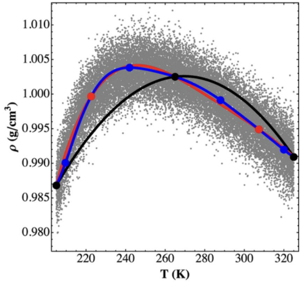


Surrogate Models – Speeding Up Expensive Calculations with ML

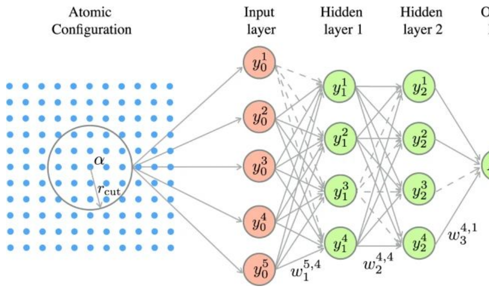
Surrogate models are fast and accurate alternatives to expensive models.



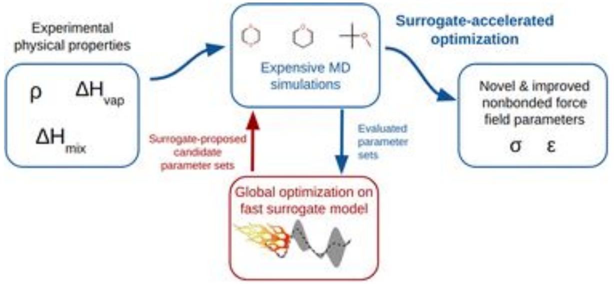
Generalized Polynomial Chaos



Neural Networks



Gaussian Processes



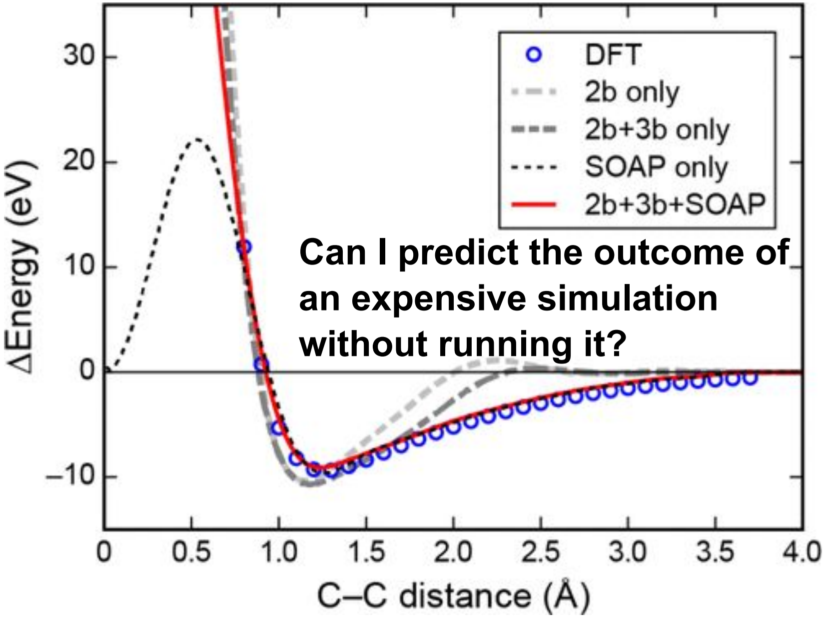
Jacobson, J. *Phys. Chem. B* 2014, 118, 28, 8190–8202

Wen, *Npj Comput. Mater.* 2020, *6*, 1– 10,

Madin [Digital Discovery](), 2023, **2**, 828-847

Surrogate Models – Speeding Up Expensive Calculations with ML

Machine Learning Potentials

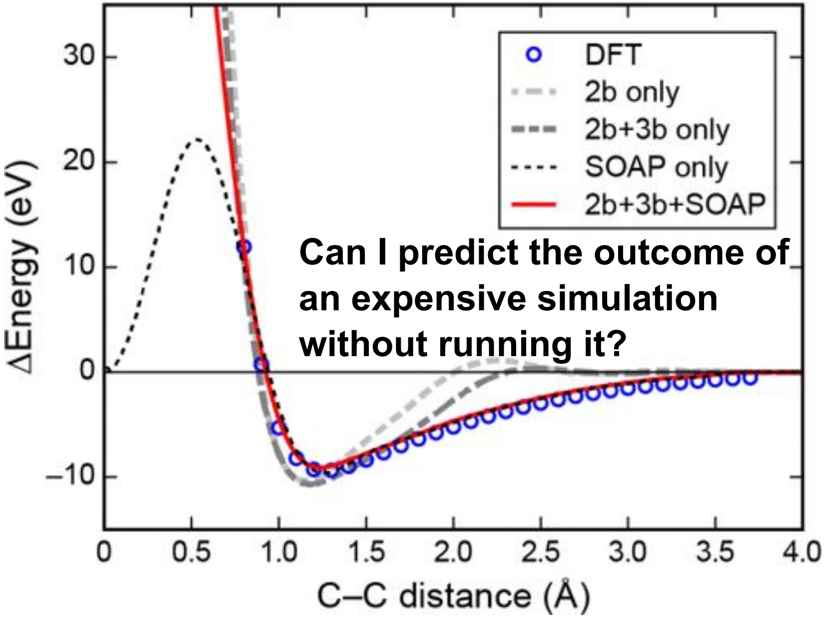


Can I predict the outcome of an expensive simulation without running it?

Chem. Rev. 2021, 121, 16, 10073-10141

Surrogate Models – Speeding Up Expensive Calculations with ML

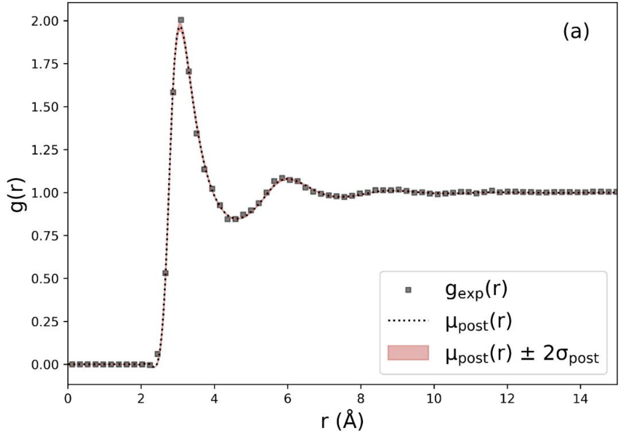
Machine Learning Potentials



Can I predict the outcome of an expensive simulation without running it?

Chem. Rev. 2021, 121, 16, 10073-10141

Estimating Classical MD Outputs with UQ



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

Shanks, B. L. et al. Accelerated Bayesian inference for molecular simulations using local Gaussian process surrogate models. *J. Chem. Theory Comput.* **20**, 3798–3808 (2024).

Project 1: Bayesian Force Field Optimization for ECC Models

Objective: Implement Bayesian UQ on ECC force field optimization in order to…

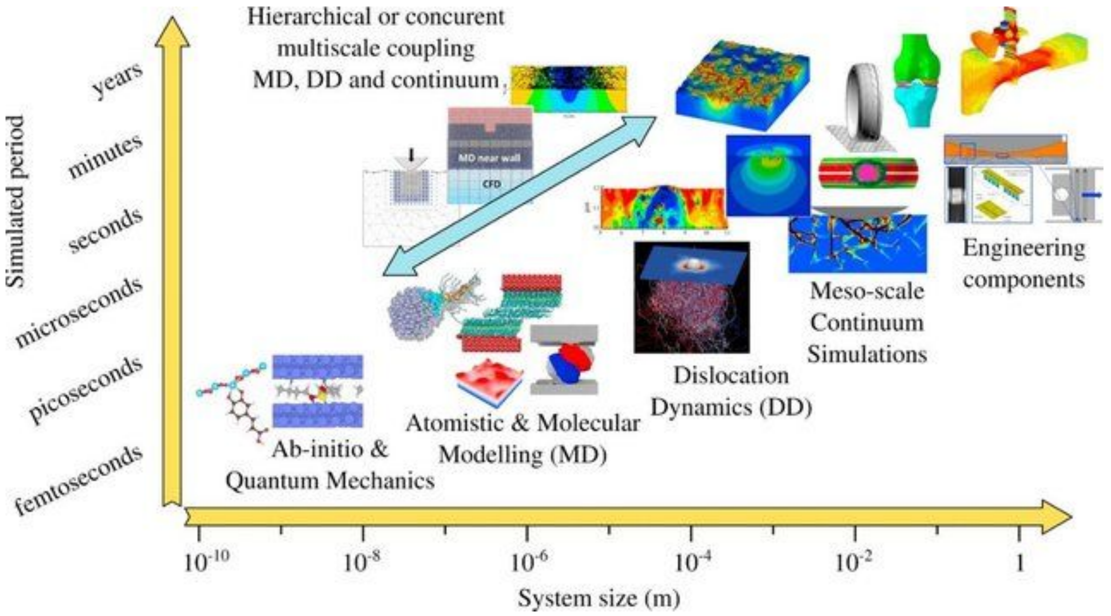
1. Identify what experiments are valuable for force field training
2. Rigorously quantify what models are the best at reproducing data
3. Quantify “trade offs” in the model parameterization (Pareto fronts)
4. Perform optimization with as few training simulations as possible (active learning)
5. Make models more interpretable and available
6. Integrate a “turn key” approach into open source software???

Project 2: Machine Learning Potentials: A Game Changer with Trade-offs

Machine learning potentials (MLPs) are surrogate models for interatomic potentials.

MLPs promise the accuracy of *ab initio* calculations at costs near classical molecular dynamics.

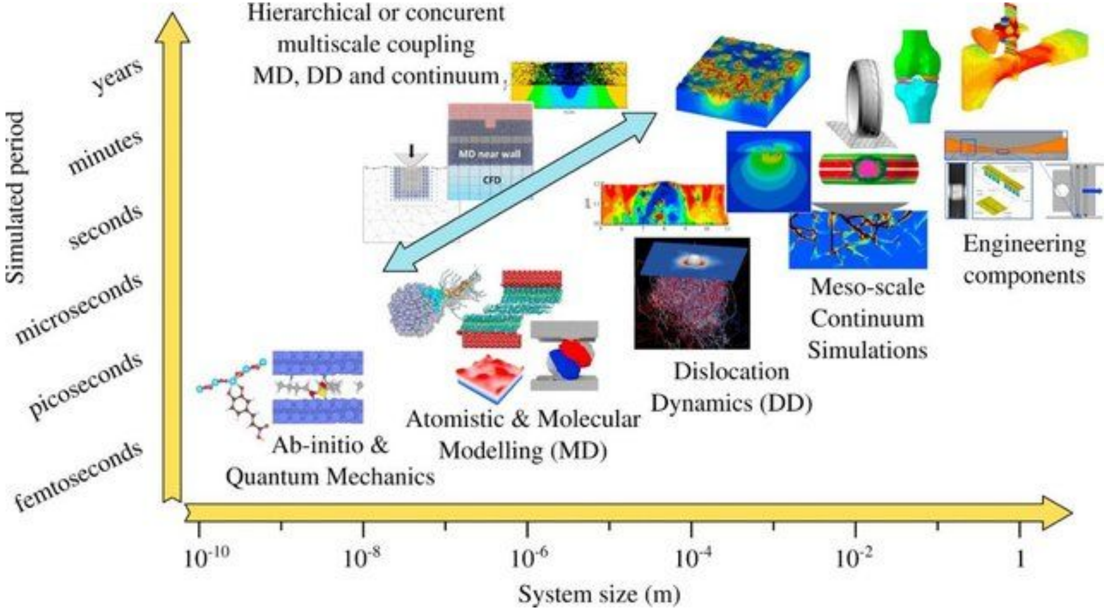
Hence, MLPs allow us to have higher accuracy forces for larger systems run for longer times.



Project 2: Machine Learning Potentials: A Game Changer with Trade-offs

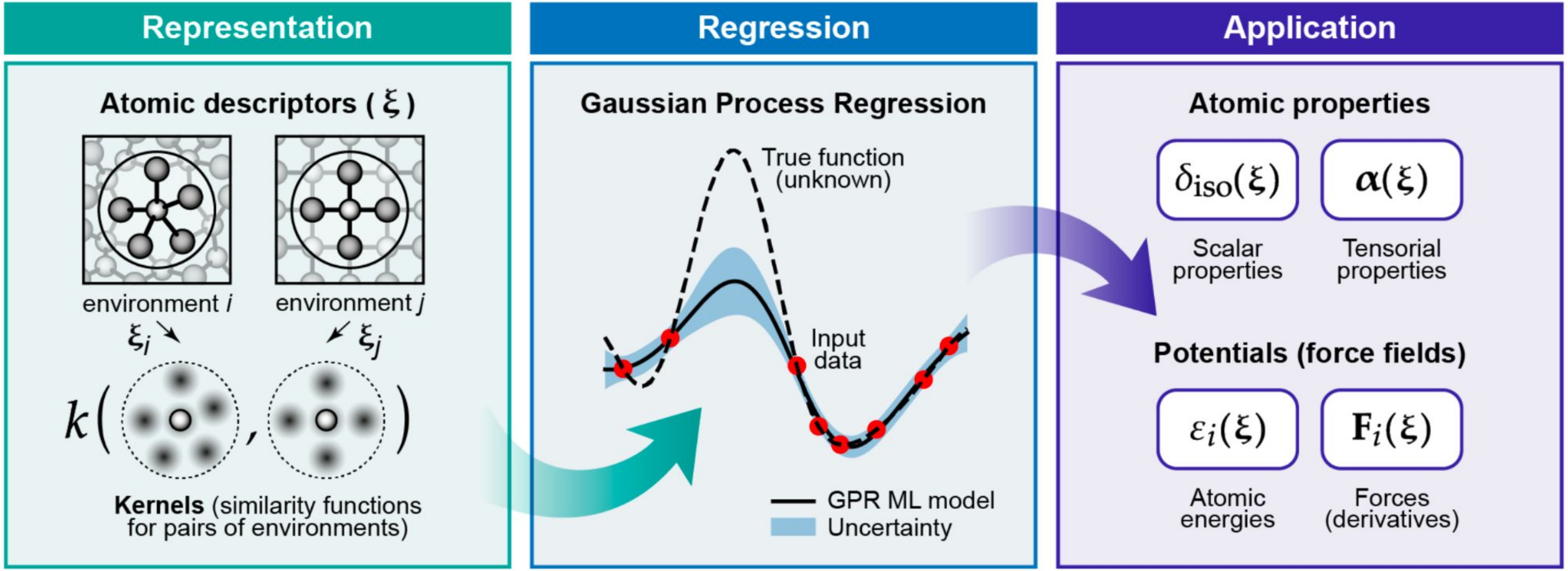
Of course, these methods come with some severe limitations, including:

- ML method selection and hyperparameter tuning
- Uncertainty quantification
- Awareness of and integration with experimental data
- Dataset creation and curation
- Computational cost of generating training data



Project 2: GAPs as Physics-Informed Machine Learning Potentials

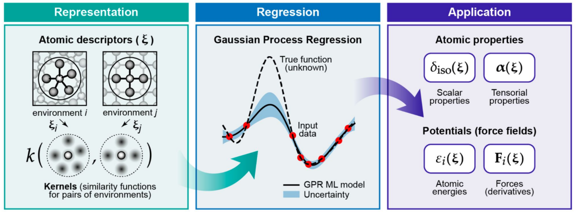
The GAP framework is already a well-established approach to learn forces from DFT calculations with UQ



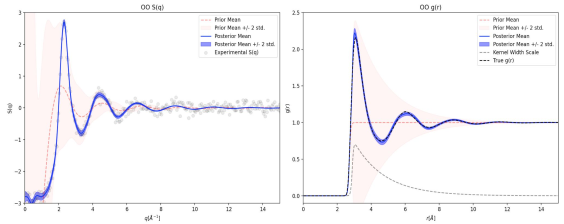
Deringer, V. L. *et al.* Gaussian process regression for materials and molecules. *Chem. Rev.* **121**, 10073–10141 (2021).

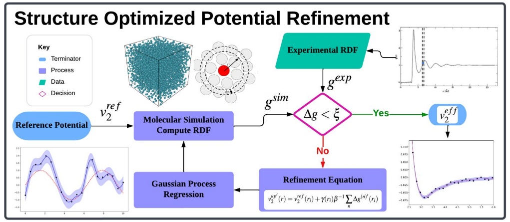
Project 2: GAPs as Physics-Informed Machine Learning Potentials

GAPs from DFT



GP Potentials from Experimental Scattering





GPs are mathematically “nice”.

As long as we can match up input and output spaces, the BCM is trivial.

Bayesian Committee Machine

The hope is our GP corrects deficiencies in the underlying DFT by training to condensed phase experimental data!

Experiment + Theory informed GP potential with UQ

Potential Synergies and Conclusions

Bayesian inference is a powerful statistical framework for uncertainty quantification

- Model optimization (force fields, DFT, partial charge selection, etc)
- Experimental data analysis
- Theoretical analysis of inverse problems
- Machine learning
- Surrogate modeling

Potential Synergies and Conclusions

Bayesian inference is a powerful statistical framework for uncertainty quantification

- Model optimization (force fields, DFT, partial charge selection, etc)
- Experimental data analysis
- Theoretical analysis of inverse problems
- Machine learning
- Surrogate modeling

Current projects

1. Bayesian optimization of ECC force field parameters (error estimation, sensitivity analysis, model comparison)
2. Uncertainty-aware, physics-informed machine learning potentials

Thank you!

If you see any potential synergies with your work and Bayesian / ML methods let's talk during the discussion section!!

Are you ML curious?
Machine Learning Advances in Molecular Physics Seminar - Thursdays @ 4pm
Email shanks.brennon@uochb.cas.cz to join the mailing list.

Webpage: https://bshanks.netlify.app