Uncertainty Quantification of Machine Learning Interatomic Potential Models

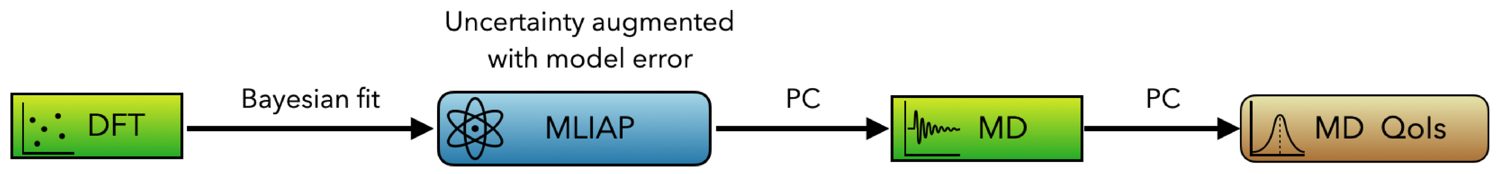
Sandia National Laboratories

Khachik Sargsyan, Logan Williams, Habib N. Najm

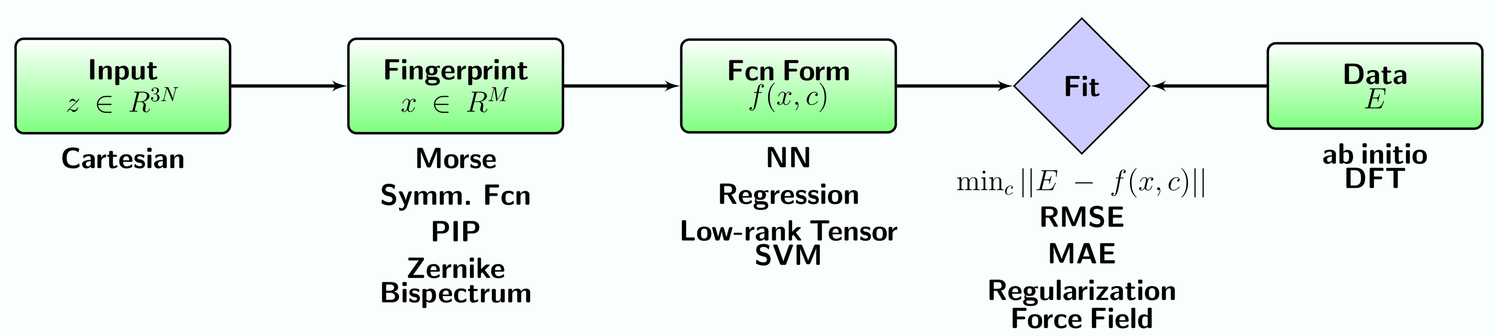
Sandia National Laboratories, Livermore, CA, USA

Big Picture

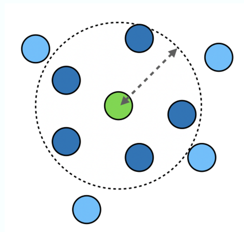
Bayesian Inference
Model Error
Polynomial Chaos Propagation



ML Interatomic Potentials (MLIAPs)



- Partition the interatomic interaction energy into individual contributions of the atoms
  $$E_{\text{total}} = \sum_{i=1}^{N} E_i$$
- Assume flexible functional forms with respect to positions of the neighboring atoms
  $$E \approx f(x, c)$$



Bayesian Inference of MLIAPs

Parameter inference

- Given a model $f(x,c)$ and data $y_i = y(x_i)$, calibrate parameters c such that $y_i \approx f(x_i, c)$
  - Linear model $y \approx Ac$ with coefficients c, or
  - NN model $y \approx NN_c(x)$ with weights/biases c.
- Weighted least-squares fit:
  $c^* = \text{argmin}_c \sum_{i=1}^{N} w_i^2 \left(f(x_i, c) - y_i\right)^2$

Likelihood function or data model

Posterior PDF, Prior PDF, Model, Data

$$\underbrace{p(c|y)}_{\text{Posterior PDF}} \propto \underbrace{p(y|c)}_{\text{Likelihood}}\underbrace{p(c)}_{\text{Prior PDF}} \propto \prod_{i=1}^{N} \exp\left(-\frac{(f(x_i,c) - y_i)^2}{2\sigma_i^2}\right)$$

Likelihood

✦ Prior contains previous knowledge or regularization

✦ Likelihood contains data noise modeling assumptions,

e.g. $y_i = f(x_i, c) + \sigma_i \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0,1)$

Data Model

Model Error

Elephant in the room: model error

$$\underbrace{y_i}_{\text{Truth}} = \underbrace{f(x_i, c)}_{\text{Model}} + \underbrace{\sigma_i \epsilon_i}_{\text{Data err.}}$$

Model ≠ Truth

Ignoring model error hurts in a few ways:

- Biased estimates of parameters c (crucial if the model is physical, and/or c is propagated through other models)
- More data leads to overconfident predictions (we become more and more certain about the wrong values of the data)

Capturing model error in data model

External correction (Kennedy-O'Hagan):

$$y_i = f(x_i, c) + \delta(x_i) + \sigma_i \epsilon_i$$

- Kennedy, O'Hagan, "Bayesian Calibration of Computer Models". *J Royal Stat Soc: Series B (Stat Meth)*, 63: 425-464, 2001.

Internal correction (embedded model error):

$$y_i = f(x_i, c + \delta(x_i)) + \sigma_i \epsilon_i$$

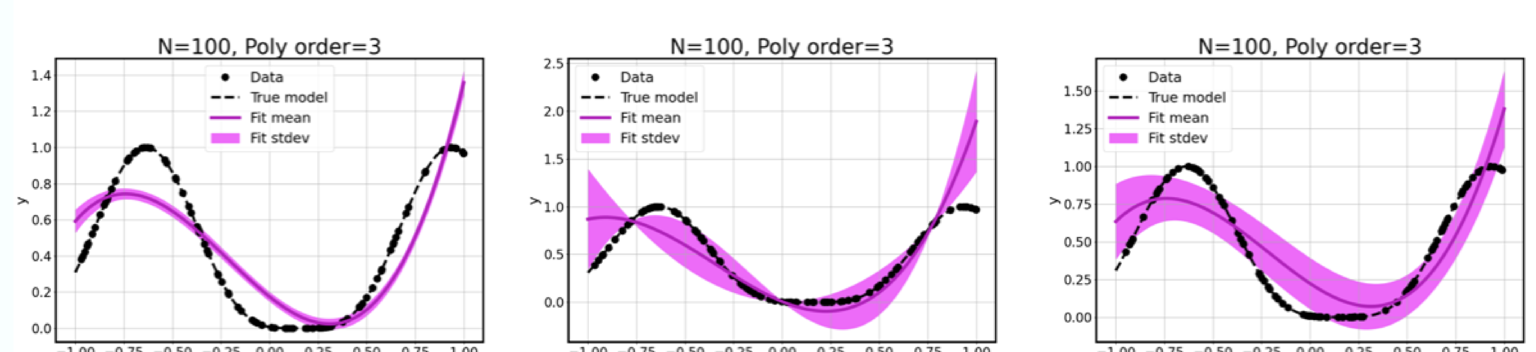
- Allows meaningful usage of calibrated model
- 'Leftover' noise term even with no data error
- Respects physics (not too relevant in our context)

- Sargsyan, Najm, Ghanem, "On the Statistical Calibration of Physical Models". *Int. J. Chem. Kinet.*, 47: 246-276, 2015.
- Sargsyan, Huan, Najm, "Embedded Model Error Representation for Bayesian Model Calibration". *Int. J. Uncert. Quantif.*, 9(4): 365-394, 2019.

Embedded model error method

- Statistical correction inside the model
  $$y_i \approx f\left(x_i, c + d\mathcal{N}(0,1)\right)$$
- Can be done non-intrusively, with a surrogate
- Jointly infer parameters of model and model error
  $$p(c, d|y) \propto p(y|c, d)p(c, d)$$
- Degenerate likelihood: needs approximations
  - Independent output approximation (IID):
    $$p(c, d|y) \propto \prod_{i=1}^{N} \exp\left(-\frac{(\mu_f(x_i,c) - y_i)^2}{2\sigma_f^2(x_i,c)}\right)$$
  - Approximate Bayesian computation (ABC):
    $$p(c, d|y) \propto \prod_{i=1}^{N} \exp\left(-\frac{(\mu_f(x_i,c) - y_i)^2 + (\sigma_f(x_i,c) - \alpha|\mu_f(x_i,c) - y_i|)^2}{2\epsilon^2}\right)$$

Classical case — Model error, IID likelihood — Model error, ABC likelihood

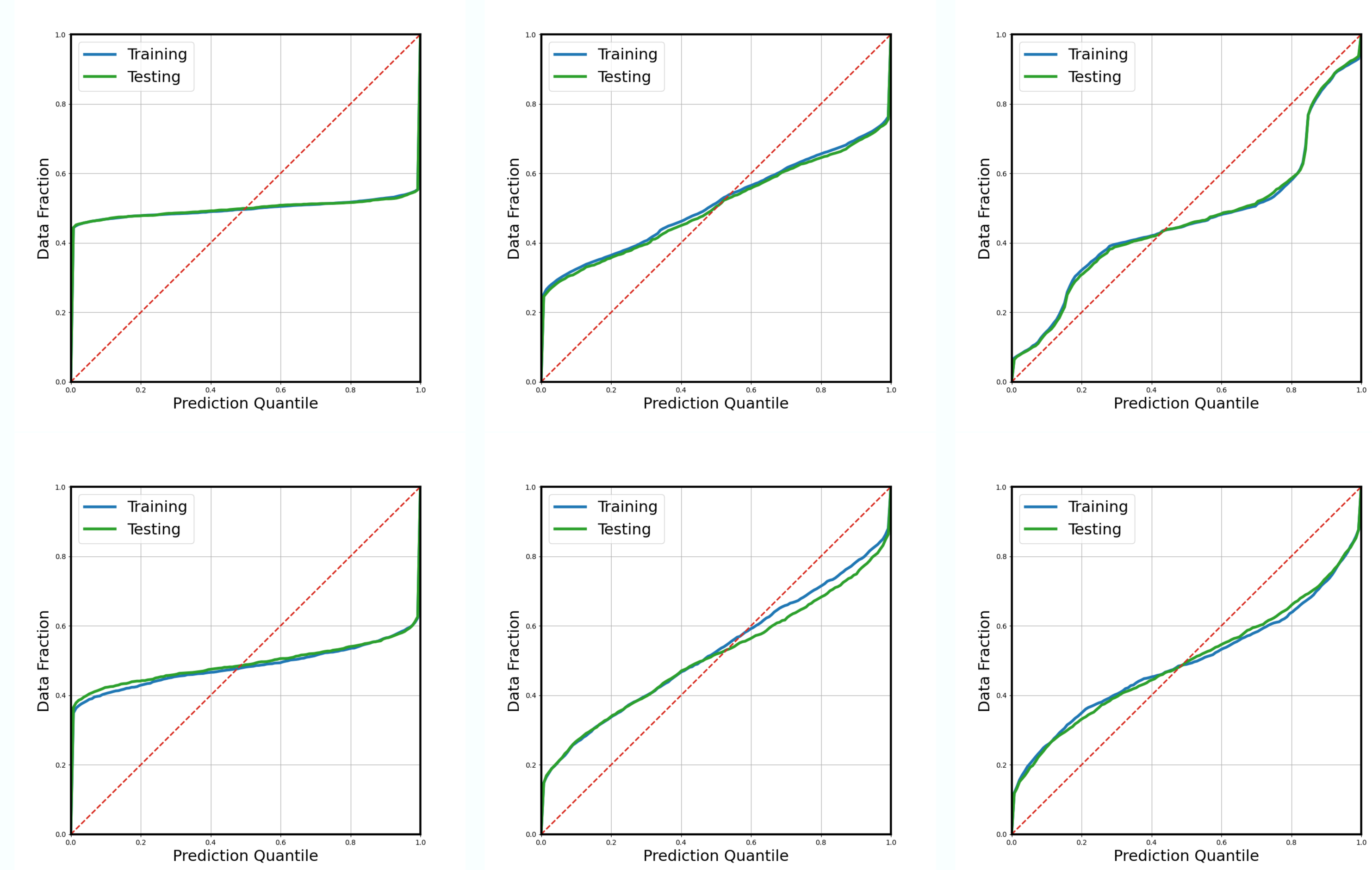


Uncertainty Validation

- We employ Spectral Neighbor Analysis Potential (SNAP) and FitSNAP (`https://github.com/FitSNAP/FitSNAP`)
- Embedded approach leads to better calibrated uncertainties enabling efficient active learning and uncertainty propagation.

Two examples with SNAP

No model error — Embedded, IID — Embedded, ABC



Uncertainty Propagation

Forward UQ via Polynomial Chaos (PC)

- Based on Bayesian MLIAP fit, construct input PC for MLIAP parameters $c = \sum_{k=0}^{K} a_k \Psi_k(\xi)$
- Sample input parameters and IAPs, $E(x) = f(x, c)$
- Obtain molecular dynamics QoIs $h = MD(E(x))$
- Build PC expansion for MD QoIs: $h = \sum_{k=0}^{K} b_k \Psi_k(\xi)$ via regression
- Evaluate QoIs statistics, compare to DFT benchmarks
- Variance-based decomposition (global sensitivity analysis) of the output PCs

Summary

- Bayesian fit of ML interatomic potentials: supervised ML
- Embedded model error with baked-in uncertainty
  - Model-error uncertainty capturing the true residual
- Polynomial chaos based uncertainty propagation through molecular dynamics

U.S. DEPARTMENT OF ENERGY

This research is funded by the U.S. Department of Energy, Office of Science, Advanced Scientific Computing Research (ASCR), and Fusion Energy Sciences (FES).

contact: Khachik Sargsyan, ksargsy@sandia.gov