SAND2022-4468C

Quantification and Propagation of Uncertainties in Machine Learning Interatomic Potentials for Molecular Dynamics

Model errors and active learning

SIAM UQ

April 15, 2022

Khachik Sargsyan, Logan Williams, Katherine Johnston, Habib Najm (SNL-CA)

Sandia National Laboratories

Acknowledgements:
Aidan Thompson, Mitch Wood,
Mary Alice Cusentino, Ember Sikorski

U.S. DEPARTMENT OF ENERGY

Funded by DOE ASCR / FES

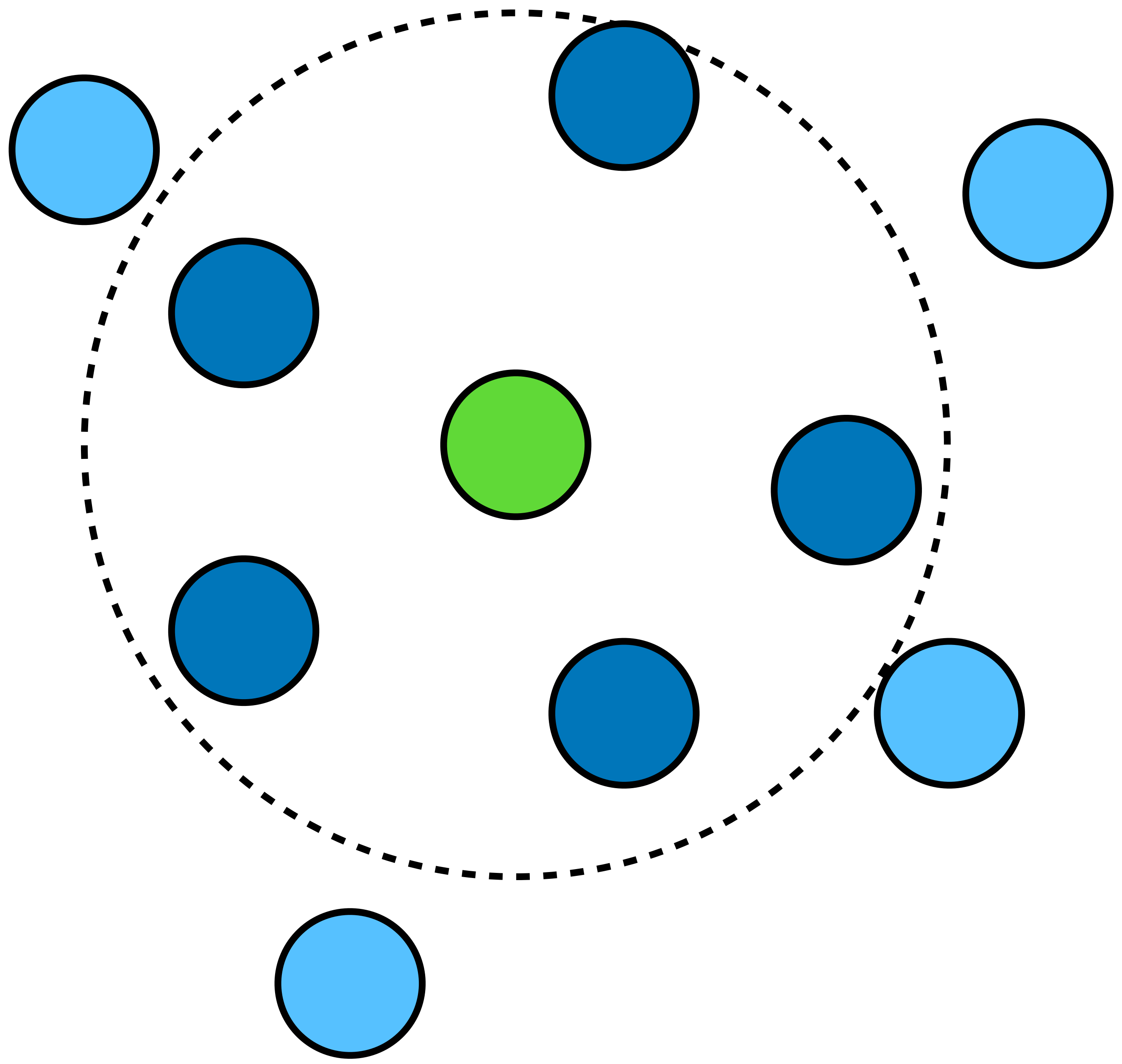
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

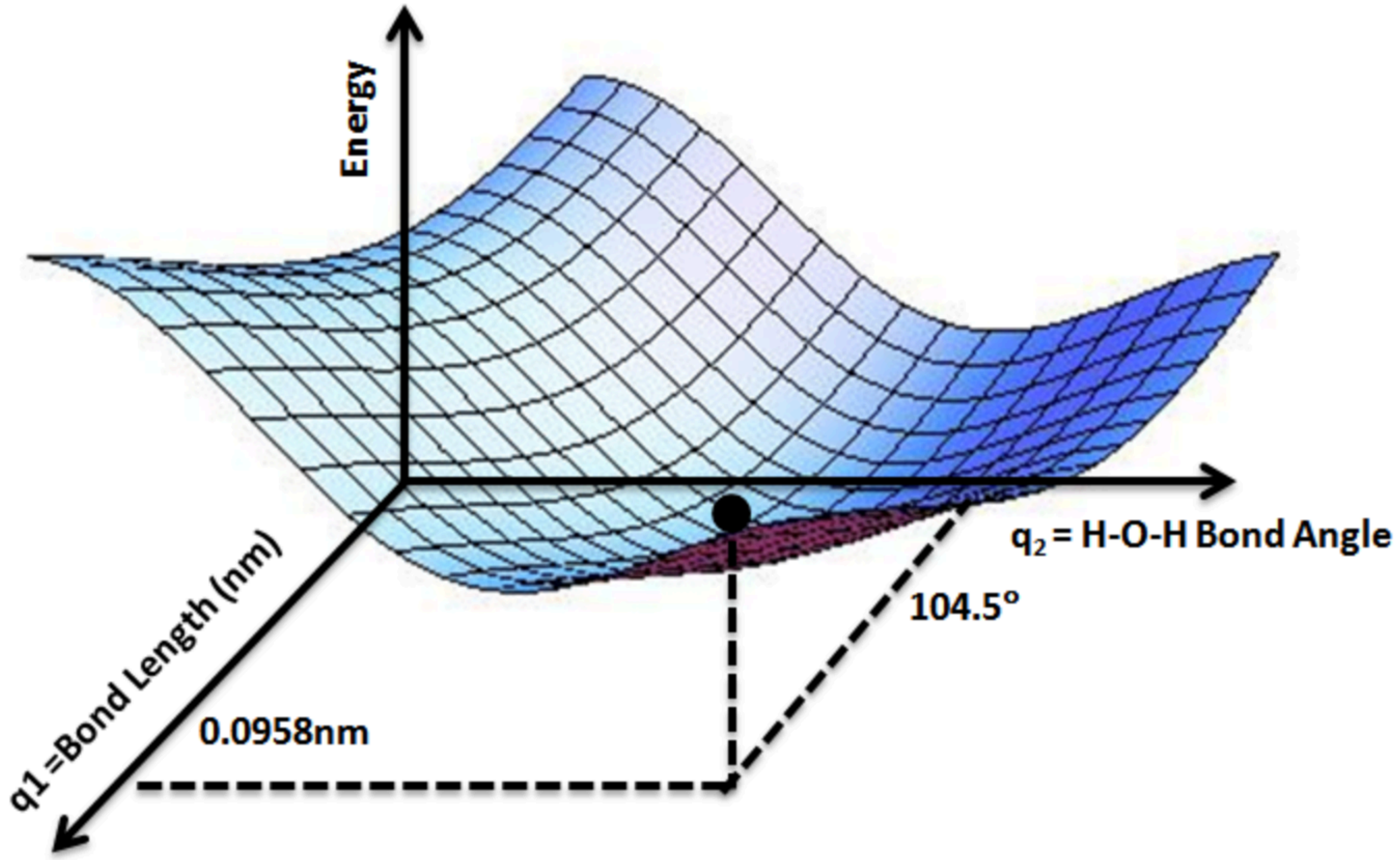
Outline

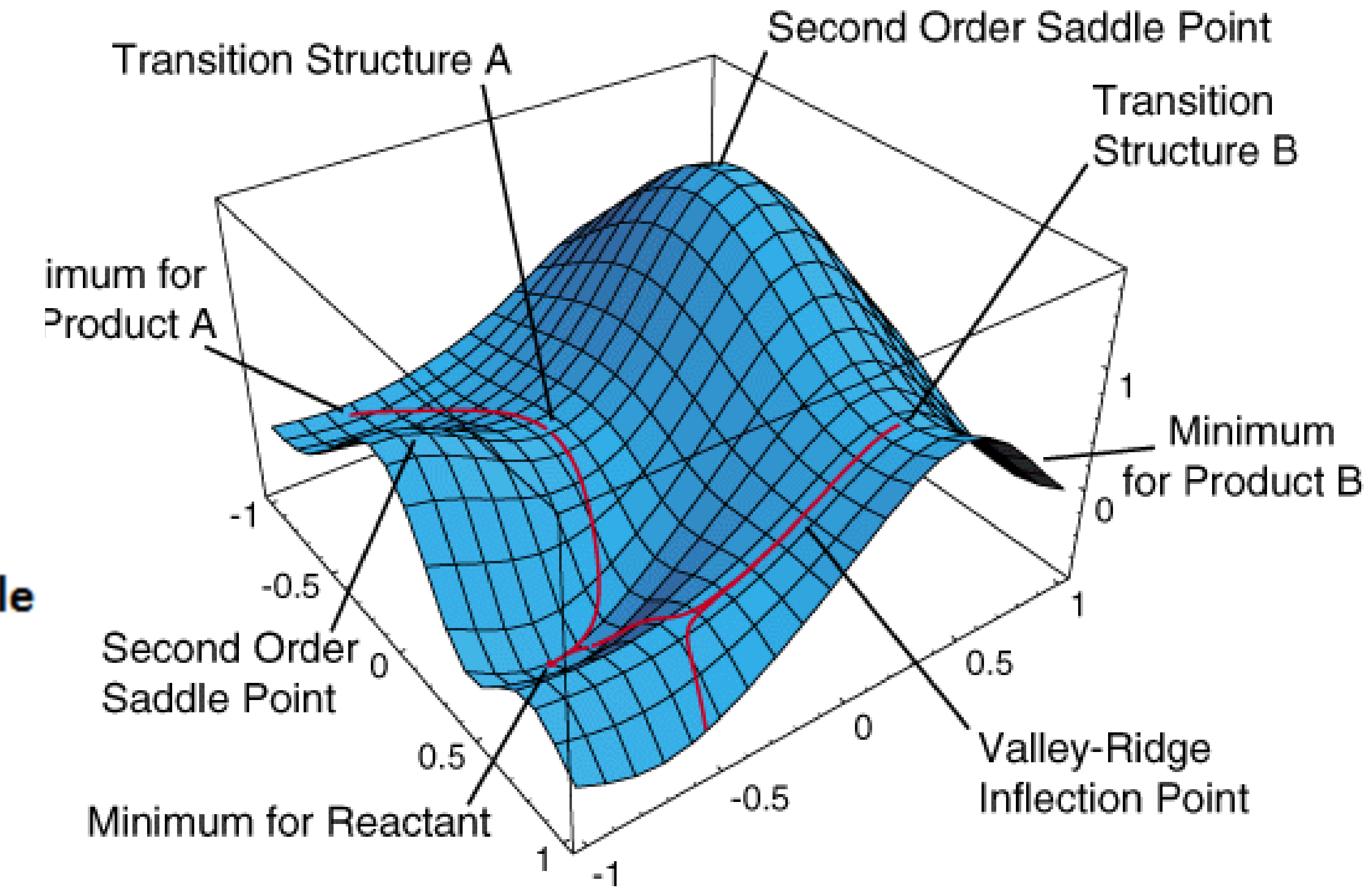
- Interatomic potentials as building blocks to approximate potential energy surfaces
- Machine learning interatomic potentials (MLIAP) - a supervised ML problem
- Active learning and need for uncertainty estimation in MLIAP construction
- (Bayesian) MLIAP hinges on proper assumptions for model-data discrepancies
- Embedded model error approach for uncertainty estimation in MLIAPs

Interatomic Potentials

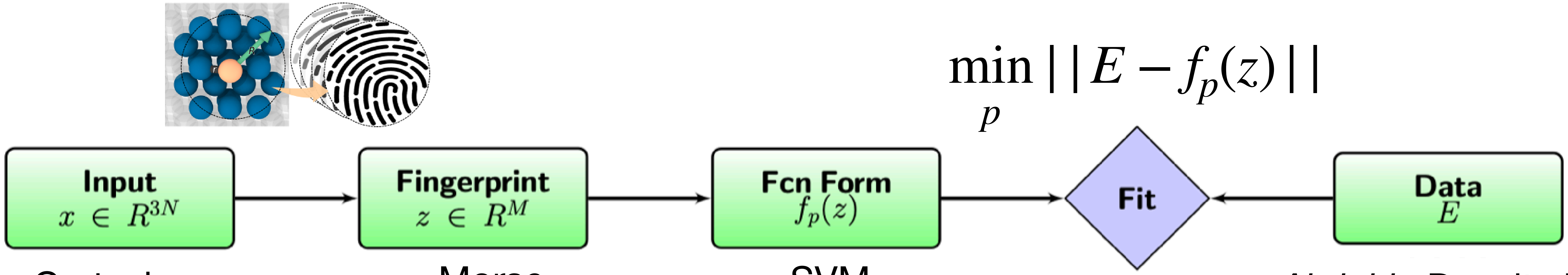
- Object of interest: potential energy E of a system defined by a configuration x , where x encapsulates coordinates of all atoms in the system
- Typically additive form. $E(x) = E_{ref} + \sum_i E(x_i) + \ldots$ using local environments







Ingredients of MLIAPs (supervised ML problem)



Cartesian

Morse
Symm. Fcns
PIP
Bispectrum
ACE
…

SVM
NN
GP
LR Tensor
Lin. Regression
…

Ab initio Density Functional Theory
include Force, Stress

- Training data (x_i, E_i) for $i = 1,\ldots,S$ and $x_i \in R^{3N}$
- Input representation, aka fingerprint, aka descriptor $x \rightarrow z(x)$
- Parametrized functional form of the approximation class $f_p(z)$
- Loss function: $\min_p \sum_{i=1}^{S} [E_i - f_p(z_i)]^2$ + regularization

State-of-the-art: largely manual and lacking systematic UQ

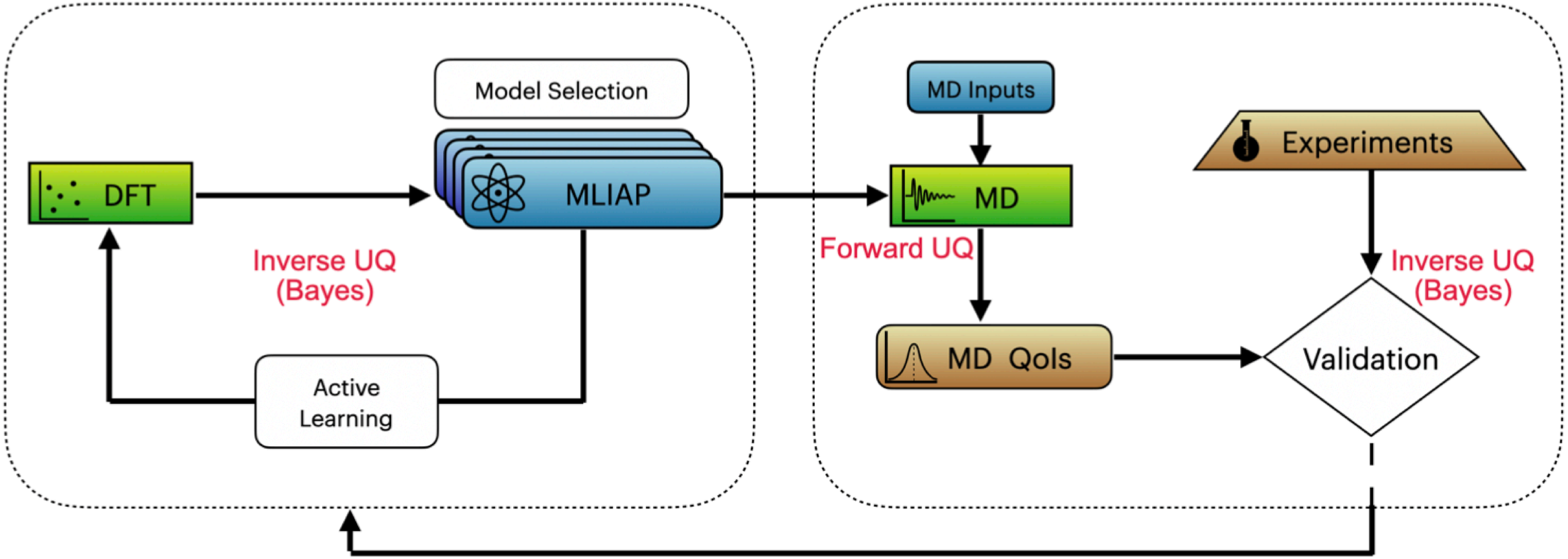
MLIAP Construction

- Good training set selection: active learning
- Fingerprint choice: invariances, symmetries
- Functional form choice: model selection
- Loss function: regularization, weighting energies and forces

MLIAP Usage

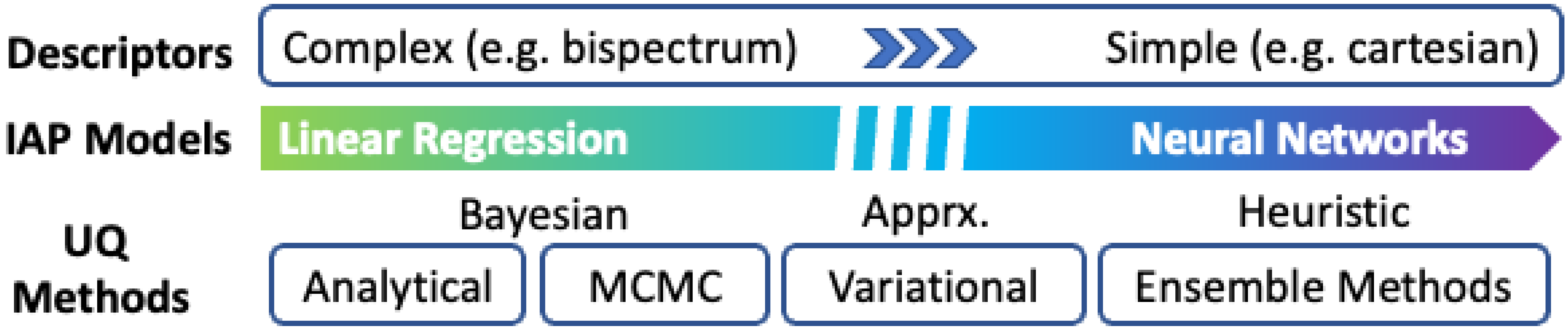
- Find reaction pathways, saddle points
- Pipe the IAPs to MD simulations

Big Picture



Bayesian inference of IAPs, model errors
Active learning

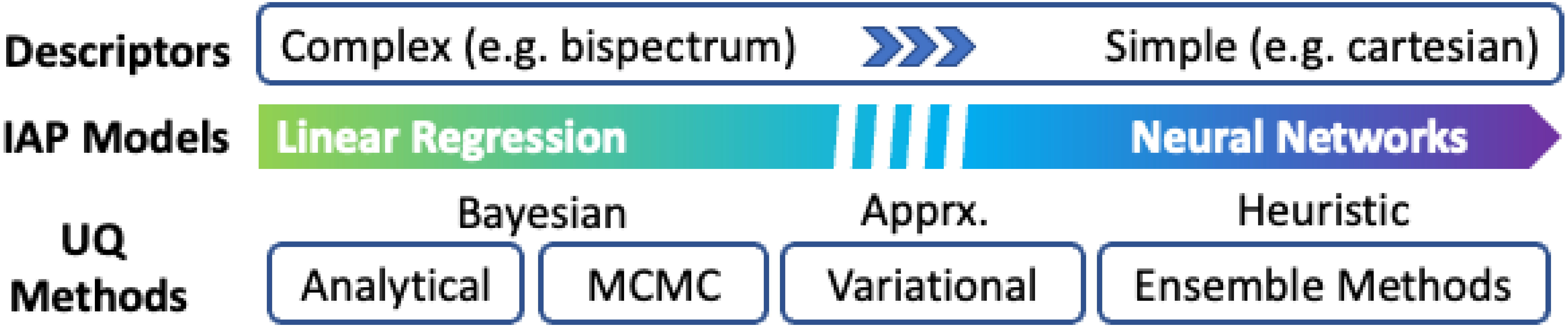
Equipping parametric fits with uncertainties



Equipping parametric fits with uncertainties

SNAP

A.P. Thompson et al. “Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials”, *Journal of Computational Physics*, 285(15), pp. 316-330, 2015.

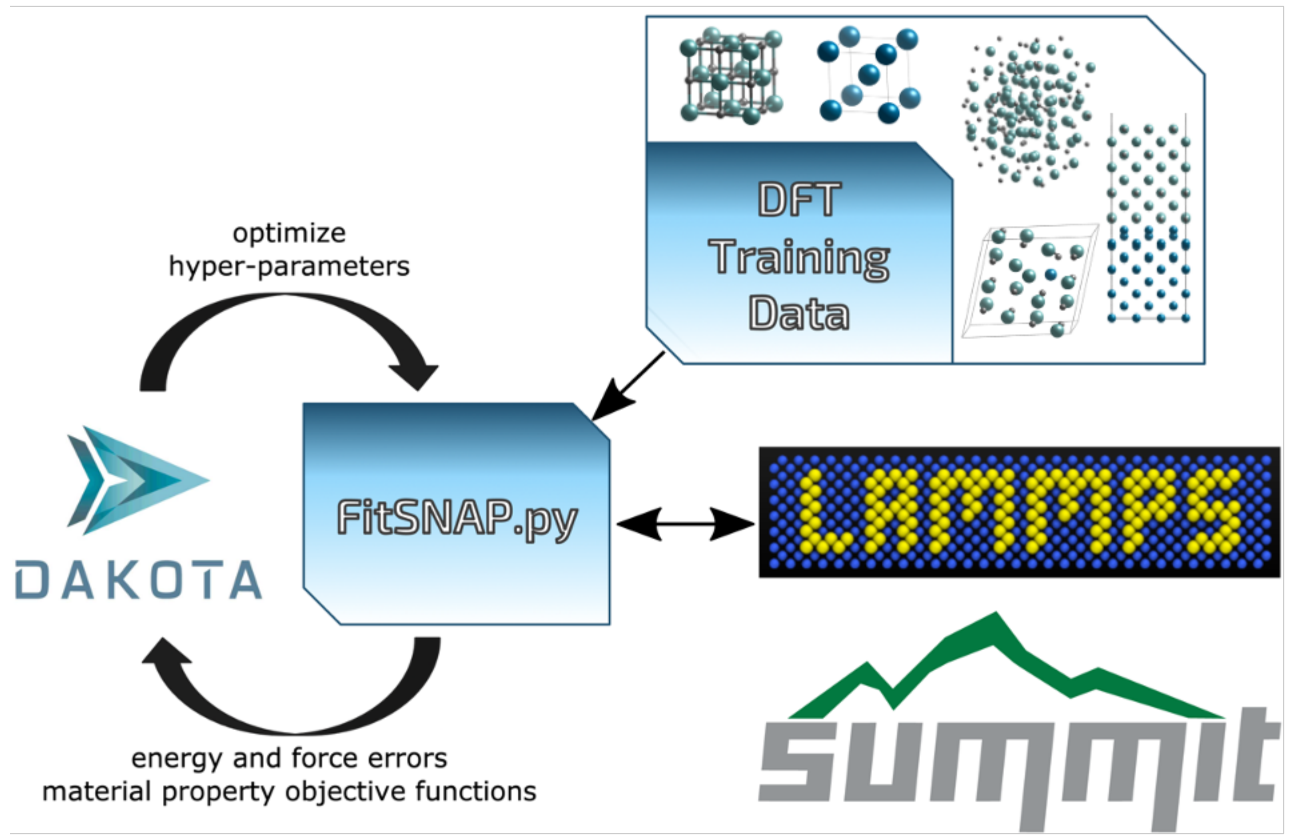


Spectral neighbor analysis potential (SNAP) details

A.P. Thompson et al. "Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials", *Journal of Computational Physics*, 285(15), pp. 316-330, 2015.

- Uses **bispectrum** as fingerprints:

- uses hyper spherical harmonics
- respects rotational, permutational, translational invariances
- incorporates forces and stresses as well
- tunable complexity/order



$$E(x) \approx \sum_k c_k B_k(x)$$

- Uses **linear regression** as model form:

- built on hyper spherical harmonics basis functions
- generalized to quadratic form as well

M. Wood and A. Thompson , "Extending the accuracy of the SNAP interatomic potential form", *Journal of Chemical Physics*, 148, 2018.

(Bayesian) Parameter Inference

- ✦ Given a model $f(x, c)$ and data $y_i = y(x_i)$, calibrate parameters c.

  - Linear model $y \approx Ac$ with coefficients c
  - NN model $y \approx NN_c(x)$ with weights/biases c

- ✦ Bayesian least-squares fit: $p(c\,|\,y) \propto p(y\,|\,c)p(c) \propto \prod_{i=1}^{N} \exp\left(-\frac{(f(x_i, c) - y_i)^2}{2\sigma_i^2}\right)$

  Corresponding data model $y_i = f(x_i, c) + \sigma_i \epsilon_i$

Elephant in the room: model is assumed to be *the* correct model behind data

Model   Data err.

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

Truth

Model $\neq$ Truth

Ignoring model error hurts in a few ways:

✦ One gets 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)

✦ More evident when there is no (observational/experimental) data error: *e.g.* DFT is data, and MLIAP is model

Posterior uncertainty does not capture true discrepancy

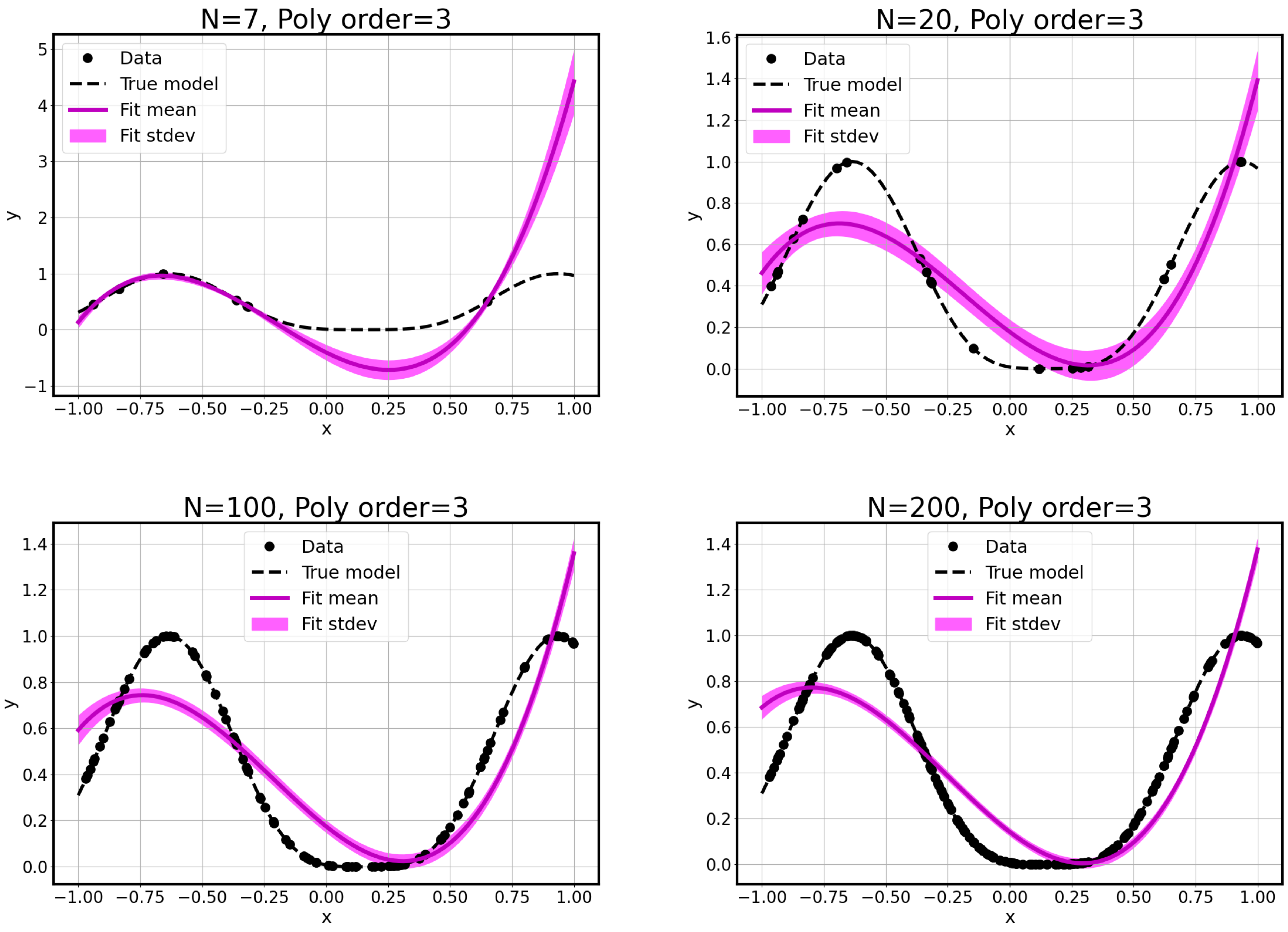
Synthetic data

$$y(x) = \sin^4(2x - 0.3)$$

Cubic fit

$$y_i \approx \sum_{k=0}^{3} c_k B_k(x)$$

More data leads to overconfident prediction



Capturing Model Error in Likelihood (a.k.a. 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 for Linear Regression Models

$$\sout{y_i \approx \sum_{k=0}^{P} c_k B_k(x) + \sigma_i \epsilon_i}$$

Note:

No formal distinction between

internal and external corrections,

but internal allows for interpretation

and model-informed error

'Embed' uncertainty in

all (or selected) coefficients

Model

Model error

$$y_i \approx \sum_{k=0}^{P} (c_k + d_k \xi_k)\, B_k(x) \quad = \quad \sum_{k=0}^{P} c_k B_k(x) + \sum_{k=0}^{P} d_k B_k(x) \xi_k$$

(still Gaussian, but correlated,

and model-informed)

Embedded Model Error: likelihood choice is challenging

Classical data model

$$y_i \approx \sum_{k=0}^{P} c_k B_k(x) + \sigma_i \epsilon_i$$

$$p(c\,|\,y) \propto \prod_{i=1}^{N} \exp\left(-\frac{(\sum_{k=0}^{P} c_k B_k(x_i) - y_i)^2}{2\sigma_i^2}\right)$$

MCMC sampling of c

Embedded model error

$$y_i \approx \sum_{k=0}^{P} (c_k + d_k \xi_k)\, B_k(x) \quad = \quad \sum_{k=0}^{P} c_k B_k(x) + \sum_{k=0}^{P} d_k B_k(x) \xi_k$$

Option 1 (IID)

$$p(c, d\,|\,y) \propto \prod_{i=1}^{N} \exp\left(-\frac{(\sum_{k=0}^{P} c_k B_k(x_i) - y_i)^2}{2\sum_{k=0}^{K} d_k^2 B_k(x_i)^2}\right)$$

MCMC sampling of c, d

or

simply optimize the posterior for c, d

Embedded Model Error: likelihood choice is challenging

Classical data model

$$y_i \approx \sum_{k=0}^{P} c_k B_k(x) + \sigma_i \epsilon_i$$

$$p(c\,|\,y) \propto \prod_{i=1}^{N} \exp\left(-\frac{\left(\sum_{k=0}^{P} c_k B_k(x_i) - y_i\right)^2}{2\sigma_i^2} \right)$$

MCMC sampling of c

Embedded model error

$$y_i \approx \sum_{k=0}^{P} (c_k + d_k \xi_k)\, B_k(x) \quad = \quad \sum_{k=0}^{P} c_k B_k(x) + \sum_{k=0}^{P} d_k B_k(x) \xi_k$$

Option 2 (ABC)

$$p(c, d\,|\,y) \propto \prod_{i=1}^{N} \exp\left(-\frac{\left(\sum_{k=0}^{P} c_k B_k(x_i) - y_i\right)^2 + \left(\sqrt{\sum_{k=0}^{P} d_k^2 B_k^2(x_i)} - \alpha \left|\sum_{k=0}^{P} c_k B_k(x_i) - y_i\right|\right)^2}{2\epsilon^2} \right)$$

Pushed forward predictive uncertainty captures the true discrepancy from the data

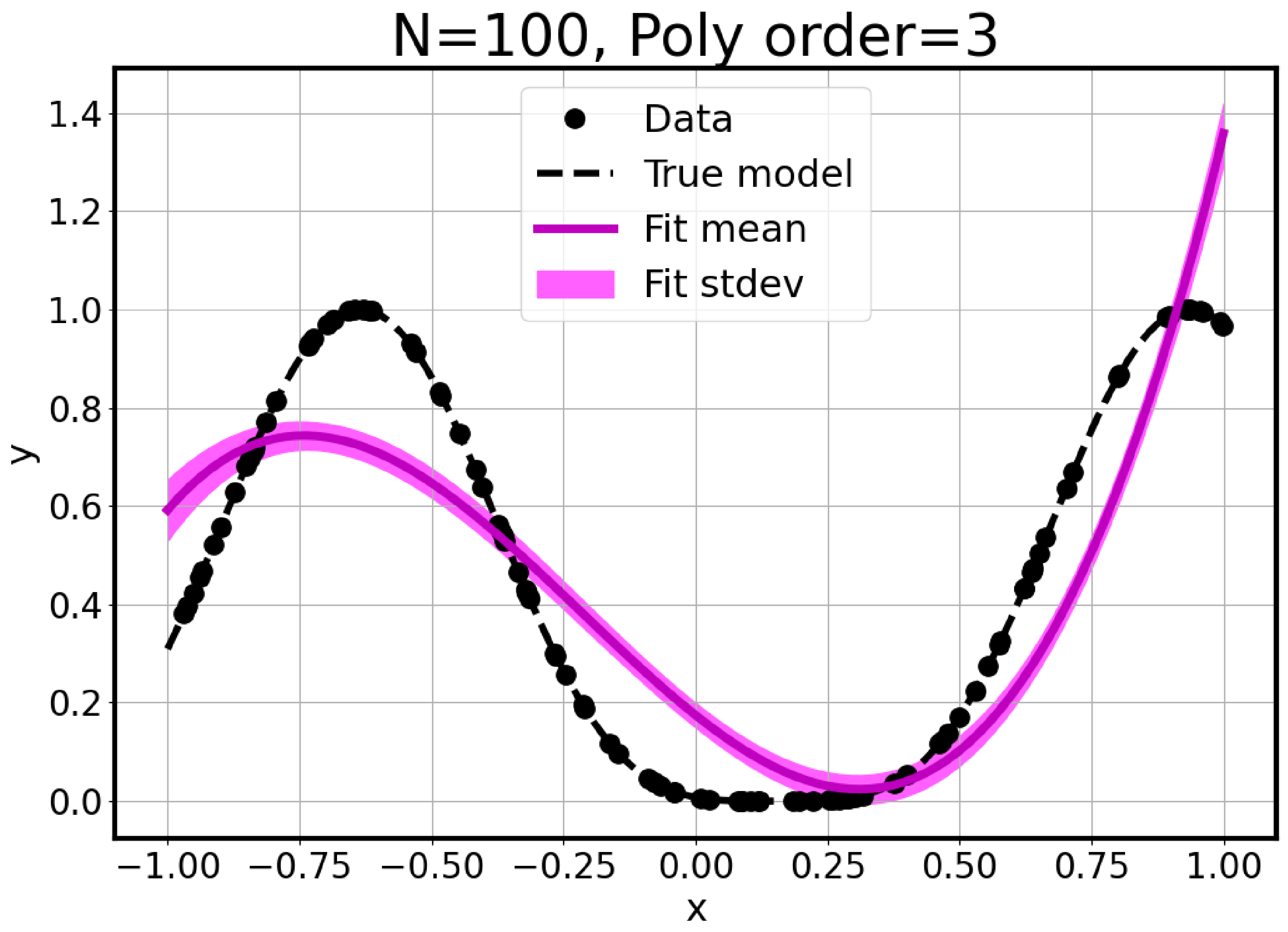
Synthetic data

$$y(x) = \sin^4(2x - 0.3)$$

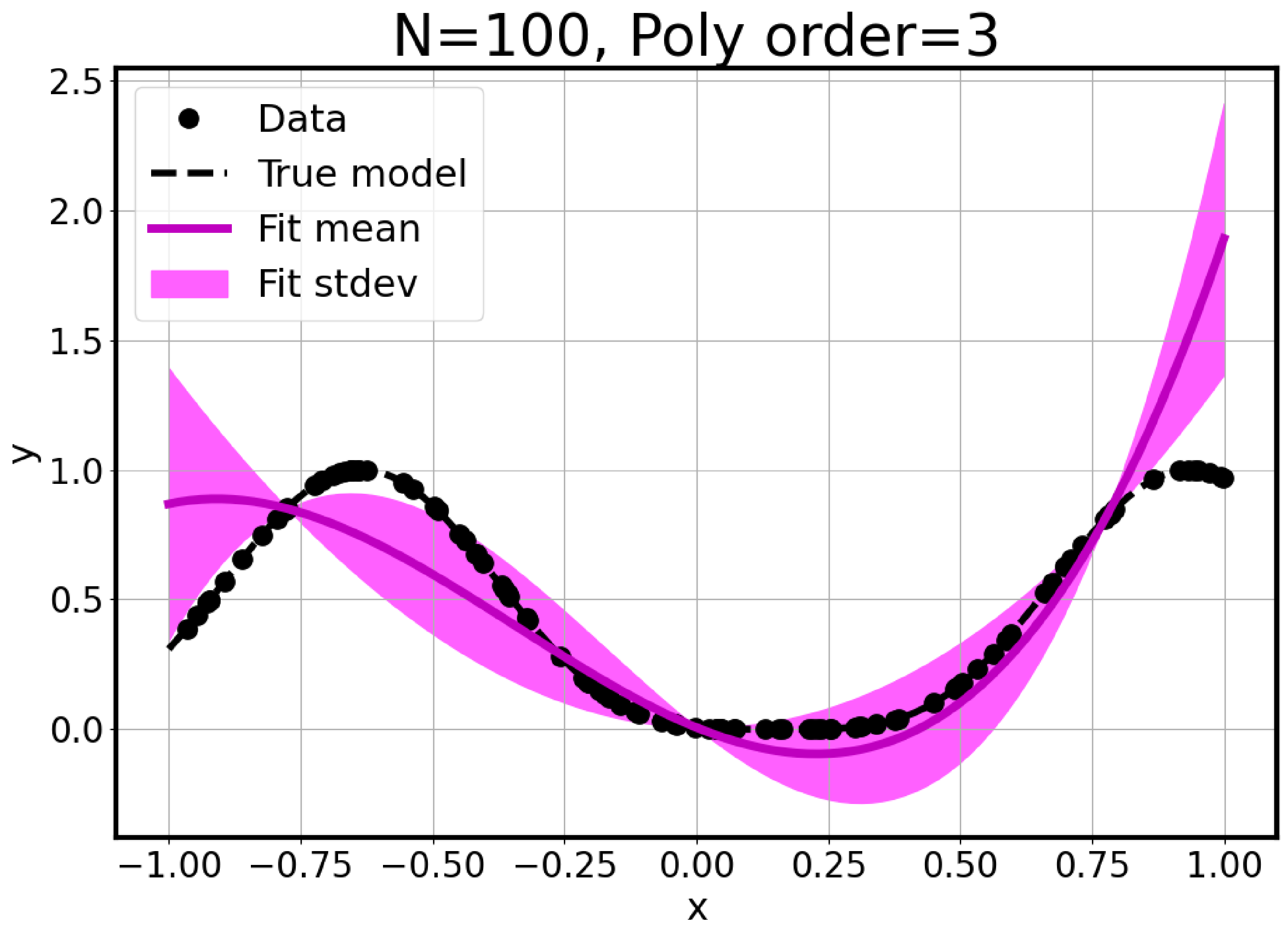
Cubic fit

$$y_i \approx \sum_{k=0}^{3} c_k B_k(x)$$

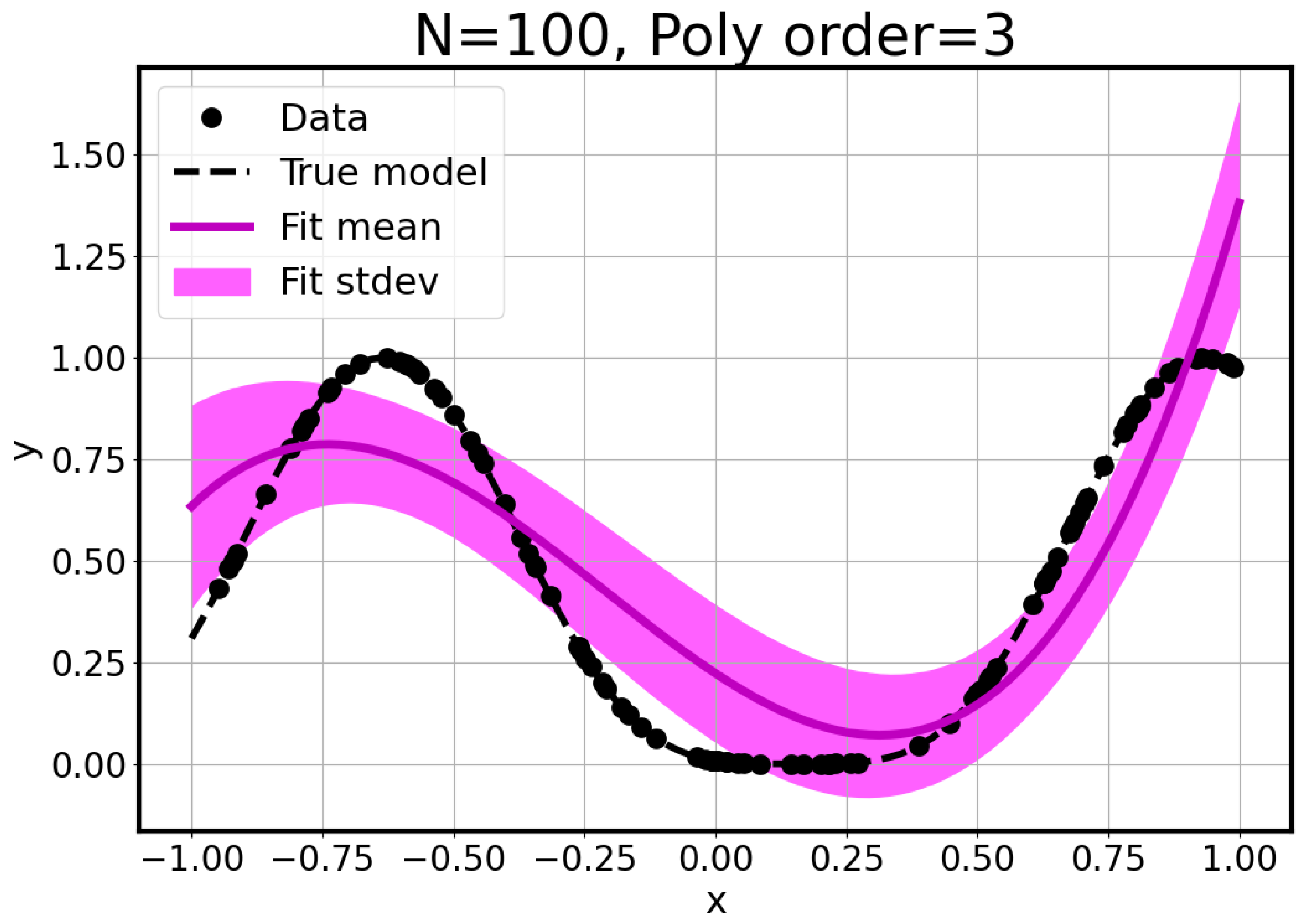
Classical case



Model error, IID likelihood

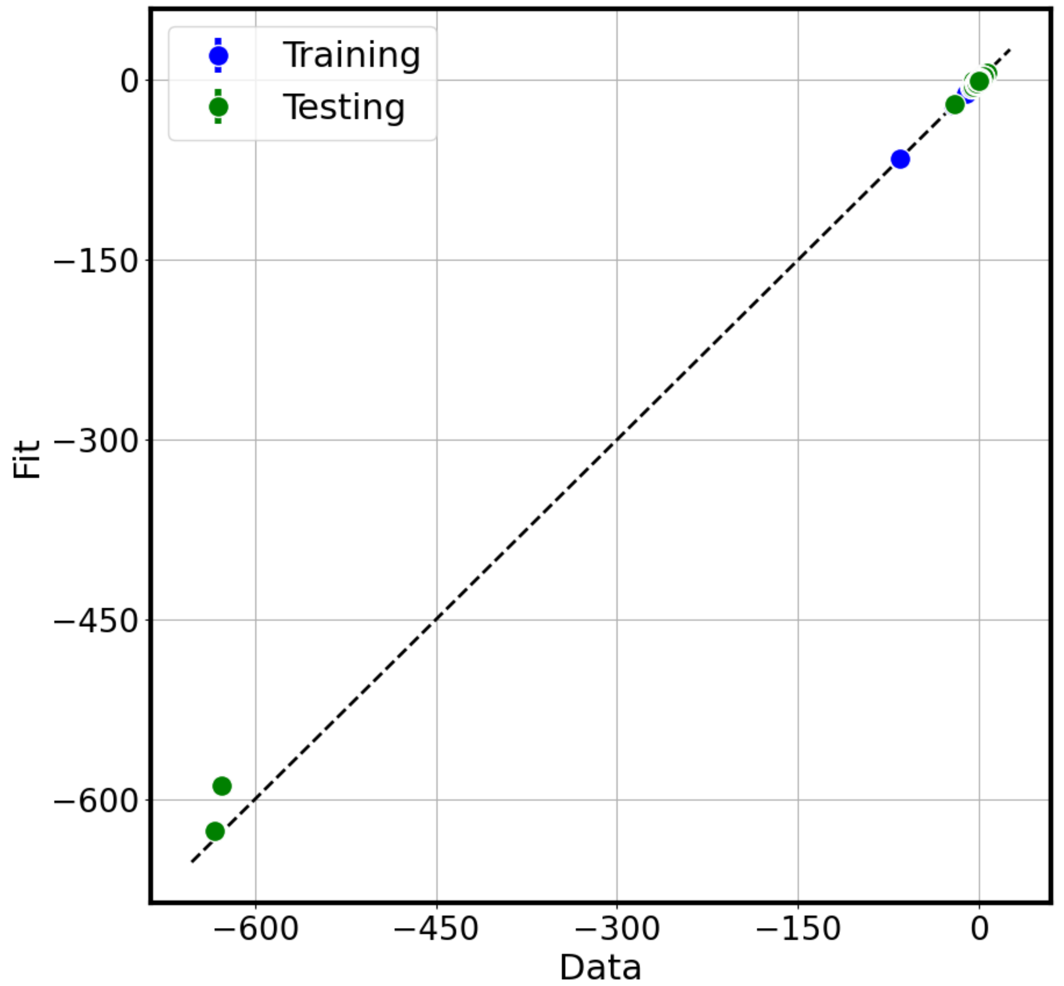


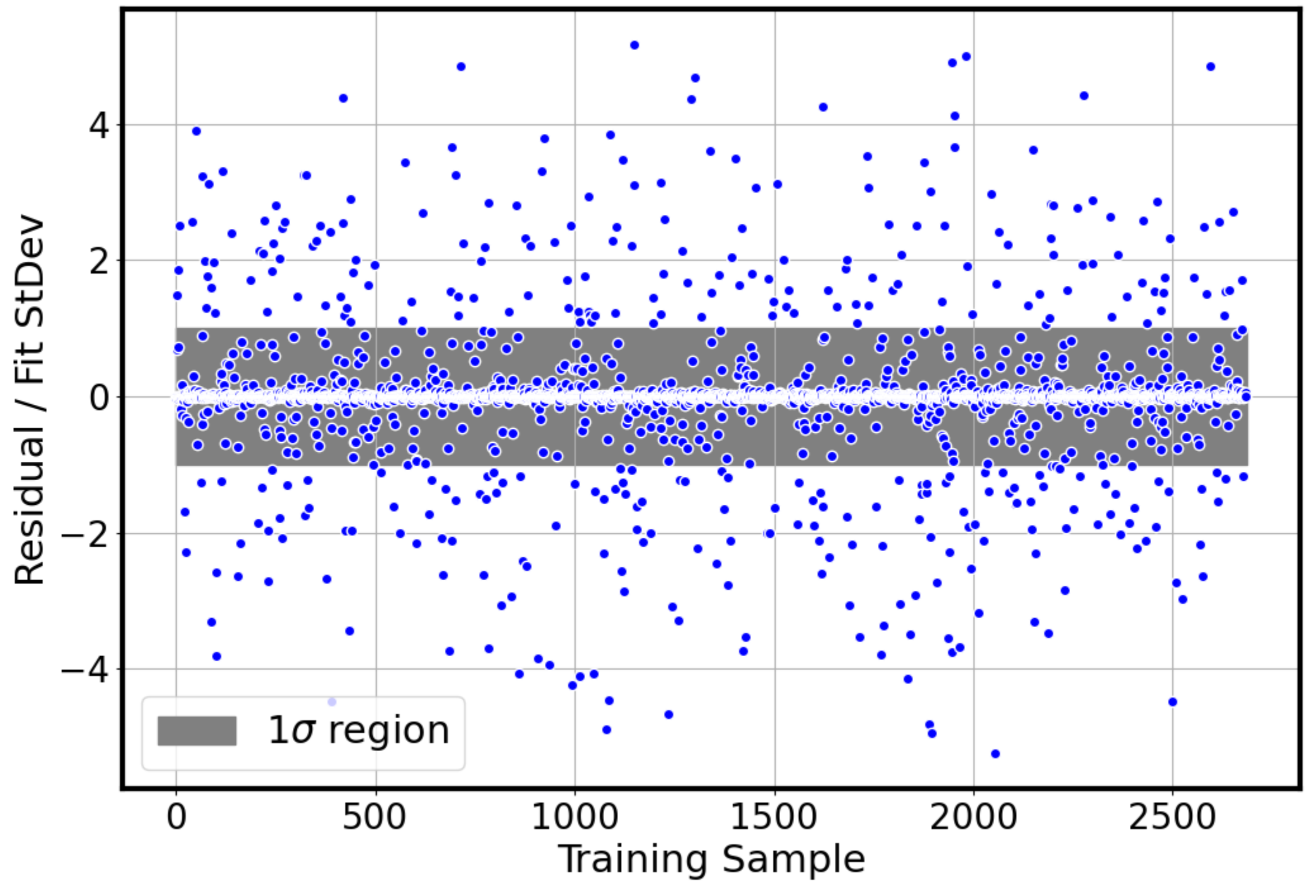
Model error, ABC likelihood

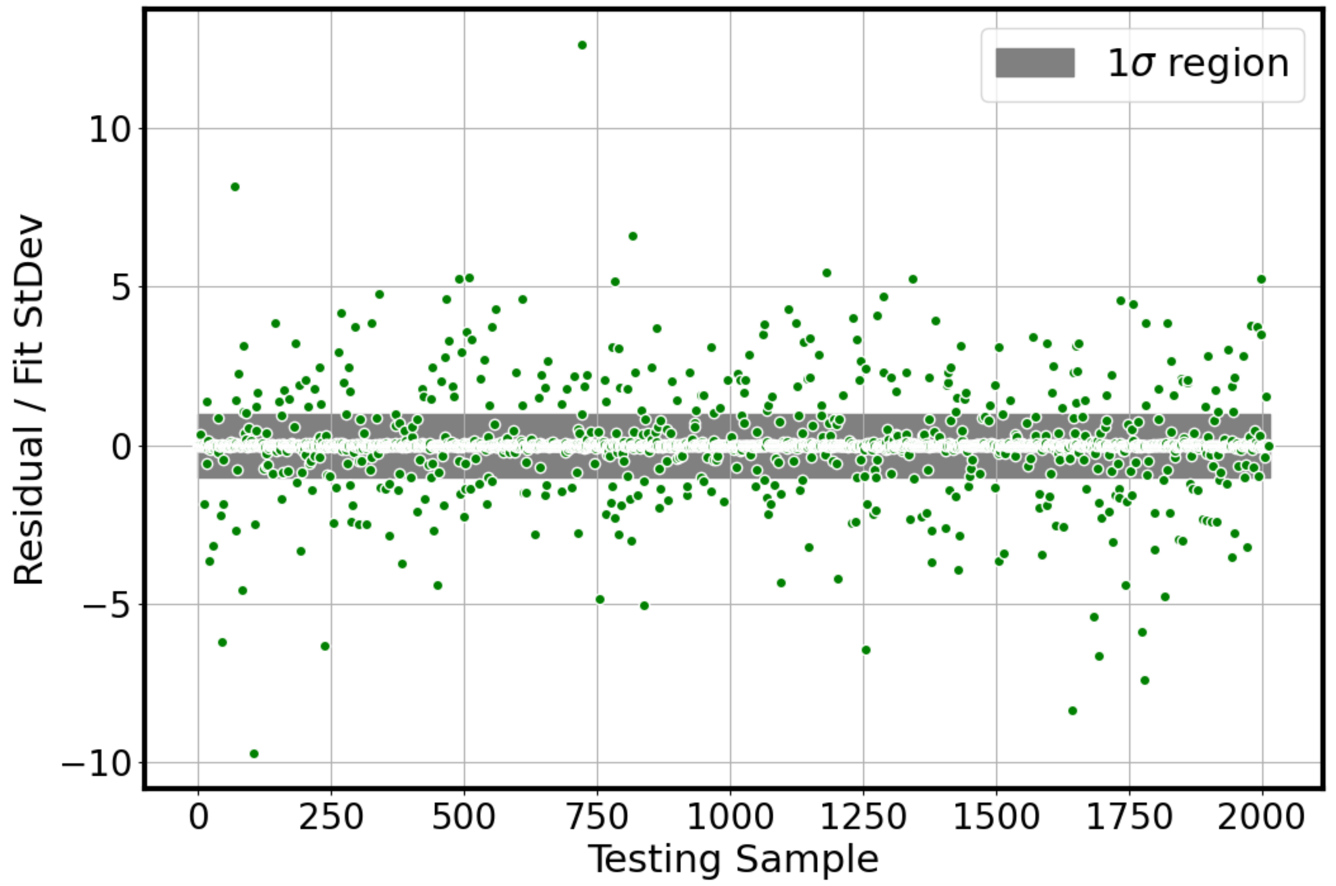


W-ZrC Dataset

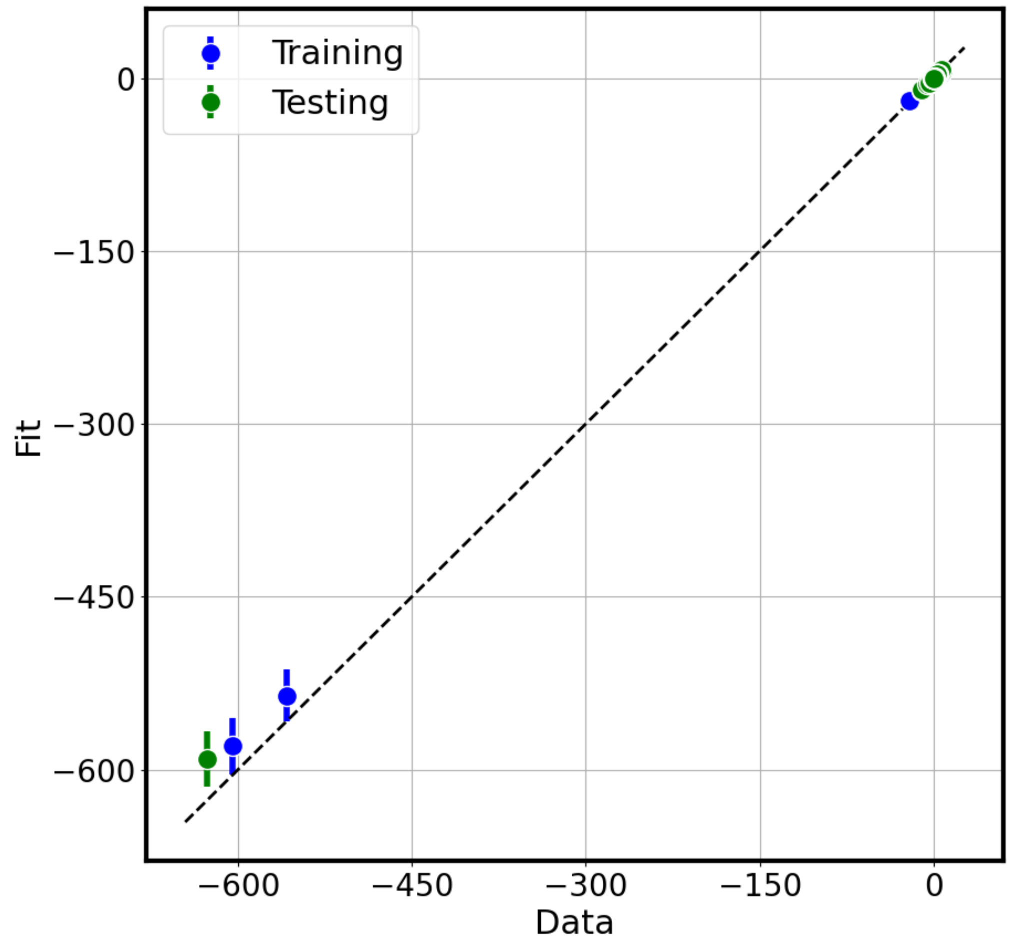
Uncertainty without model error

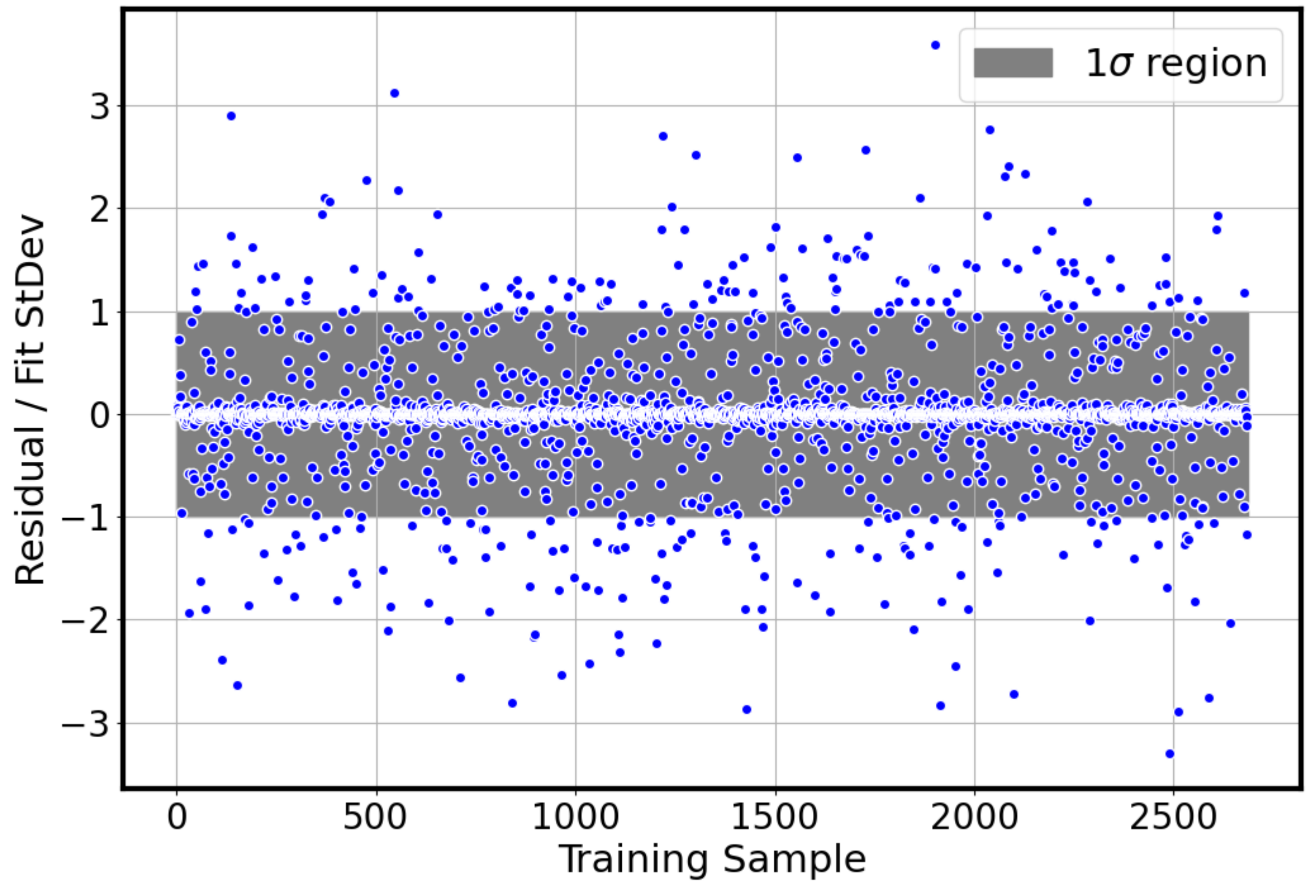


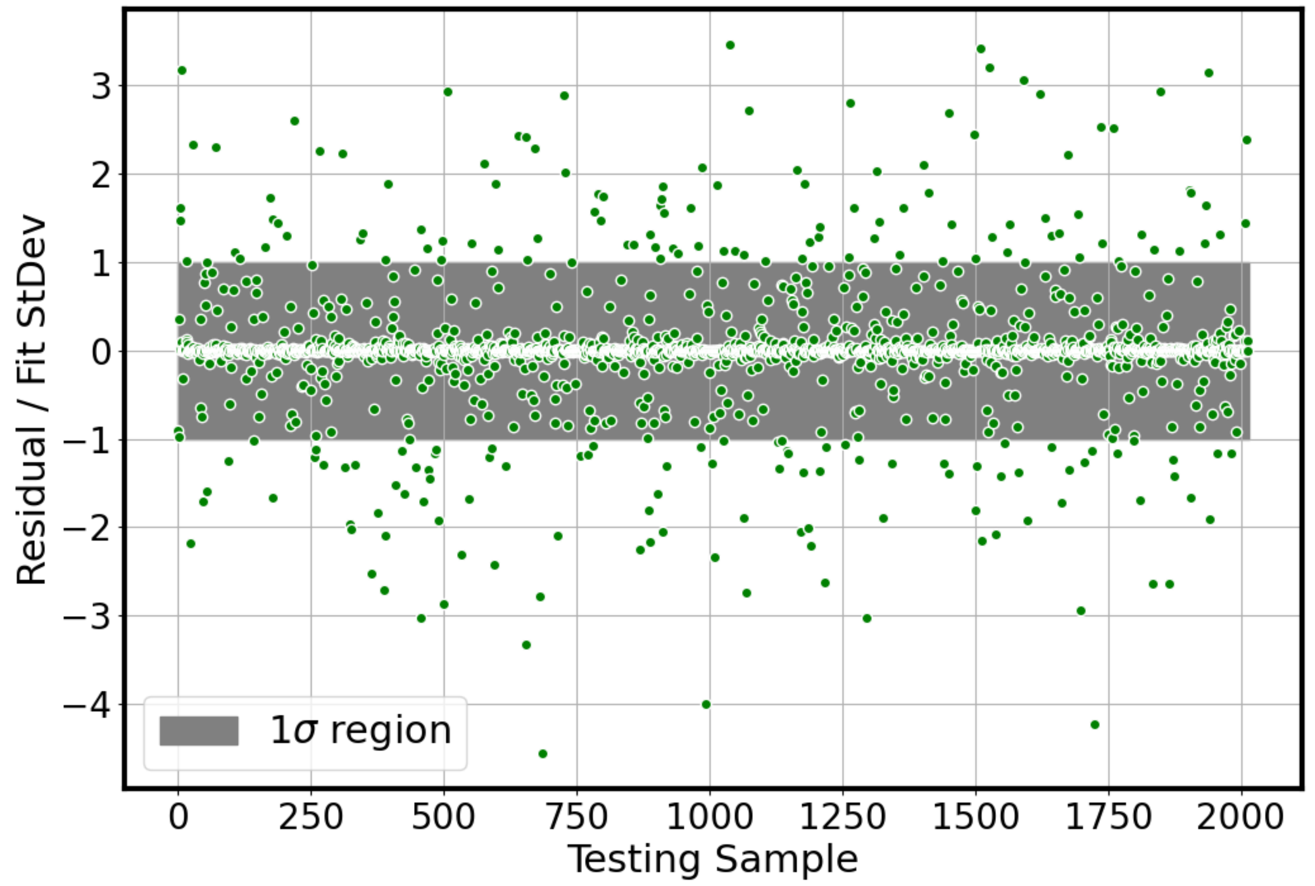




Uncertainty with model error







Several challenges/choices

- Embedding type: *e.g.*

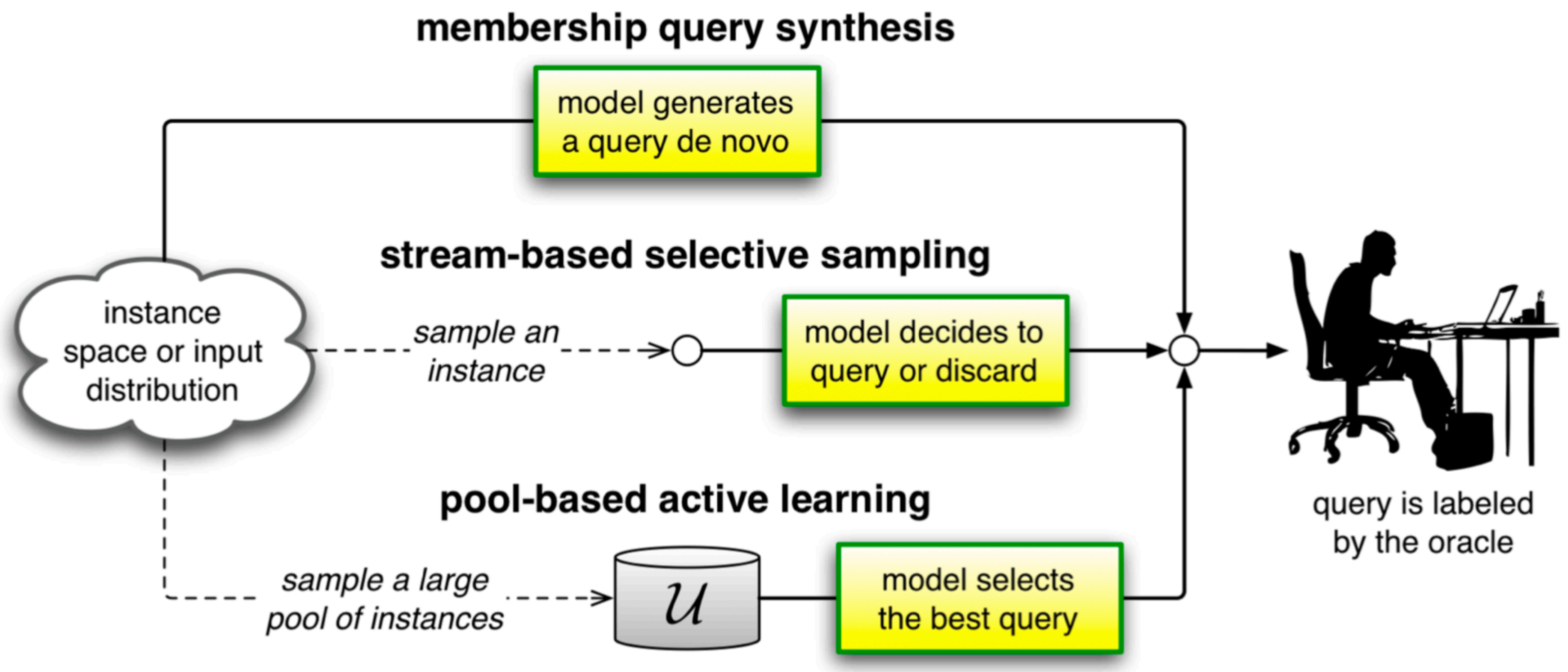
<u>additive</u> $y_i \approx \sum_{k=0}^{P} (c_k + d_k \xi_k)\, B_k(x)$ or <u>multiplicative</u> $y_i \approx \sum_{k=0}^{P} (c_k + c_k d_k \xi_k)\, B_k(x)$

- Degenerate (Gaussian) likelihoods: resort to approximate Bayesian computation (ABC) or independent (IID) assumptions

- Difficult posterior PDFs for MCMC, choice of priors for embedding parameters
- Which coefficients to embed the model error in?
- Connect predictive uncertainty and the residual error with an extrapolation metric
- Weighting between energies, forces and stresses
- Major challenge: data sizes are large, linear algebra chokes

Active Learning: motivation

- Choose the training samples adaptively
- Achieve greater accuracy with fewer training samples
- In conventional ML, minimize human effort of labeling images
- For us, minimize the number of *ab initio* QM calculations
- (aka optimal experimental/computational design)

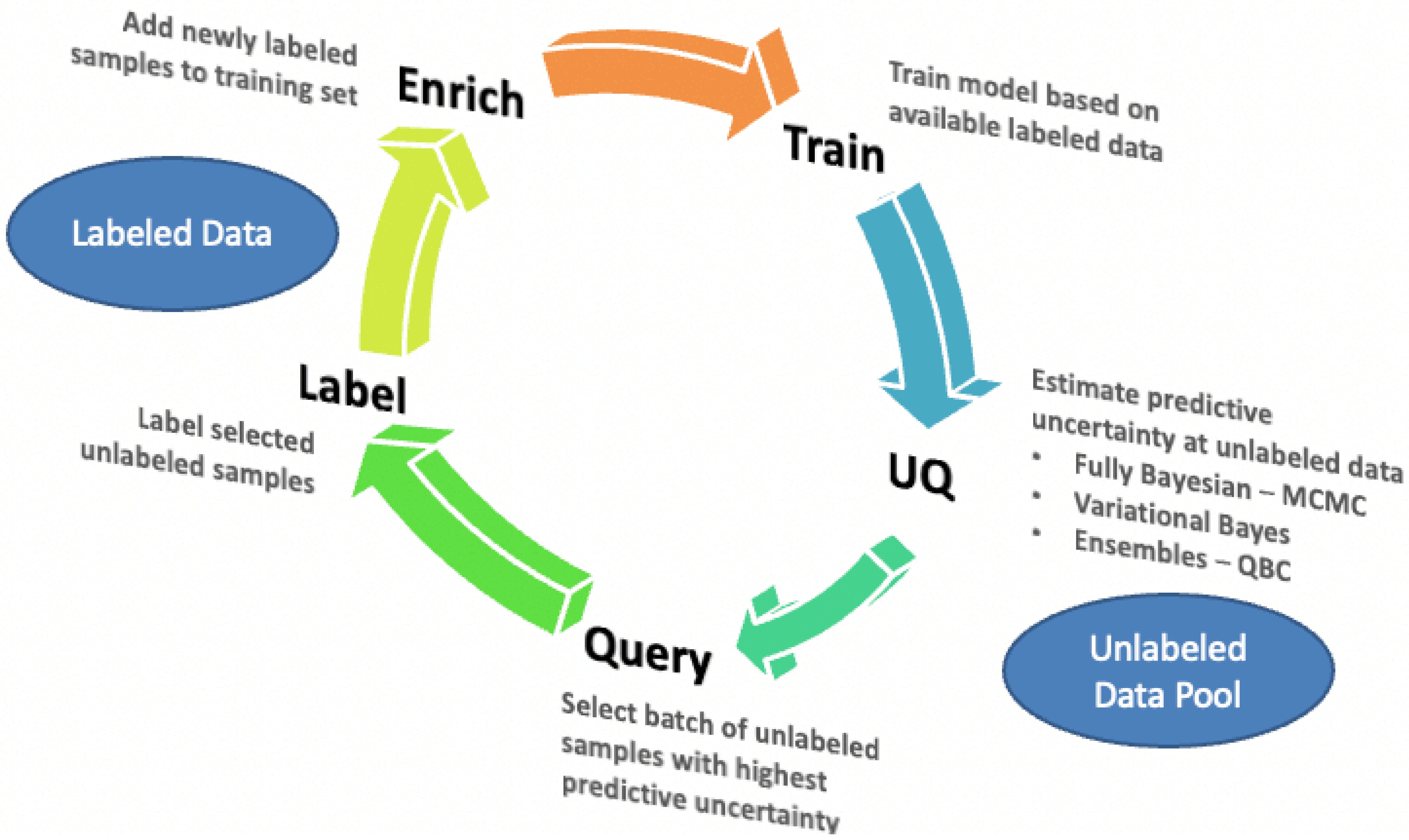


Detect and query extrapolative (high-uncertainty?) configurations on-the-fly and get QM data for those.

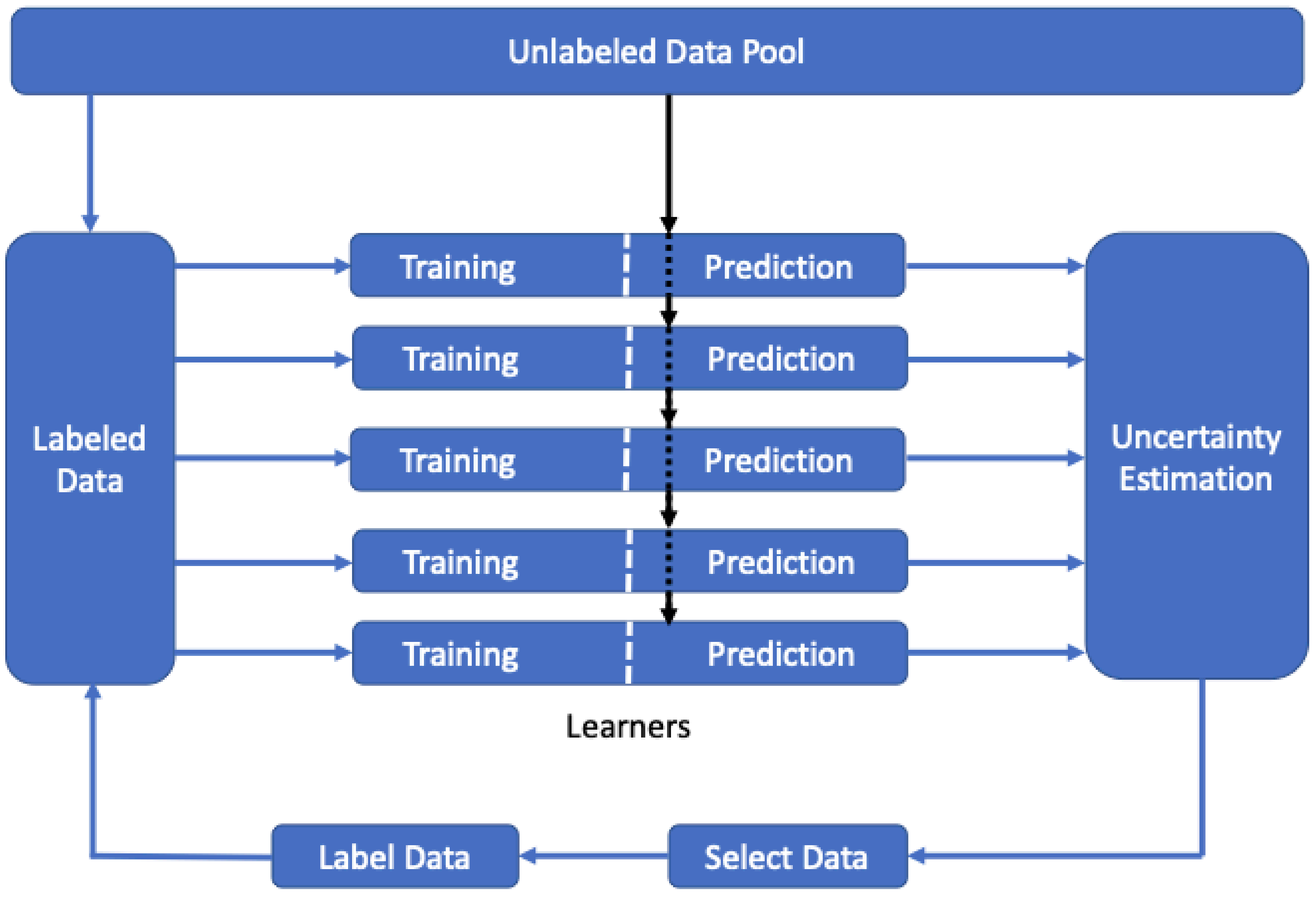
Key: query strategy, whether to query QM or not. If such decision can be made reliably, then one does not need to start with a very good training set.

[B. Settles, "Active learning literature survey", Computer Sciences Technical Report 1648, University of Wisconsin-Madison, 2009]

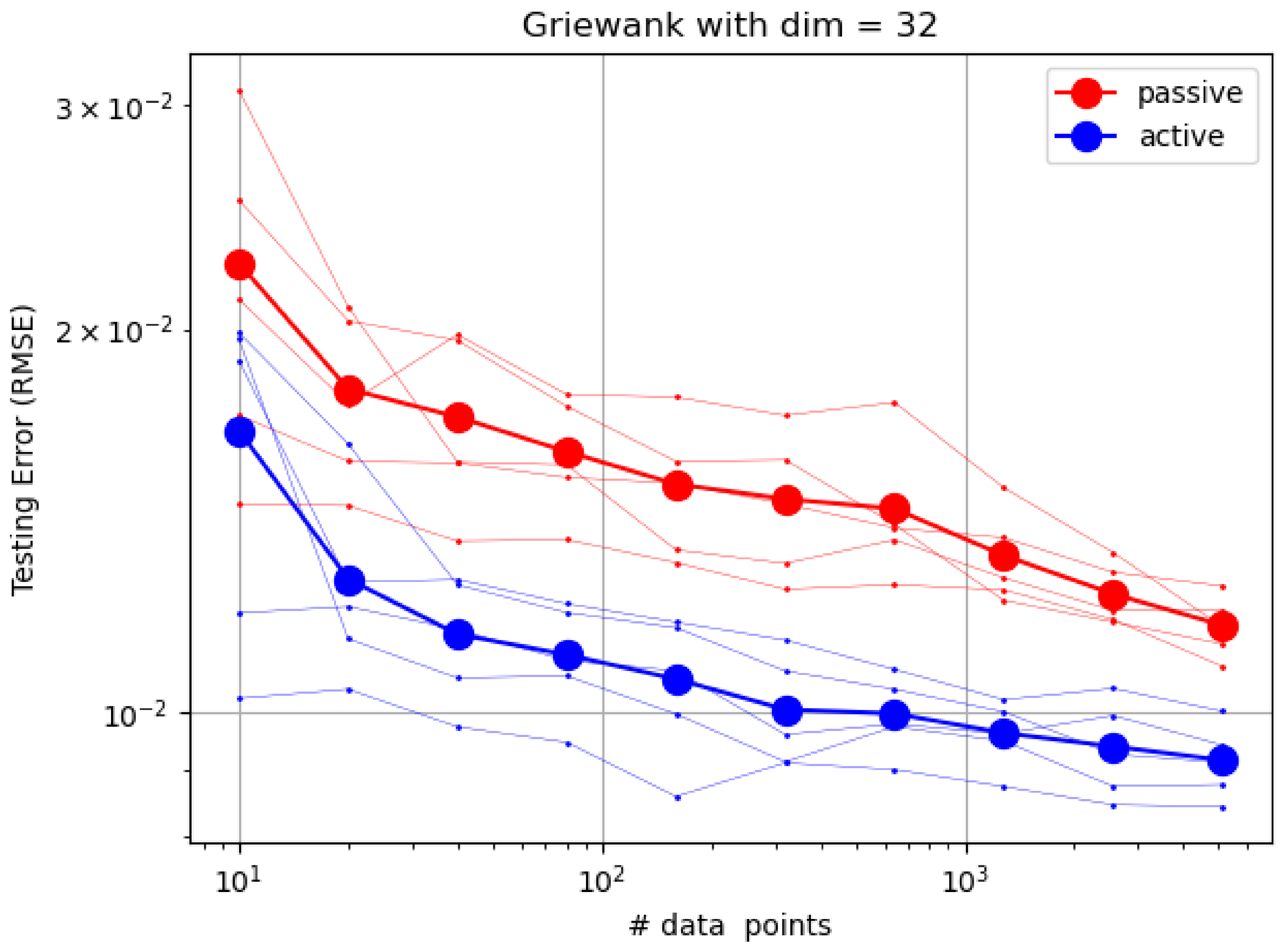
Active Learning Loop



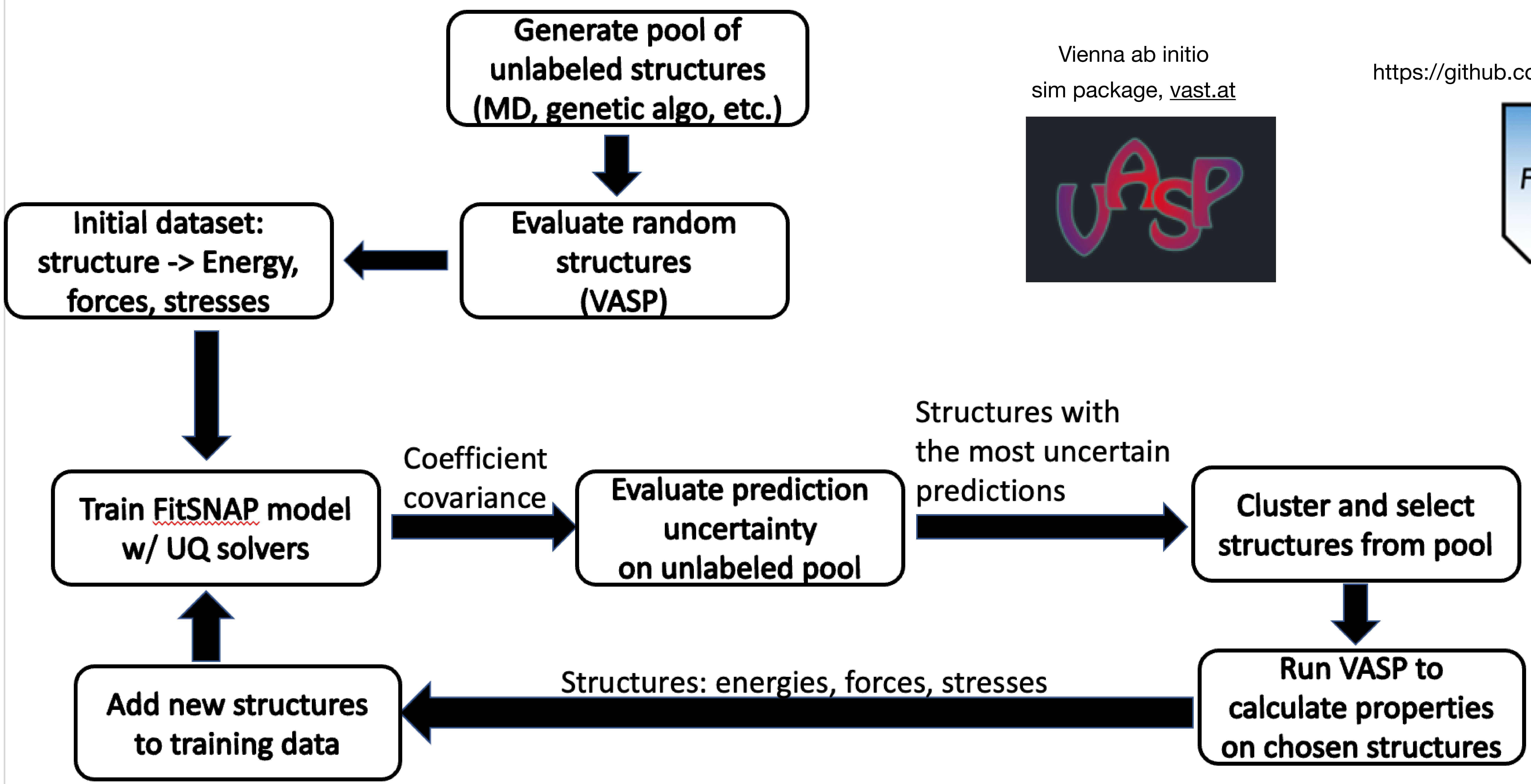
Active Learning: Query-by-Committee (QBC)



- Start with a training set of N points
- Launch K learners, each with fN training points (f=0.8)
- Evaluate the learners' performance at all points in the pool
- Select training points from the pool that correspond to the highest 'disagreement' and add them to the training set



Active Learning: current workflow



Vienna ab initio
sim package, vast.at

https://github.com/FitSNAP/FitSNAP

Summary

- Embedded **model error** for Bayesian inference of MLIAPs
  - Leads to data model with baked-in uncertainty
  - Meaningful model-error uncertainty capturing the true residual
  - Choices to make: priors, likelihoods, MCMC sampler, where to embed…
- Initiating a workflow for **active learning** via QBC
  - Anchored in uncertainty estimation, even if heuristic
  - Promising initial results
  - Choices to make: query strategy, UQ method, metric of ‘newness’…