Uncertainty Quantification and Active Learning in Atomistic Computations

Habib N. Najm

Sandia National Laboratories, Livermore, CA

Error Control in first-principles modelling

CECAM-HQ-EPFL, Lausanne, Switzerland
June 20-24, 2022

Acknowledgement

This work was supported by:

- DOE Office of Basic Energy Sciences, Div. of Chem. Sci., Geosci., & Biosci.
- DOE Office of Advanced Scientific Computing Research (ASCR) Scientific Advanced Computing (SciDAC) program
- DOE Office of Fusion Energy Sciences

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & En wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear contract DE-NA0003525. The views expressed in the article do not necessarily represent the views of the U.S. United States Government.

Outline

Scope

- Focus: potential energy surface (PES) / inter-atomic surrogate constructions
  - PES/IAP surrogate representation and model str
  - PES/IAP surrogate parameter estimation/calibra

  with emphasis on errors and uncertainties
- Multiple motivations
  - Uncertainty quantification (UQ) and global sensi molecular dynamics (MD) computations
  - Minimizing errors in fitted PES for downstream u
    - *e.g.* exploration, reaction rate computation
- UQ importance is recognized in IAP-based atomisti

  Frederiksen et al. 2004; Angelikopoulos et al. 2012; Duff et al. 2018; Longbottom and Wilkins 2019; Patrone et al. 2020; Vandermause et al. 2020; Zhou and Foiles 2017; Chen et al. 2021; Armstrong et al. 2013.
- We ignore errors/uncertainties in quantum chemistry (Q density functional theory (DFT) computations

  FYI: Wang et al. 2021; Yang et al. 2021; Prager et al. 2013; Klippenstein et al. 2011; Harding et al. 2011

Elements of a PES/IAP Surrogate Constr

- There are two key elements of any PES/IAP surrogat
  - A feature vector that summarizes system geome
    - For example $\boldsymbol{v}(\boldsymbol{x};\boldsymbol{\theta}) : \mathbb{R}^{3N} \to \mathbb{R}^n$
    - Where $\boldsymbol{\theta}$ is a parameter vector
  - A function that maps the feature vector to poten

$$E(\boldsymbol{x}) \approx f(\boldsymbol{v}(\boldsymbol{x};\boldsymbol{\theta});\boldsymbol{w})$$

where $\boldsymbol{w}$ is a parameter vector

- The functional forms of $\boldsymbol{v}()$ and $f()$ are important modeling ch
- Constructions should natively satisfy
  - translational, rotational, and permutational invariances
  - Permutational invariance: swapping two same-t does not alter the system potential energy

Examples of PES/IAP constructions

- Embedded Atom Model (EAM) Daw & Baskes, 1984
  - $E_i = \frac{1}{2}\sum_{j\neq i}\phi_{\alpha\beta}\left(r_{ij}\right) + F_\alpha\left(\sum_{j\neq i}\rho_{\alpha\beta}\left(r_{ij}\right)\right)$
- Symmetry Function descriptors Behler & Parrinello, 2007
  - Geometry descriptors used with Neural Network
  - Atomic Environment Vector (AEV) construction Smith et al, 2017
- Gaussian Approximation Potential (GAP) Bartók et al, 2010
  - Geom. descriptor, neighbor density bispectrum,
- Coulomb matrix descriptor, geometry and molecular Rupp et al, 2012
- Spectral Neighbor analysis potential (SNAP) Thompson et al, 2015
  - Similar to GAP, but with linear regression
- Atomic Cluster Expansion (ACE) Drautz, 2019
  - Generalizes/subsumes many earlier constructio
- 3D Graph Convolutional Network (3DGCN) descript Cho & Choi, 2019
  - Geometry representation in terms of molecular
- Generalized deep metric learning with GNNs Nat et al, 2020

Our PES surrogate requirements

- Translational and rotational invariance
  - Rely on internal coordinates:
    - Distances between pairs of atoms
    - Angles formed among sets of 3 atoms
- Descriptor size independent of number of molecule
  - An atom-centric feature vector construction
  - Represent local geometry around each atom
- Permutational invariance
  - Additive superposition of atomic contributions t
- Sufficient complexity to fit data from different atom systems in a range of geometric conformations
- Generalizability
  - Controlled approximation accuracy outside the t

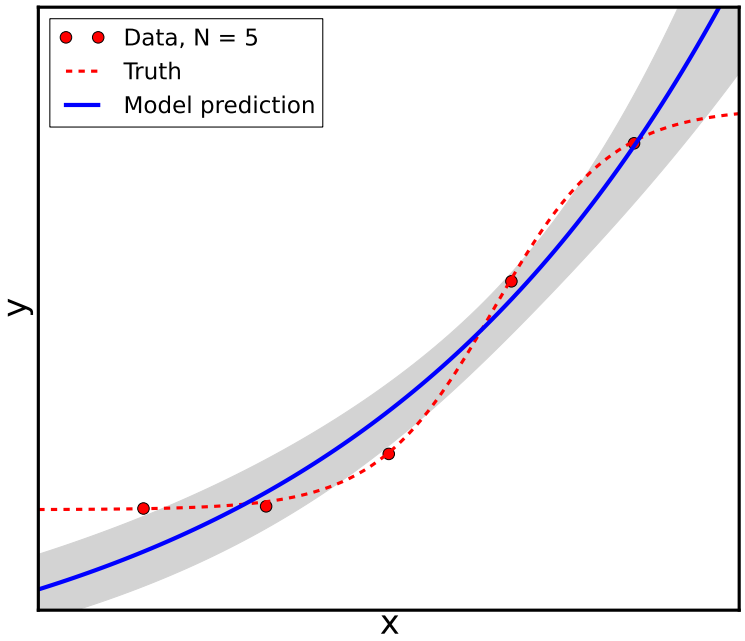
PES/IAP Surrogate Calibration

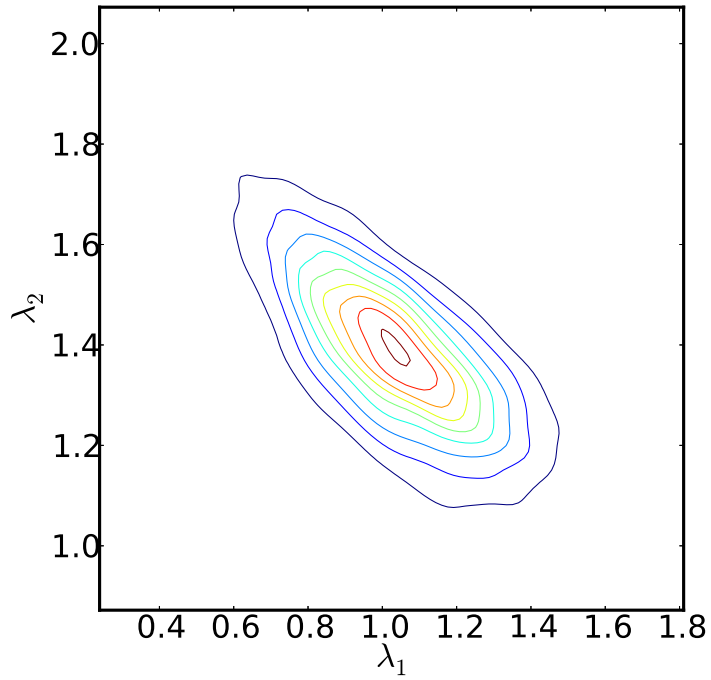
- Least squares regression, with some means of regula
  dominant means of PES/IAP surrogate calibration
- Given large number of parameters, and limitations o
  data sizes, one expects some uncertainty in fitted m
- Bayesian methods provide robust means of estim at
  uncertainties, as well as model comparisons/selecti
- In a computational setting, use Markov Chain Monte
  to sample from the Bayesian posterior
  - MCMC is feasible in high-dimensions when deali
    that are linear in parameters
  - It has severe computational limitations in high-d
    nonlinear models
  - A range of heuristic/approximate methods exist i
    challenging setting

Dealing with Model Error in UQ

- All models are wrong in principle
- Models of physical systems rely on
  - Presumed theoretical framework
  - Mathematical formulation
- Practical models of complex physical systems rely o
  - Simplifying assumptions
  - Numerical discretization of governing equations
  - Computational software & hardware
- model error is frequently non-negligible
- Estimating model error is useful for
  - model comparison & validation
  - model improvement & scientific discovery
  - reliable computational predictions

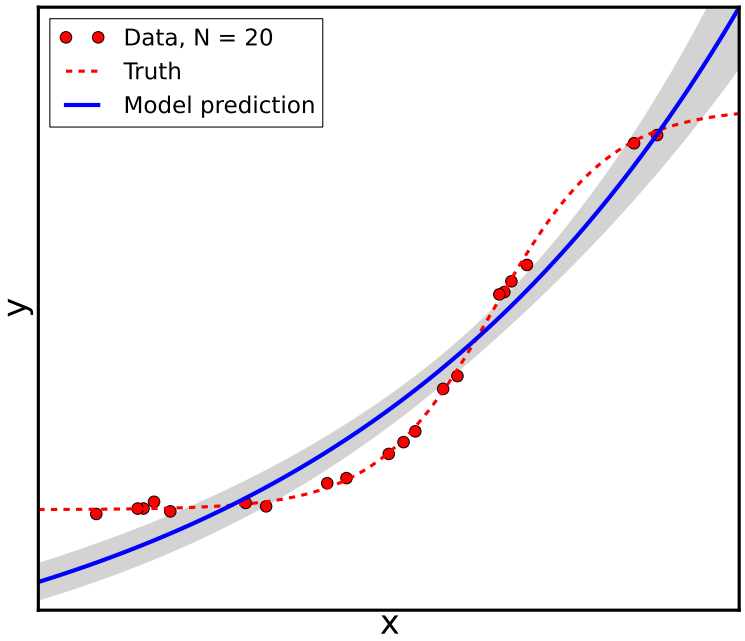
Challenges with Model Calibration due t

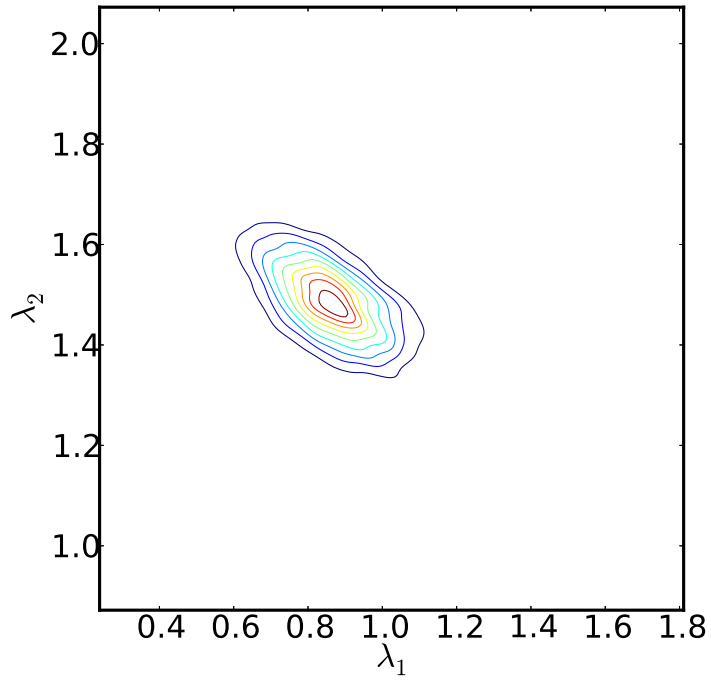




- Conventional parameter estimation: $y_{data} = f(x, \lambda) + \epsilon_d$
- Additional data results in reduced parameteric poste
- One gets more confident about predictions with the
- Predictive uncertainty in calibrated model has no uti
- Ignoring model error leads to irrelevant predictive er

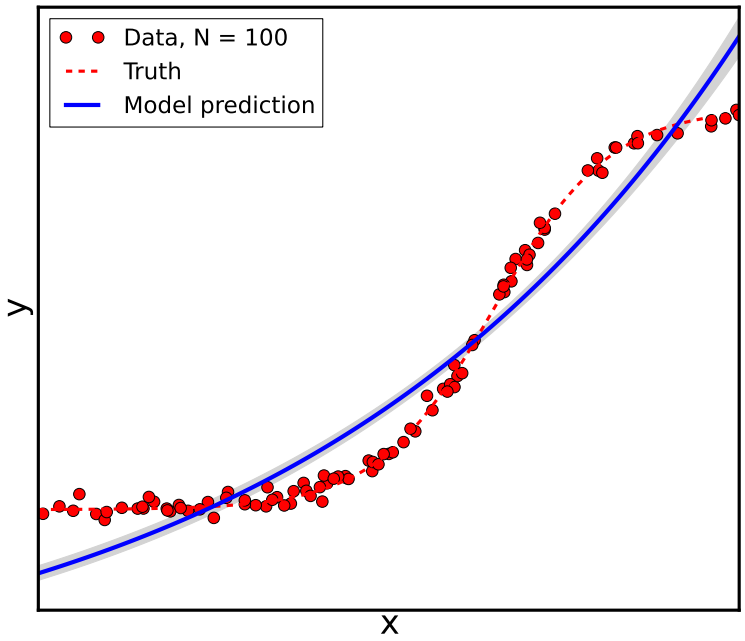
Challenges with Model Calibration due t

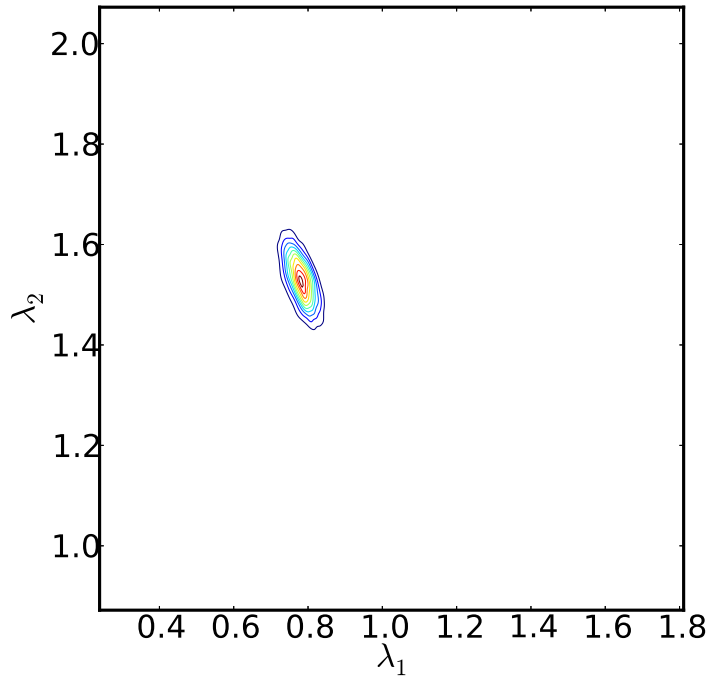




- Conventional parameter estimation: $y_{data} = f(x, \lambda) + \epsilon_d$
- Additional data results in reduced parameteric poste
- One gets more confident about predictions with the
- Predictive uncertainty in calibrated model has no uti
- Ignoring model error leads to irrelevant predictive er

Challenges with Model Calibration due t





- Conventional parameter estimation: $y_{data} = f(x, \lambda) + \epsilon_d$
- Additional data results in reduced parameteric poste
- One gets more confident about predictions with the
- Predictive uncertainty in calibrated model has no uti
- Ignoring model error leads to irrelevant predictive er

Statistical modeling of model error

Error framework:

Measurements: $y_{data} = y_{truth} + \epsilon_d$

Model predictions: $y_{truth} = y_{model} + \epsilon_m$

Thus: $y_{data} = y_{model} + \epsilon_m + \epsilon_d$

Error modeling – example

Model: $y_{model} = f(x, \lambda)$

Data Error: $\epsilon_d \sim N(0, \sigma^2)$

Model Error: $\epsilon_m \sim GP(\mu(x), C(x, x'))$

Model calibration:

Estimate model parameters λ along with those of ϵ_m, ϵ_d

Kennedy & O'Hagan 2001, Higdon et al. 2004, Bayarri et al. 2007

Challenges – Physical Models

- Arbitrary choice of statistical model (i.e. GP) spatial structure does not take the physical model into acct
  - Potential violation of implicit constraints in physi
  - *e.g.* incompressible flow: $\nabla \cdot u = 0$
- Difficulty in disambiguation of model & data error
- Calibration of model error on measured observable quality of other model predictions

Model Error Embedding

- Embed model error in specific submodel phenomen
  - a modified transport or constitutive law
  - a modified formulation for a material property

Berliner 2003

Sargsyan et al. 2019; Hakim 2018, Huan et al. 2018; Sargsyan 2015

- Pros:
  - Allows placement of model error term in location modeling assumptions and approximations are m
    - as a correction or high-order term
    - as a possible alternate phenomenology
  - explore if it can explain discrepancy on observab
  - naturally preserves model structure and associat
- Cons:
  - complex likelihood for general nonlinear $f(x, \lambda, \epsilon_m)$, $p(y|\lambda)$

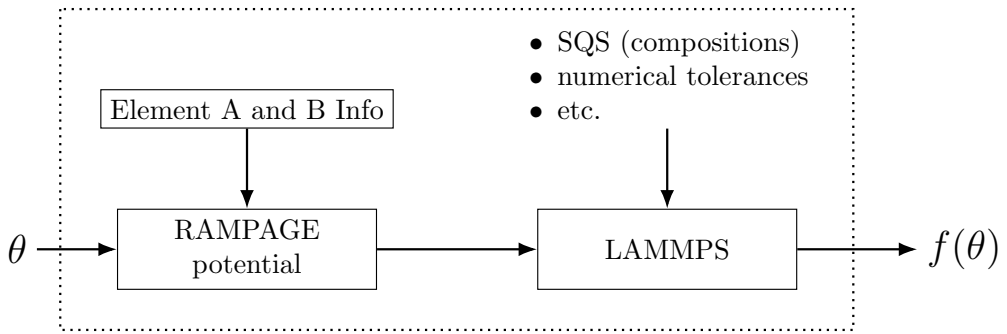
Case Study: Bayesian Calibration of IAPs

Work with:

- Arun Hegde and Cosmin Safta (Sandia Nat. Labs.)
- Elan Weiss and Wolfgang Windl (Ohio State U.)

Goals:

- Focus on Embedded Atom Model (EAM) potential fo
  - (Au,Cu)
- Bayesian calibration of EAM potential with DFT data
- Forward propagation of IAP uncertainty in MD simul LAMMPS



EAM potential

- RAMPAGE/EAM potentials for binary alloys (Ward 2
- Two elements (types A, B)

Given atom i of type $\alpha \in \{A, B\}$, and atom j of type $\beta \in \{A, B\}$, we have

$$
\begin{aligned}
E_i &= \frac{1}{2} \sum_{j \neq i} V_{\alpha,\beta}(r_{ij}) + F_\alpha \left(\sum_{j \neq i} \rho_{\alpha,\beta}(\rho_{ij}) \right) \\
\rho_{BA}(r) &= r^6 \left(\exp(-S_A r) + 2^9 \exp(-2 S_A r) \right) \\
\rho_{AB}(r) &= r^6 \left(\exp(-S_B r) + 2^9 \exp(-2 S_B r) \right) \\
V_{AB}(r) &= D \left(\exp\left(-2\alpha \left(r - r_{eq}\right)\right) - 2 \exp\left(-\alpha \left(r - r_{eq}\right)\right) \right)
\end{aligned}
$$

with parameters $(r_{eq}, D, \alpha, S_A, S_B)$.

Calibration problem setup

- Predictive model: $y_i = f_i(x;\theta)$ where

$$\begin{aligned} x &\in \{\text{discrete set of compositions}\} \\ \text{QoI}_i &\in \{\text{lattice constant, } C_{11}, C_{12}, C_{44}, \text{bulk modulus, mixing enthalpy}\} \\ \theta &= (r_{eq}, D, \alpha, S_A, S_B) \end{aligned}$$

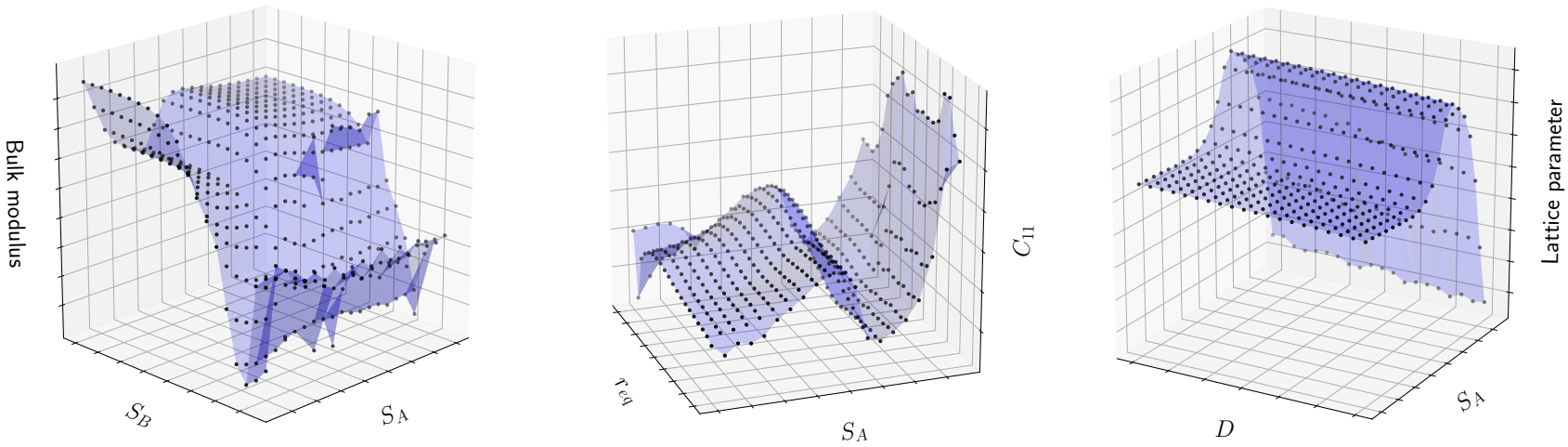
- Data model:

$$\begin{aligned} y_i(x) &= f_i(x;\theta) + \epsilon_i(x) \\ \epsilon_i(x) &\sim \begin{cases} \mathcal{N}\left(0, \sigma^2 f_i(x;\theta)^2 + \tau^2\right) & \text{for QoI} = \text{mixing enthalpy} \\ \mathcal{N}\left(0, \sigma^2 f_i(x;\theta)^2\right) & \text{otherwise} \end{cases} \end{aligned}$$

- 102 calibration targets: 6 QoIs at each of 17 compositions
- DFT data generated for each target quantity
- We use uniform priors on $(r_{eq}, D, \alpha, S_A, S_B, \ln\sigma, \ln\tau)$

Model Surrogates

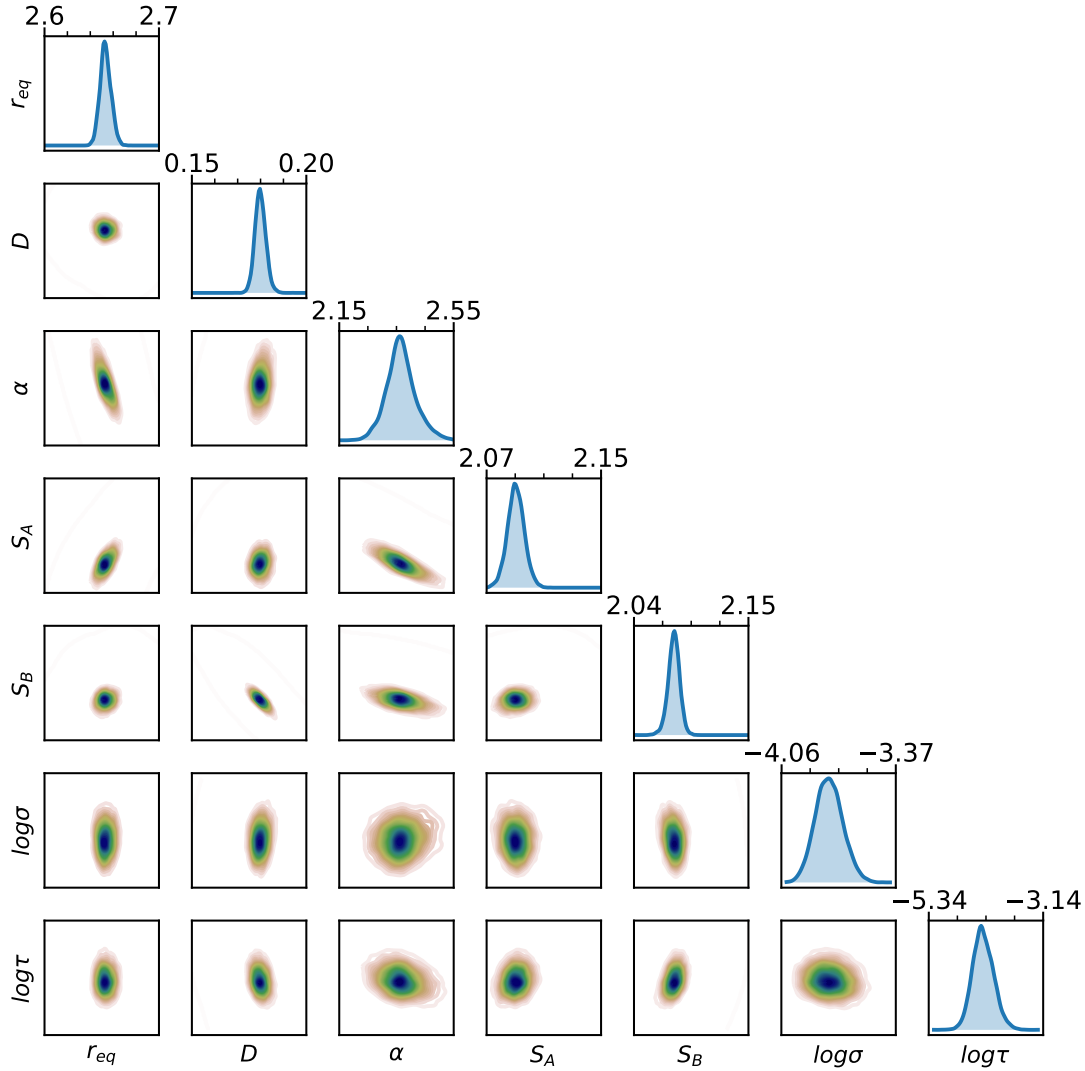
- LAMMPS expense necessitates use of surrogates fo
- Model outputs exhibit complex behavior, and patho



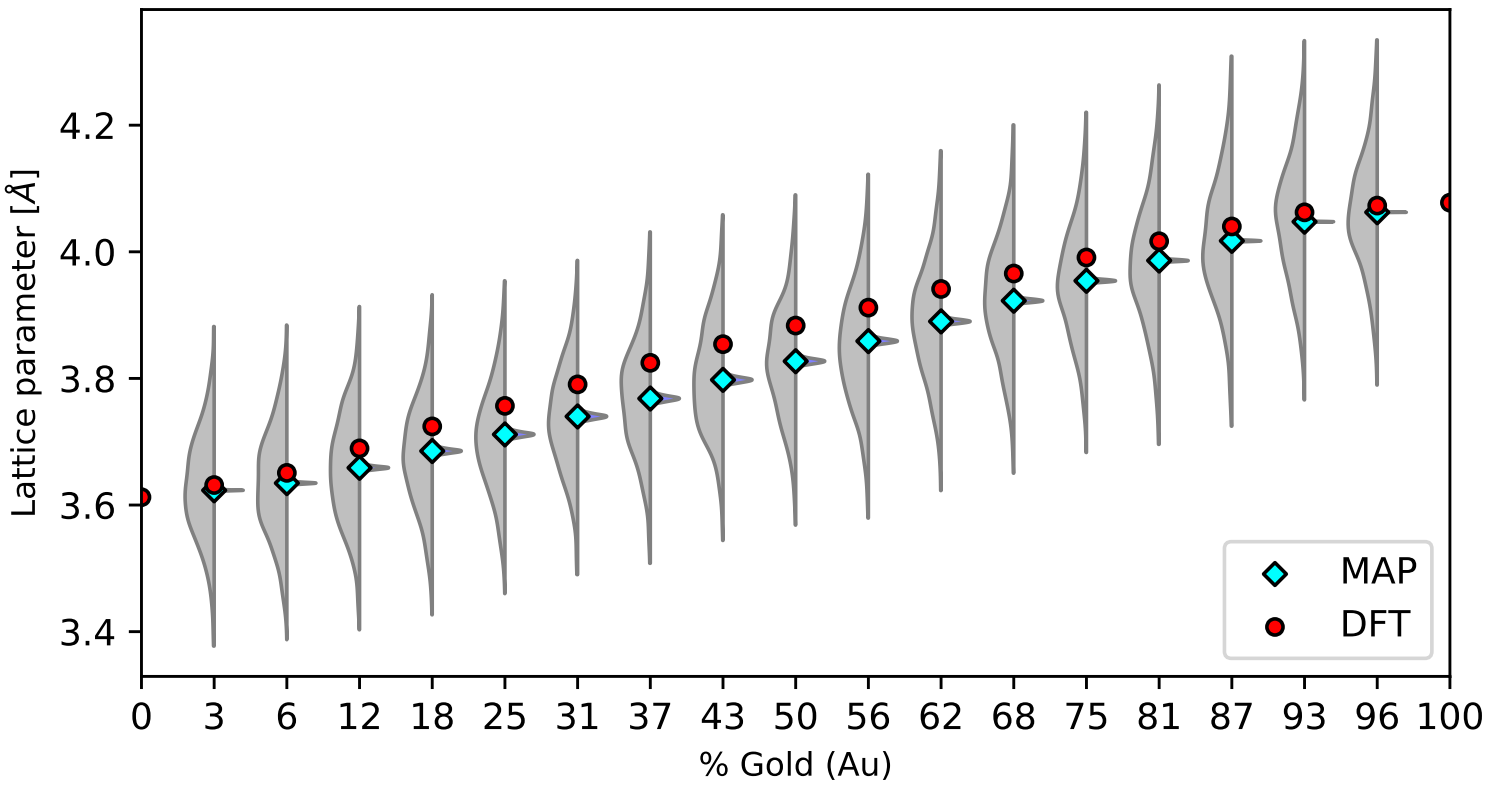
- We build Gaussian process surrogates for each of the
  - superior accuracy relative to Legendre polynomi
- Using randomly sampled train/test data, we adaptiv surrogate hypercube to ensure
  - target surrogate accuracy
  - surrogate hypercube contains posterior samples

Posterior on model parameters

- Marginal 1D poste
  are nearly Gaussi
- Posterior parame
  uncertainties are
  with COV 2%
- Marginal 2D post
  reveal pairwise
  correlations on v
  parameters



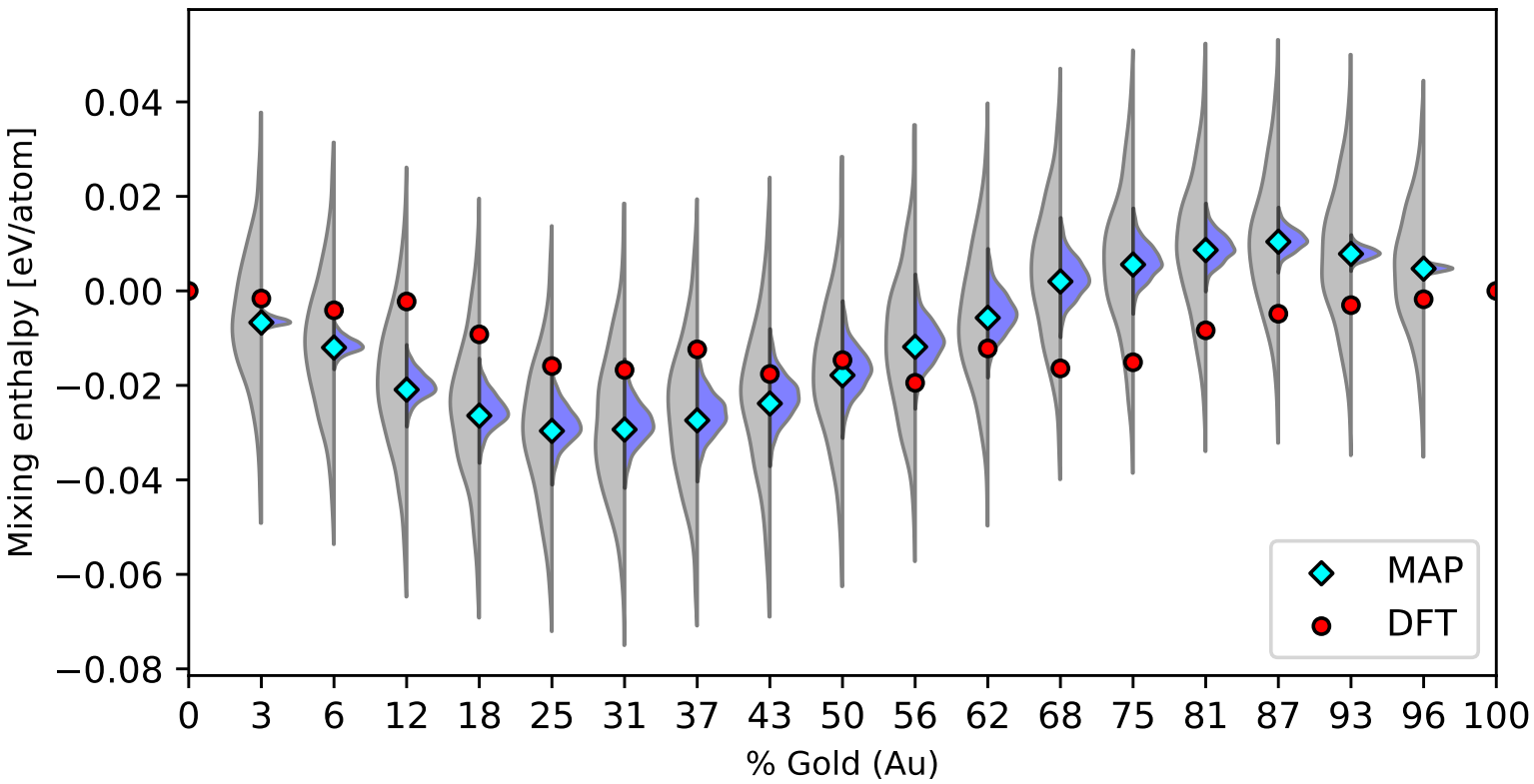
Posterior on model parameters



Push-forward posterior (PFP, right) and Posterior predictive (PP, le

- Significant model errors observed with various com
- PFP width small w.r.t. discrepancy between MAP pr
- PP does encompass these discrepancies on average

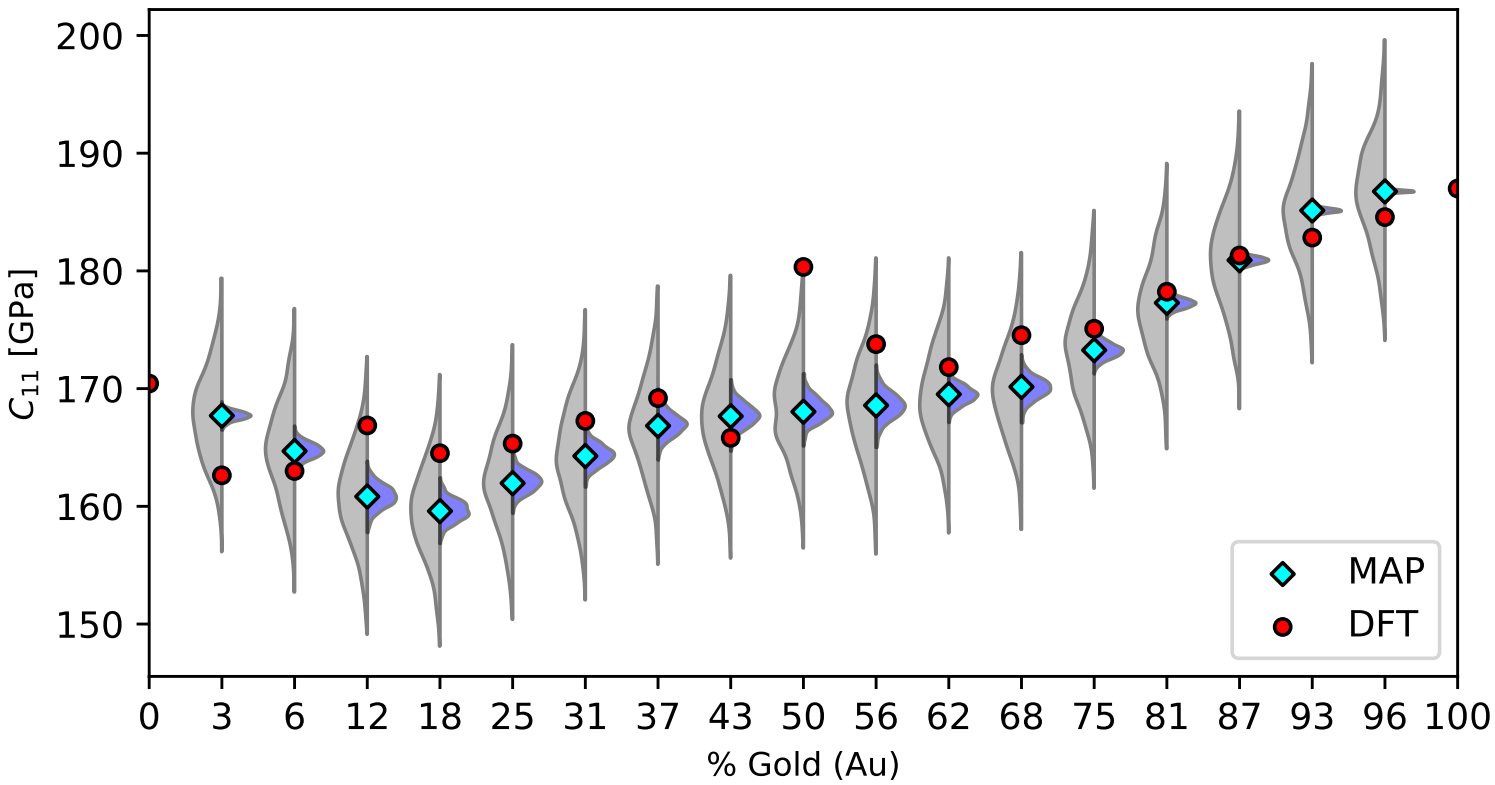
Posterior on model parameters



Push-forward posterior (PFP, right) and Posterior predictive (PP, le

- Significant model errors observed with various com
- PFP width small w.r.t. discrepancy between MAP pre
- PP does encompass these discrepancies on average

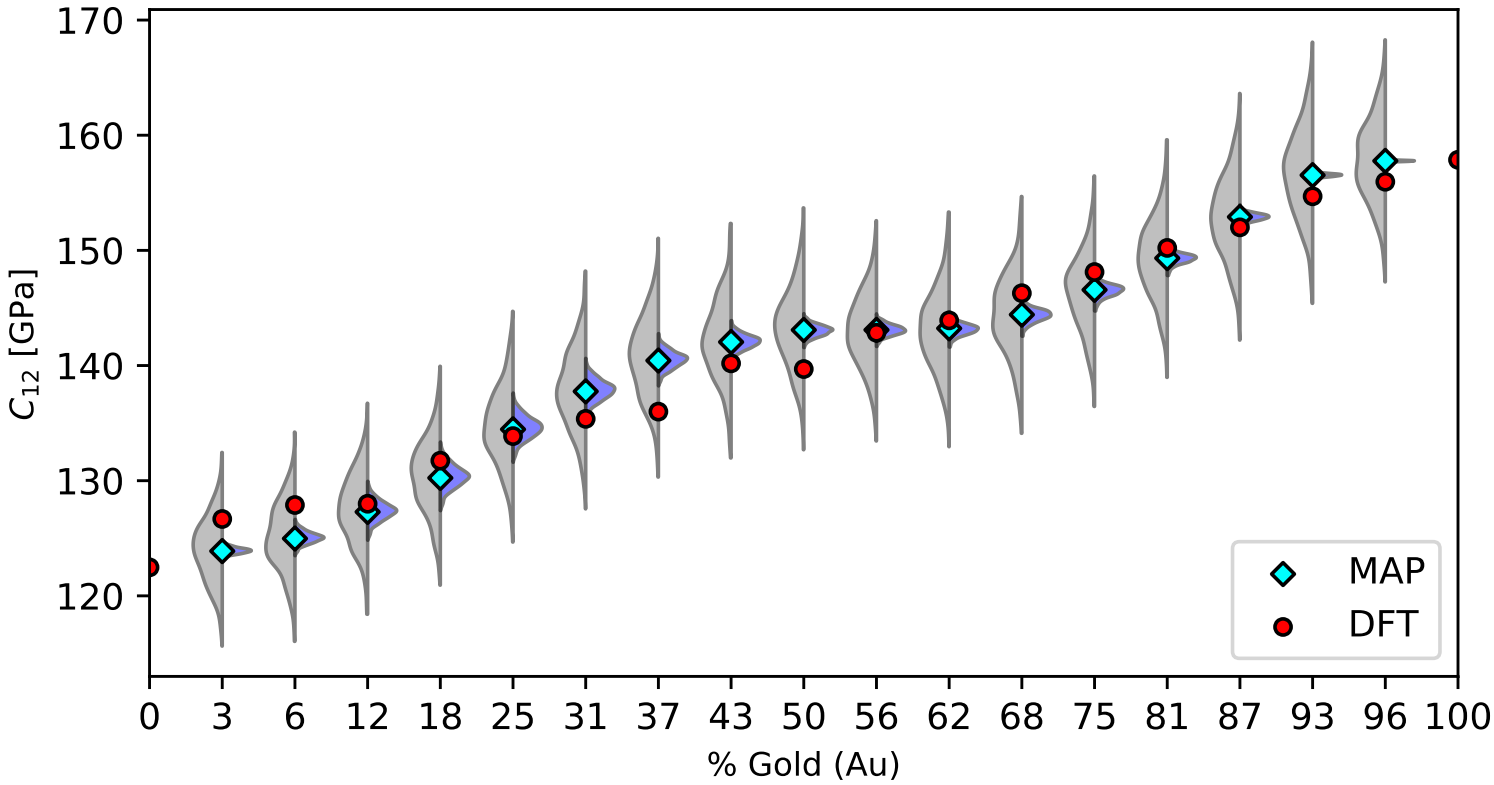
Posterior on model parameters



Push-forward posterior (PFP, right) and Posterior predictive (PP, le

- Significant model errors observed with various com
- PFP width small w.r.t. discrepancy between MAP pre
- PP does encompass these discrepancies on average

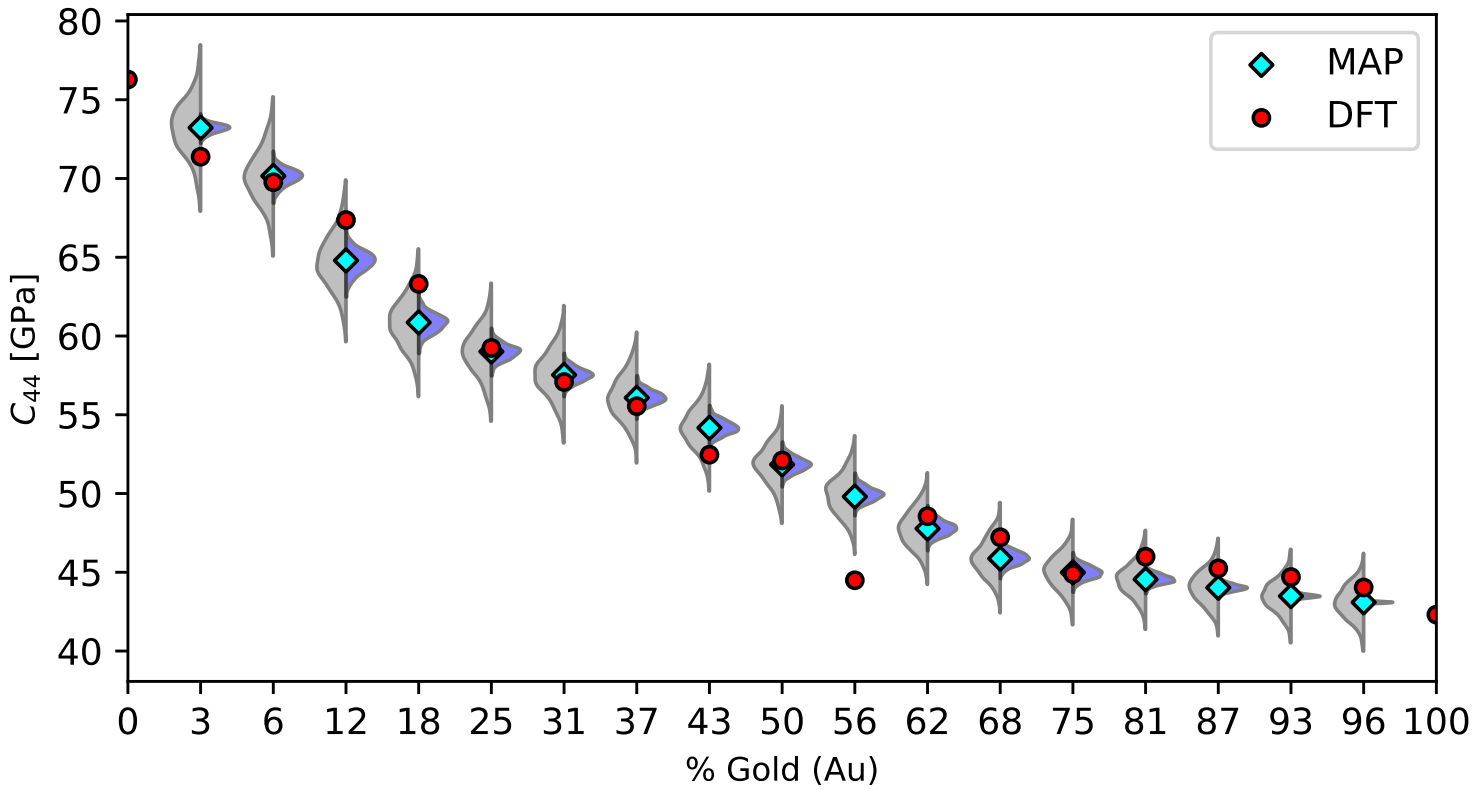
Posterior on model parameters



Push-forward posterior (PFP, right) and Posterior predictive (PP, le

- Significant model errors observed with various com
- PFP width small w.r.t. discrepancy between MAP pre
- PP does encompass these discrepancies on average

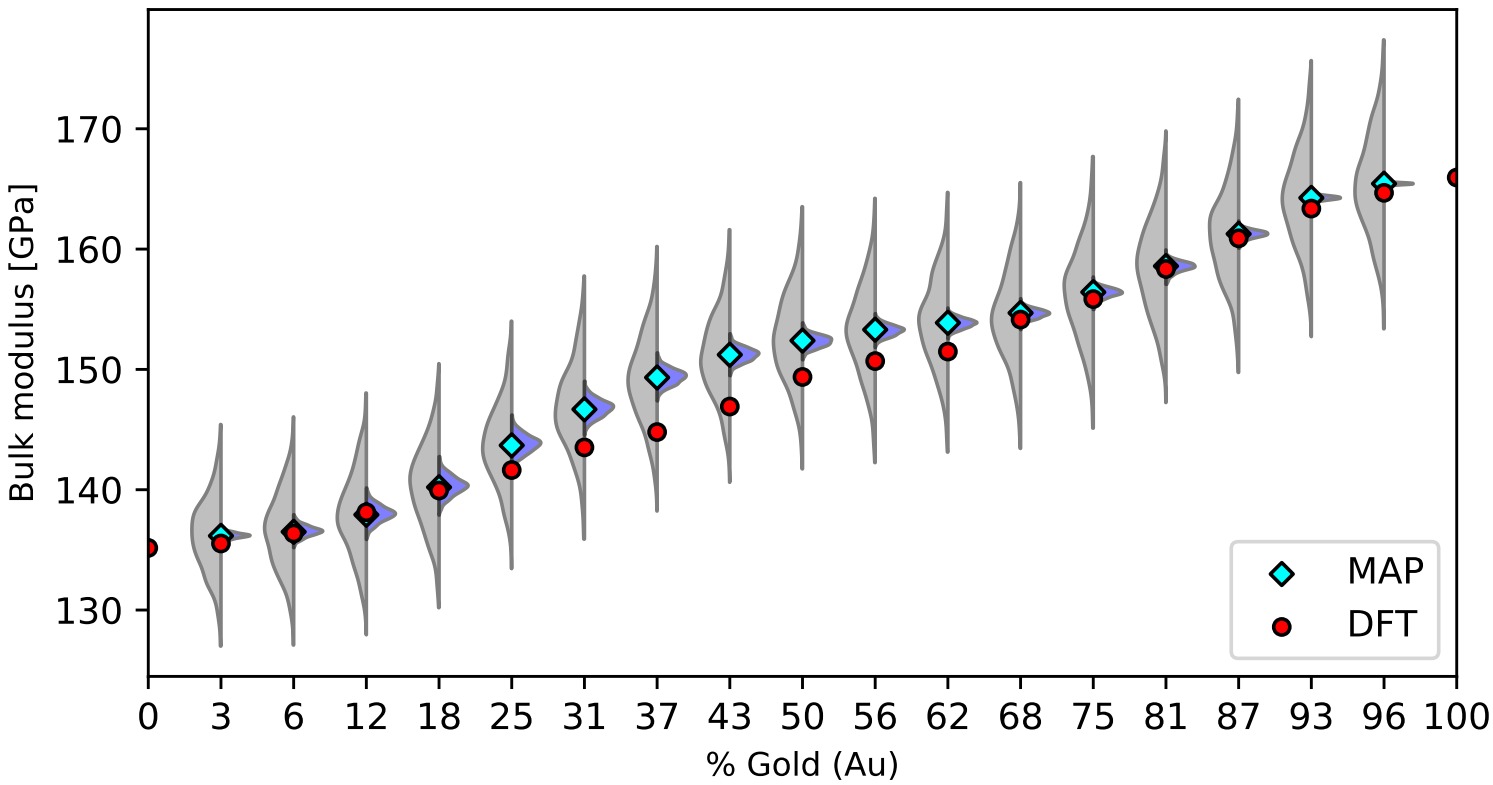
Posterior on model parameters



Push-forward posterior (PFP, right) and Posterior predictive (PP, le

- Significant model errors observed with various com
- PFP width small w.r.t. discrepancy between MAP pre
- PP does encompass these discrepancies on average

Posterior on model parameters



Push-forward posterior (PFP, right) and Posterior predictive (PP, le

- Significant model errors observed with various com
- PFP width small w.r.t. discrepancy between MAP pre
- PP does encompass these discrepancies on average

Conclusions – Binary Alloy Study

- As may be expected, the simplicity of the EAM pote
  to sigificant discrepancies in the MAP model predict
- These discrepancies are captured on average by the
  predictive (PP) across all QoIs/compositions
- At the same time, the PP uncertainty is much larger t
  observed discrepancy
- Improvements to this simple additive model-error c
  using QoI-specific GP corrections (ala Kennedy & O'
  reduce this mistach
- Further improvements, using embedded model erro
  will also improve the PP on other QoIs, e.g. forces

- Data curation as well as improved automated conve
  would help reduce data pathologies, improve surrog
  enhance overall fidelity of the Bayesian calibration

Case Study: PES for Hydrocarbon Molec

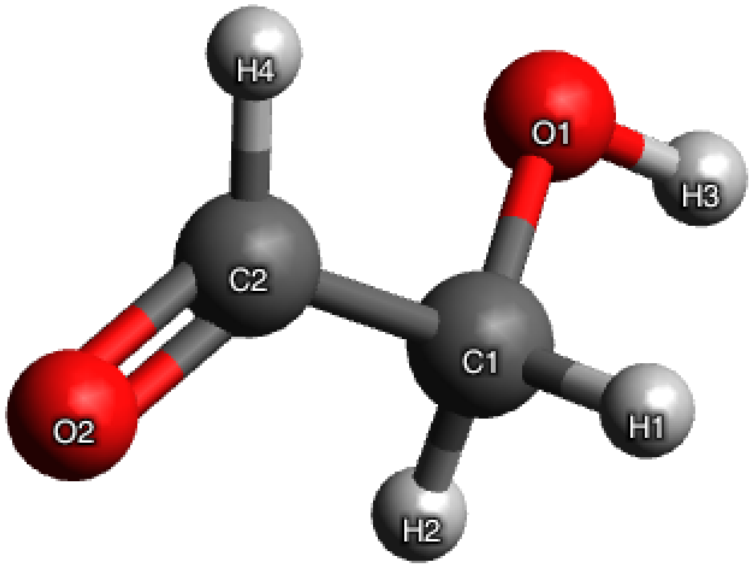
Work with:

Yoona Yang[1], Carles Martí[1], Judit Zádor[1], and Michael Eldred[2]
Sandia National Laboratories
Livermore[1], CA & Albuquerque[2], NM

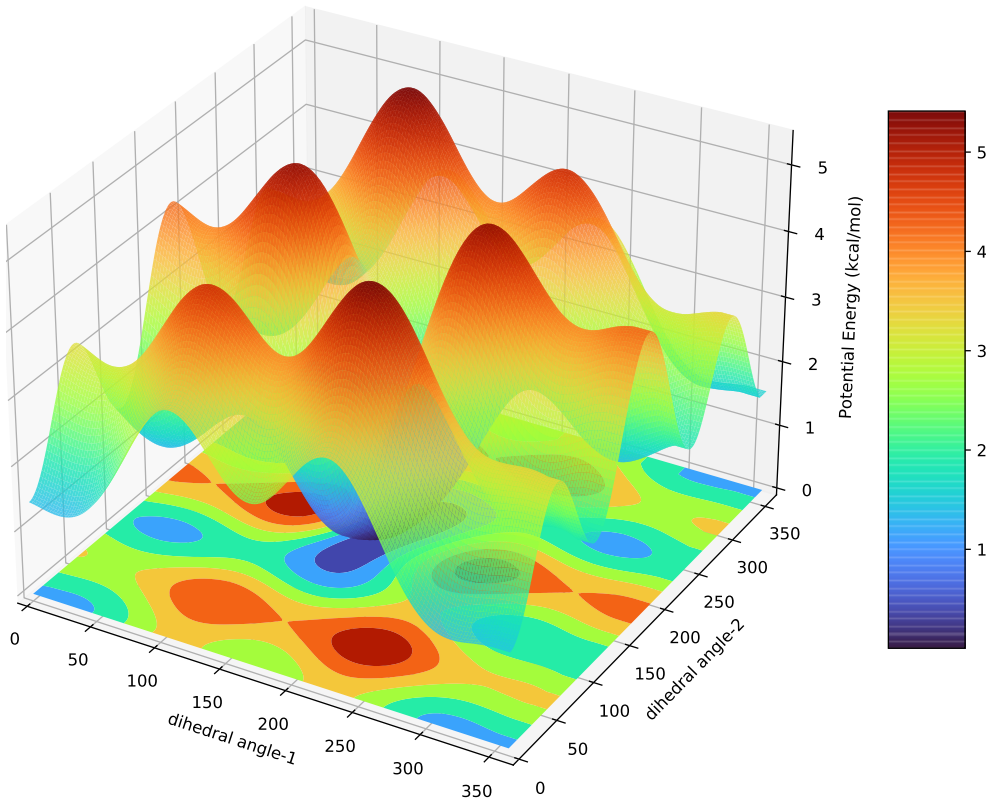
Goals:

- Focus on neural network (NN) PES with feature vect
  Behler & Parrinello symmetry functions
- Atomic environment vector (AEV) construction
- Active learning using Query by Committee (QBC)
- Force training
- Multifidelity training

PES for Glycolaldehyde



Molecular structure

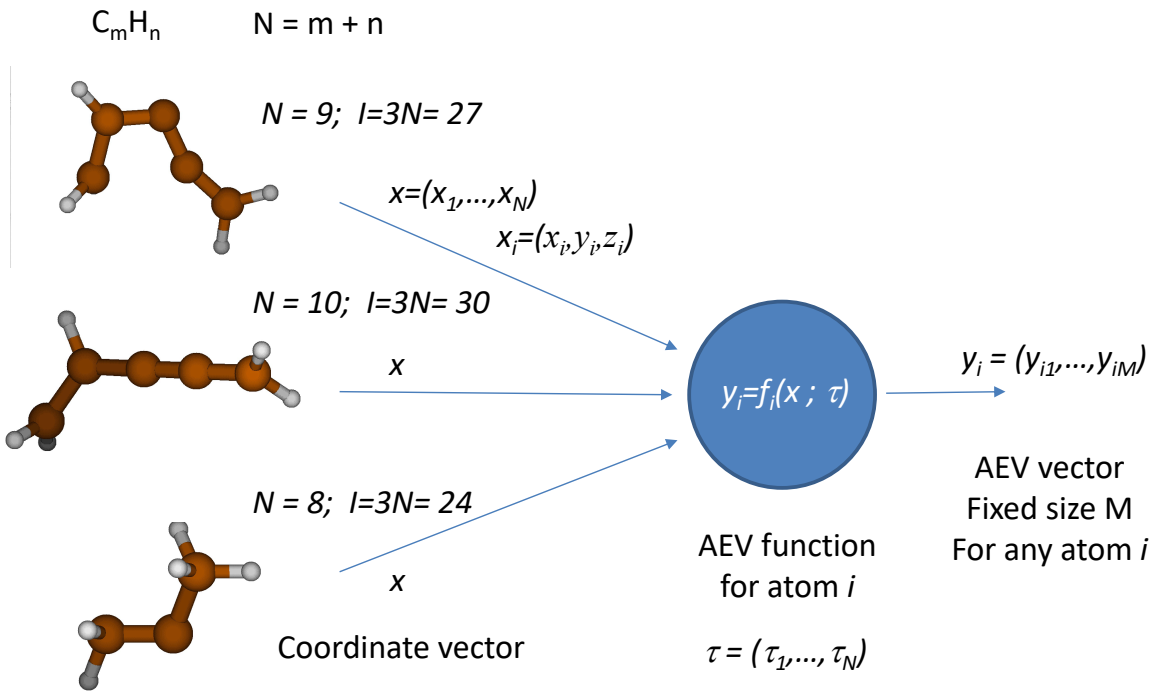


PES of glycolaldehyde with variation in the angles (O2-C2-C1-O1,C2-C1-O1-H3

Symmetry Function AEV construction

The neighborhood geometry around each atom is repre
fixed length atomic environment vector (AEV) descript

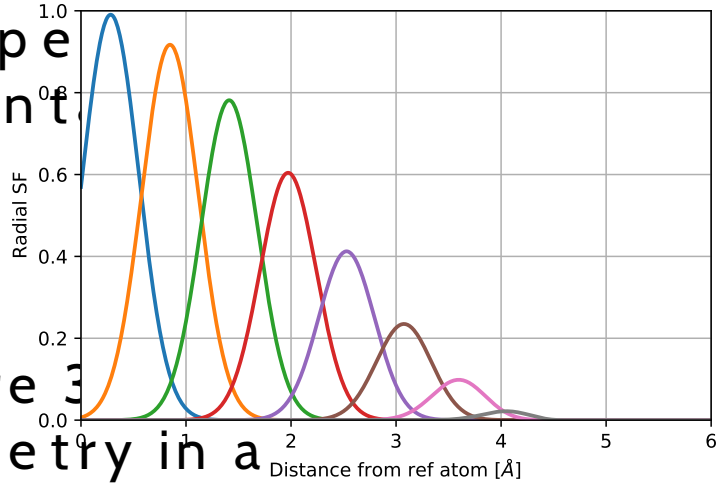
- AEV length is independent of the number of atoms i
- It's defined by specifying:
  - The set of allowed atomic types, e.g. {C, H, O}
  - The specified symmetry functions



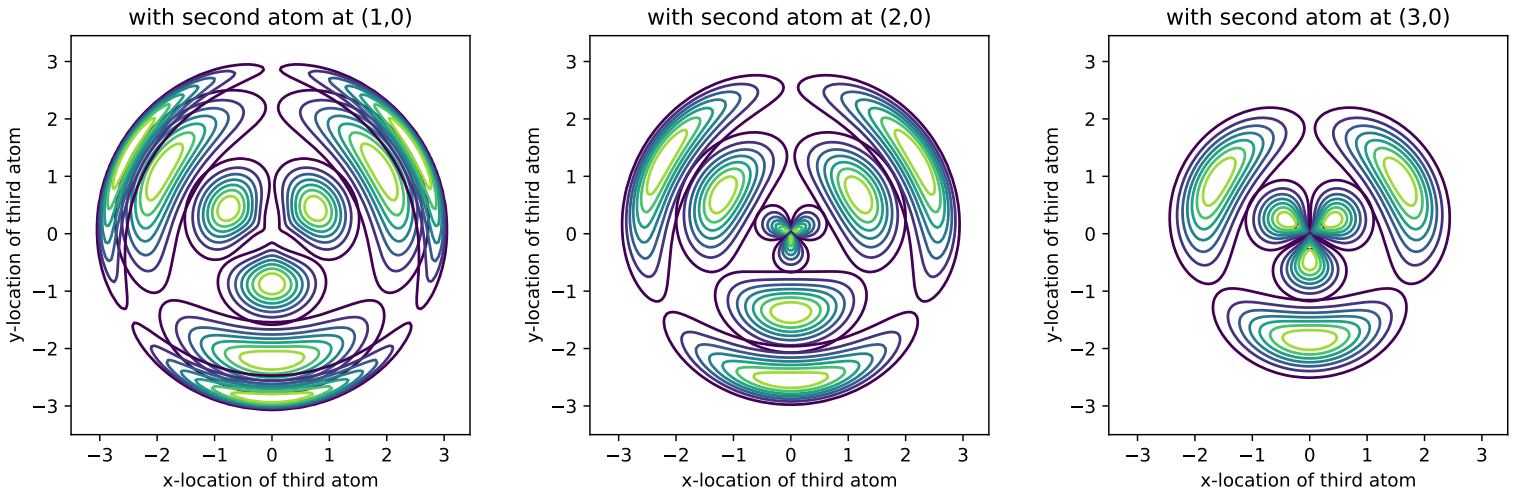
Symmetry Functions capture localradia

- RadialSFs capture atom-type neighborhood indep of orient
  - Pair-wise interactions

$$y^{\tau}_{R,i} = \bigcup_{\rho,\eta} \sum_{j \in S_\tau, j \neq i} e^{-\eta(R_{ij}-\rho)^2} f_c(R_{ij})$$

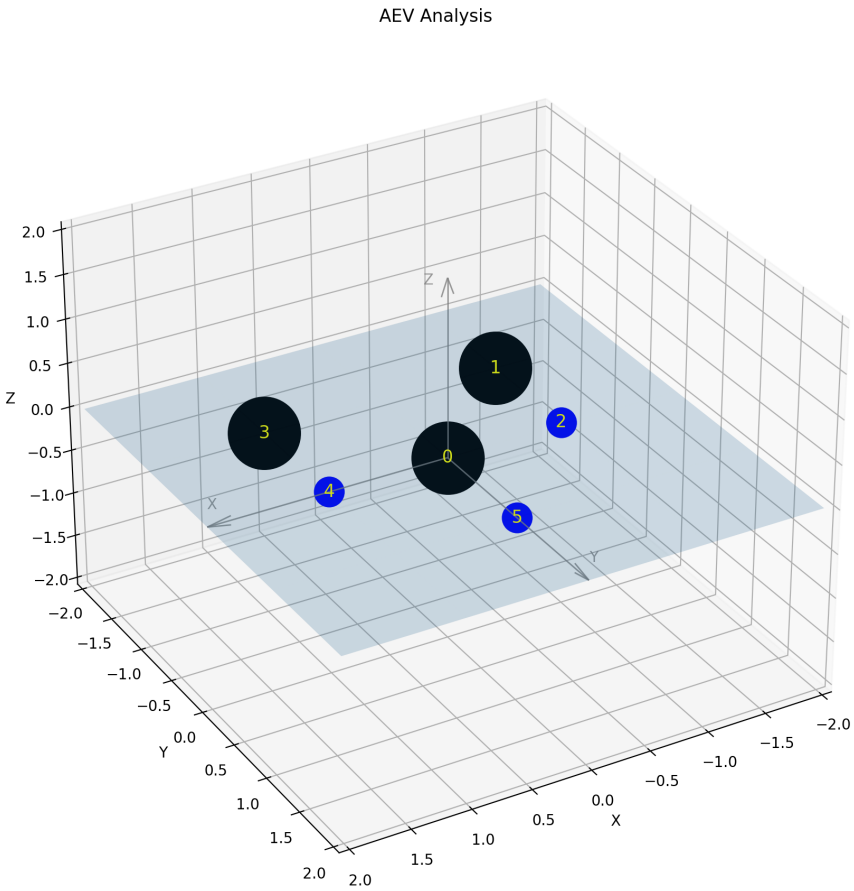


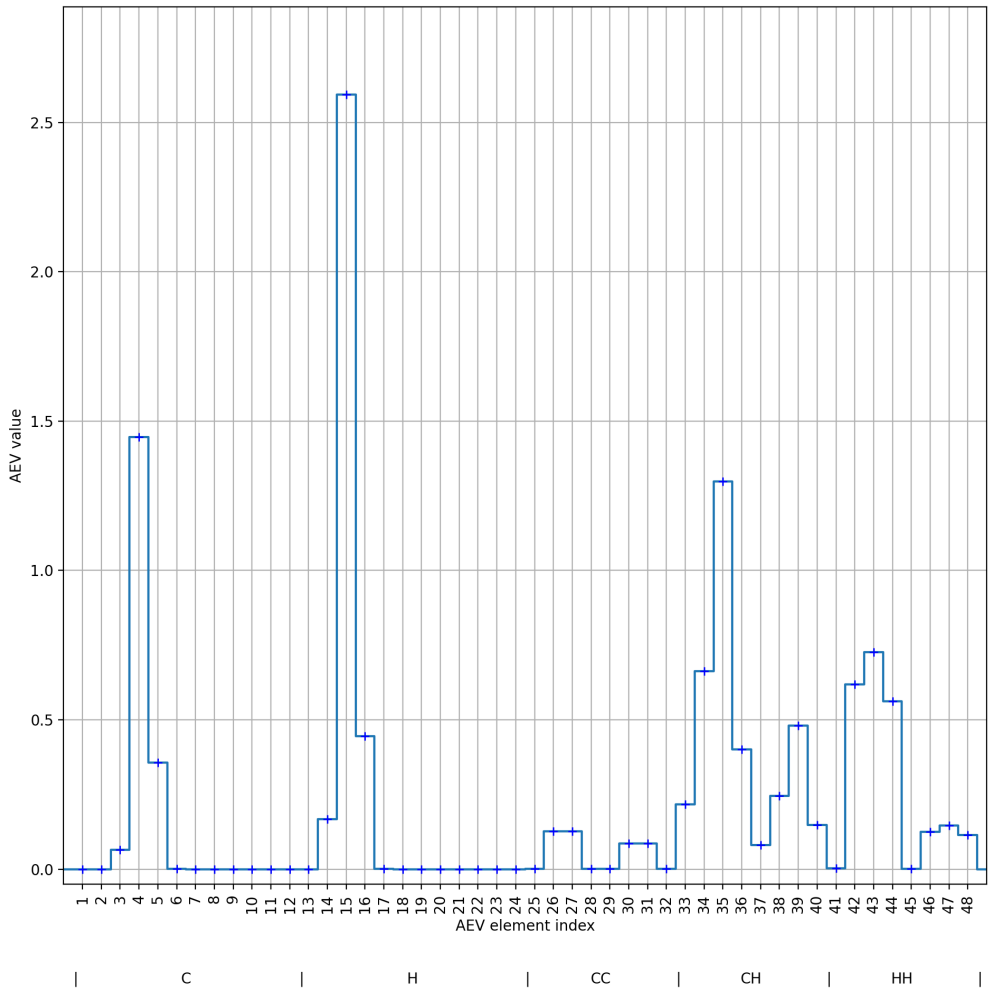
- AngularSFs similarly capture 3 type-specific angular geometry in a plane, with radialweighting



e.gforatominm a{C,H}system $y_i = (y^{\mathsf{C}}_{R,i}, y^{\mathsf{H}}_{R,i}, y^{\mathsf{C,C}}_{A,i}, y^{\mathsf{C,H}}_{A,i}, y^{\mathsf{H,H}}_{A,i}) \in \mathbb{R}^M$

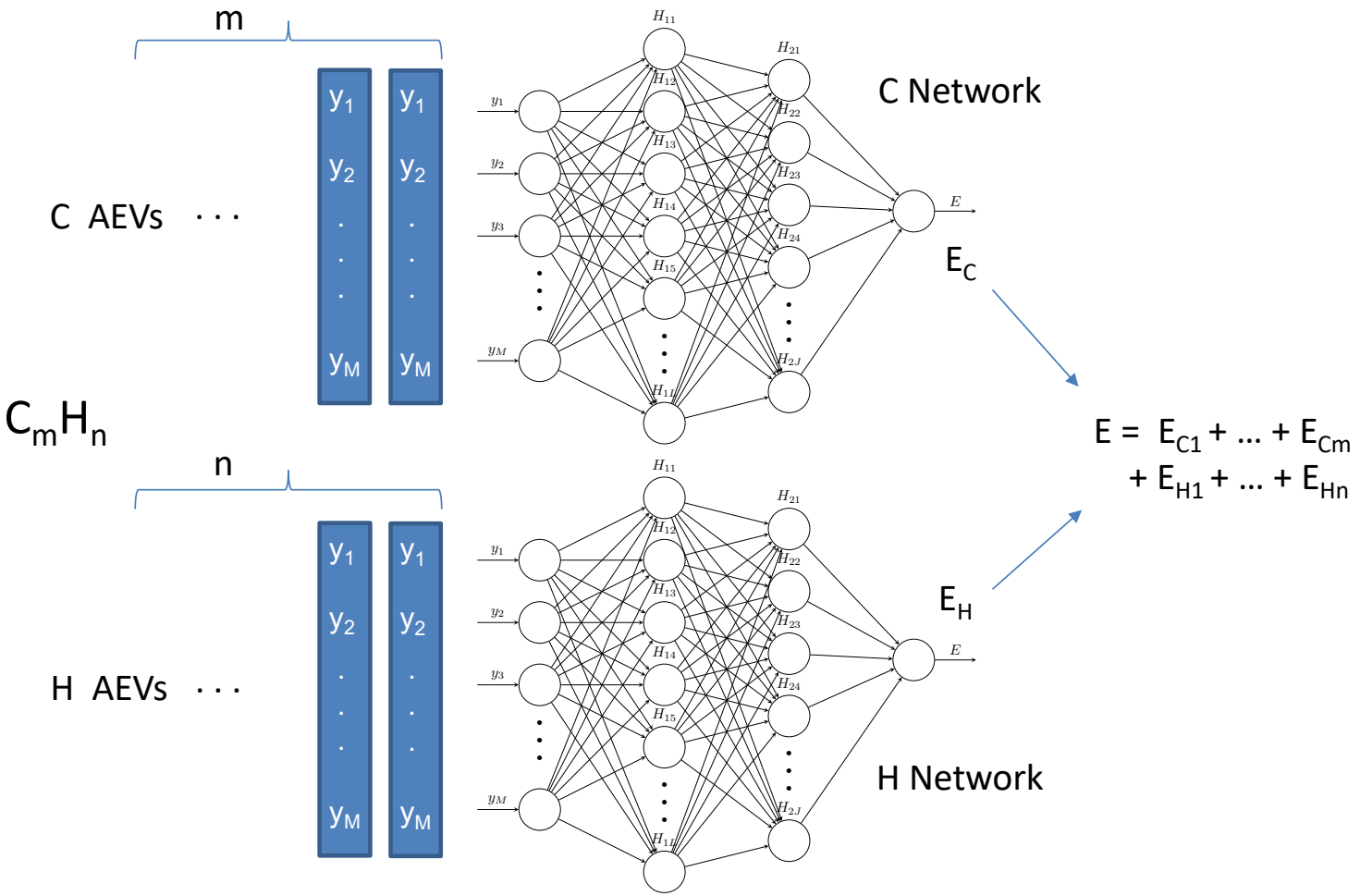
AEV illustration for one atom in a 6-atom





PLAY

NN PES setup for C_mH_n molecule in a {C,H} sys



Potential energy is the sum of outputs of atomic-type-specifi

One data point is (C_mH_n Cartesian-coords, E),

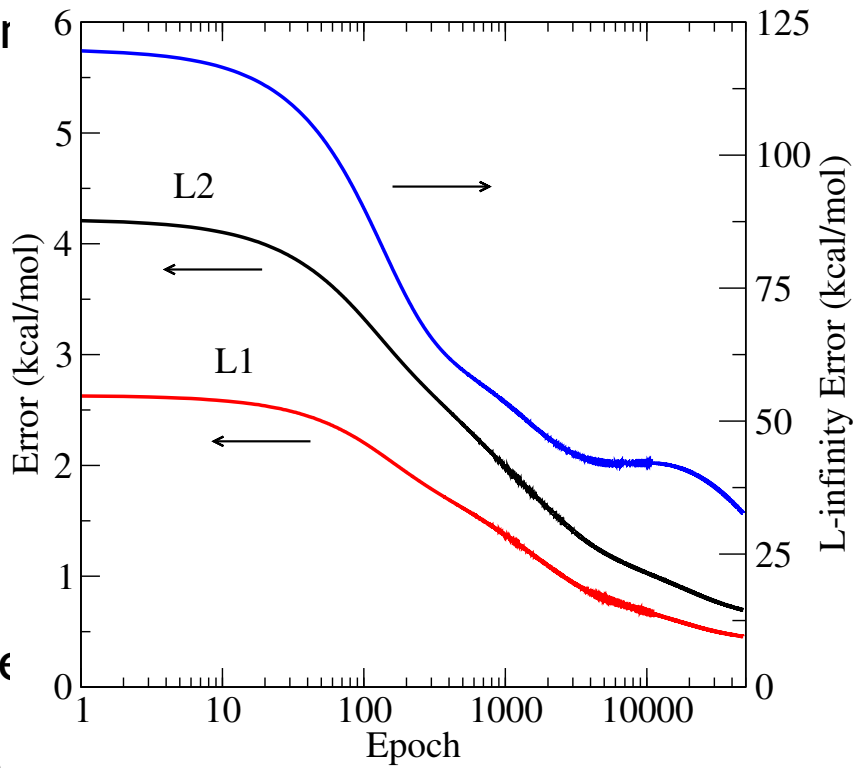
Feed forward NN MLPs; Gaussian activation functions; using PyTorch

Construction follows Smith et al, 2017.

Sample training for a molecular {C,H} da

62,829 samples of {C,H} r

C_2H_6	8469
C_3H_4	7200
CH_4	5400
C_3H_6	20160
C_2H_4	5760
C_3H_8	12960
C_2H_2	2880

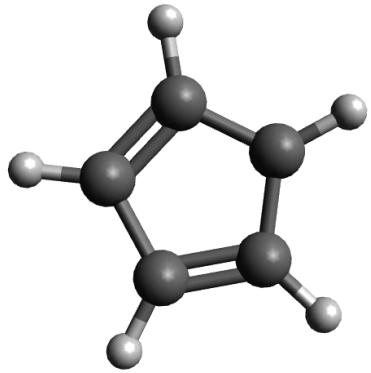


SGD Training Error Decay

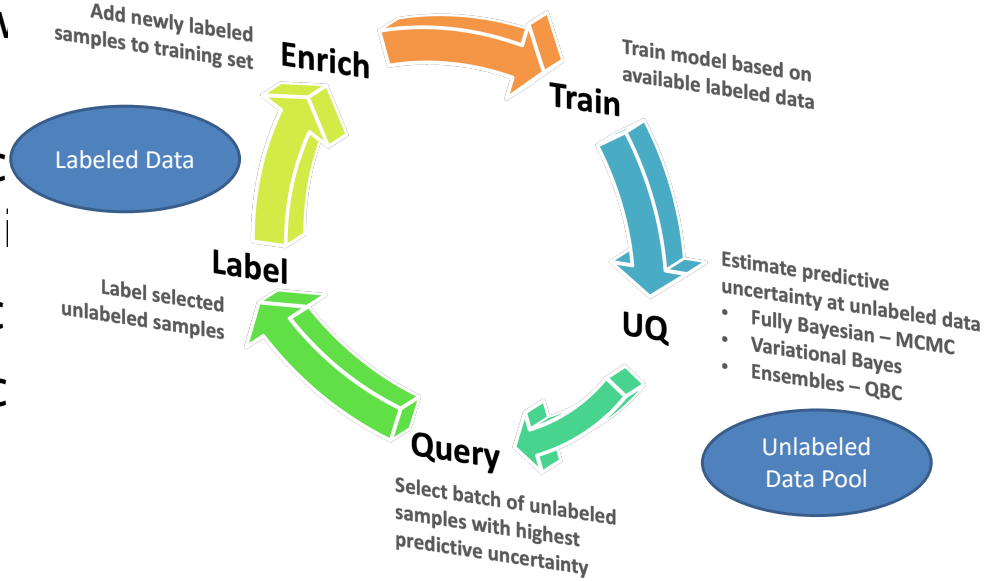
- Data randomly shuffle
- Training:80%, testing
- 256-long AEV
- NN complexity: 3 hidden layers w (128,128,64) neur
- Activation: Gaussian
- Total no. of 2-NN parameters: $2\times\sum_{\ell=1}^{L_{\text{Hidden}}} N_\ell(N_{\ell-1}+1) = 115,328$

Learning the PES of C_5H_5

- Above database: small changes aroun
- We focus on multiple basins & transiti
- Generate training data relying on:
  - PES exploration with KinBot github.com/zadorlal
  - Normal mode random sampling
  - Active learning using query by Committee (QBC)
  - DFT computations using QChem www.q-chem.com



- Training with ADAM, w
  test-error is used to in
  - the learning rate sc
  - the stopping criteri
- Target error below 1 kc
  - "chemical" accurac



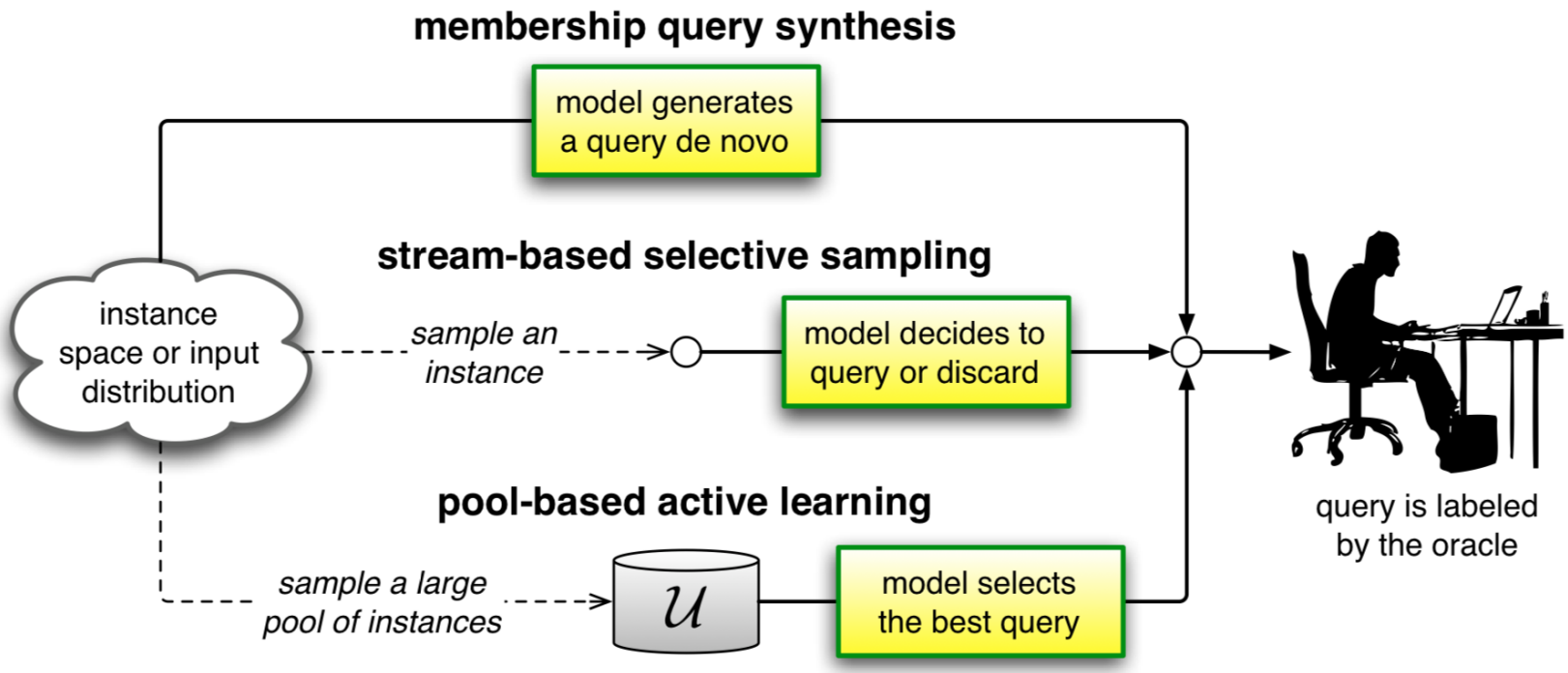
Active Learning (AL)

- Goal: improve model fit by adding data where teste
- Since the truth away from training data is unknown, some *estimate* of the prediction error
- Uncertainty in the prediction from the fitted model surrogate for this error
- Ideally, the model can be fitted using Bayesian infer push forward posterior uncertainty used as the error
- While this is generally feasible in low-dimensional a models, it is unfeasible in deep NN (DNN) models
- A range of methods have been used to provide appr estimates of predictive uncertainty in DNNs

  Gawlikowski et al, arXiv: 2107.03342, 202

- We will focus here on ensemble methods employin randomized weight initialization and data subsettin

Active Learning: selection of training configurations

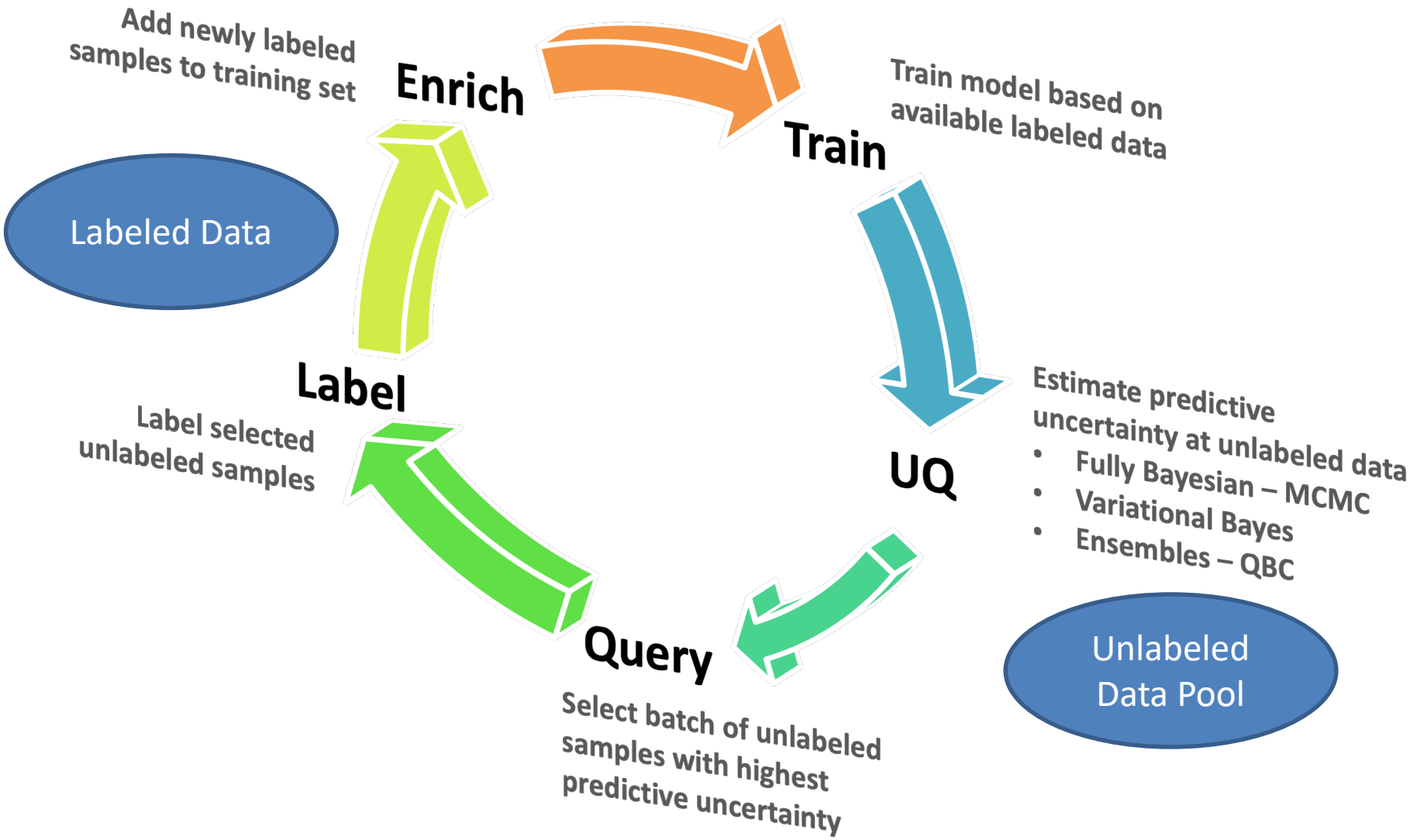


B. Settles, "Active learning literature survey", Computer Report 1648, University of Wisconsin-Madison,

Query Strategies: almost all rely on some form of uncertai

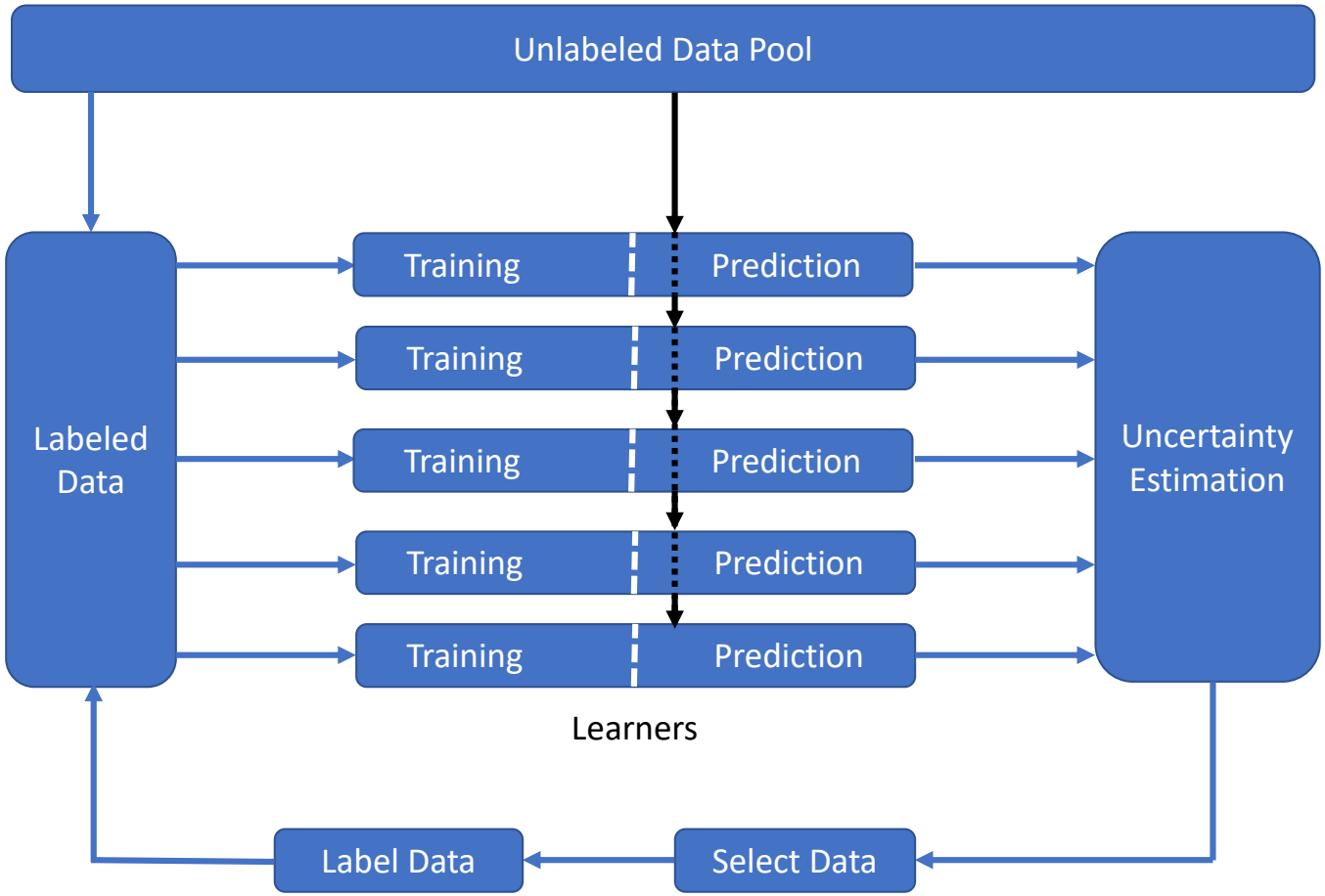
- **Uncertainty sampling:** active learner queries the instanc which it is least certain how to label.
- **Query-by-committee:** committee of competing models, a query about which they most disagree. Need a mea disagreement.
- **Expected model change:** query would lead to greates change, e.g. largest gradient length.
- **Variance Reduction and Fisher Information Ratio:** variance component of generalization error estim at Information)
- **Estimated error reduction:** the expected future erro would result if some new instance x is labeled and ad set, and then select the instance that minimizes tha

Pool Based UQ-Informed Active Learnin



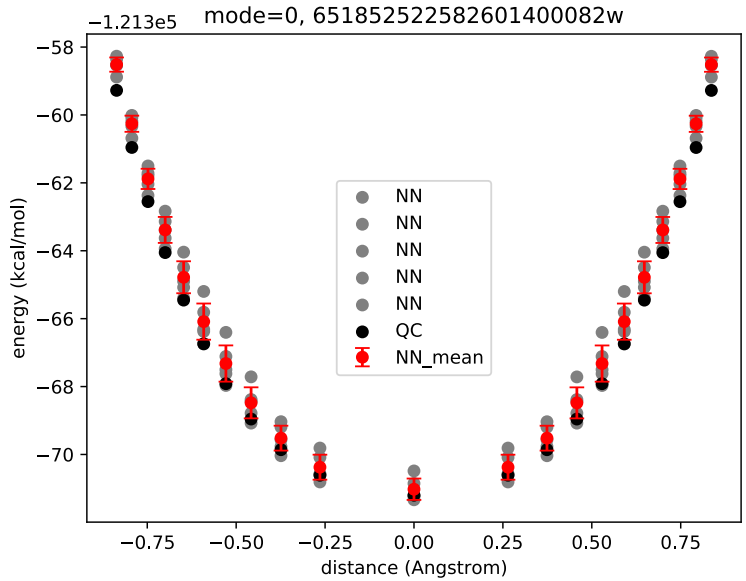
Query By Committee (QBC)

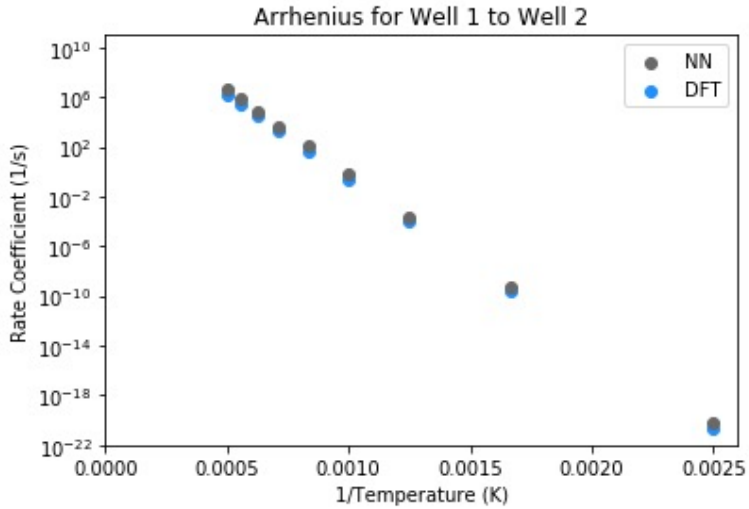
- Train M learners, $N_1, N_2, \ldots, N_M$
- Randomize initial weights and data subsetting among
- Use scatter of M predictions as a surrogate for predictive
- Learner-average prediction provides a robust low-e



Two-Basin H_5C_5 System fitting

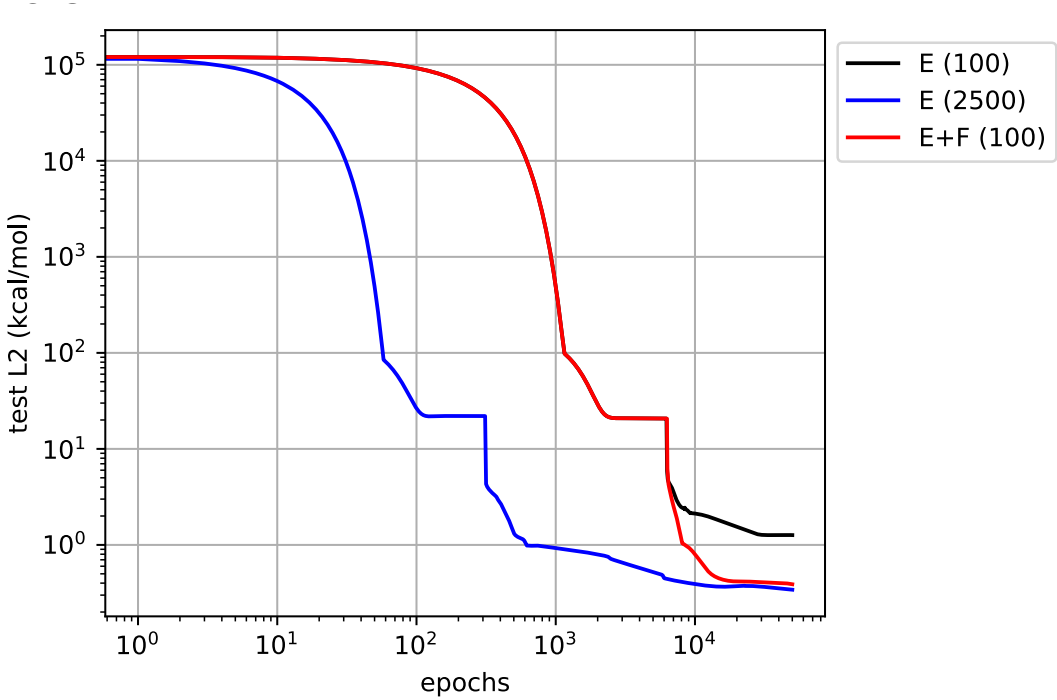
- Two-basins and the connec
  reaction path over the trans
  state
- Training using QBC active le
  with 5 NNs
  ∼30K data points
- DFT test data is generated l
  normal mode random sam.....
- Data is well captured by the
  prediction from 5 trained N
  - Tests data in basins, as w
    transition state regions
- Reaction rates estimated u
  NNPES or DFT agree over a
  range of temperature





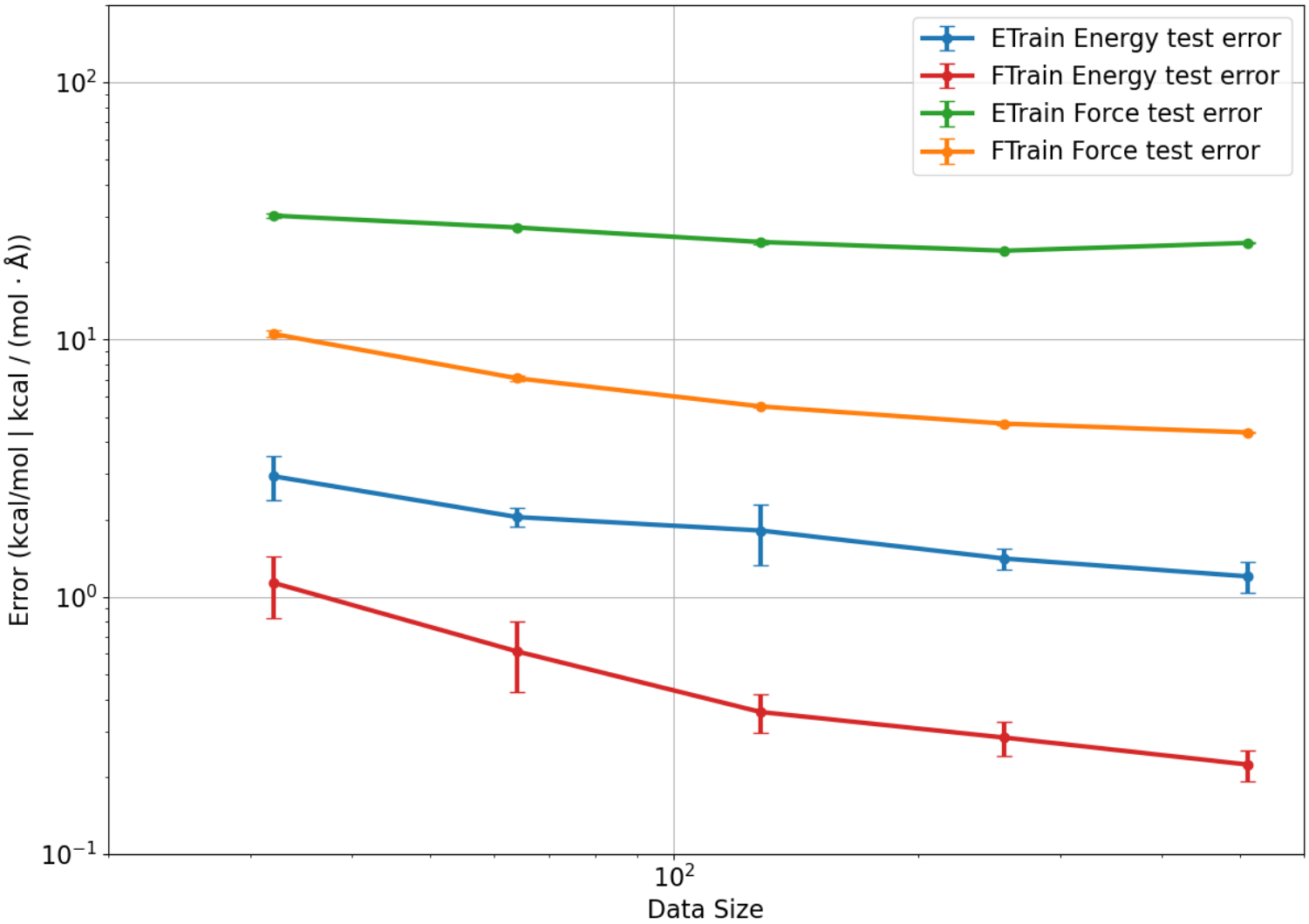
Advantages of force training – How to train

- DFT provides energy and forces ($-\nabla E(x)$)
- Force data adds significant information on the PES shape
- In principle should reduce data needs for a given
- Training on both energies and forces is a significant
  - Optimizer needs the NN & AEV Hessian
  - Used C++ & automatic differentiation
- We find 25× reduction in data needs for the same tested



NN: (128,128,64) Gaussian

Advantages of force training



- ∼10 x reduction in test error due to force training

Closure

- Described AEV-NN models as potential energy surf
  atomistic modeling
- The formulation of more expressive descriptors is ar
  area of work
  - Graph CNNs coupled with internal coordinates a
    candidates going forward
- We demonstrated the accuracy of a trained NN PES
  computations on a subset of the potential energy la
  - Used UQ informed pool based QBC active learni
    selection
- We see force training as providing significant saving
  number of DFT/QC runs
  - Using force training does entail more data per str
    more costly training

Case Study: Refractory Alloys

Work with:

Logan Williams[1], Krishnachik Sargsyan[1],
Mitchell Wood[2], Mary Alice Cusentino[2], and Aidan Thompson[2]
Sandia National Laboratories
Livermore[1], CA & Albuquerque[2], NM

Goals:

- UQ in MD simulations of refractory alloys
  - relevant in fusion reactor wall-materials design
- Bayesian calibration of SNAP potentials with DFT da
- Forward UQ in MD simulations
- Attribution of errors and uncertainties

Spectral neighbor analysis potential (SN

- Uses *bispectrum* as fingerprints
  - hyperspherical harmonics
  - respects rotational, permutational, translational
  - incorporates energy, force, and stress data
  - tunable complexity/order

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

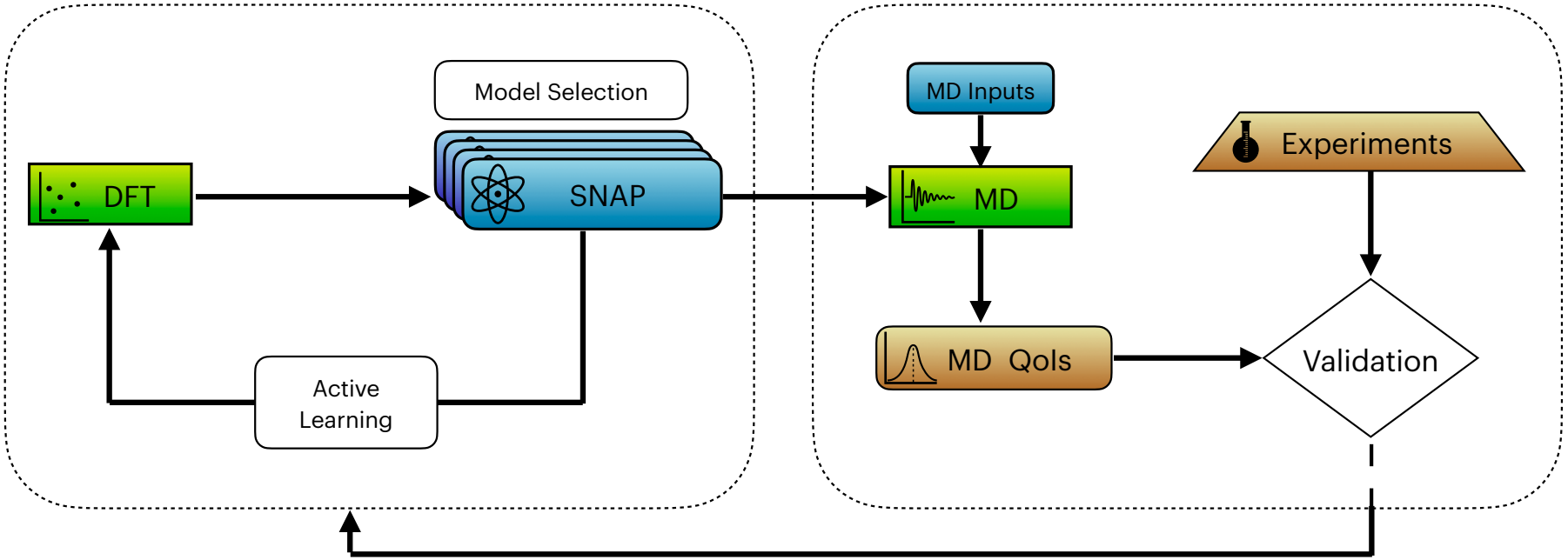
Thompson, J. Comput. Phys. 285 (2015)

- Uses linear regression
  - generalized to quadratic form as well

Wood, J. Chem. Phys. 150 (2019)

- We will target computations of alloys involving Zr, W
  – under a range of conditions
- Will use VASP for DFT computations & LAMMPS for

Utilization of UQ

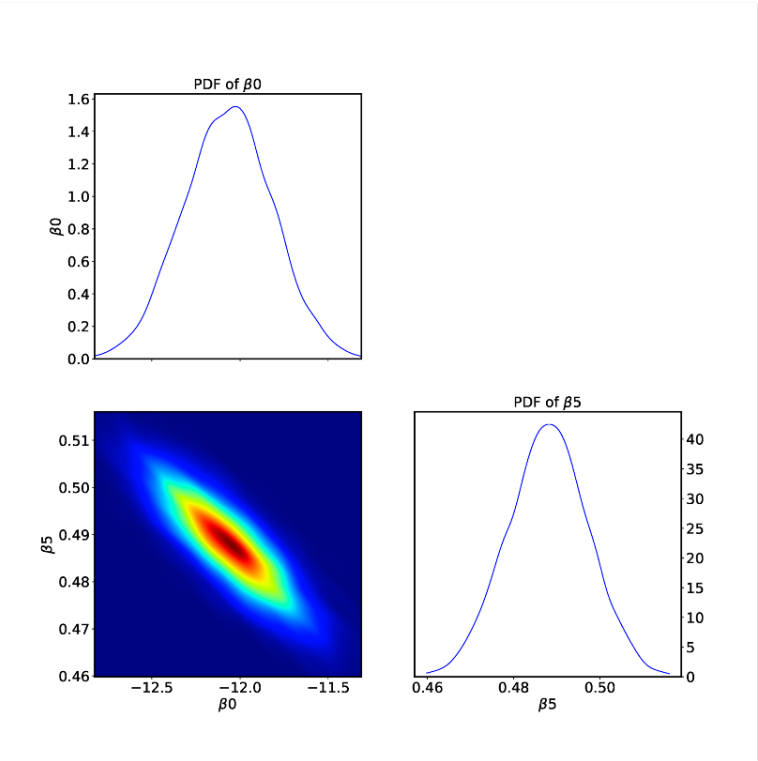


Bayesian Inference

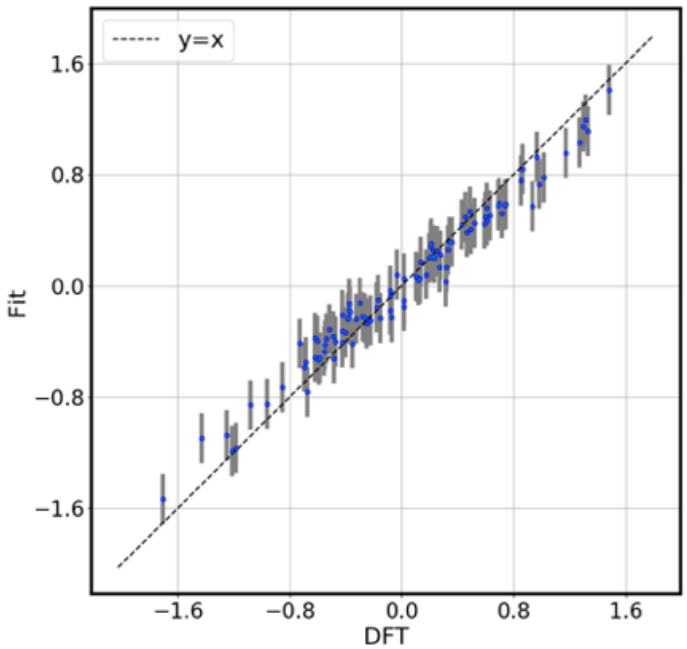
Forward UQ

Bayesian Estimation of SNAP IAP Param

- Demonstrated Bayesian inference of linear SNAP p
  with adaptive Markov chain Monte Carlo (AMCMC) u
- Posterior density quantifies uncertainty in SNAP pa
- Uncovered correlations among uncertain parameter

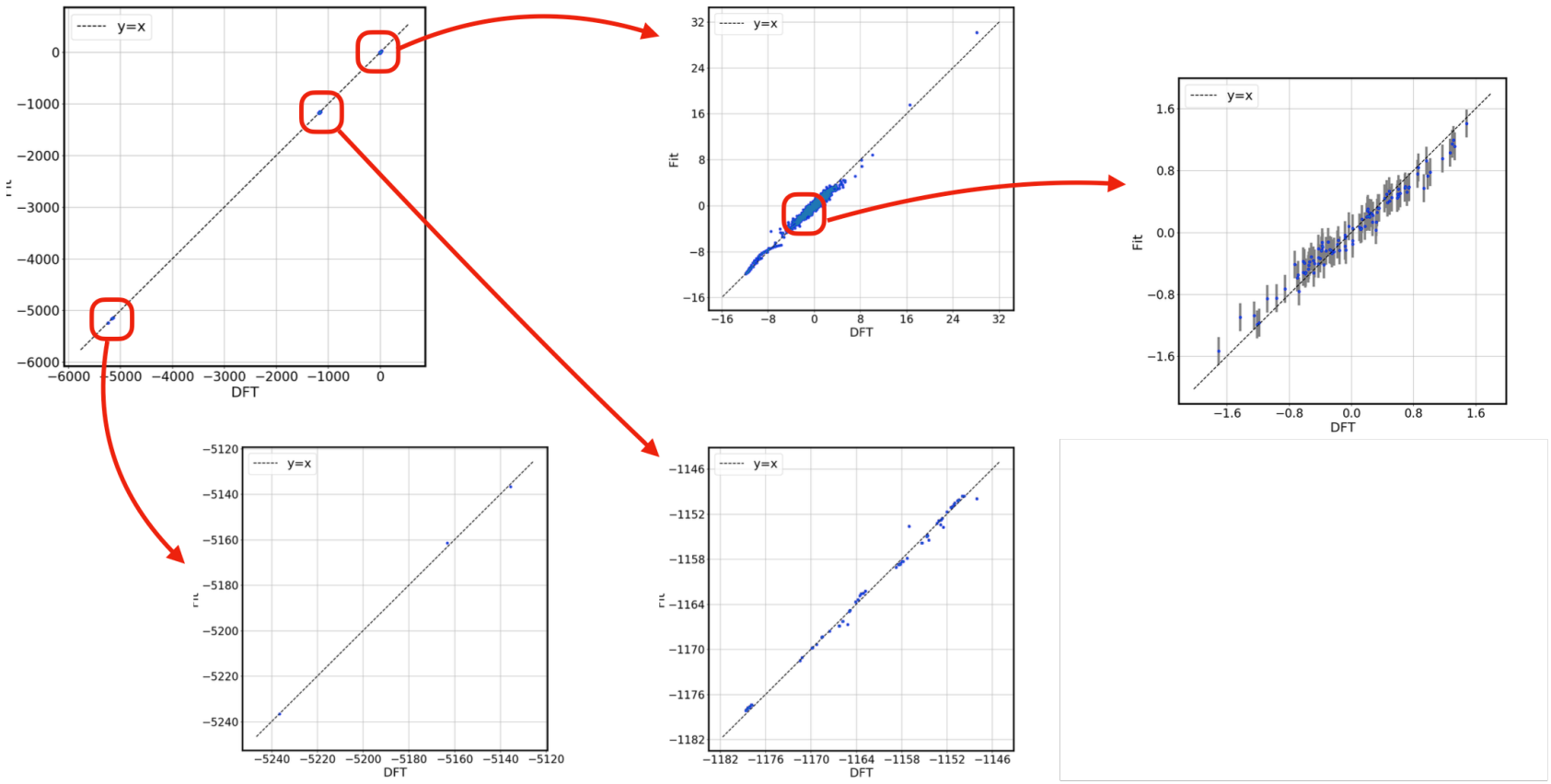


Marginal posterior PDFs on SNAP model parameters, showing uncertainty and correlation



SNAP fit with posterior predictive uncertain
DFT data

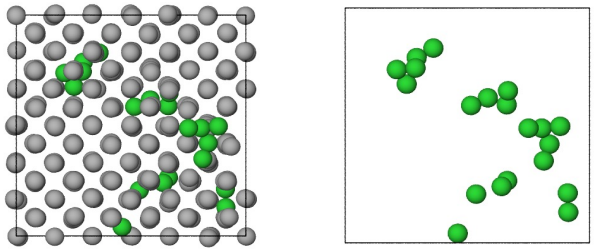
Bayesian Estimation of SNAP IAP Param



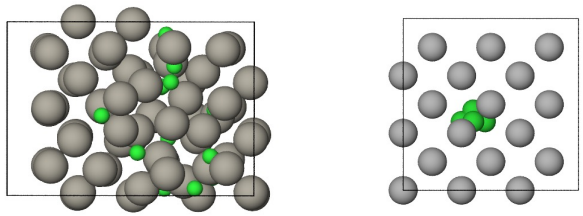
- Posterior predictive uncertainty spans discrepancy w undersome, but not all, conditions
- Complex discrepancy landscape, and non-uniform
- Explore customized data weighting, and modelerro

Training set selection is crucial

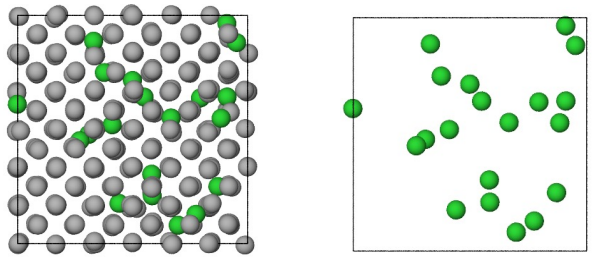
- Configurations chosen for training influence results
- Example: W-H (tungsten/hydroge



- **Top:** Initial IAPs resulted in hydroger in bulk tungsten, which should not



- **Mid:** Additional training data was generated and added to training set
- **Bot:** Including these specific configu prevented unphysical hydrogen cl

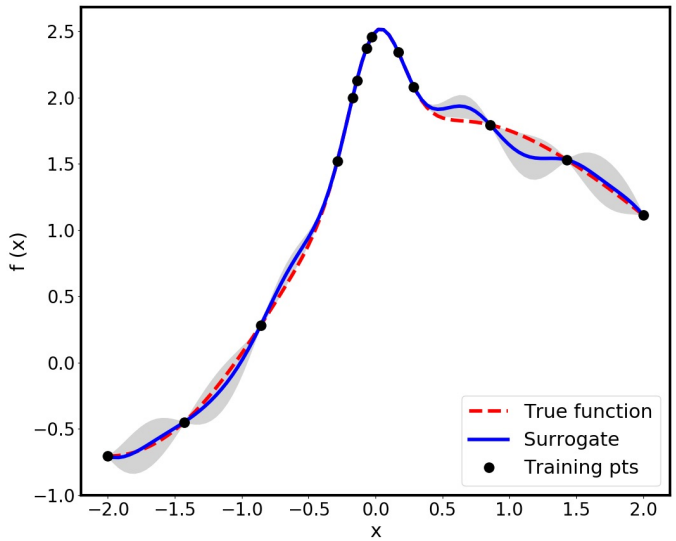


Tungsten: Grey, Hydrogen: Green

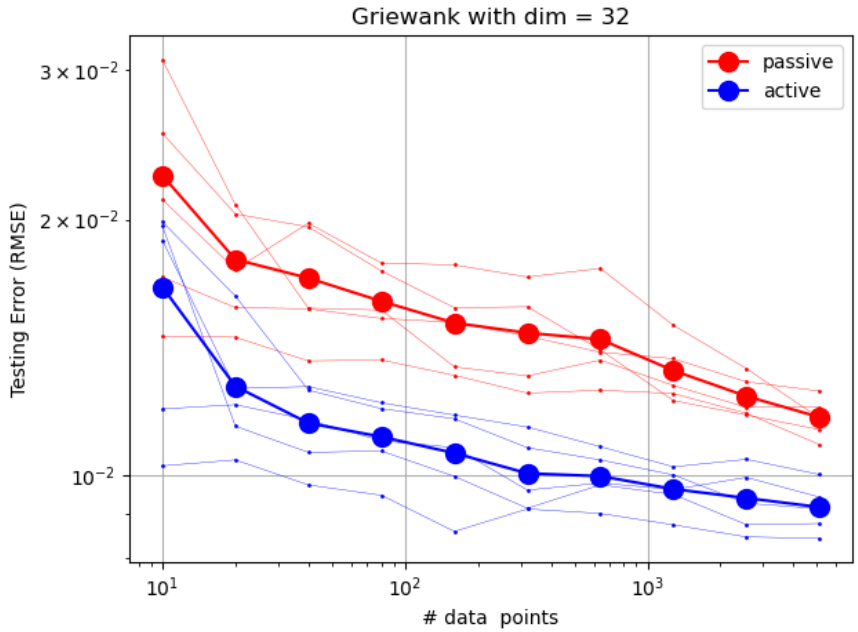
Results from Mary Alice Cusentino (SNL), u

Active Learning – Progress

- Explored different means of estimation of predictive
  - Full Bayesian solution – analytical & MCMC
  - Variational inference
  - Ensemble methods – query by committee (QBC)
- Developed active learning capability using QBC
  - Train multiple learners w/ randomized data subsets & i
  - Estimate predictive uncertainty using scatter in learne
  - Use average prediction as best estimate

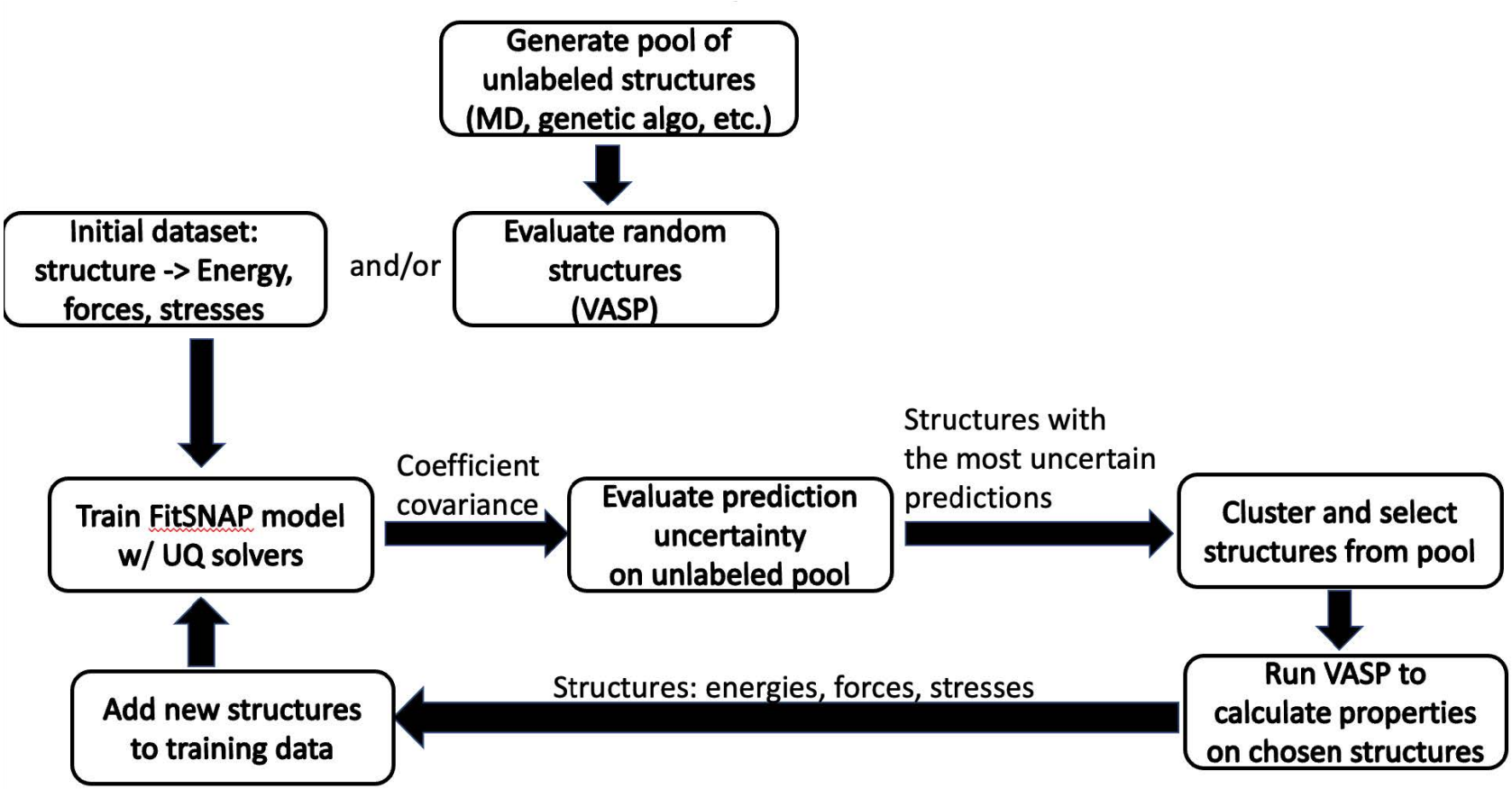


Predictive uncertainty is high in high error in fitted function



AL demonstration on NN learning of m problem in 32 dimensions

Active Learning Implementation



Preliminary Observations: AL in HEA: Ta

- Energy, force, and stress data included
- Pool includes:
  - Ab-initio MD (AIMD) data at $T = 300 - 3200$K
  - Elastic, Volumetric, and Uniaxial strain data
  - Surface data with different orientations 100, 111,
- The starting training data had some highly decompr Strain structures,
- The first set of structures chosen were all either Volu Uniaxial Strain structures
  - Almost all Volumetric and Uniaxial strain structu the entire time were for compressed structures
- After this, the AL algorithm chooses almost exclusiv from a high-temperature (3200 K) AIMD set

Refractory Alloys Closure

- Outlined results with Bayesian estimation of SNAP r coefficients with DFT data
- Results highlight the need for model error represent
  - More about this in Thursday talk by Khachik Sargs
- We are working with pool based UQ-informed AL re Bayesian posterior uncertainty in SNAP coefficients
- Preliminary results with AL on high entropy alloy are confidence in the performance of the construction
- We will include model error uncertainty accounting

Conclusions and Comments

- Discussed three case studies with different PES/IAP
  - Embedded atom model
  - NN+Atomic environment vector built on symme
  - Spectral neighbor analysis potential
- Outlined utility of UQ and active learning
- Parametric IAPs and pre-designed feature vectors w degree of model error that needs to be accounted fo
  - Graph CNNs, operating directly on internal coord interesting path forward in molecular systems
- UQ will play an important role in error attribution in model selection
- Forward UQ in MD will necessitate the use of multifi methods