Hyperactive Learning (HAL)

Building minimal training databases for data-driven interatomic potentials from scratch

Cas van der Oord, Dávid Kovács, Gábor Csányi (University of Cambridge)

Matthias Sachs (University of Birmingham)

Christoph Ortner (University of British Columbia)

Introduction

- Atomic Cluster Expansion (ACE)

  Bayesian view, posterior, committees

- Minimal databases

  ‘How much’ data needed to describe bulk, amorphous and liquid properties (Si)?

- Hyperactive Learning (HAL)

  Biasing Active Learning (AL) towards uncertainty

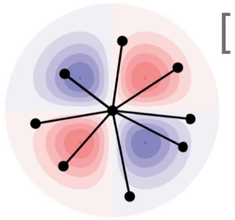
- Results

  Polymer (polyethylene glycol PEG) and alloys (AlSi10, AlSi10Mg, FeCrMnNiCo, MoNbTaW)

Atomic Cluster Expansion (ACE)

- Systematically improvable set of polynomials fitting atomic energies

$$E_{tot} = \sum_i E_i(R)$$



[1]

$$R = \{\mathbf{r}_{ij}\}_{j=1}^{N}$$

- Expressing as a linear model

$$E_i(R) = \sum_i c_i \Phi_i(R) = \mathbf{c} \cdot \mathbf{\Phi}(R)$$

PES symmetries:
- translation
- permutation
- rotation

- Linear model → (Bayesian) Linear regression

Bayesian Linear Regression

- Statistical linear problem

$$\mathbf{y} = \mathbf{\Psi c} + \epsilon$$

- Bayesian approach: posterior ∝ prior x likelihood
- Prior:

$$p(\mathbf{c}\,|\,\alpha) = \mathcal{N}(\mathbf{c}\,|\,0, \alpha^{-1}\mathbf{I})$$

Bayesian Ridge Regression (BRR)

$$p(\mathbf{c}\,|\,\alpha) = \mathcal{N}(\mathbf{c}\,|\,0, A^{-1})$$

$$diag(A) = \{\alpha_1, \ldots, \alpha_{N_{basis}}\}$$

Automatic Relevance Determination (ARD)

- Likelihood:

$$p(\mathbf{y}\,|\,\mathbf{c}, \beta) = \prod_{i=1}^{N_{obs}} \mathcal{N}(y_i\,|\,\mathbf{c} \cdot \mathbf{\Phi}(R_i), \beta^{-1})$$

- Hyperparameter optimisation: α, β

Posterior

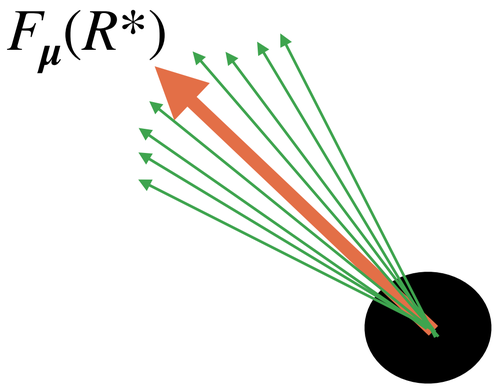
- Sample committee members $\{\mathbf{c}^j\}_{j=1}^k$ from posterior

$$p(\mathbf{c}\,|\,\mathbf{y},\alpha,\beta) = \mathcal{N}(\mathbf{c}\,|\,\boldsymbol{\mu},\boldsymbol{\Sigma}) \qquad \begin{aligned}\boldsymbol{\mu} &= \beta\boldsymbol{\Sigma}\boldsymbol{\Psi}^T\mathbf{t} \\ \boldsymbol{\Sigma}^{-1} &= \alpha\mathbf{I} + \beta\boldsymbol{\Psi}^T\boldsymbol{\Psi}\end{aligned}$$

- (Cheap) committee predictions

$$F^{\boldsymbol{\mu}}(R^*) = \boldsymbol{\mu}\cdot\nabla\boldsymbol{\Phi}(R^*) \qquad F^j(R^*) = \mathbf{c}^j\cdot\nabla\boldsymbol{\Phi}(R^*)$$

- Relative force uncertainty $U_i(R^*)$

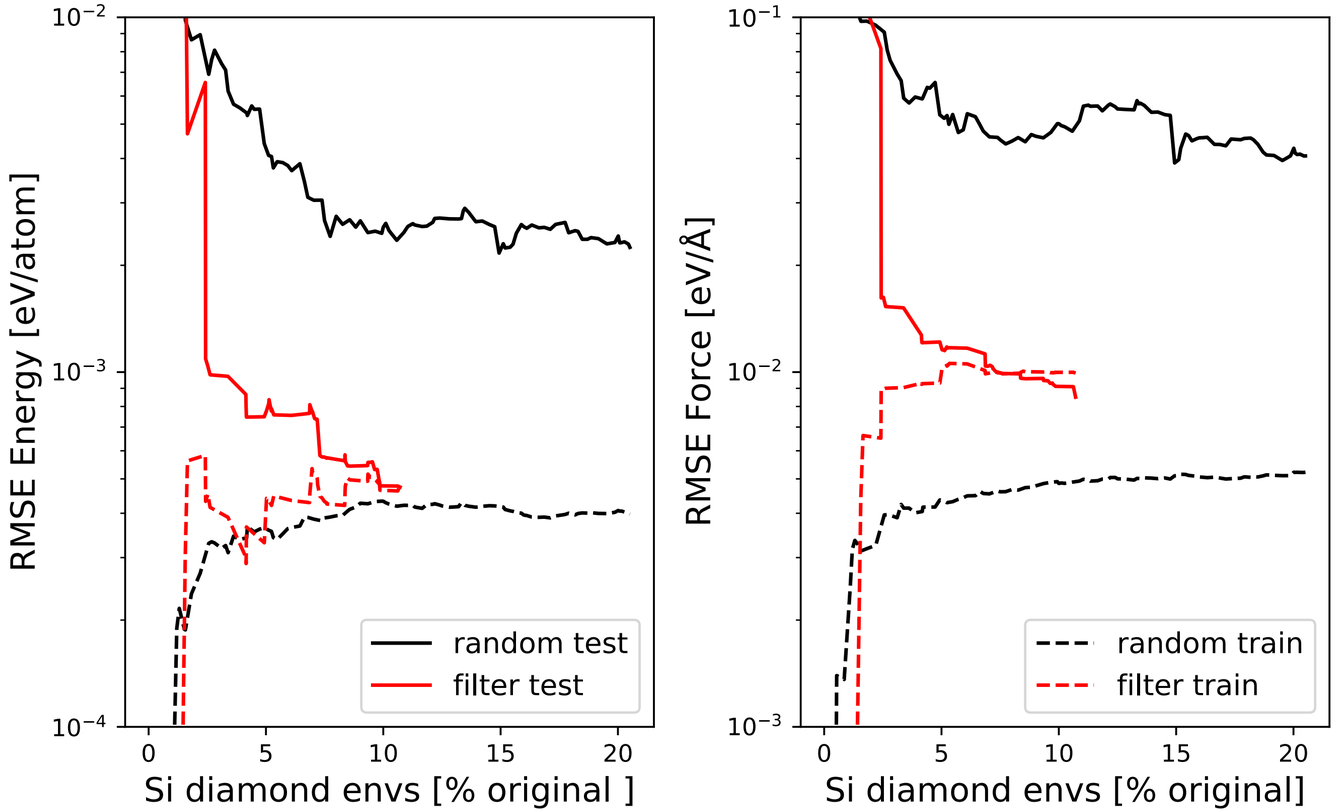


$$U_i(R^*) = \frac{\frac{1}{k}\sum_{j=1}^k \|F_i^j(R^*) - F_i^{\boldsymbol{\mu}}(R^*)\|}{\|F_i^{\boldsymbol{\mu}}(R^*)\| + \delta F}$$

- local
- relative
- regularised (δF)

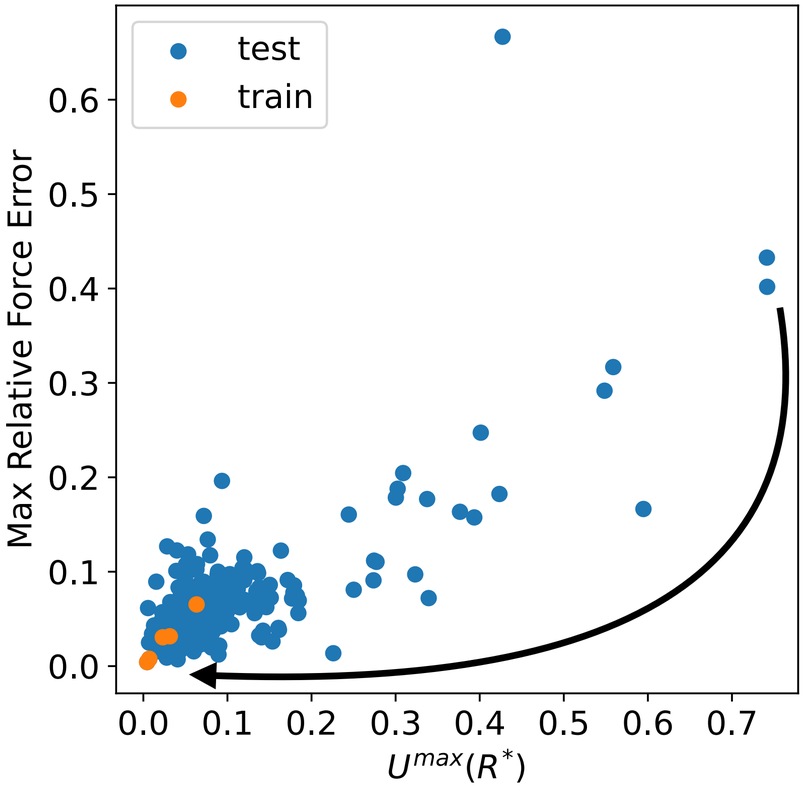
Minimal database

- Si diamond database: 489 configurations, 16807 environments [2]
- Aim: assemble minimal database keeping accurate properties
- Greedy algorithm: select test configuration based on $U^{max} = \max U_i(R^*)$

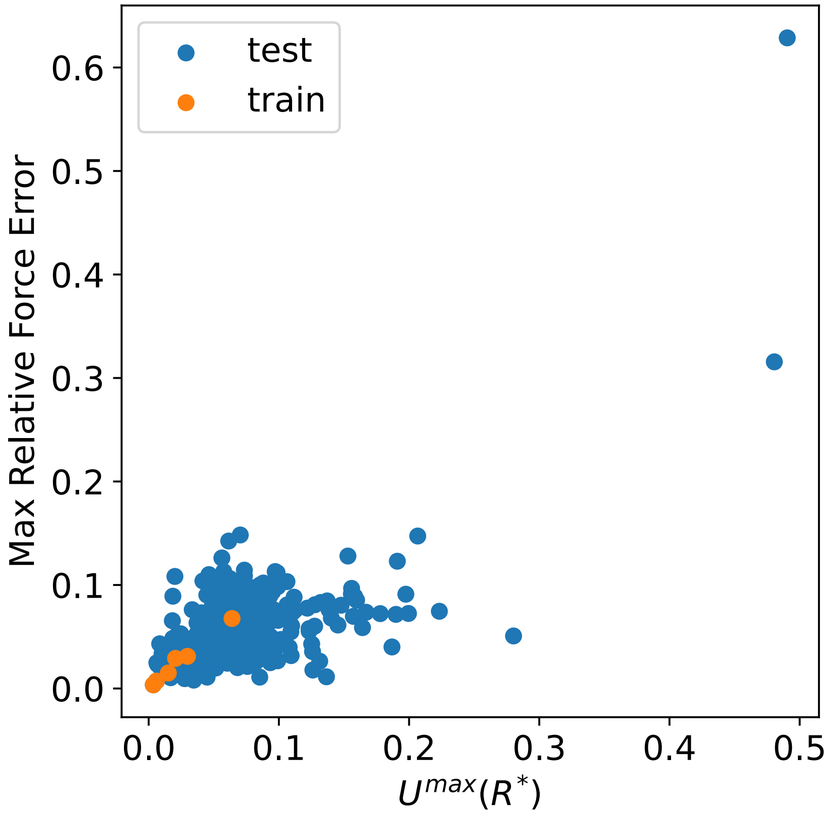


Error correlations

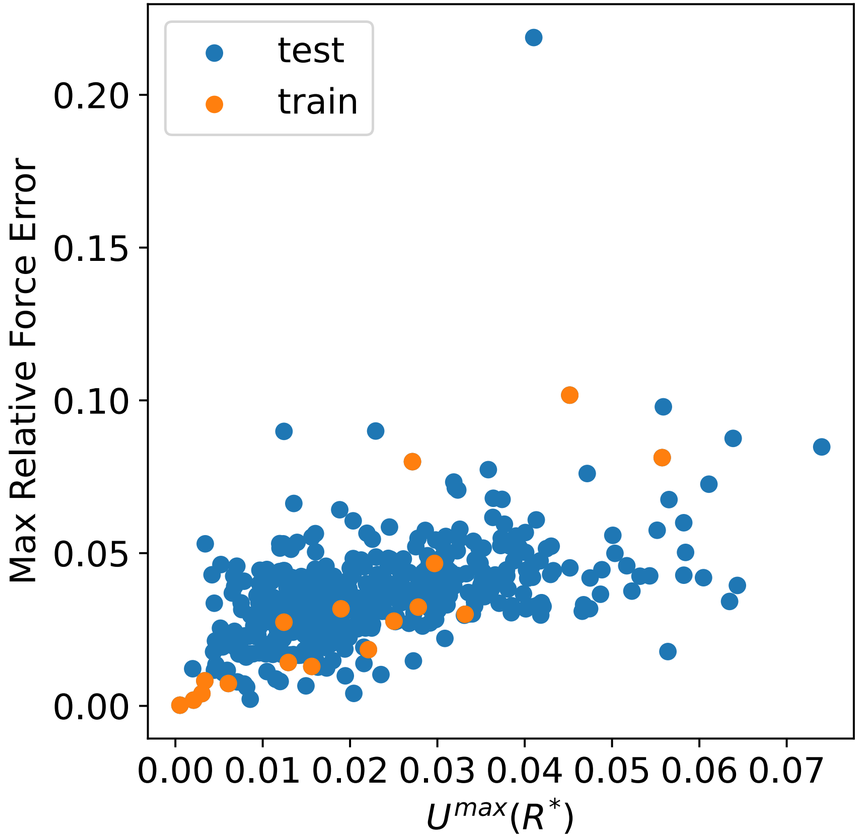
Iteration 4



Iteration 10

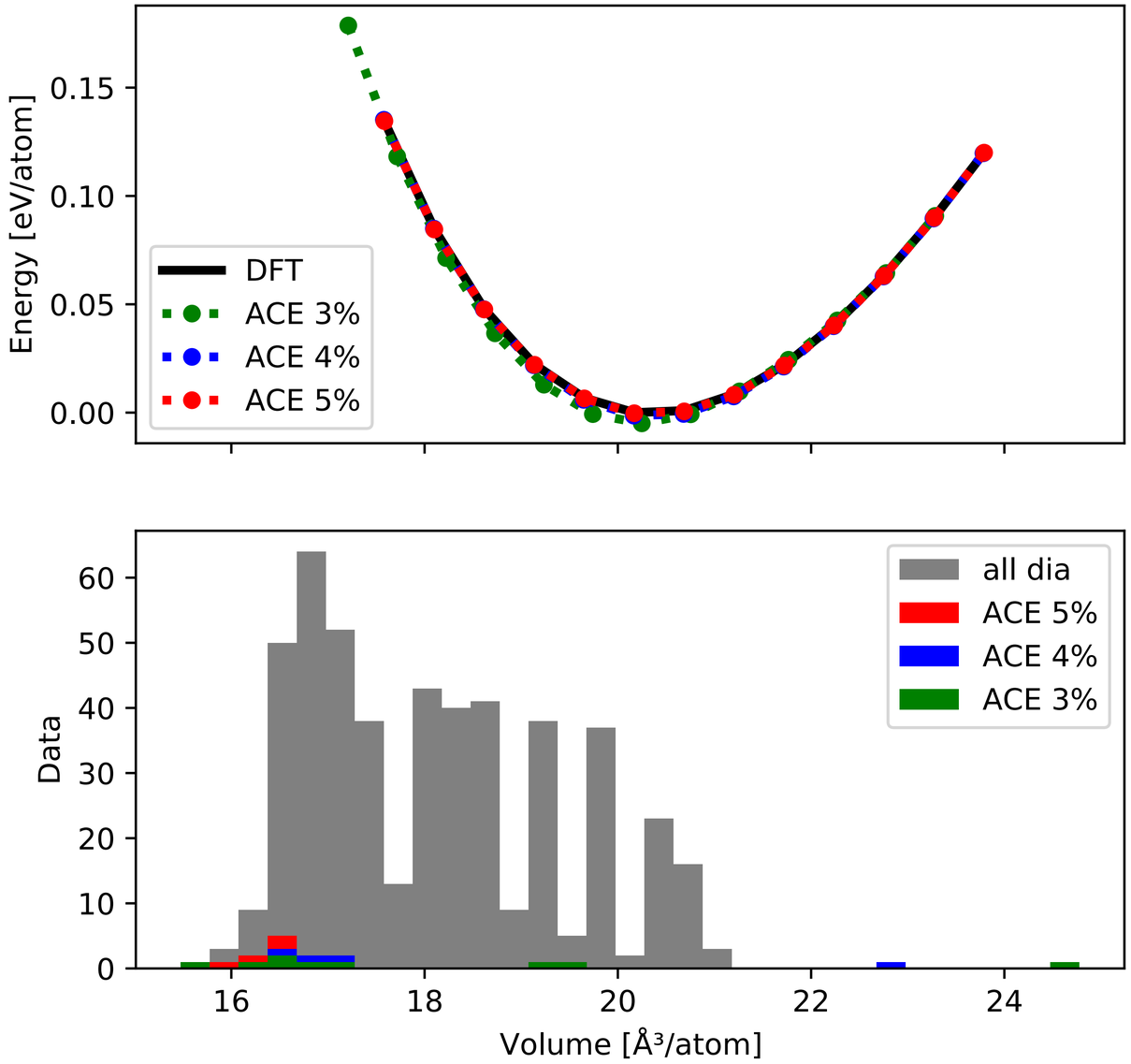


Iteration 16

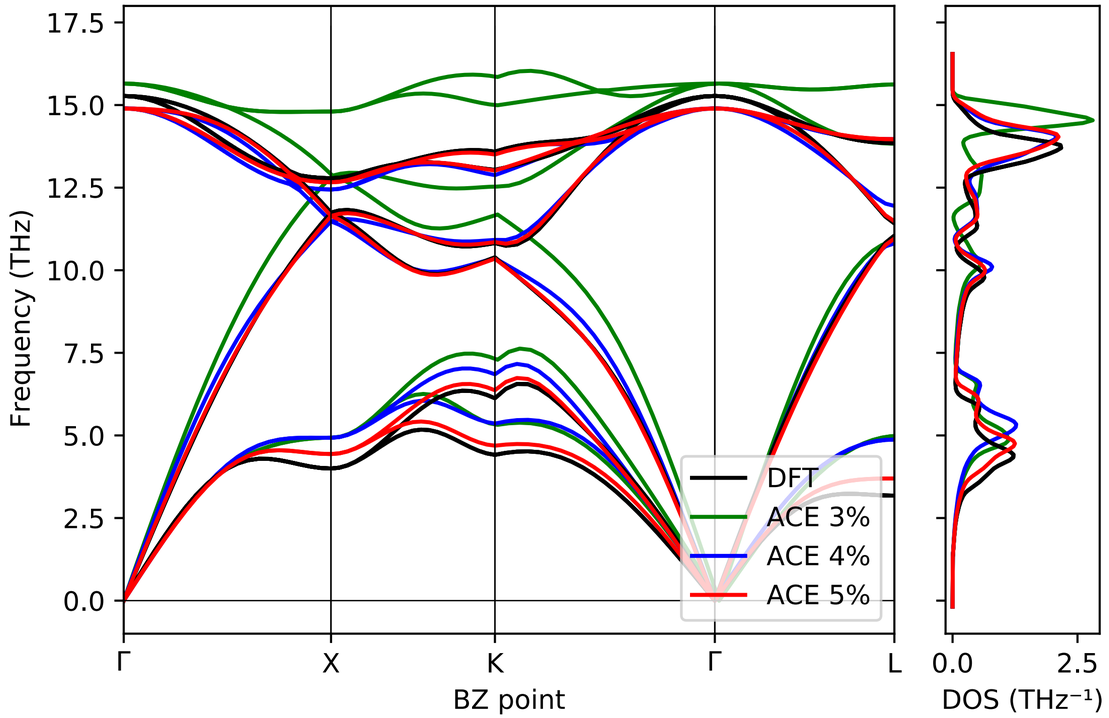


Si diamond properties

Energy/Volume



Phonons

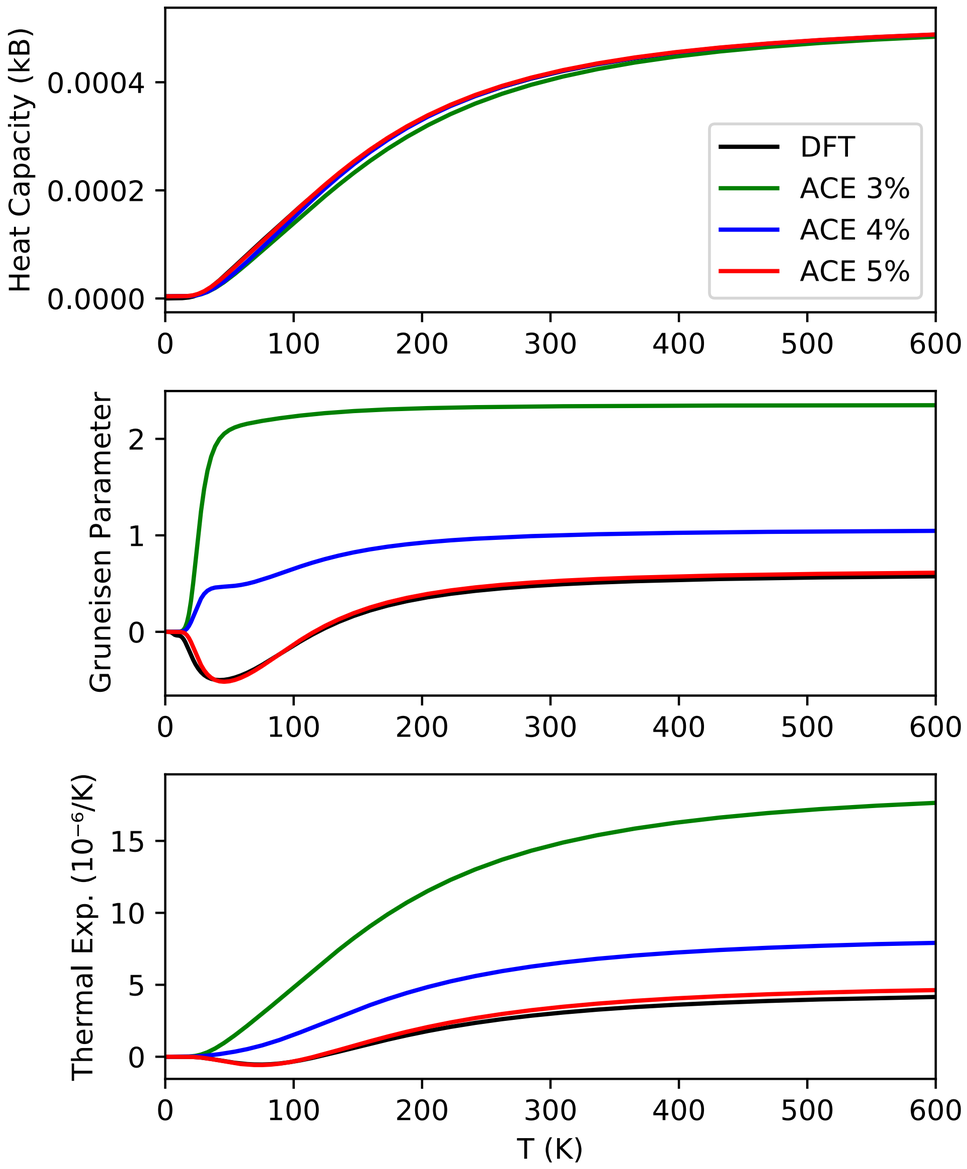


Elastic Constants

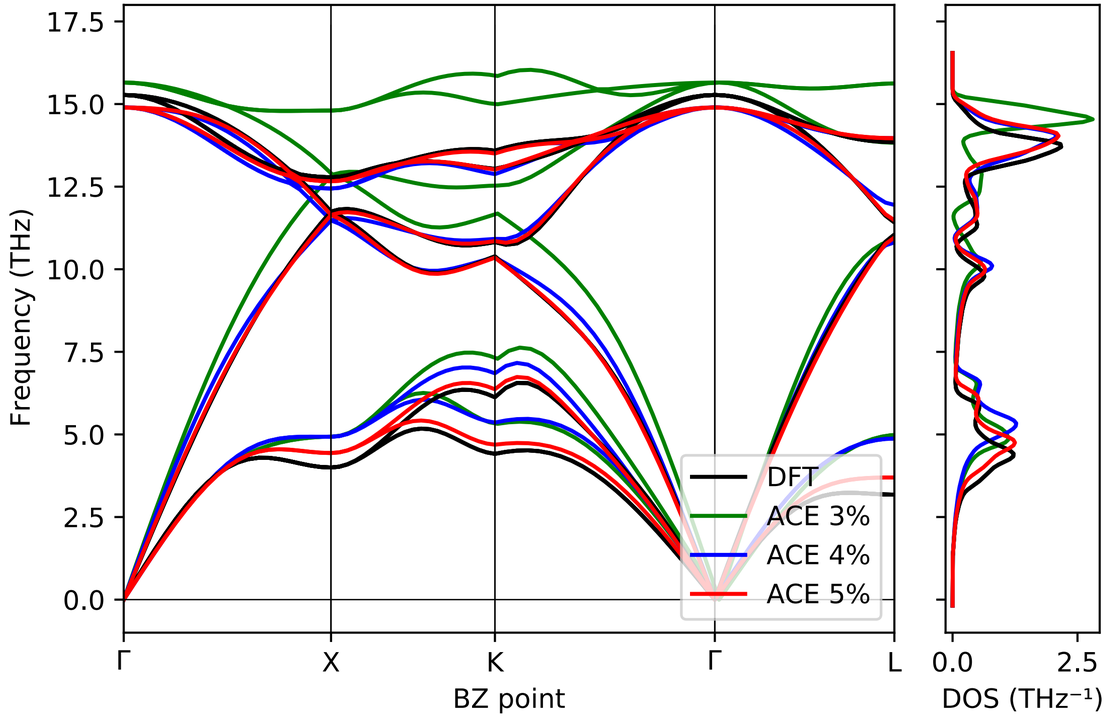
	B [GPa]	c_{11} [GPa]	c_{12} [GPa]	c_{44} [GPa]
ACE 3%	98.2 [+19%]	188.1 [+28%]	53.3 [+6%]	79.7 [+8%]
ACE 4%	84.2 [+2%]	159.8 [+9%]	46.4 [-8%]	75.7 [+3%]
ACE 5%	82.5 [0%]	148.7 [+1%]	49.3 [-2%]	73.7 [+1%]
DFT	82.6	147.2	50.3	73.1

Si diamond properties

Thermal Properties



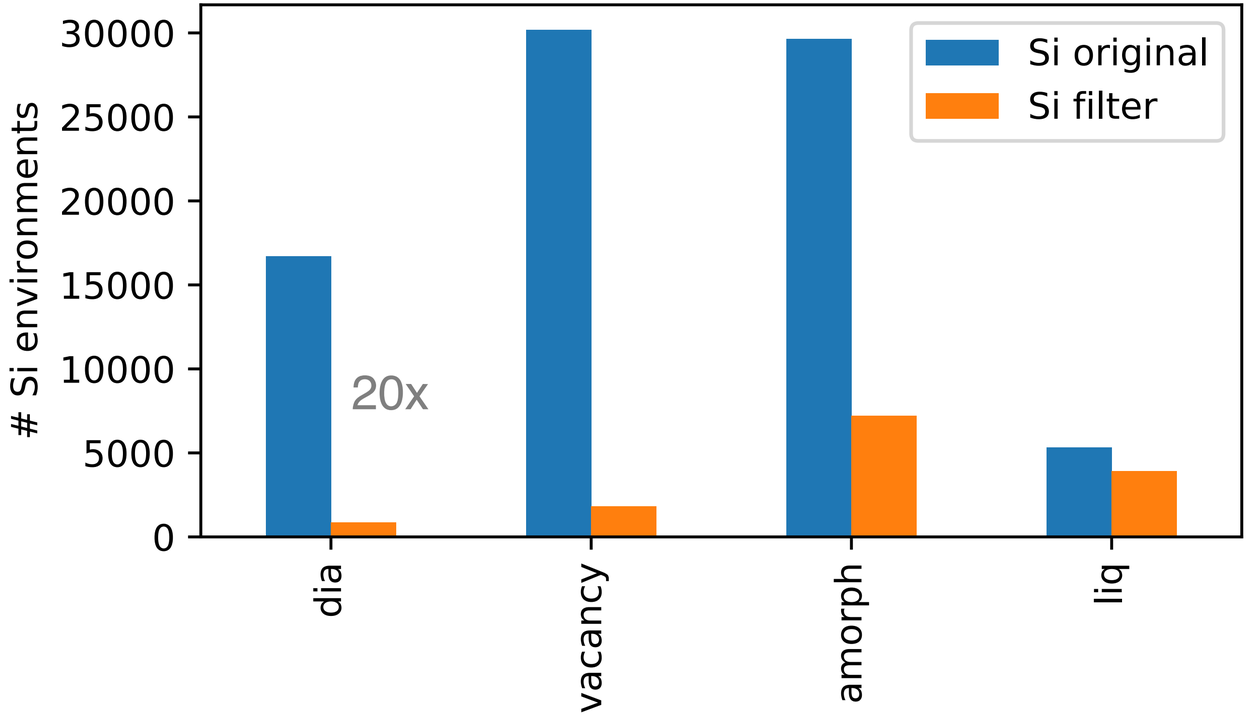
Phonons



Elastic Constants

	B [GPa]	c_{11} [GPa]	c_{12} [GPa]	c_{44} [GPa]
ACE 3%	98.2 [+19%]	188.1 [+28%]	53.3 [+6%]	79.7 [+8%]
ACE 4%	84.2 [+2%]	159.8 [+9%]	46.4 [-8%]	75.7 [+3%]
ACE 5%	82.5 [0%]	148.7 [+1%]	49.3 [-2%]	73.7 [+1%]
DFT	82.6	147.2	50.3	73.1

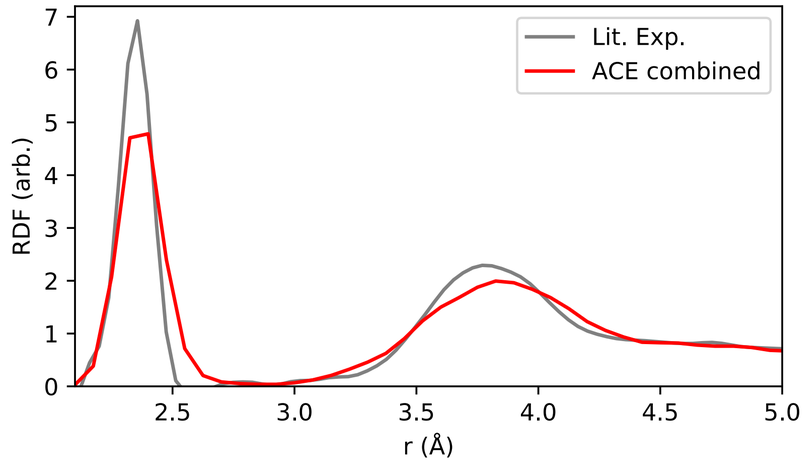
Si database



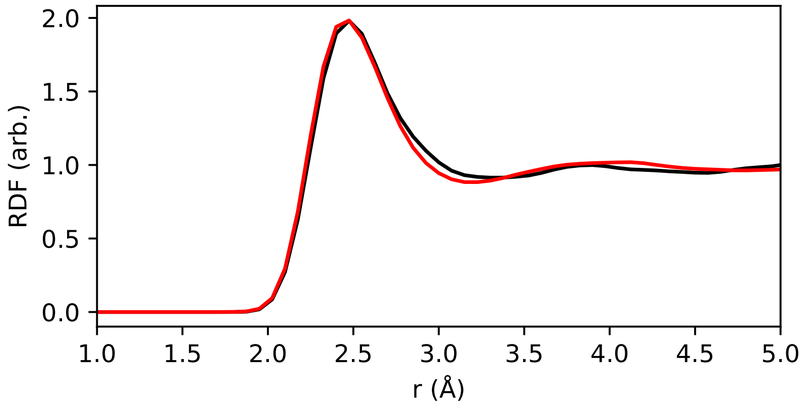
- Si diamond reduced from 489 to 17 configurations
- More diverse configuration types require more data

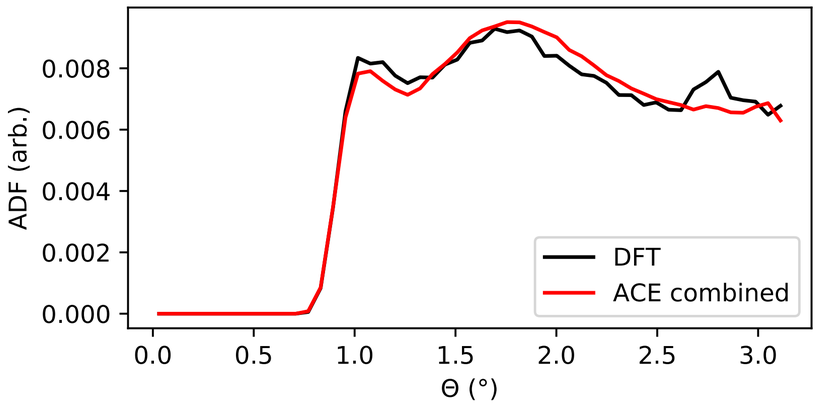
How to build minimal databases from scratch?

Amorphous properties



Liquid Properties





Hyperactive Learning (HAL)

- HAL potential energy surface $\tilde{E}(R^*)$

$$\tilde{E}(R^*) = E(R^*) - \tau\sigma(R^*)$$

- Taking derivative

$$\nabla\sigma(R^*) = \frac{\nabla\sigma^2(R^*)}{\sigma(R^*)} = \frac{2\mathbf{\Phi}(R^*)^T\mathbf{\Sigma}\nabla\mathbf{\Phi}(R^*)}{\sigma(R^*)}$$

committee approximation $\{\mathbf{c}_j\}_{j=1}^k$

- Biasing force $F^{\hat{\sigma}^2}(R^*)$

$$F^{\hat{\sigma}^2}(R^*) := \nabla\hat{\sigma}^2(R^*) = \frac{2}{k}\sum_{j=1}^{k}\left(E^j(R^*) - E^{\boldsymbol{\mu}}(R^*)\right)\left(F^{\boldsymbol{\mu}}(R^*) - F^j(R^*)\right)$$

Hyperactive Learning (HAL)

- Equations of motion

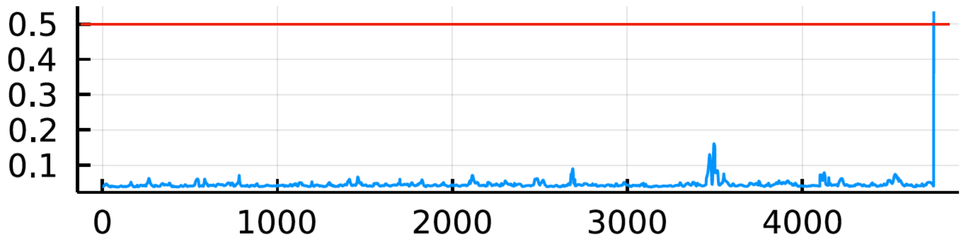
$$F(R) = F^{\mu}(R) + F^{\gamma} + \tau F^{\sigma}(R) = M\dot{V}(t)$$

$$V(t) = \dot{R}(t)$$

- Adaptive biasing τ_r to keep $\tau_r F^{\sigma}(R)$ constant relative to $F^{\mu}(R)$
- Apply softmax to $U_i(R)$ and trigger DFT once threshold U^{tol} is reached

$$U^{softmax}(R) = \max f^{softmax}(U_i(R))$$

$U^{softmax}(R)$



U^{tol}

$$f^{softmax}(x)_i = \frac{e^{x_i}}{\sum_{i=1}^{N} e^{x_i}}$$

HAL: PEG-2

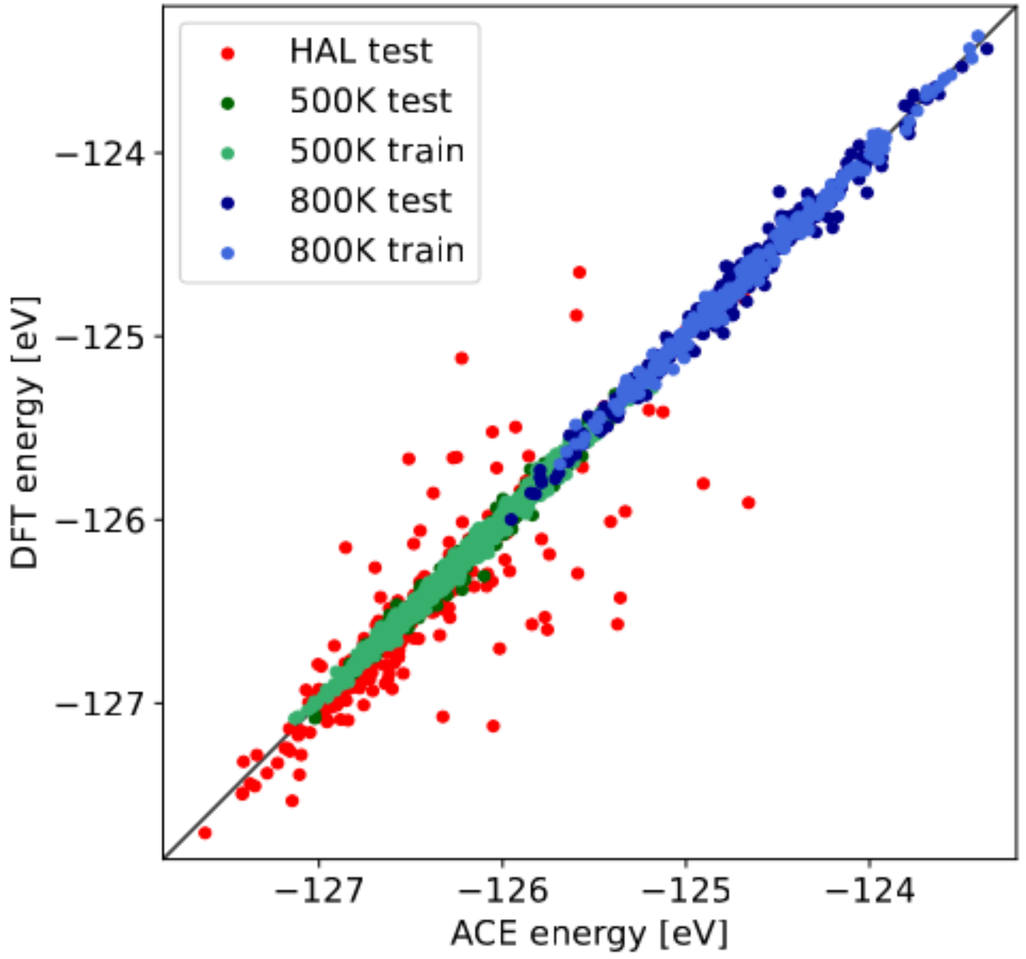
- Initial 20 configuration PEG-2 database ($C_4H_{10}O_2$)
- “Simple” ACE model (correlation order 3, polynomial degree 10)
- Benchmark AL vs HAL in achieving 1 million stable MD steps

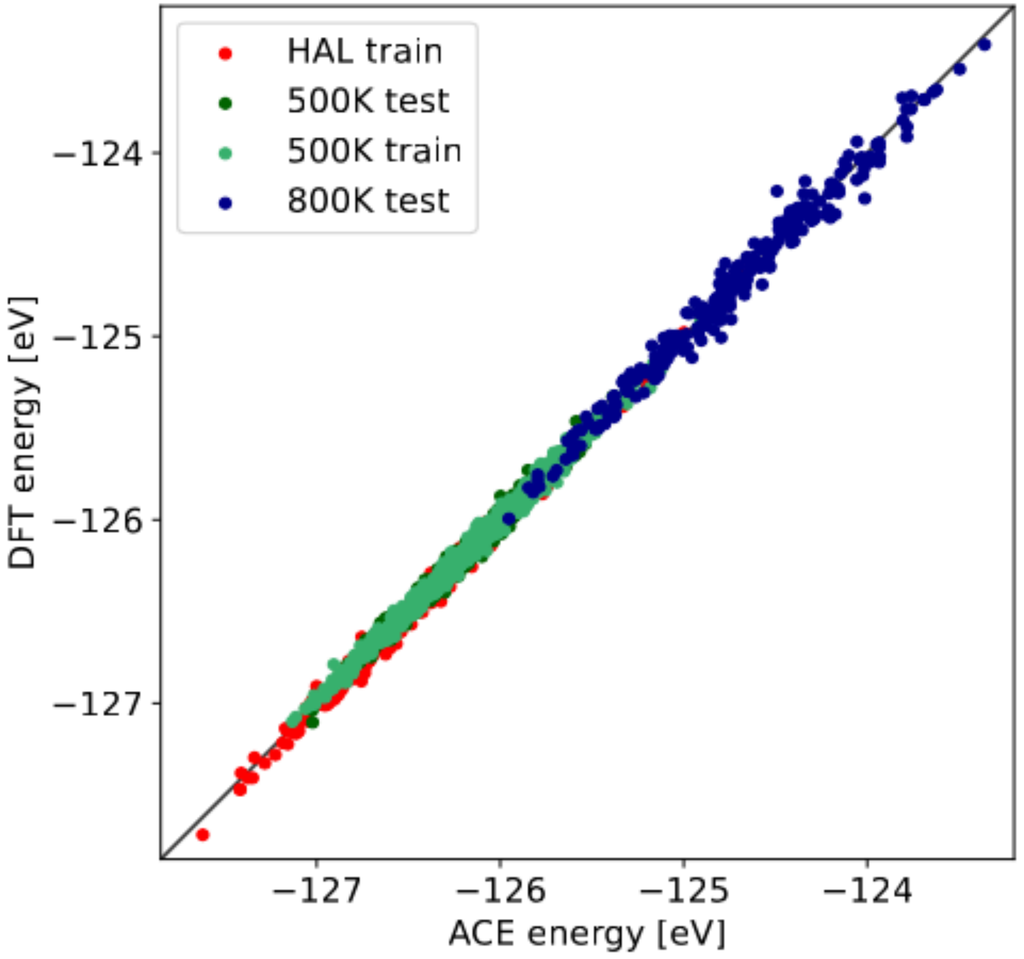
T	τ_r	Iterations	Timesteps
300K	0.0	33	4.67×10^6
	0.25	21	6.61×10^4
400K	0.0	30	1.66×10^6
	0.25	24	3.95×10^4

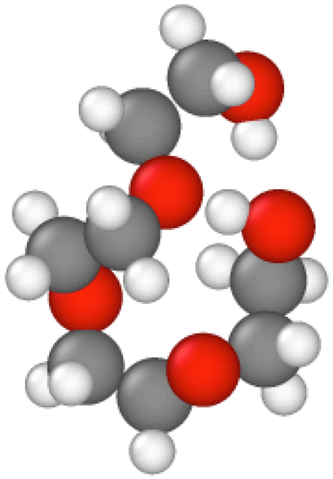
- HAL achieves MD stability faster than AL

HAL: PEG-4

- HAL finds large errors at 500K



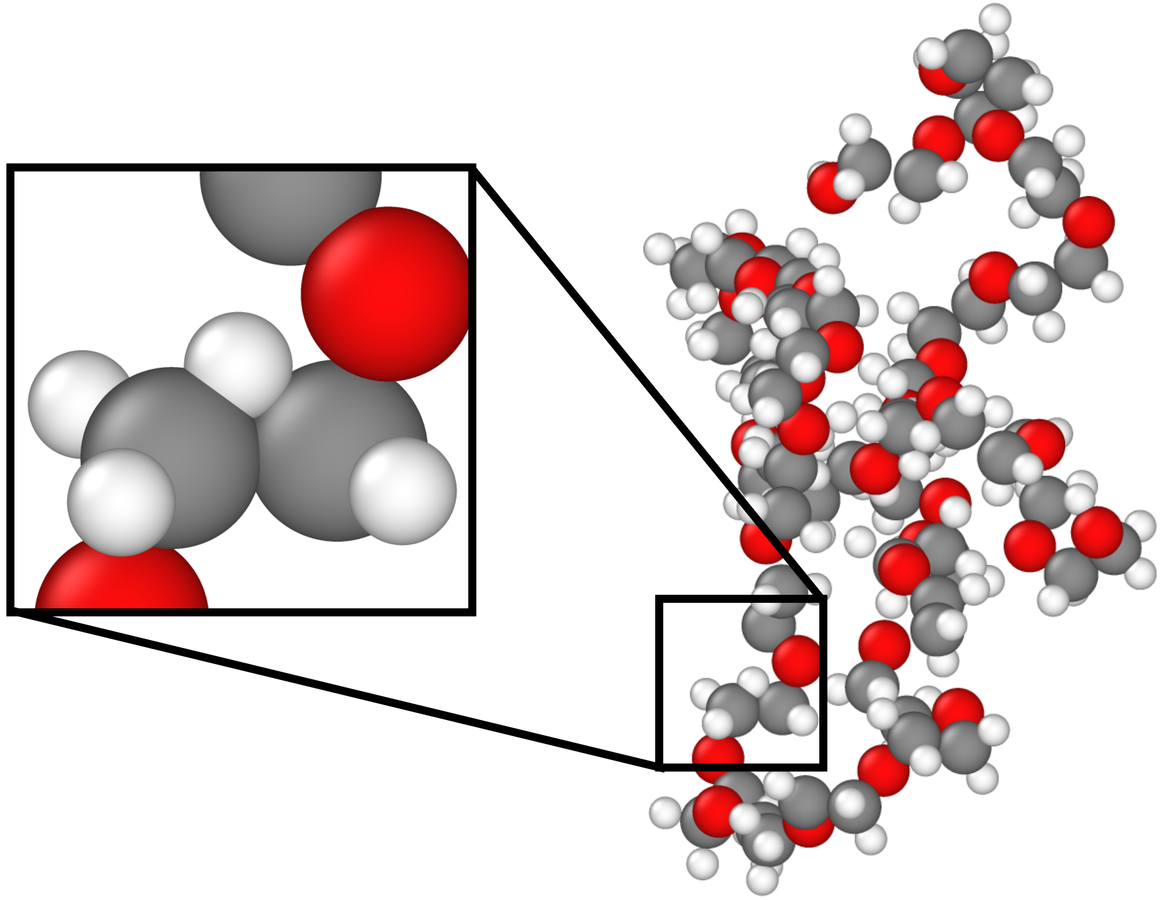




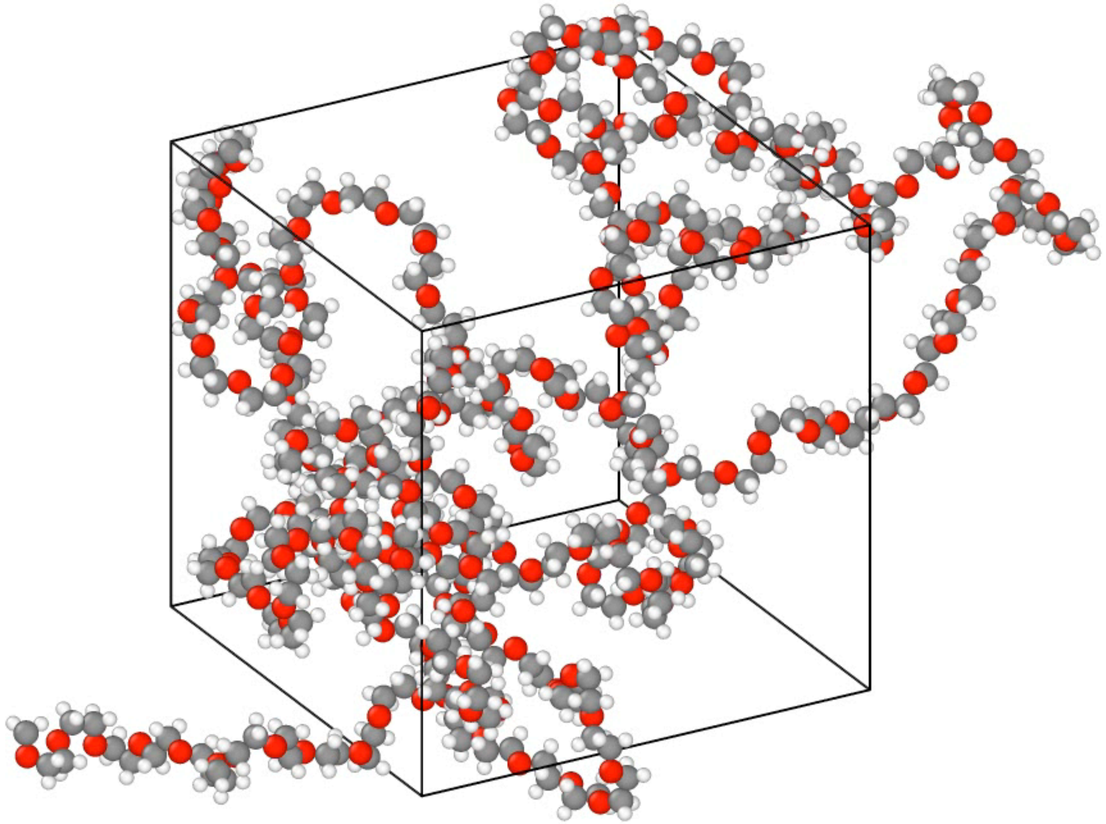
HAL config

HAL: PEG-n

- Add HAL DFT data up to PEG-32

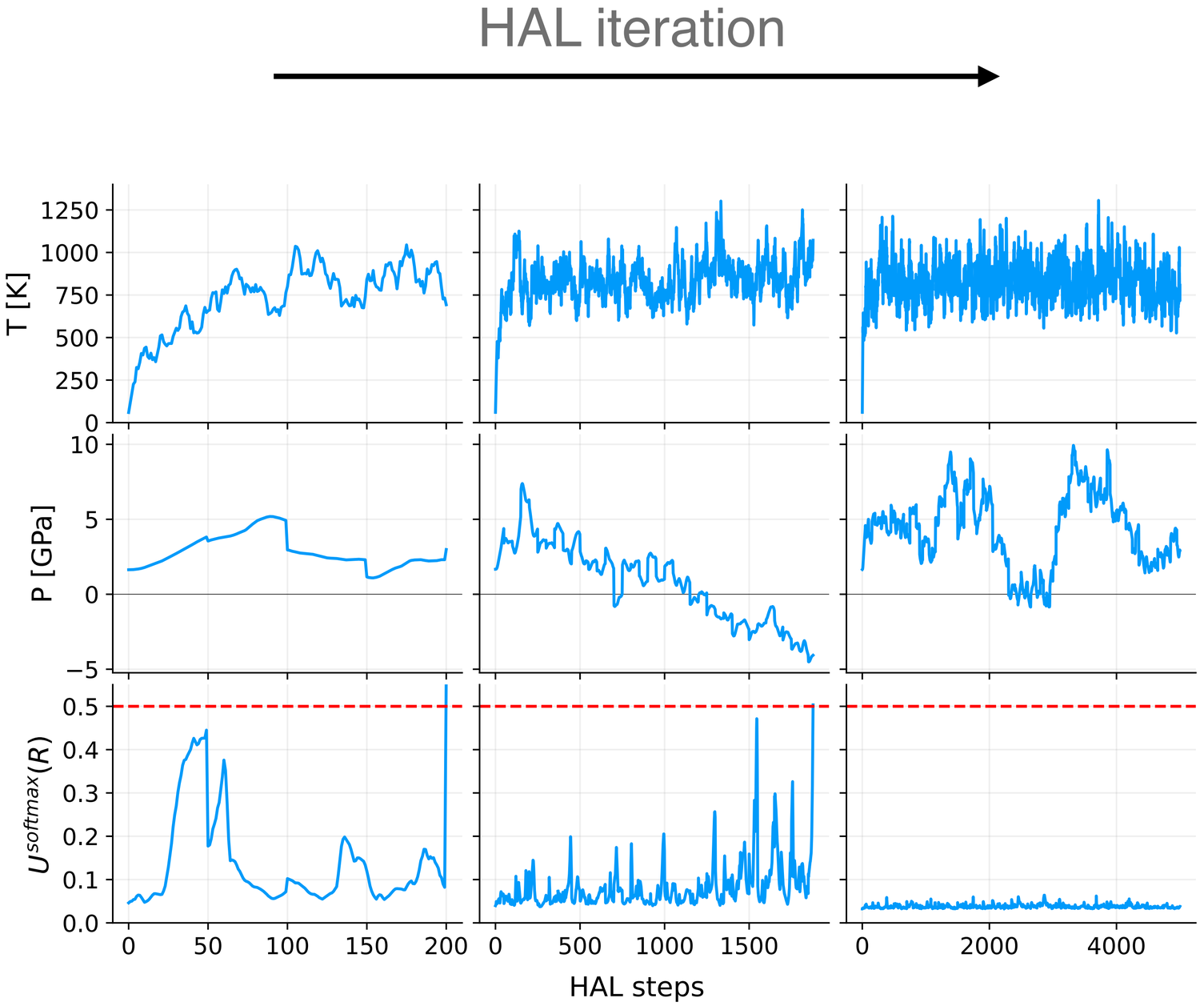


- Test: 700K PEG-200 MD (PBC)

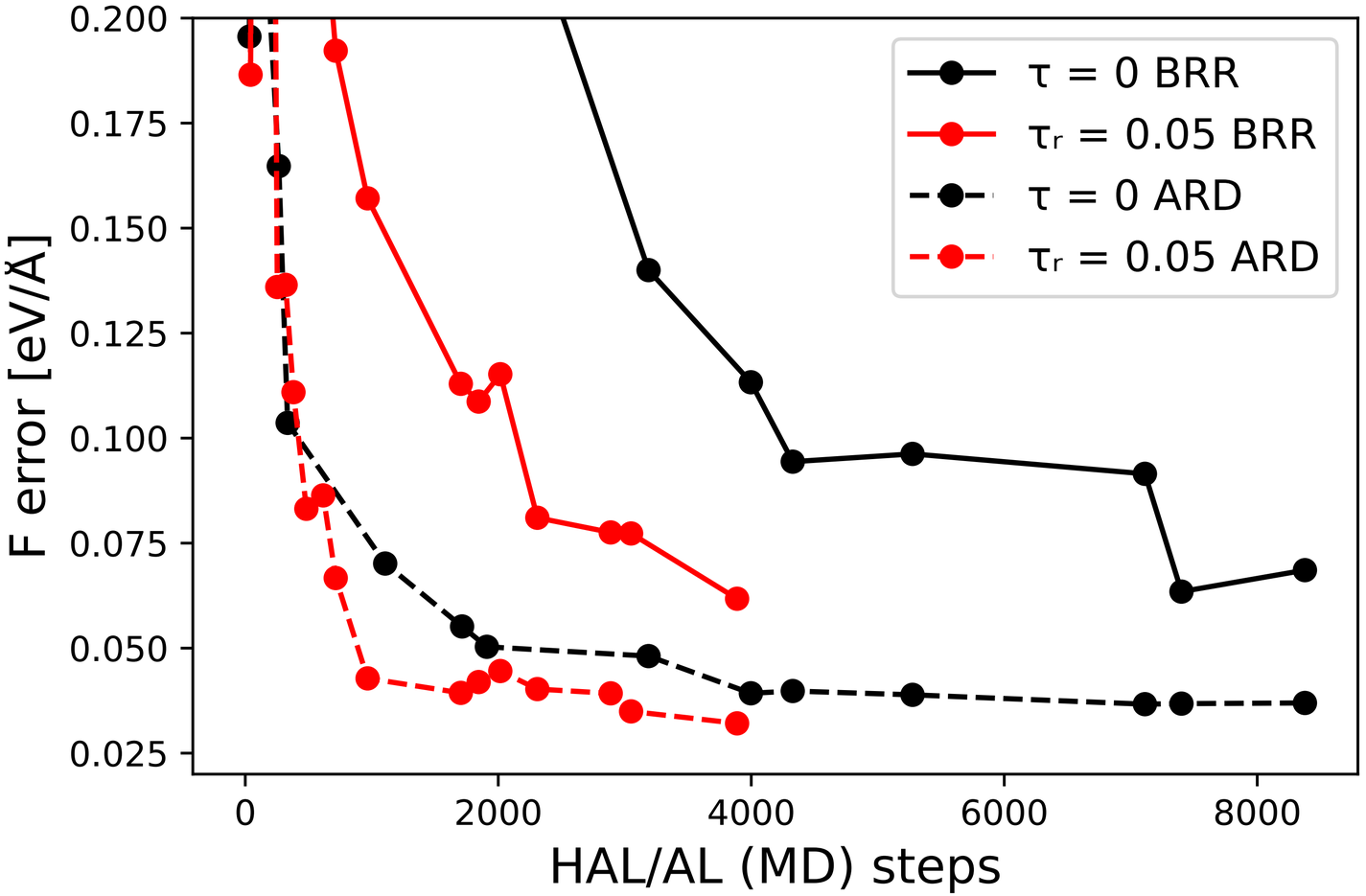


HAL: AlSi10

- Initial database (32-atom):
  - 10 random alloys
- 32 committee members
- Random alloy HAL:
  - swap/volume MC steps
  - thermostat at 800K
- Liquid HAL:
  - thermostat at 3000K
  - barostat at 0.1 GPa



HAL: AlSi10 random alloy

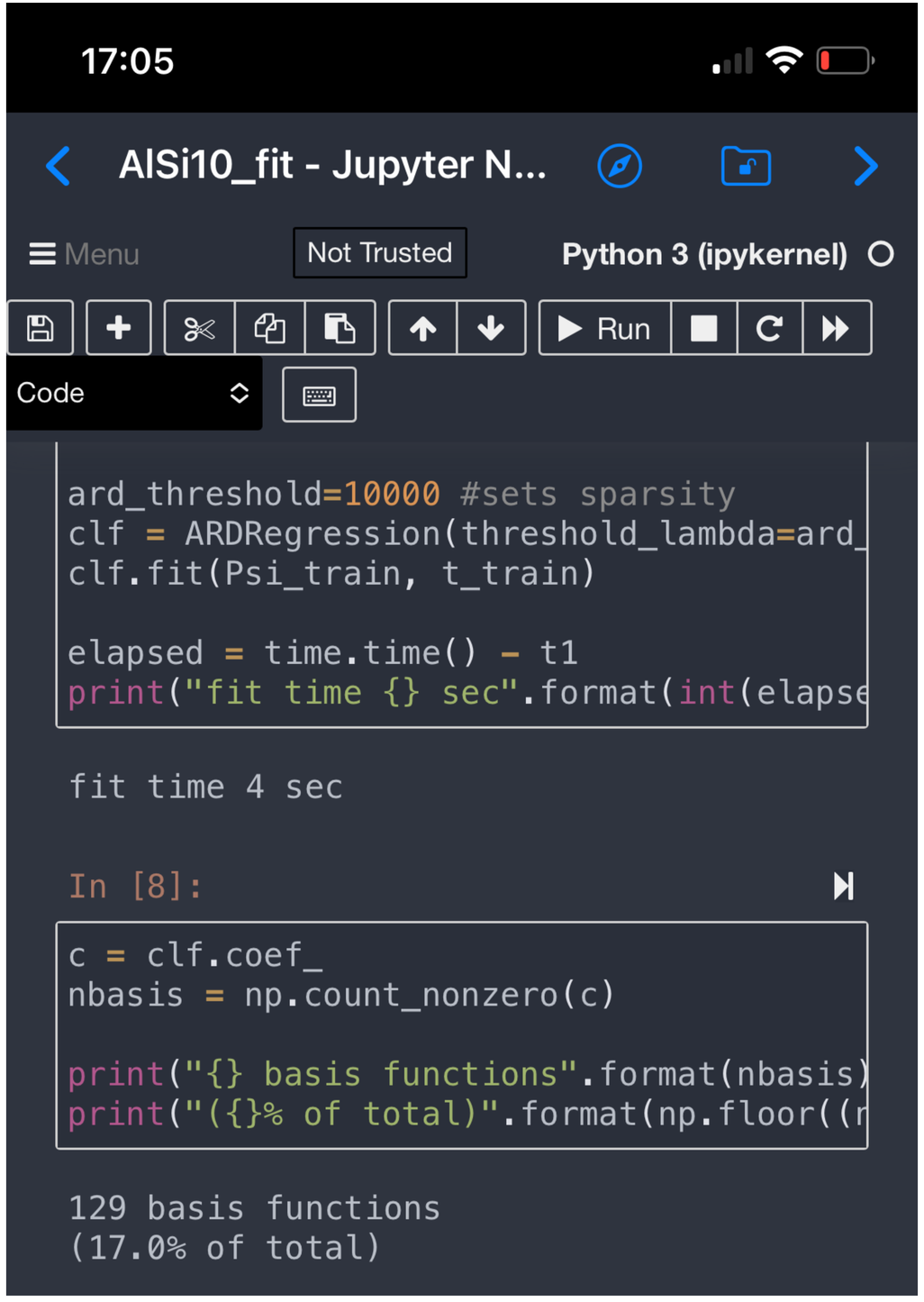


- 800K alloy HAL test database
- Test set error convergence
- Accelerate learning curve:
    - biasing (HAL)
    - sparsity (ARD)

AlSi10

- HAL database (32-atom):
  - random: 42 configs
  - liquid: 46 configs
- Design matrix Ψ (9064 x 723)
- ARD fit (α' is relevance threshold)

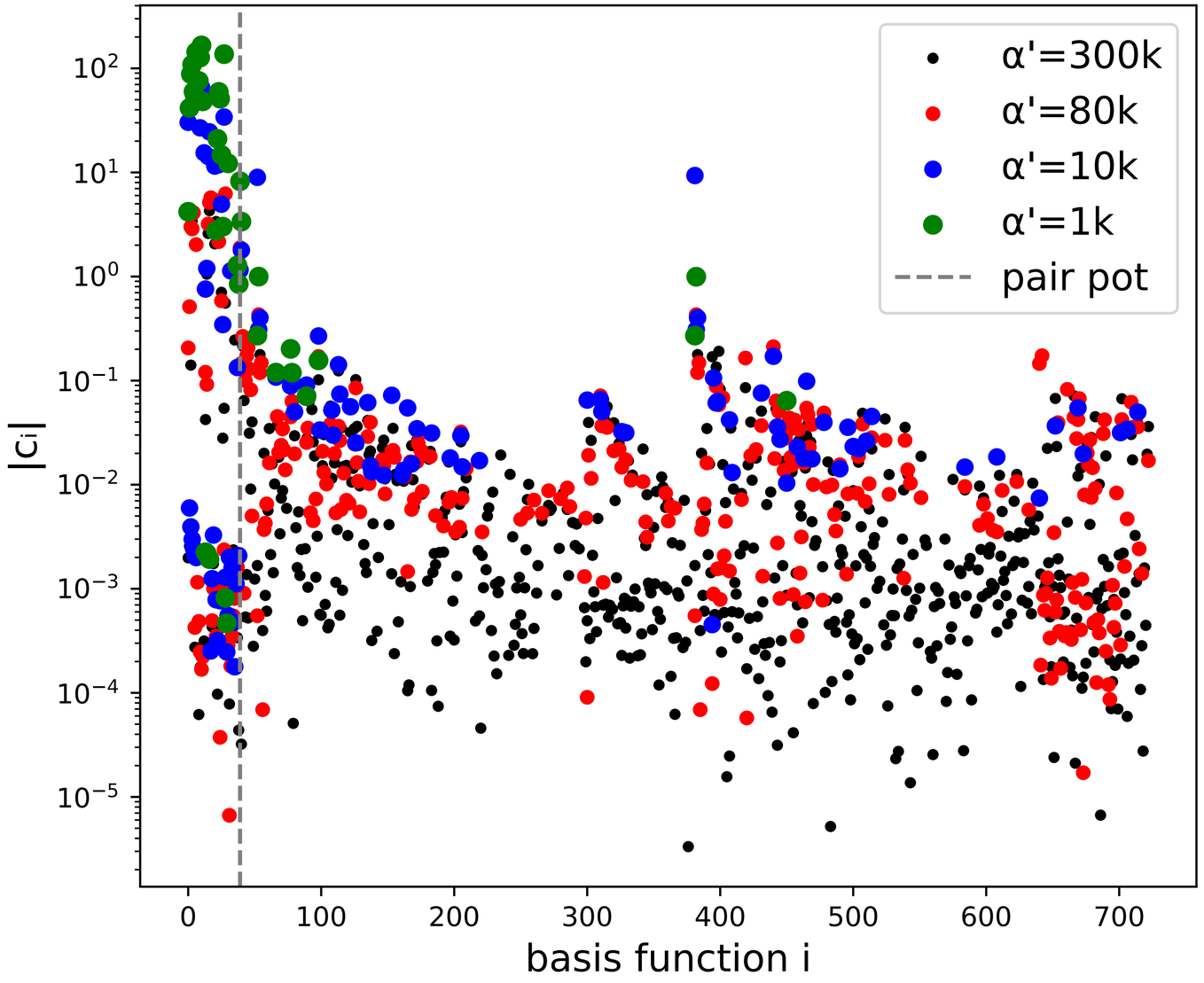
α'	N_{basis}	Fit (s)	Training Error E	Training Error F	Test Error E	Test Error F
1k	38	1.58	7.693	0.135	8.006	0.147
10k	116	3.16	4.199	0.095	6.229	0.104
80k	295	16.86	2.401	0.080	5.131	0.089
300k	621	62.16	1.869	0.074	5.188	0.095



AlSi10

- HAL database (32-atom):
  - random: 42 configs
  - liquid: 46 configs
- Design matrix Ψ (9064 x 723)
- ARD fit (α' is relevance threshold)

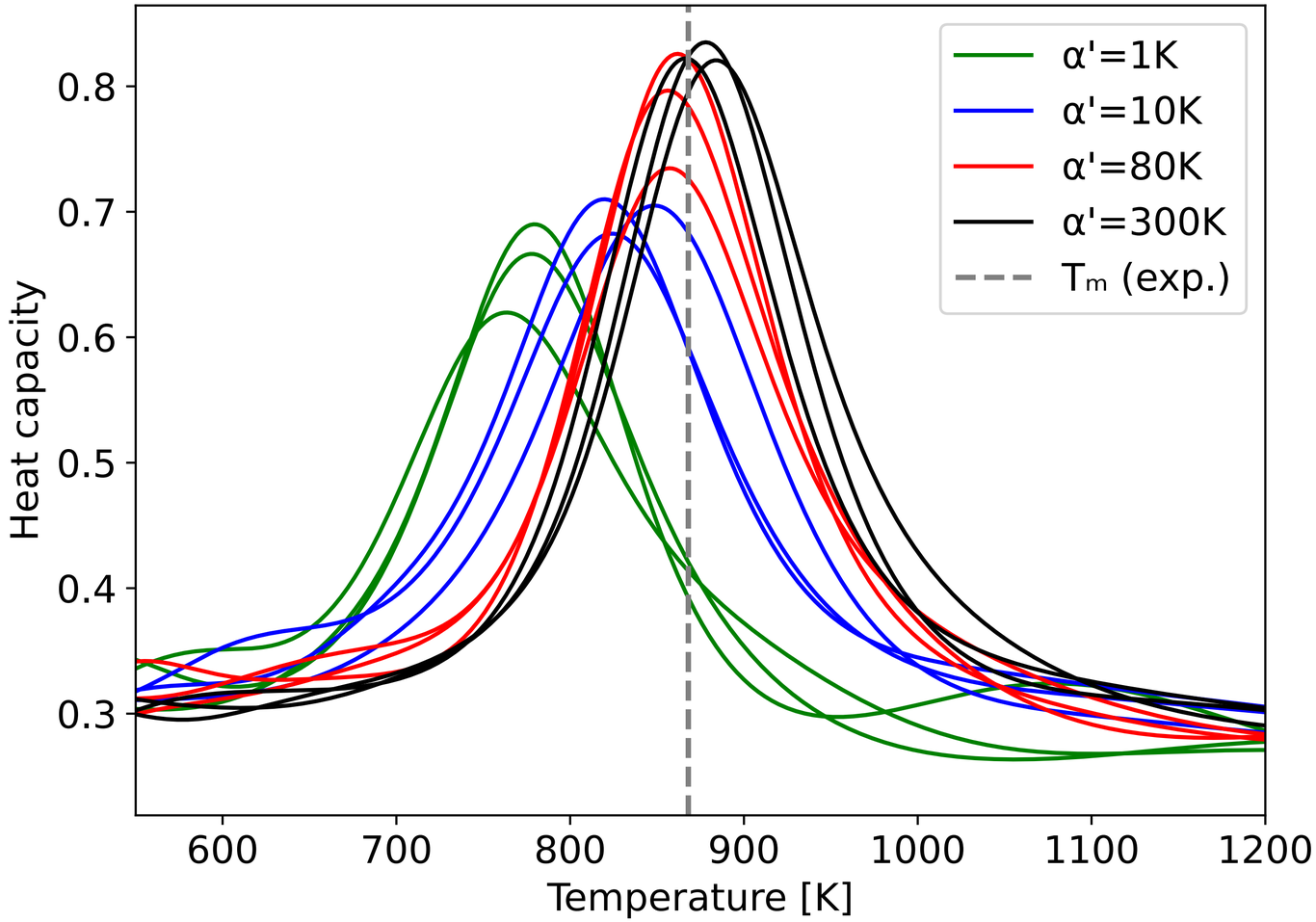
α'	N_{basis}	Fit (s)	Training Error E	Training Error F	Test Error E	Test Error F
1k	38	1.58	7.693	0.135	8.006	0.147
10k	116	3.16	4.199	0.095	6.229	0.104
80k	295	16.86	2.401	0.080	5.131	0.089
300k	621	62.16	1.869	0.074	5.188	0.095



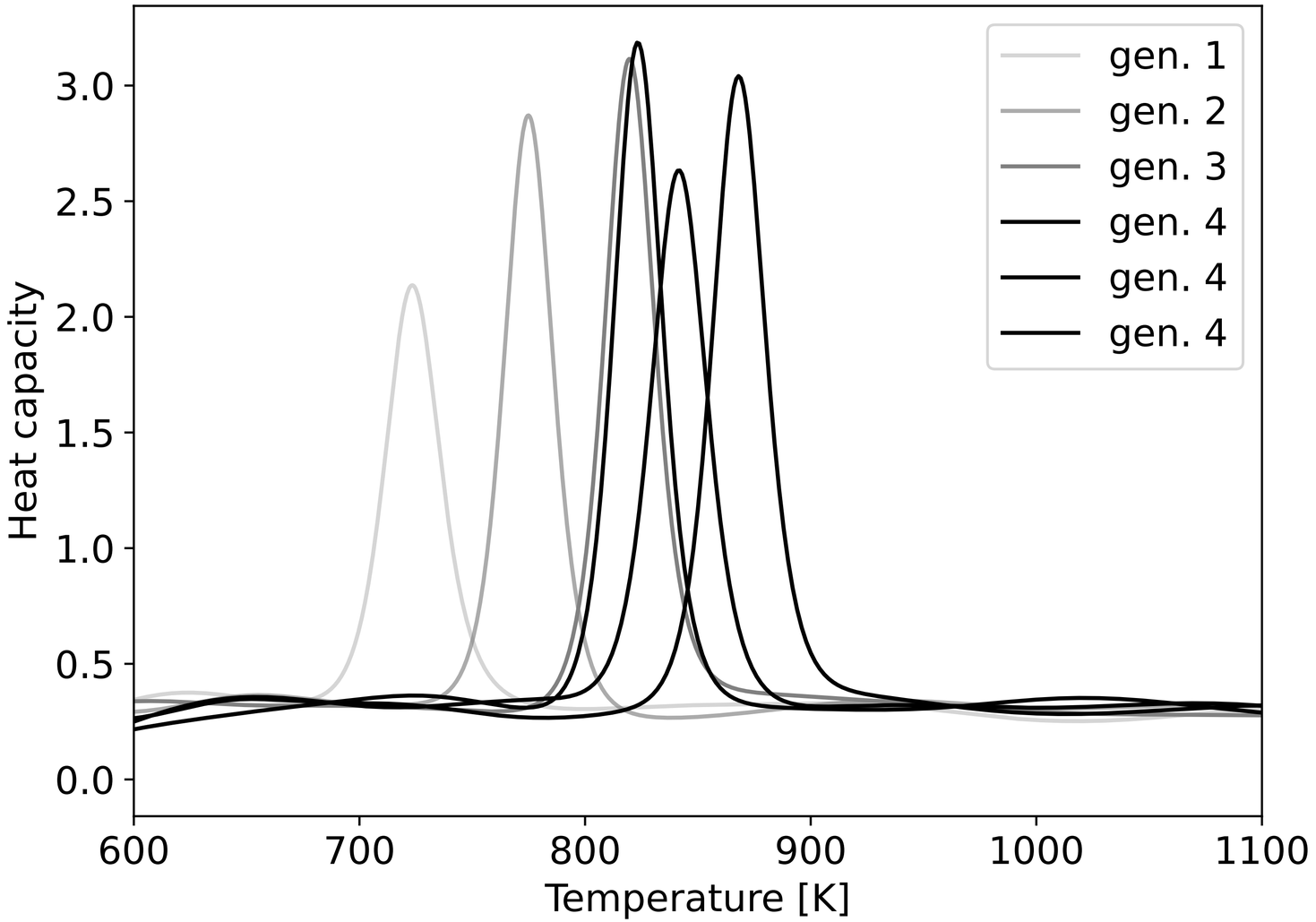
AlSi10

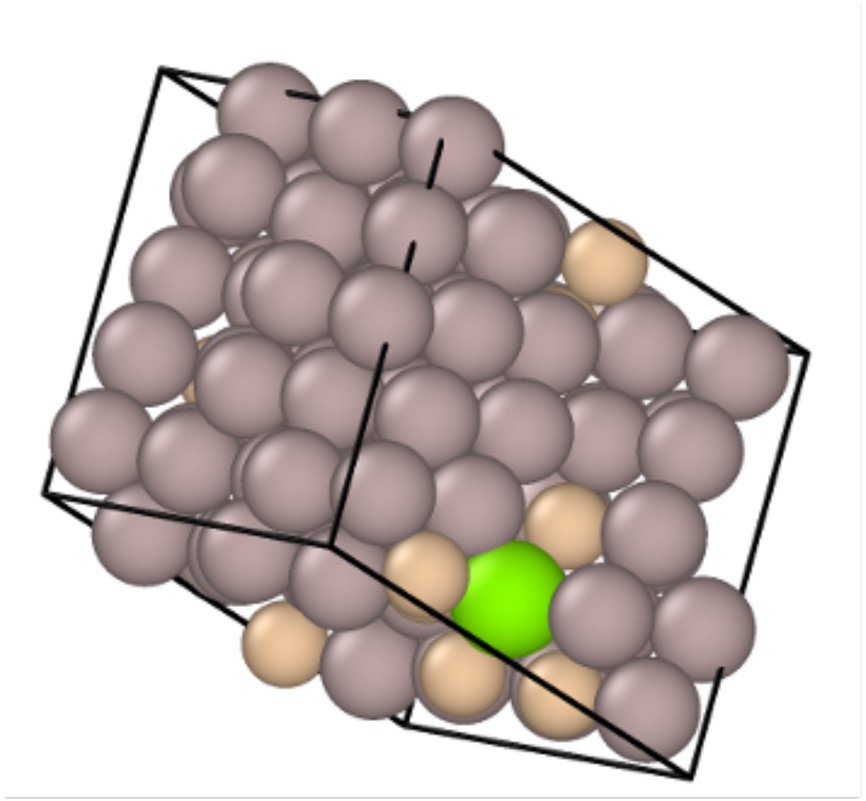
- HAL database (32-atom):
  - random: 42 configs
  - liquid: 46 configs
- Design matrix Ψ (9064 x 723)
- ARD fit (α' is relevance threshold)

α'	N_{basis}	Fit (s)	Training Error E	Training Error F	Test Error E	Test Error F
1k	38	1.58	7.693	0.135	8.006	0.147
10k	116	3.16	4.199	0.095	6.229	0.104
80k	295	16.86	2.401	0.080	5.131	0.089
300k	621	62.16	1.869	0.074	5.188	0.095



AlSi10Mg

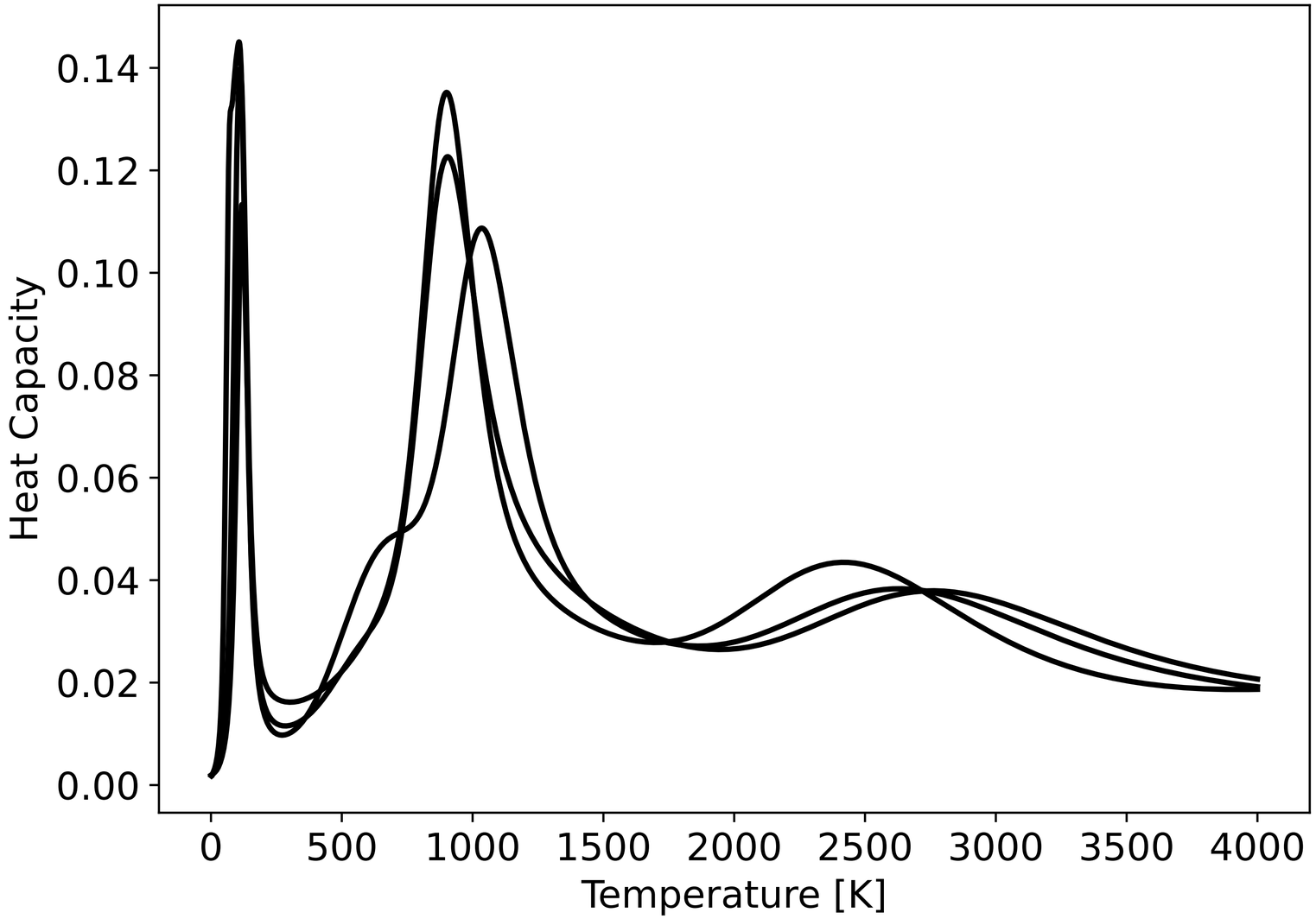




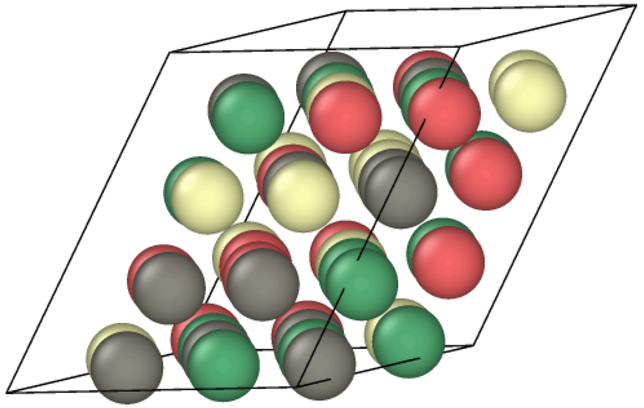
Guinier-Preston zone

T_m is 868K

MoNbTaW

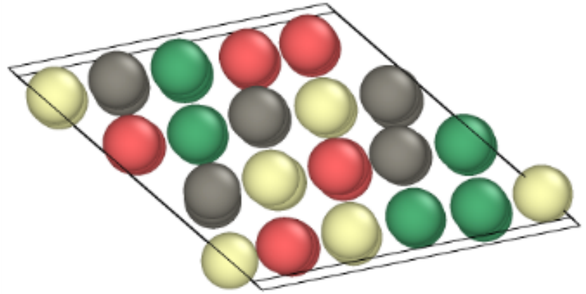


- Alloy properties
- HAL database:
  - 97 random (bcc) alloy
  - 151 liquid alloy
- Order/disorder transitions

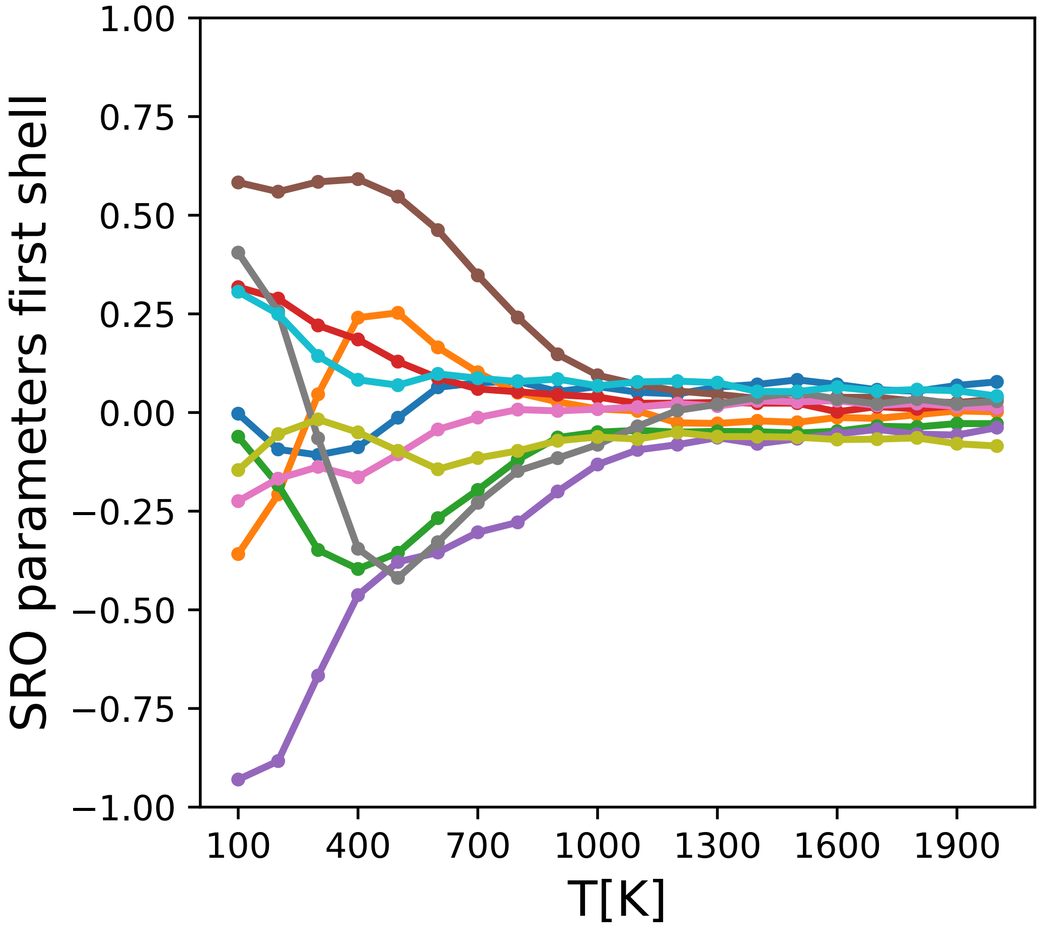


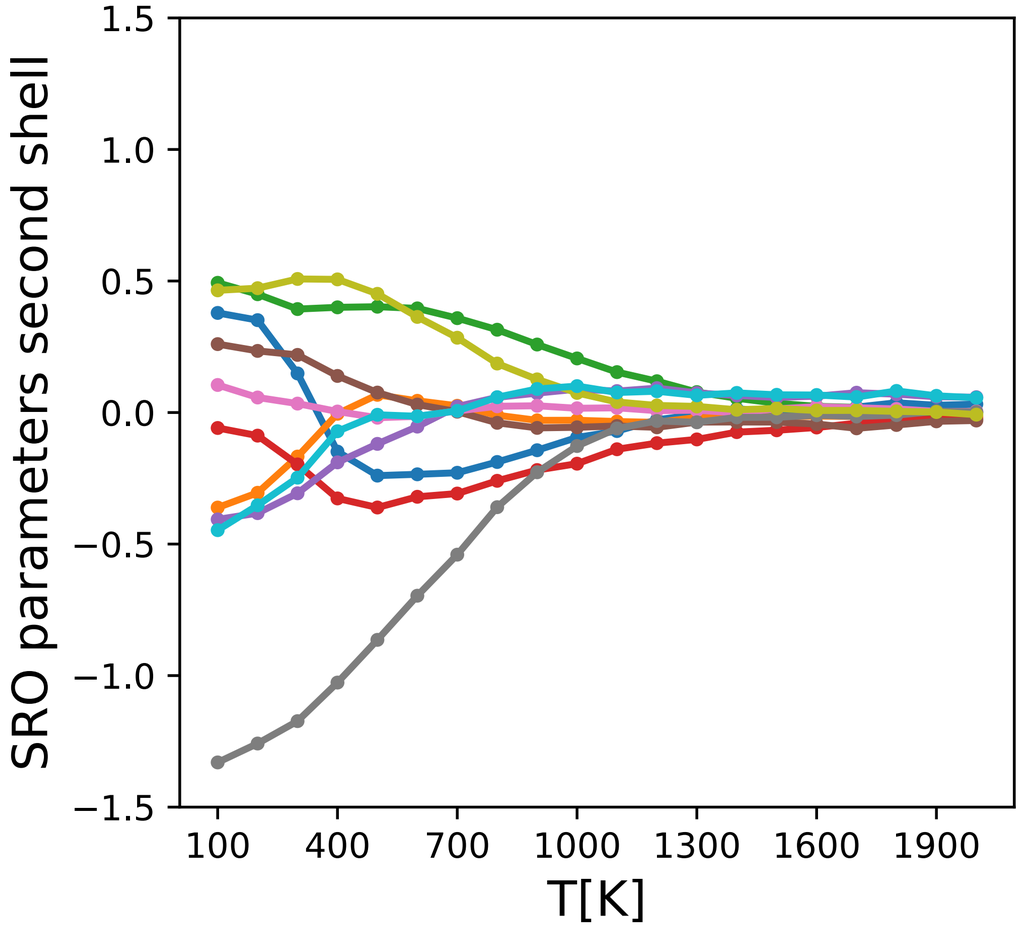
disordered structure

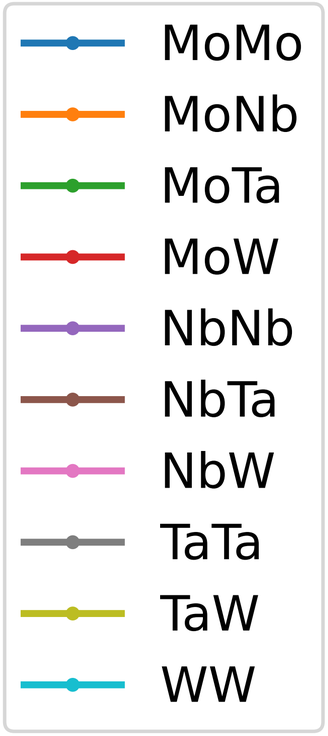
MoNbTaW



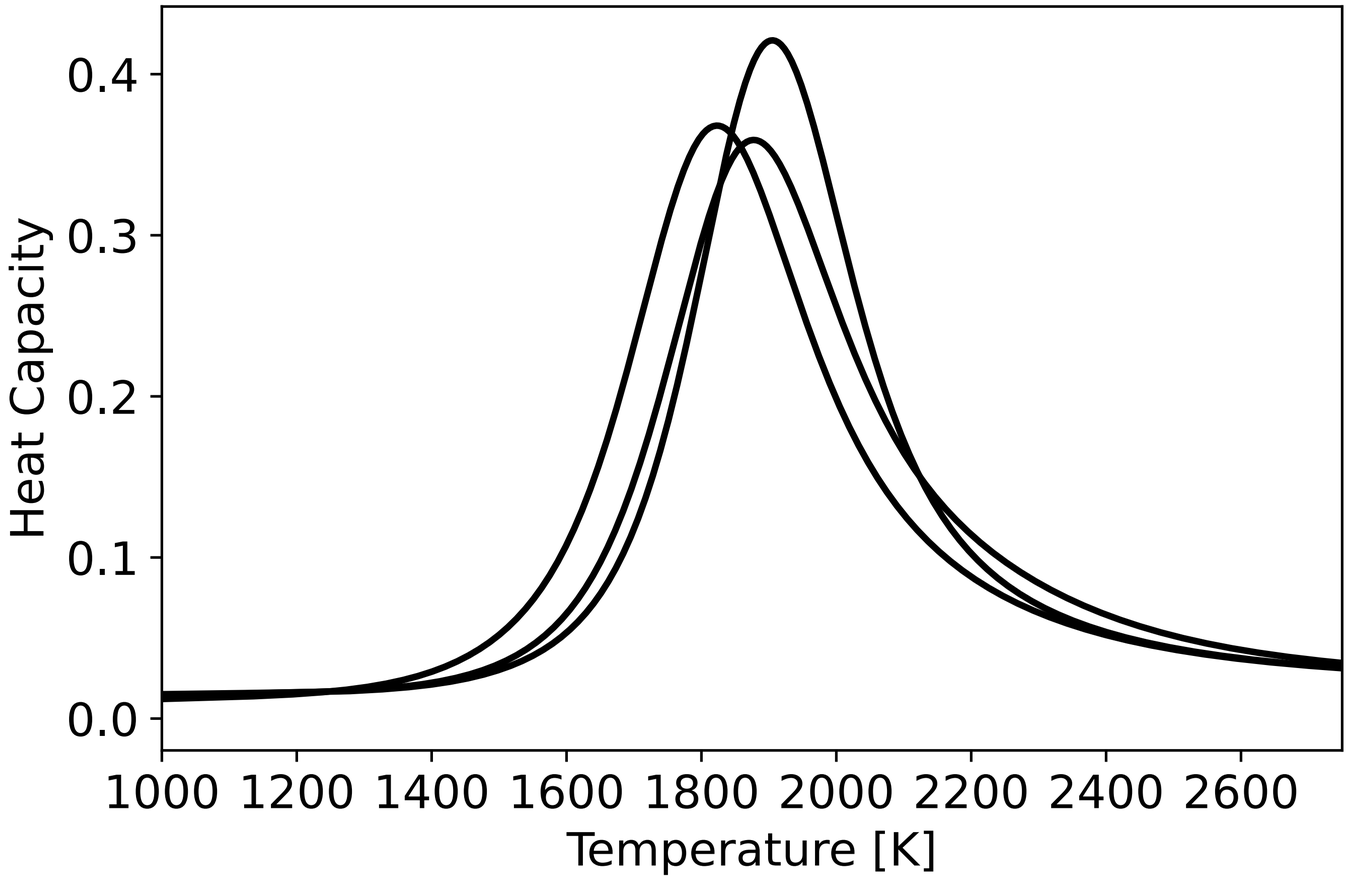
ordered structure







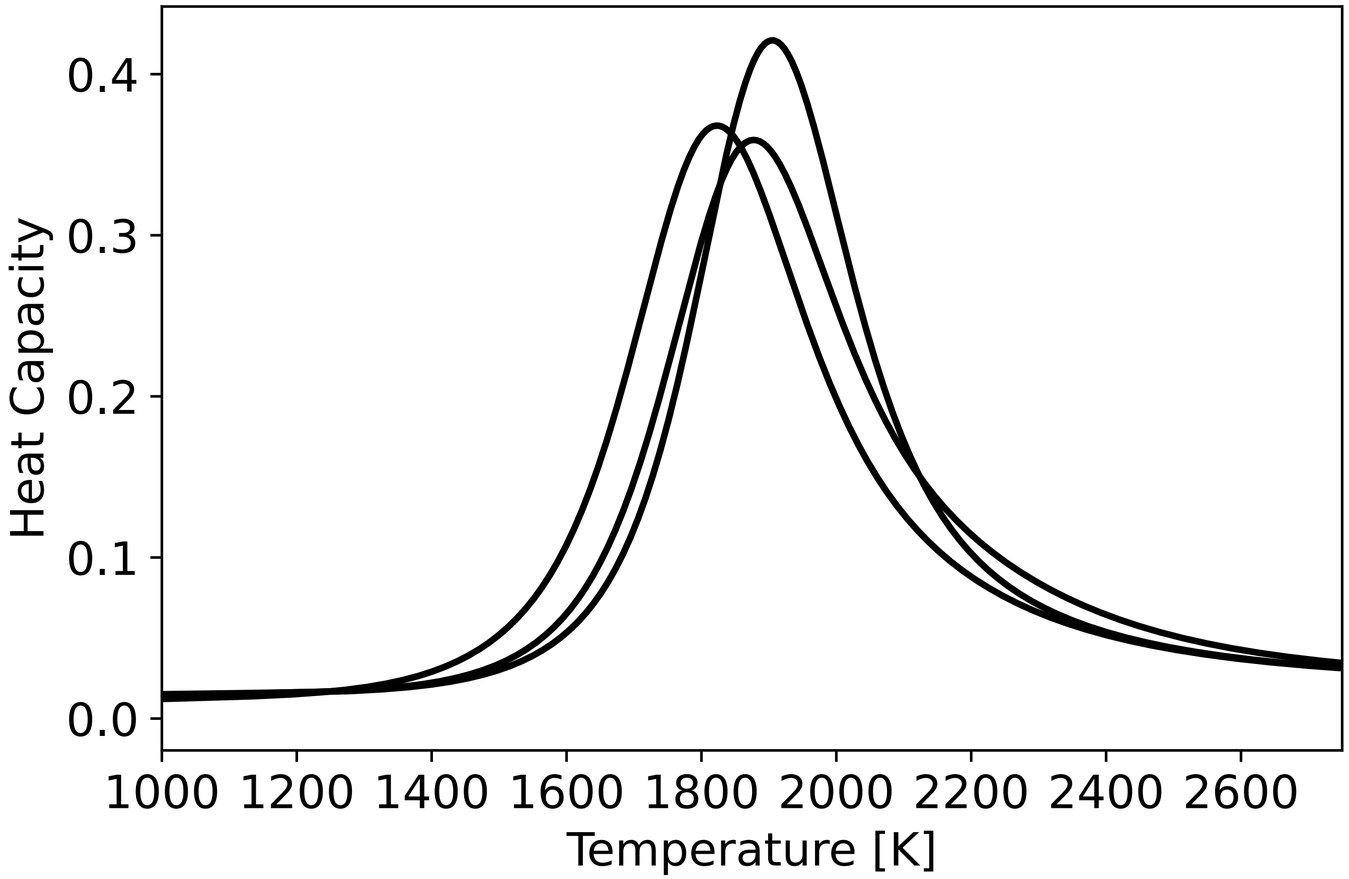
FeCrMnNiCo

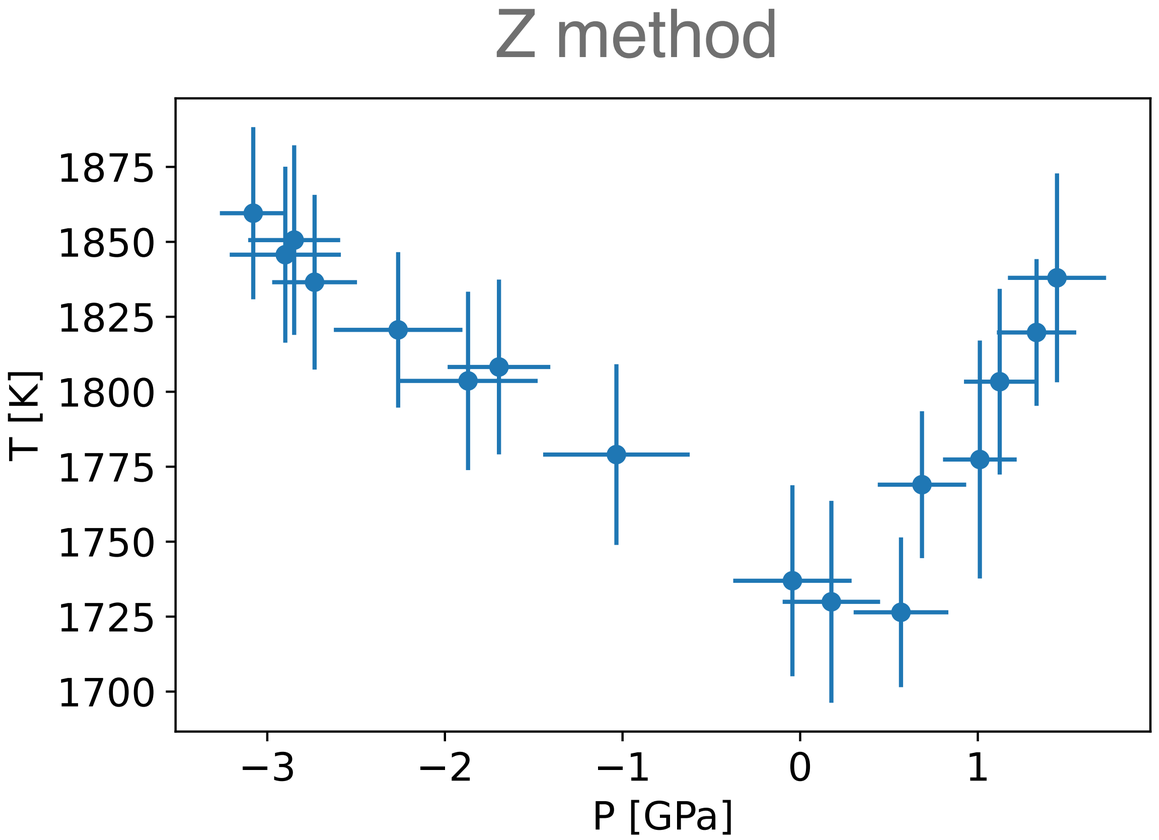


- HAL database (50-atom):
  - 75 (equiatomic/fcc) random
  - 168 liquid
- Total: 243 configs (~12K envs)
- Fitting: 2:34h (ARD)
- 13K basis functions (from 18K)
- Performance: 2.37 ms/atom
- Training Errors:

configuration type	E [meV/atom]	F [eV/Å]	V [meV]
FeCrMnNiCo liquid	11.29	0.323	130.03
FeCrMnNiCo random	6.17	0.244	79.82

FeCrMnNiCo





- Experimental T_m is ~1350C or 1623K

Conclusion

- Linear ACE models
- Sample committee members $\{\mathbf{c}_j\}_{j=1}^{k}$ from posterior
- Cheap committee evaluations
- Relative force uncertainty metric $U_i(R^*)$
- Reduce existing database using greedy algorithm
- HAL: bias MD towards uncertainty
- Softmax for increased sensitivity on emerging uncertainty
- Build minimal databases automatically and efficiently for polymers and alloys

Sources

- [1] Ralf Drautz, Atomic cluster expansion for accurate and transferable interatomic potentials, Phys. Rev. B, Condens. Matter, 99 (1), 2019
- [2] Albert P. Bartók, James Kermode, Noam Bernstein, and Gábor Csányi, Machine Learning a General-Purpose Interatomic Potential for Silicon, Phys. Rev. X **8**, 041048, 2018