


# Recent progress in icet and hiphive: Force constant and cluster expansions in Python

Paul Erhart

F. Eriksson, M. Ångqvist, E. Fransson,
J. M. Rahm, P. Tanner, J. Brorsson (Chalmers, Gothenburg)

P. Rozycko, T. H. Rod (ESS-DMSC, Copenhagen)

A. Hashemi, Z. Fan (Aalto, Finland)

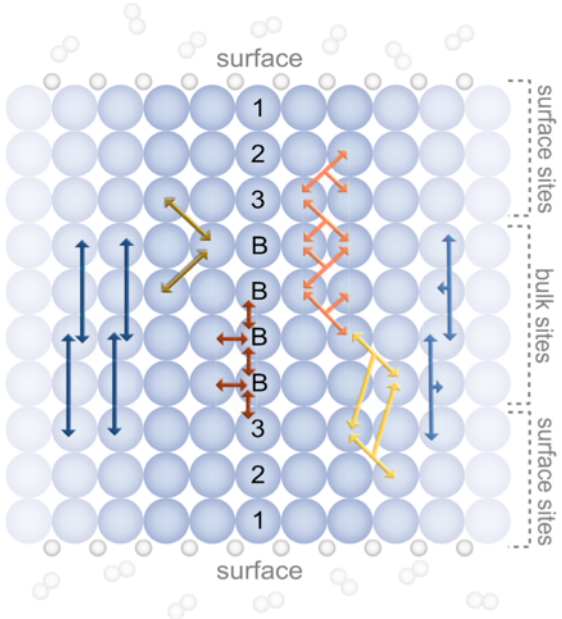
# Game plan

icet

**Cluster expansions (CEs)**

→ Dealing with strain
- Constituent strain formalism

→ Dealing with low symmetry systems
- Merging orbits
- Bayesian regression

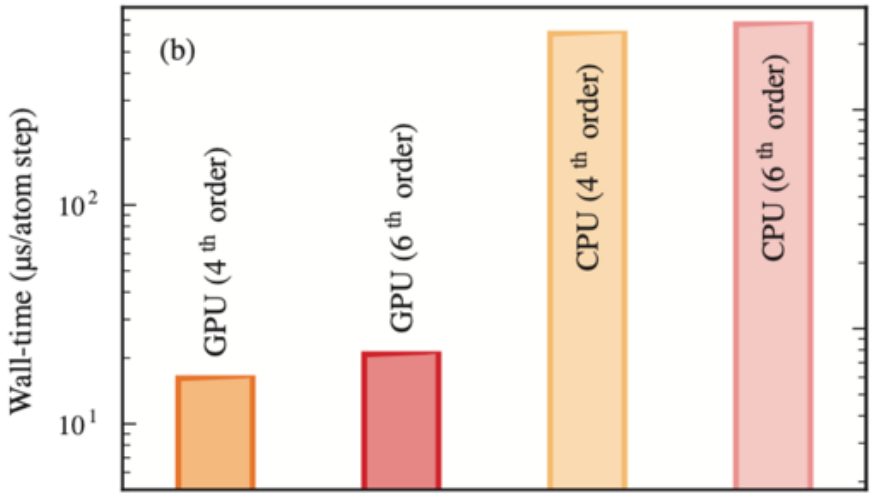



hiphive

**Force constant (FC) expansions**

→ Efficacy of regression schemes

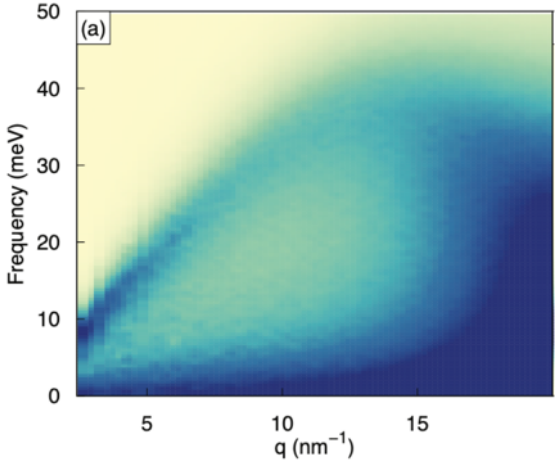
→ GPU-accelerated MD with ab-initio accuracy



# Prelude: Other packages related to ase

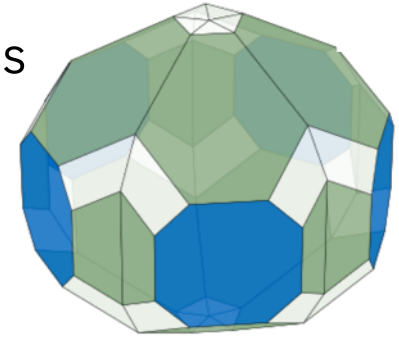
dynasor

Dynamical structure factors and correlation functions




wulffpack

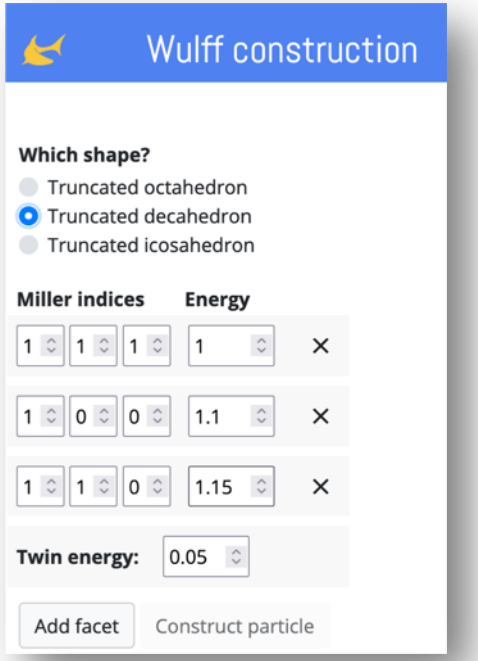
Wulff constructions



sharc

Web applications

- Special quasi-random structures
- Wulff constructions
- and more




https://materialsmodeling.org/software

icet

# Alloy cluster expansions

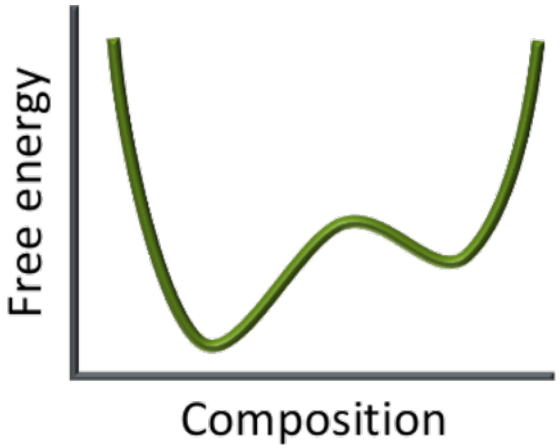
# Alloy cluster expansions

Expansion of the total energy in site occupations

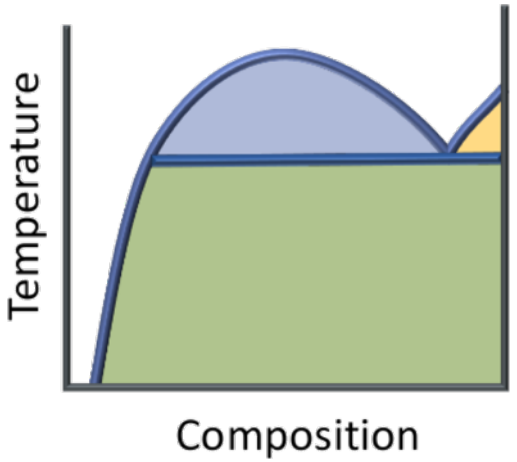
$$E(\boldsymbol{\sigma}) = E_0 + \sum_i J_i \sigma_i + \sum_{ij} J_{ij} \sigma_i \sigma_j + \sum_{ijk} J_{ijk} \sigma_i \sigma_j \sigma_k + \dots$$

$\boldsymbol{\sigma}$ Occupation vector

Free energies



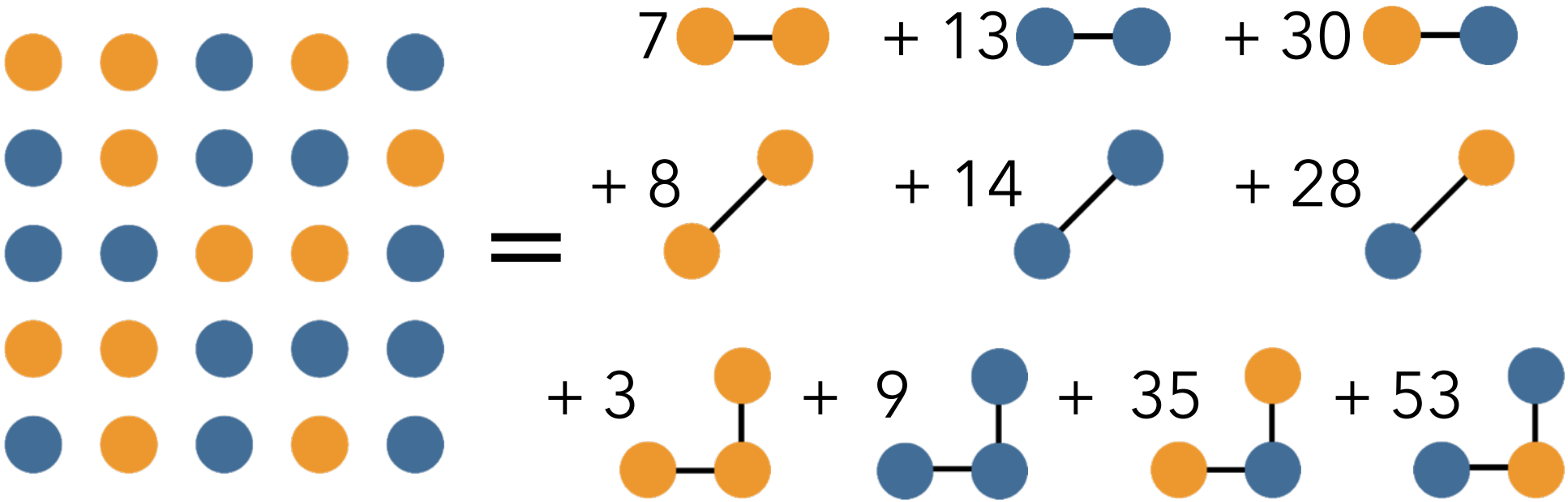

Phase diagrams




Extracting the effective cluster interactions (ECIs) $J_\alpha$ is a linear problem

$$E(\boldsymbol{\sigma}) = E_0 + \sum_\alpha m_\alpha J_\alpha \left\langle \Pi_\alpha(\boldsymbol{\sigma}) \right\rangle_{\alpha'}$$

# Finding the ECIs is a linear problem




+ longer-ranged pairs, triplets, quadruplets etc

Cluster vector → = [1.0, 0.12, –0.2, –0.12, –0.12, …]

Constant

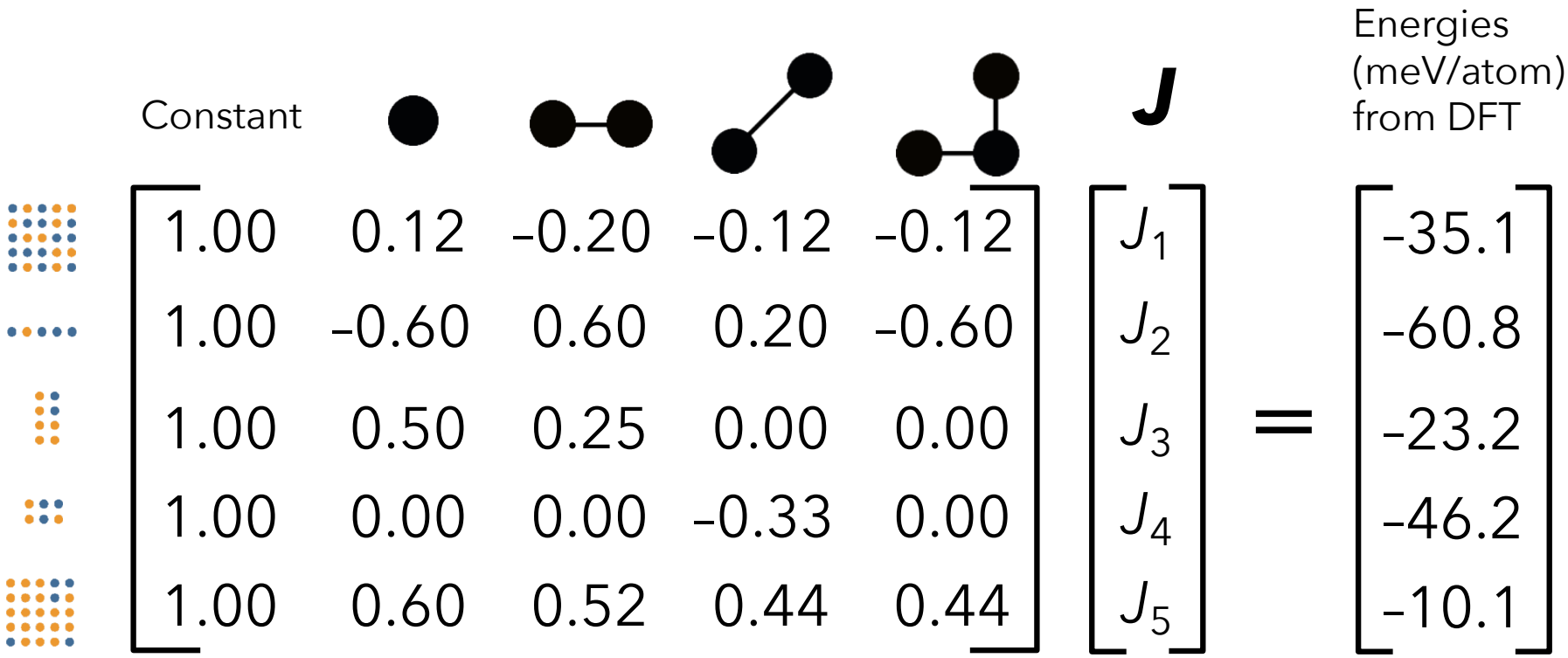
Concentration

Nearest neighbors

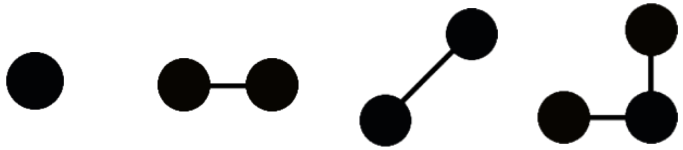
2nd nearest neighbors

First triplet

# Finding the ECIs is a linear problem



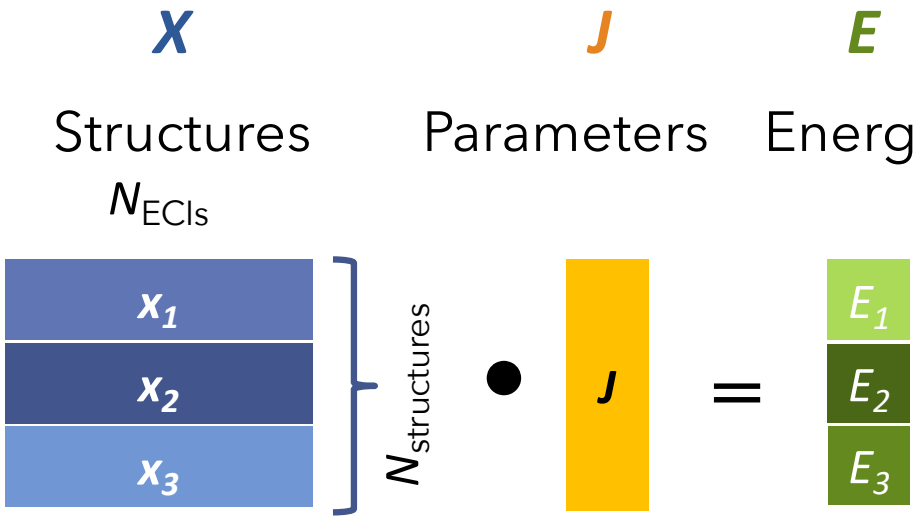

Solution: $\boldsymbol{J}^{\mathsf{T}} = [-37.8, 34.1, -9.9, 25.4, 2.67]$



# Finding the ECIs is a linear problem

## General procedure

1. Compile structures into one fit matrix ***X***
2. Solve the linear problem ***X J*** = ***E***




## How?

Least-squares (OLS)

$$\boldsymbol{J}_{\text{opt}} = \arg\min_{\boldsymbol{J}} \|\boldsymbol{X}\boldsymbol{J} - \boldsymbol{E}\|_2$$

LASSO

$$\boldsymbol{J}_{\text{opt}} = \arg\min_{\boldsymbol{J}} \|\boldsymbol{X}\boldsymbol{J} - \boldsymbol{E}\|_2 + \lambda_1 \|\boldsymbol{J}\|_1$$

Ridge regression

$$\boldsymbol{J}_{\text{opt}} = \arg\min_{\boldsymbol{J}} \|\boldsymbol{X}\boldsymbol{J} - \boldsymbol{E}\|_2 + \lambda_2 \|\boldsymbol{J}\|_2$$

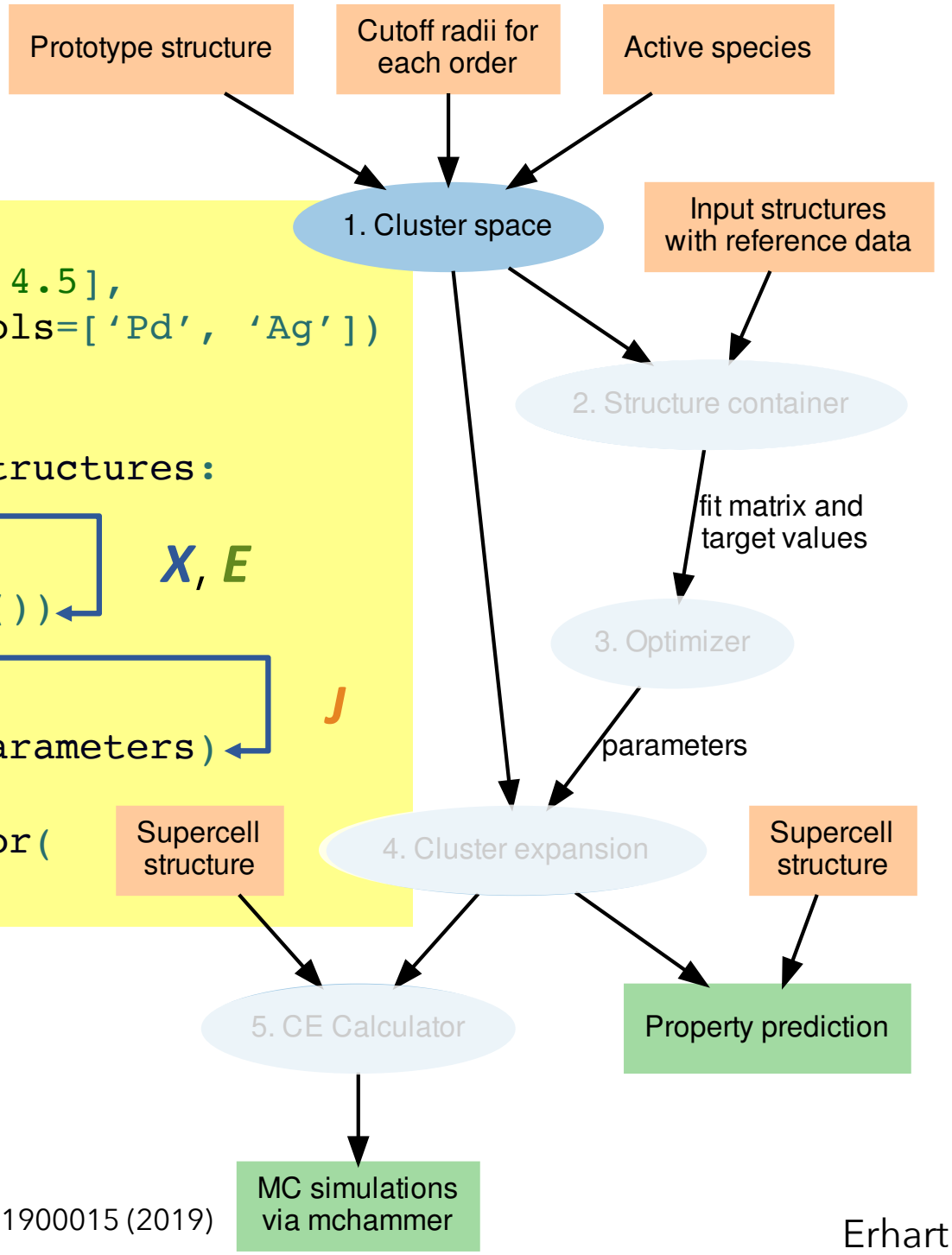
Bayesian

$$p\left(\boldsymbol{E}|\boldsymbol{J}, \boldsymbol{X}, \sigma^2\right) \propto \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{X}\boldsymbol{J} - \boldsymbol{E})^2\right)$$


[1] Mueller & Ceder, PRB **80**, 024103 (2009)
[2] Ångqvist *et al.*, Adv. Sim. Theo. **2**, 1900015 (2019)

# icet Workflow

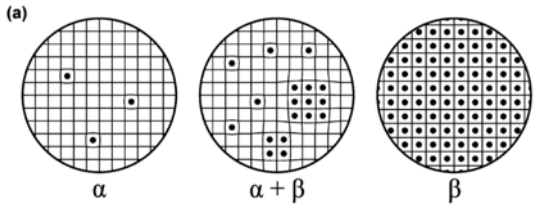
```
cs = ClusterSpace(ideal_cell,
                  cutoffs=[6.0, 4.5],
                  chemical_symbols=['Pd', 'Ag'])

sc = StructureContainer(cs)
for atoms in list_of_training_structures:
    sc.add_structure(atoms)

opt = Optimizer(sc.get_fit_data())
opt.train()

ce = ClusterExpansion(cs, opt.parameters)

calc = ClusterExpansionCalculator(
           ce, supercell)
```

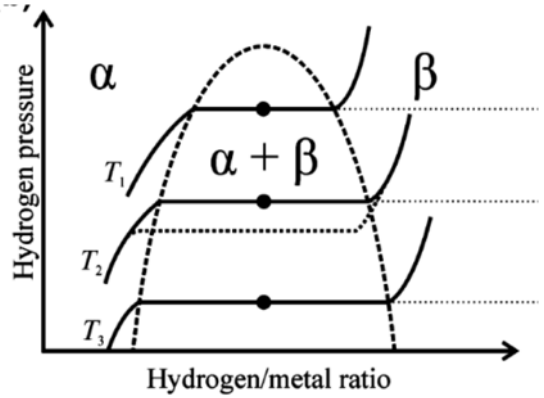




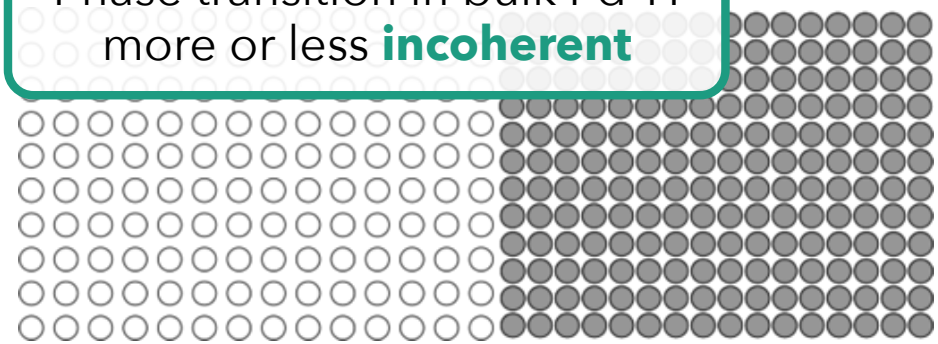

[1] Ångqvist *et al.*, Adv. Sim. Theo. **2**, 1900015 (2019)

# Coherent vs incoherent phase equilibrium



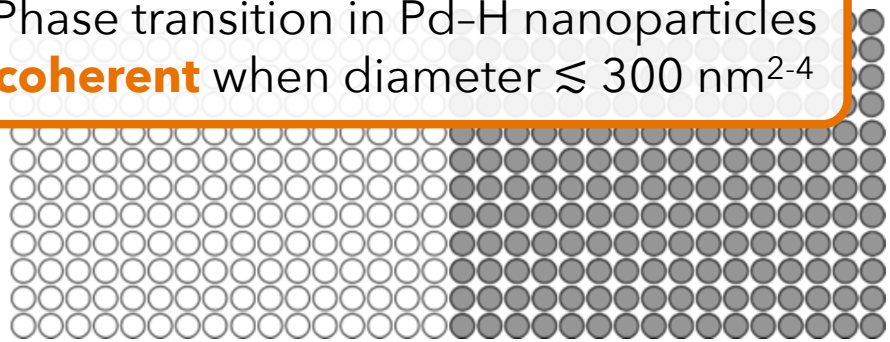

Phase transition in bulk Pd–H more or less **incoherent**



$$F(c) = cF(1) + (1-c)F(0) + \text{interface}$$

Barrier to be overcome with **nucleation and growth**

Phase transition in Pd–H nanoparticles **coherent** when diameter $\lesssim$ 300 nm[2-4]



$$F(c) \approx cF(1) + (1-c)F(0) + \text{strain} + \text{interface}$$

Barrier scales with **volume** (macroscopic)[1]


[1] Schwarz and Khachaturyan, *Acta Materialia* **54**, 313 (2006).
[2] Griessen, Strohfeldt, and Giessen, *Nature Materials* **15**, 311 (2016).
[3] Ulvestad *et al.*, *Nature Materials* **16**, 565 (2017).
[4] Narayan *et al.*, *Nature Communications* **8**, 1 (2017).

# Dealing with strain: CE with constituent strain [1]

Split energy in short and long-ranged part

→ Better models for size-mismatched systems

short-ranged (chemistry) long-ranged (strain)

$$E(\sigma) = E_{\mathrm{CE}}(\sigma) + \sum_{k} \Delta E_{\mathrm{CS}}(\hat{\boldsymbol{k}}, c) F(\boldsymbol{k}, \sigma)$$

constituent strain energy

$$\Delta E_{\mathrm{CS}}(\hat{\boldsymbol{k}}, c) = \min_{a_{\mathrm{SL}}} \left[ (1-c) \Delta E^{\mathrm{epi}}_{\mathrm{Pd}}(a_{\mathrm{SL}}, \hat{\boldsymbol{k}}) + c \Delta E^{\mathrm{epi}}_{\mathrm{PdH}}(a_{\mathrm{SL}}, \hat{\boldsymbol{k}}) \right]$$

Parametrized using interface calculations

$$F(\boldsymbol{k}, \sigma) = |S(\boldsymbol{k}, \sigma)|^2 \, e^{-|\eta \boldsymbol{k}|^2} \Big/ 4c(1-c)$$

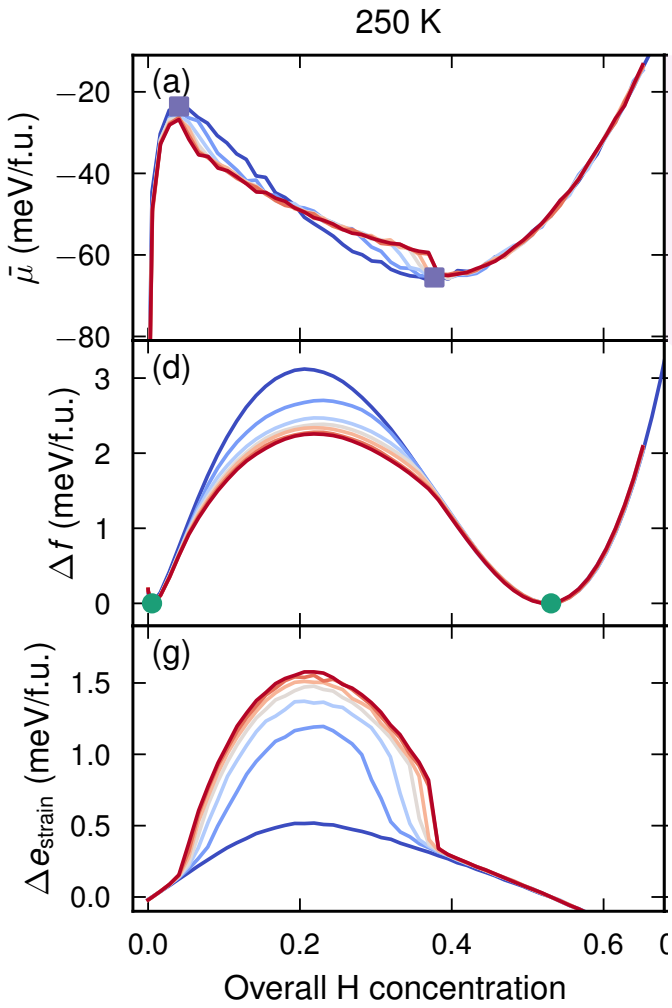
measure of the extent of configuration $\sigma$ matching a concentration wave with $\boldsymbol{k}$

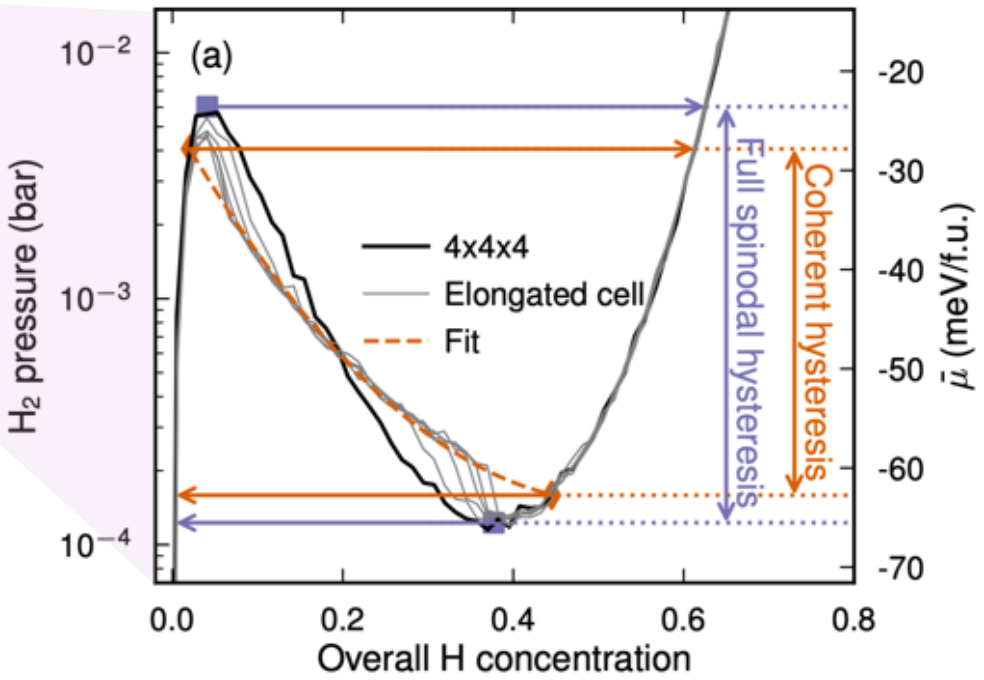


[1] Ozolins *et al.*, PRB **57**, 4816 (1998)

# How about sampling?

CE with strain combined with variance-constrained SGC ensemble [1]
for sampling system via MC as a **continuous function of composition**



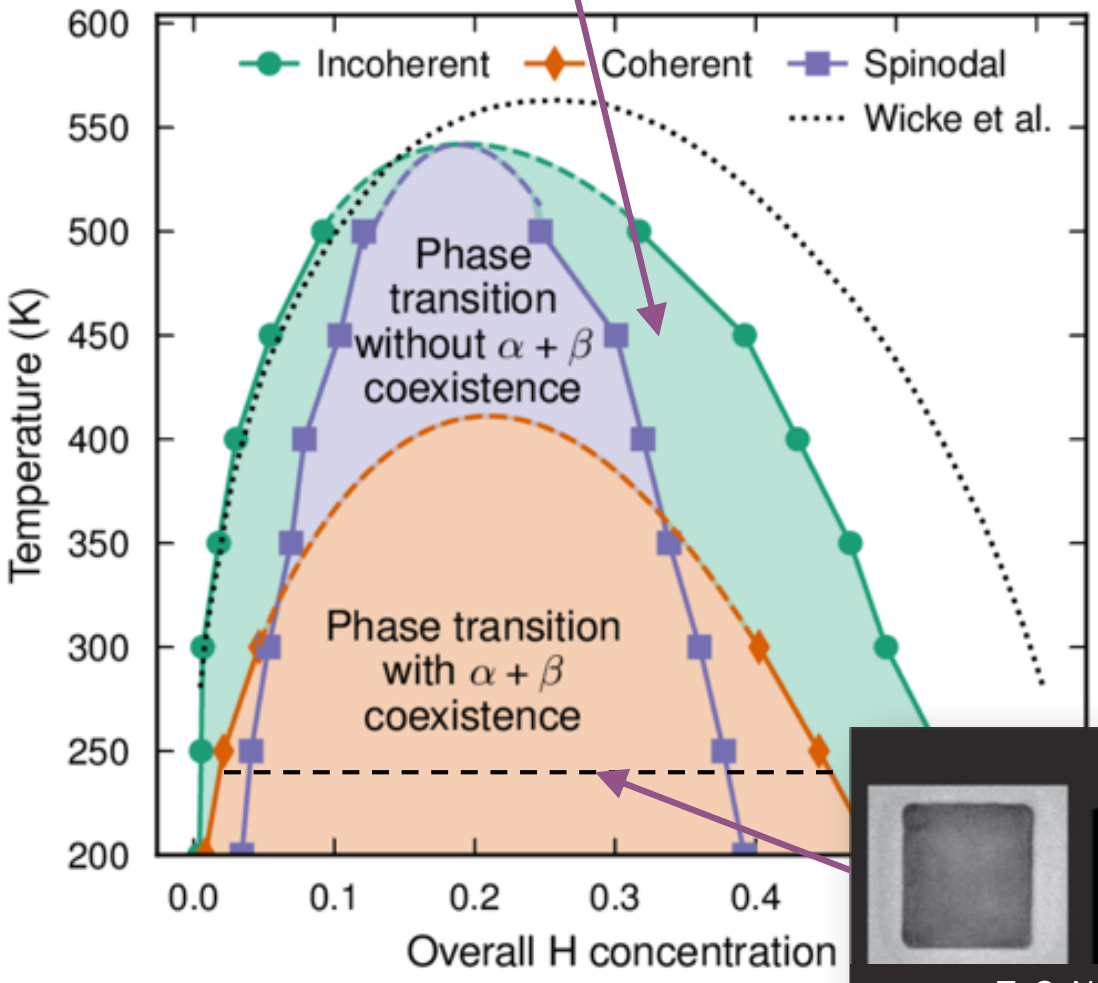




We can access both coherent and incoherent binodal as well as the incoherent spinodal [2]


[1] Sadigh & PE, PRB **86**, 134204 (2012)
[2] Rahm *et al.*, arxiv:2110.07883 (2021)

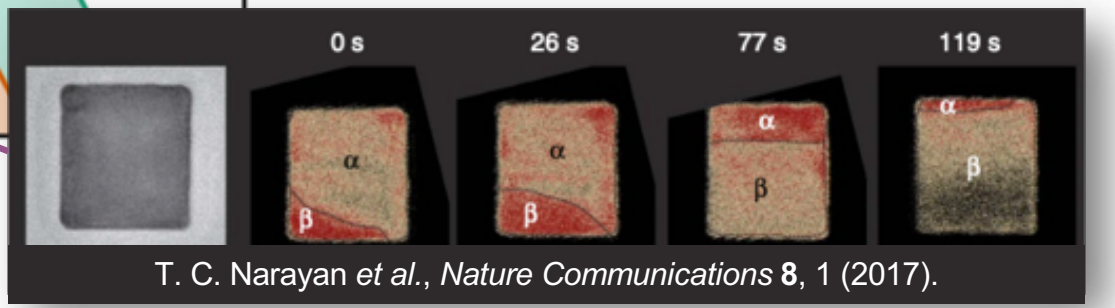
# Application to Pd:H

No visualizations here




## Conclusions

- Coherent phase diagram can be calculated with a modified cluster expansion and carefully analyzed Monte Carlo simulations.
- The phase transition changes character above approximately 400 K.
- Can extract hysteresis and interface energies quantitatively



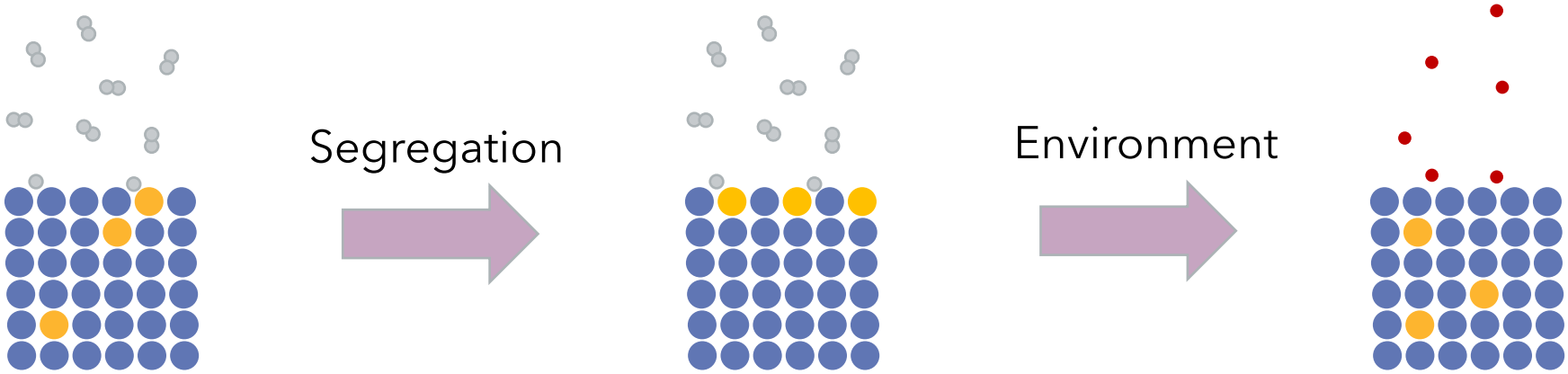

T. C. Narayan *et al.*, *Nature Communications* **8**, 1 (2017).




[1] Rahm *et al.*, arxiv:2110.07883 (2021)

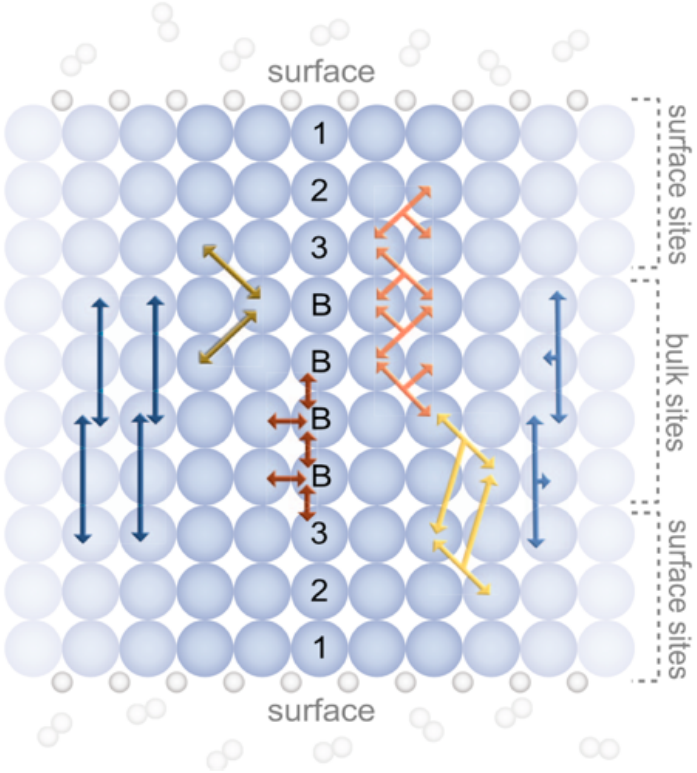
# Dealing with low symmetry

Example: Surface (PdAu in $H_2$ atmosphere)




Challenge for CEs due to symmetry reduction → **Exploit local symmetries**

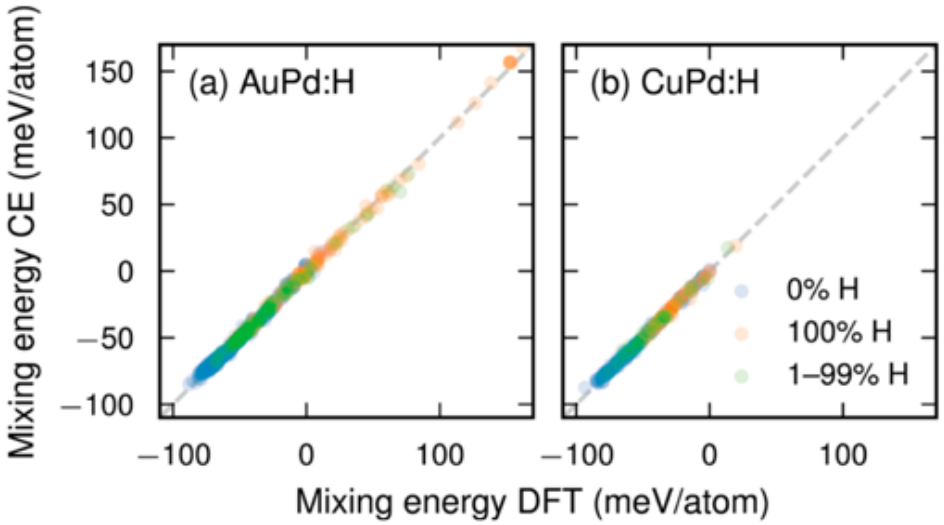
1. Merging orbits
2. Bayesian regression






[1] Ekborg-Tanner and PE
J. Phys. Chem. C **125**, 17248 (2021)

# Dealing with low symmetry: Merging orbits

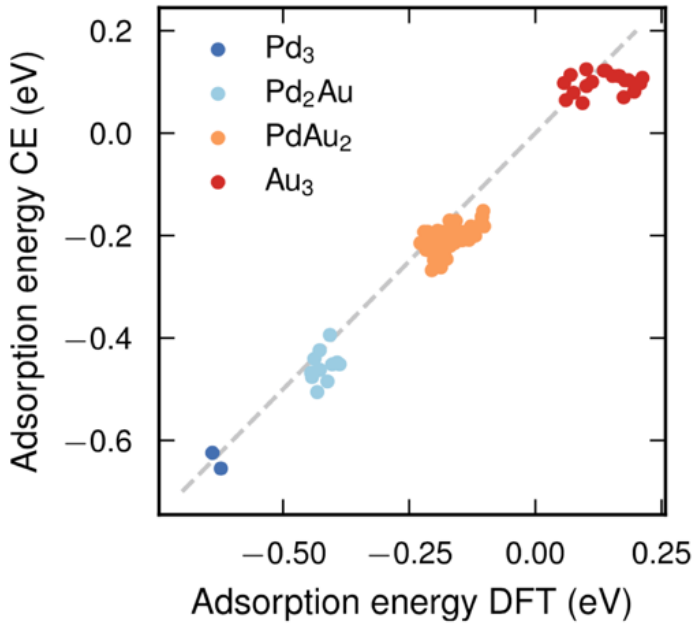


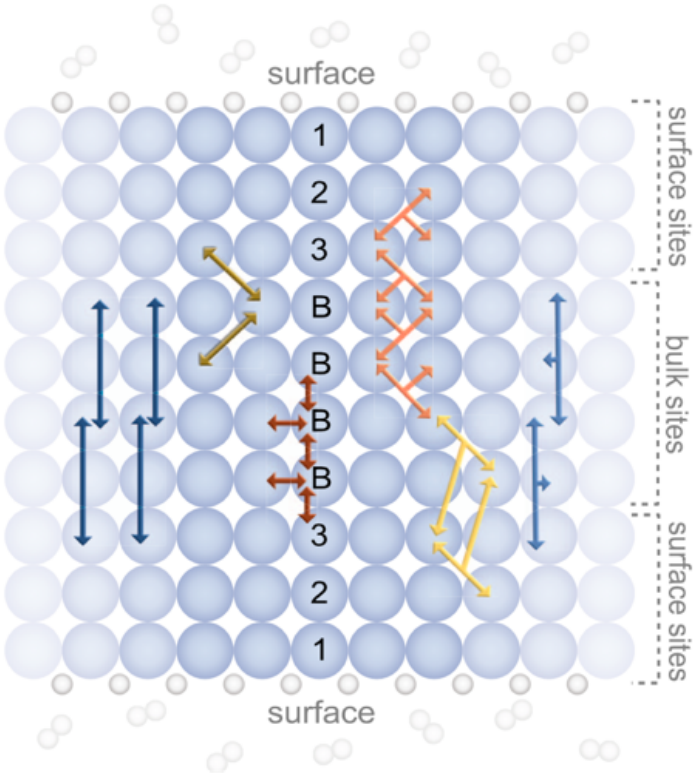

3 surface layers then bulk

Orbits based on heuristic similarity criterion

267 ECIs → 97 ECIs

Intractable → Doable





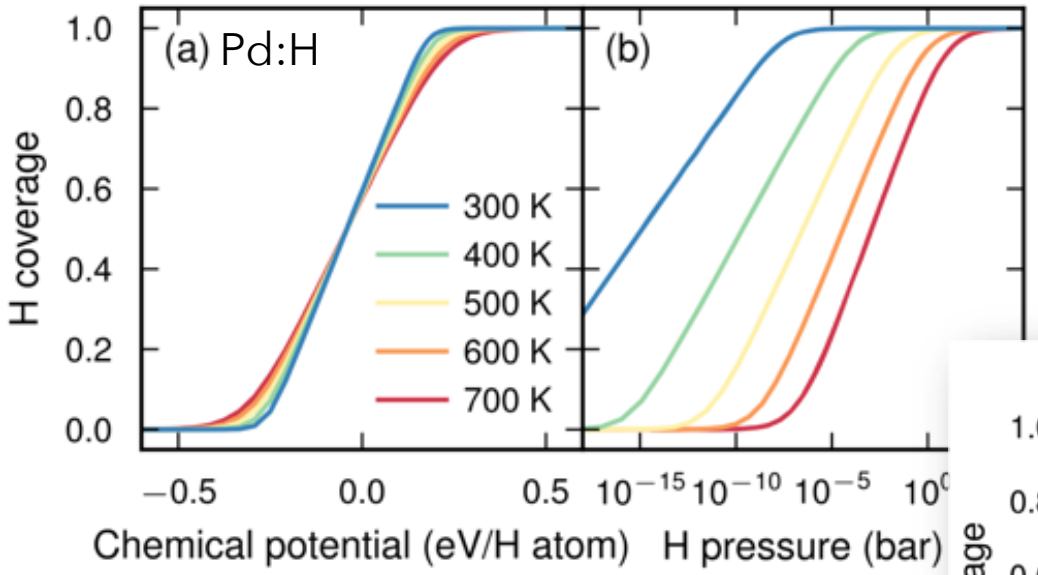

[1] Ekborg-Tanner and PE
J. Phys. Chem. C **125**, 17248 (2021)

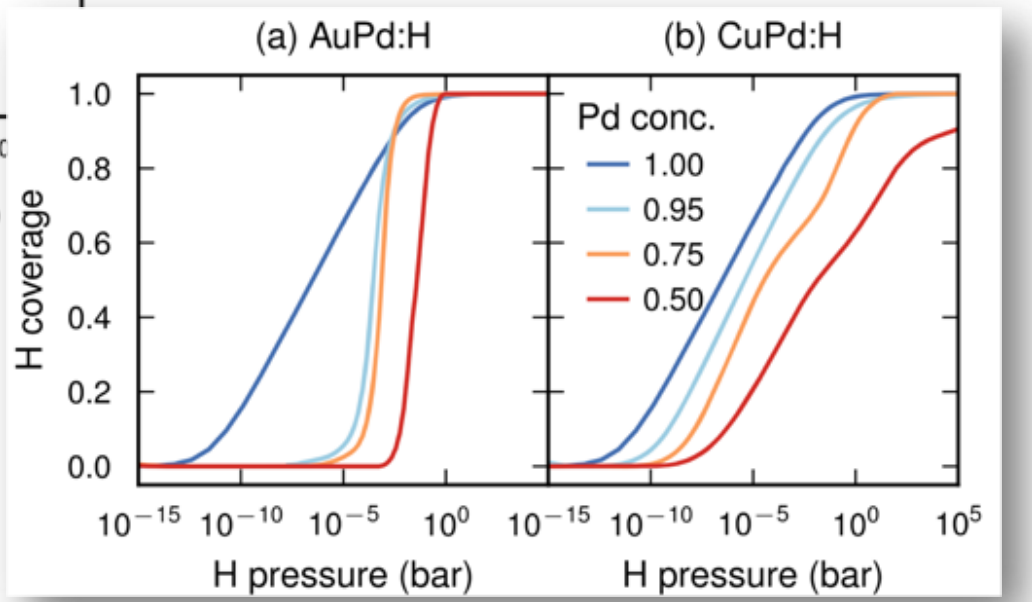
# From chemical potential to partial pressure

$$\mu_{H_2}(T, p_{H_2}) = \mu^{\circ}_{H_2}(T) + k_B T \ln \frac{p_{H_2}}{p^{\circ}_{H_2}} \quad \Rightarrow \quad p_{H_2} = \exp\left(\frac{2\tilde{\mu}_H + 2E_{corr} + \Delta P(T)}{k_B T}\right)$$

Chemical potential in SGC-MC simulations

Conversion from mixing energies and corrections for DFT errors







[1] Ekborg-Tanner and PE
J. Phys. Chem. C **125**, 17248 (2021)

# Low symmetry: Bayesian regression

## Solving the linear CE system

Least-squares (OLS)

$$\boldsymbol{J}_{\text{opt}} = \arg\min_{\boldsymbol{J}} \|\boldsymbol{X}\boldsymbol{J} - \boldsymbol{E}\|_2 \longrightarrow \boldsymbol{J}_{\text{opt}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{E}$$

Ridge regression

$$\boldsymbol{J}_{\text{opt}} = \arg\min_{\boldsymbol{J}} \|\boldsymbol{X}\boldsymbol{J} - \boldsymbol{E}\|_2 + \lambda_2\|\boldsymbol{J}\|_2 \longrightarrow \boldsymbol{J}_{\text{opt}} = \left(\mathbf{X}^T\mathbf{X} + \lambda_2\mathbf{1}\right)^{-1}\mathbf{X}\boldsymbol{E}$$

General case [1]

$$\boldsymbol{J}_{\text{opt}} = (\mathbf{X}^T\mathbf{X} + \boldsymbol{\Lambda})^{-1}\mathbf{X}^T\boldsymbol{E}$$

Inverse covariance matrix

```
sc = StructureContainer(cs)
for atoms in list_of_training_structures:
    sc.add_structure(atoms)
X, E = sc.get_fit_data()
```



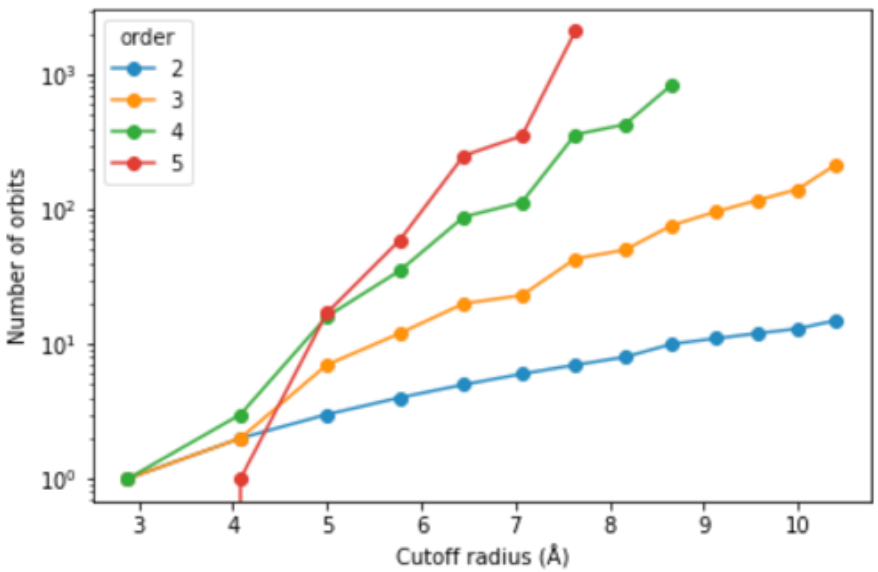
[1] Mueller and Ceder, Phys. Rev. B **80**, 024103 (2009)

# Impose physical insight through priors

Insight 1: ECIs decay with distance *r* and order *n* → diagonal terms

$\Lambda_{aa} \rightarrow \infty$ exclude orbit *a*

$\Lambda_{aa} \rightarrow 0$ and $\Lambda_{a\beta} \rightarrow 0$ recovers OLS




$$\Lambda_{\alpha\alpha} = \gamma_1(\gamma_2 r + \gamma_3 + 1)^{\gamma_4 n + \gamma_5}$$ additional hyperparameters

**CV-RMSE (for AuCu CE in MSc course)**

23 meV/atom → 16 meV/atom → 12 meV/atom

OLS Ridge Covariance



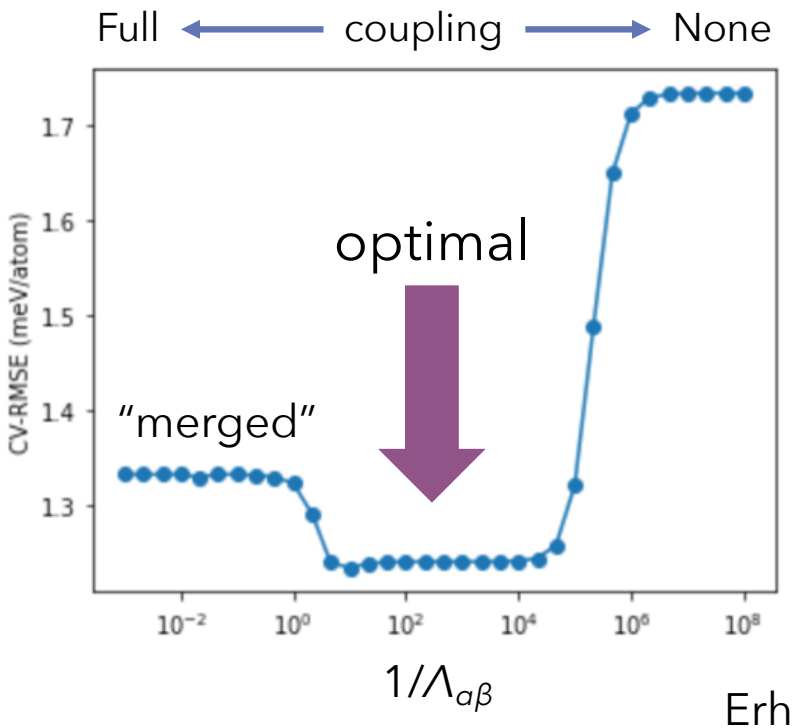
[1] Mueller and Ceder, Phys. Rev. B **80**, 024103 (2009)

# Impose physical insight through priors

Insight 2: ECIs of similar orbits should be similar → off-diagonal terms

$\Lambda_{\alpha\beta} \rightarrow 0$ no coupling between $\alpha$ and $\beta$
$\Lambda_{\alpha\beta} \rightarrow \infty$ force ECIs for $\alpha$ and $\beta$ to be identical

Simple example:
All off-diagonal terms identical




**Some possible strategies**

- Manually force ECIs to be identical → merging approach
- Use similarity measure to set $\Lambda_{\alpha\beta}$
- Couple ECIs to bulk values, e.g., in a surface case



[1] Mueller and Ceder
Phys. Rev. B **80**, 024103 (2009)

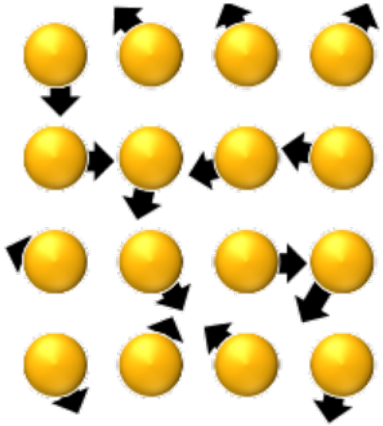
hiphive

Force constant expansions

# Force constant expansions

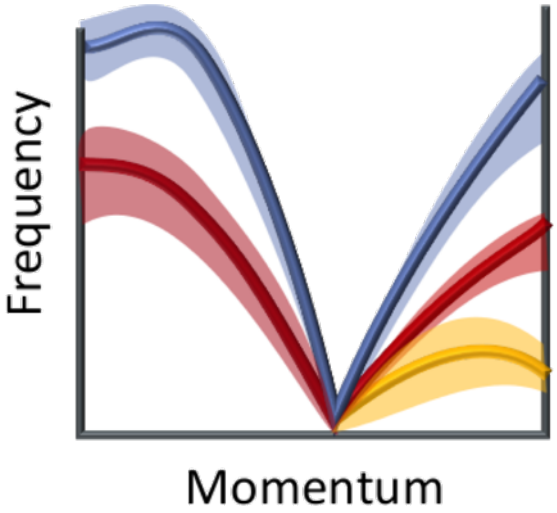
Taylor expansion of the total energy

$$E = E_0 + \sum_i \Phi_i u_i + \frac{1}{2!}\sum_{ij} \Phi_{ij} u_i u_j + \frac{1}{3!}\sum_{ijk} \Phi_{ijk} u_i u_j u_k \ldots$$



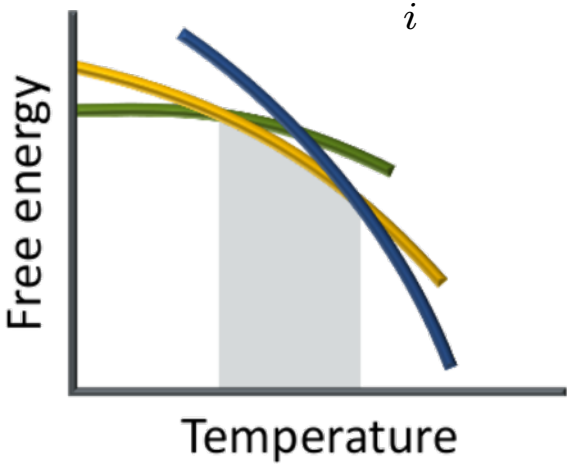
Phonon dispersions

$$\Phi_{ij} \rightarrow D_{\alpha\beta}(\boldsymbol{q}) \rightarrow \omega_k(\boldsymbol{q})$$




Free energies

$$F_{vib} = k_B T \sum_i \ln \omega_i$$

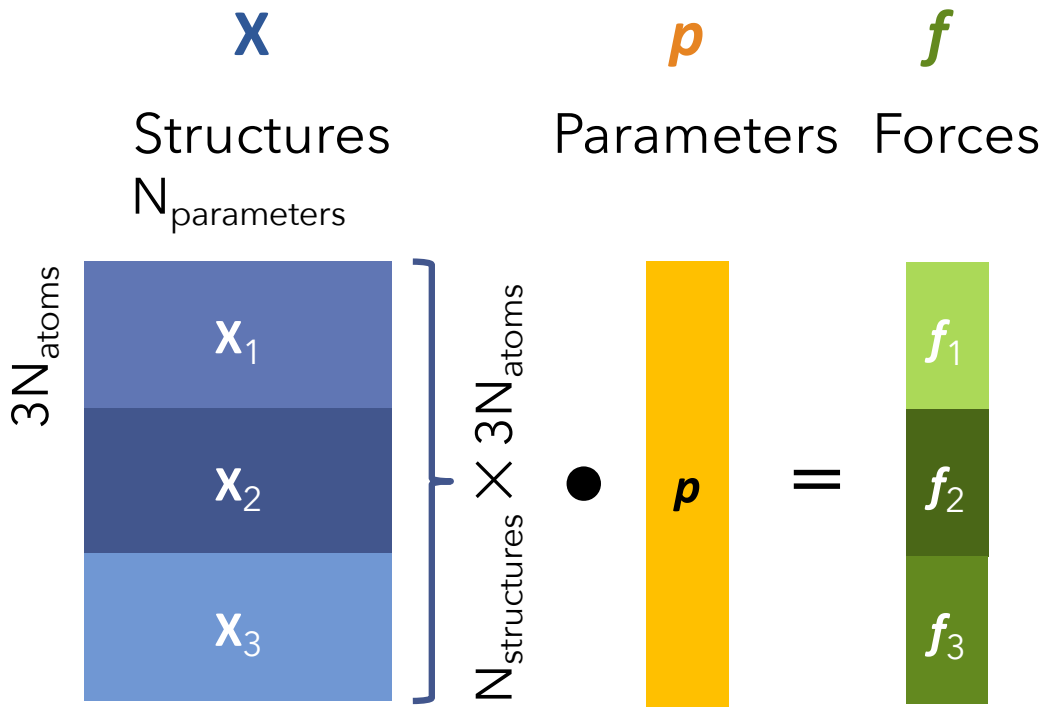



Need efficient means to extract Φ matrices

# FC expansions also a linear problem

## General procedure

1. Compile multiple structures into one (large) fit matrix **X**
2. Solve the linear problem **X** ***p***=***f***




**How?**

Least-squares (OLS)

$$\boldsymbol{p}_{\text{opt}} = \arg\min_{\boldsymbol{p}} \|\mathbf{X}\boldsymbol{p} - \boldsymbol{f}\|_2$$

LASSO

$$\boldsymbol{p}_{\text{opt}} = \arg\min_{\boldsymbol{p}} \|\mathbf{X}\boldsymbol{p} - \boldsymbol{f}\|_2 + \lambda_1 \|\boldsymbol{f}\|_1$$

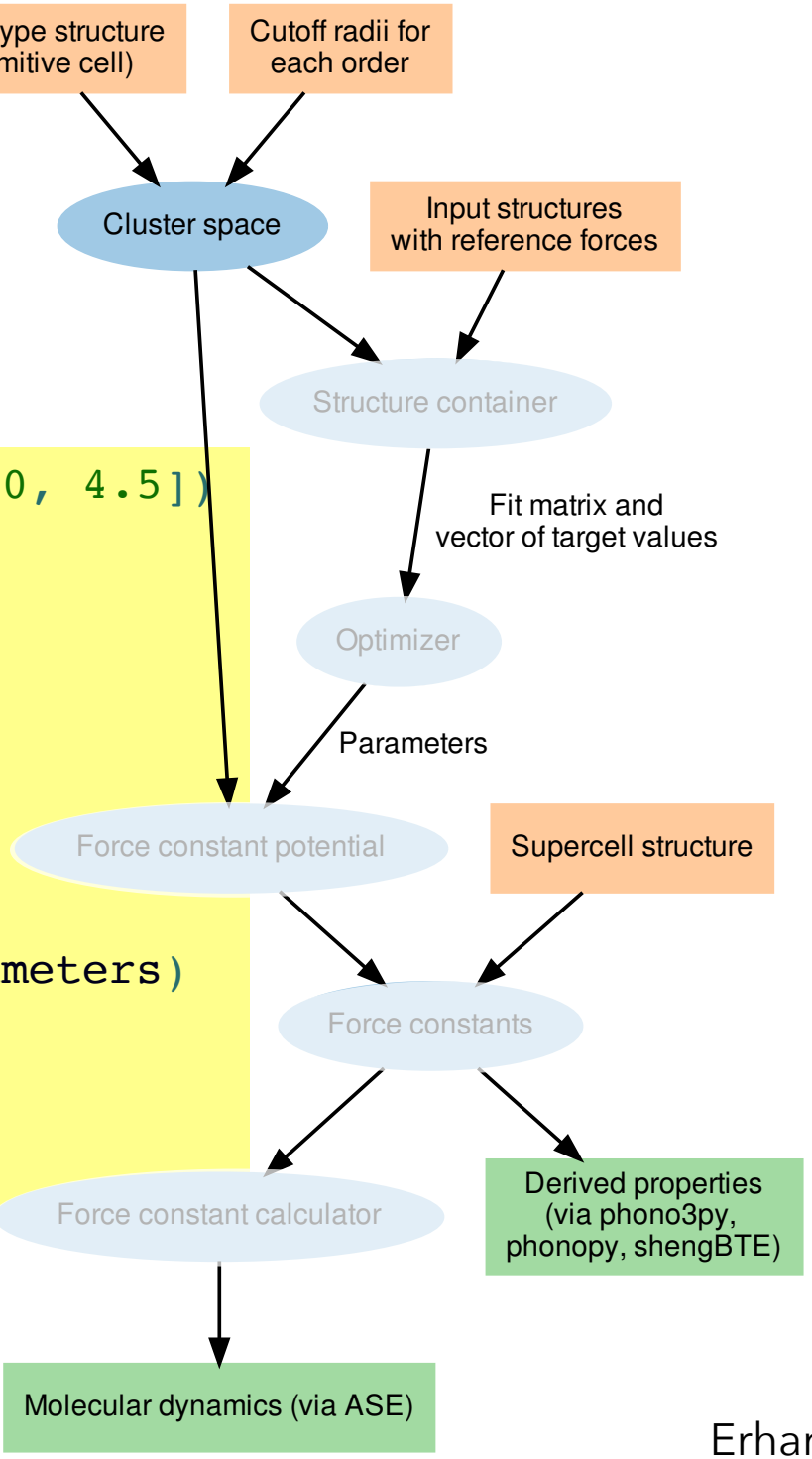
Ridge regression

$$\boldsymbol{p}_{\text{opt}} = \arg\min_{\boldsymbol{p}} \|\mathbf{X}\boldsymbol{p} - \boldsymbol{f}\|_2 + \lambda_2 \|\boldsymbol{f}\|_2$$

# Implementation

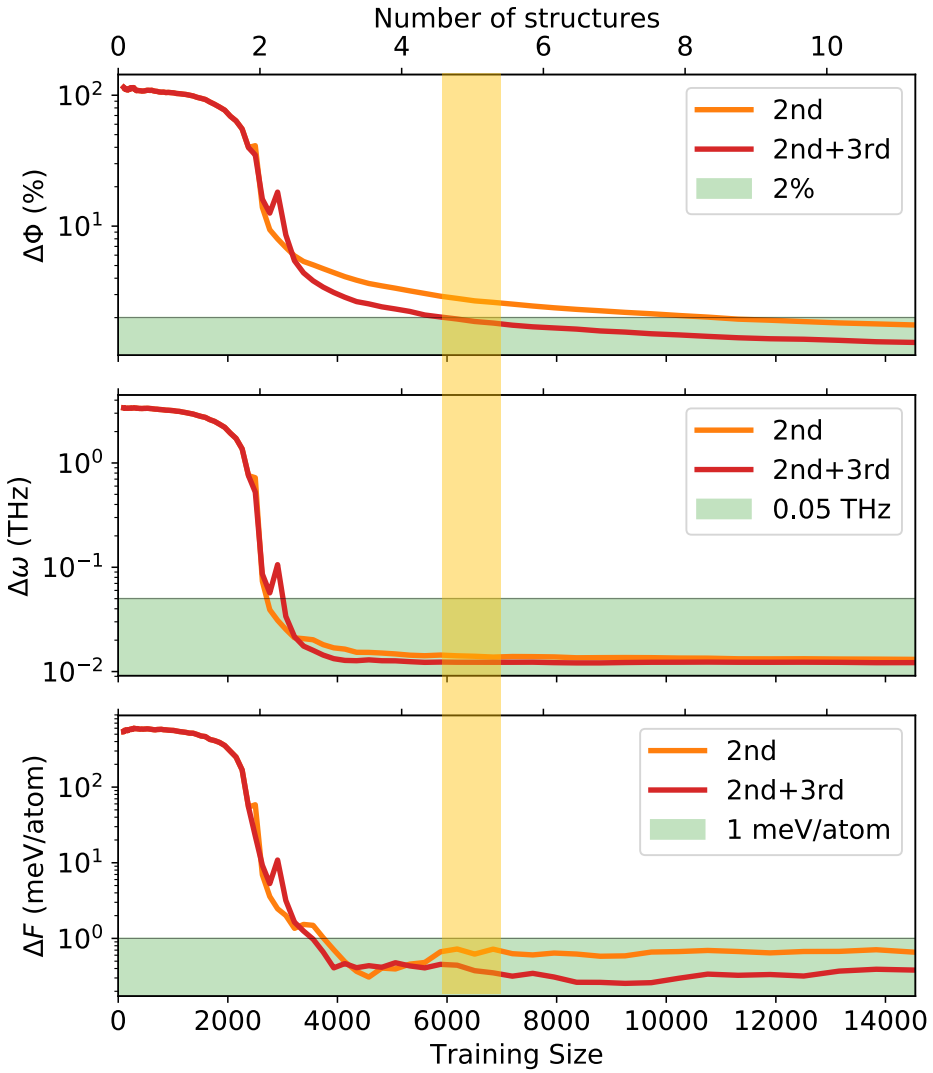
1. Cluster decomposition
2. Structure decomposition
3. Parameter optimization

```
cs = ClusterSpace(ideal_cell, cutoffs=[6.0, 4.5])

sc = StructureContainer(cs)
for atoms in list_of_training_structures:
    sc.add_structure(atoms)

opt = Optimizer(sc.get_fit_data())
opt.train()

fcp = ForceConstantPotential(cs, opt.parameters)

fcs = ForceConstants(fcs, supercell)

calc = ForceConstantCalculator(fcs)
```

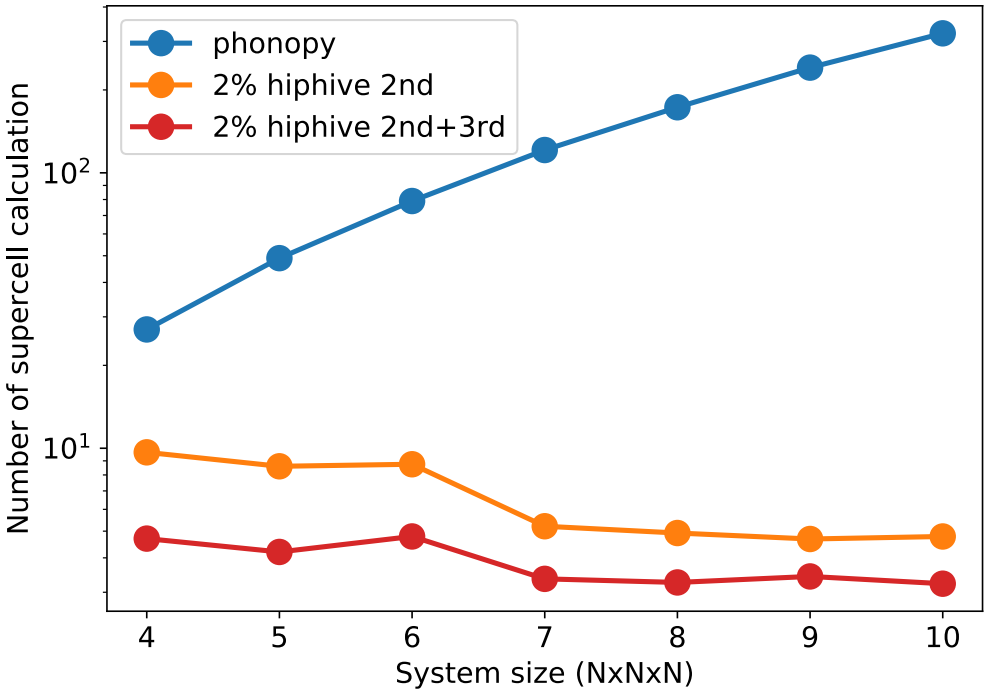


# Efficiency 2nd-order: Vacancy in BCC-Ta

## Convergence with training size




## Comparison with phonopy



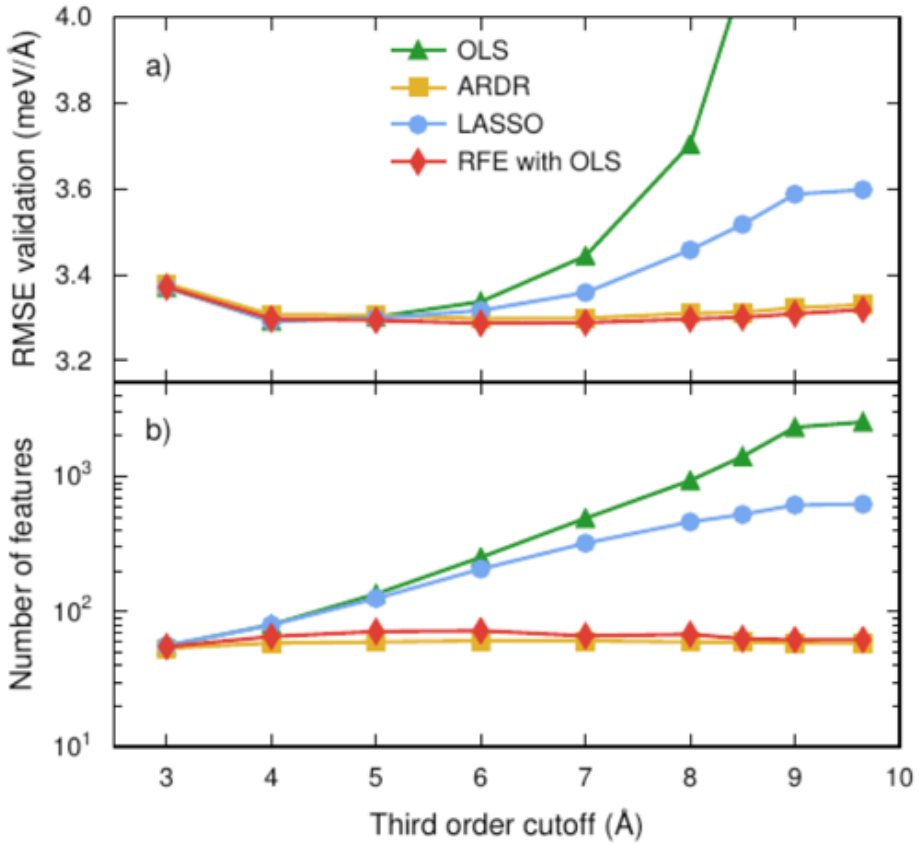

Approach avoids unnecessary size-scaling and maximizes exploitation of available data

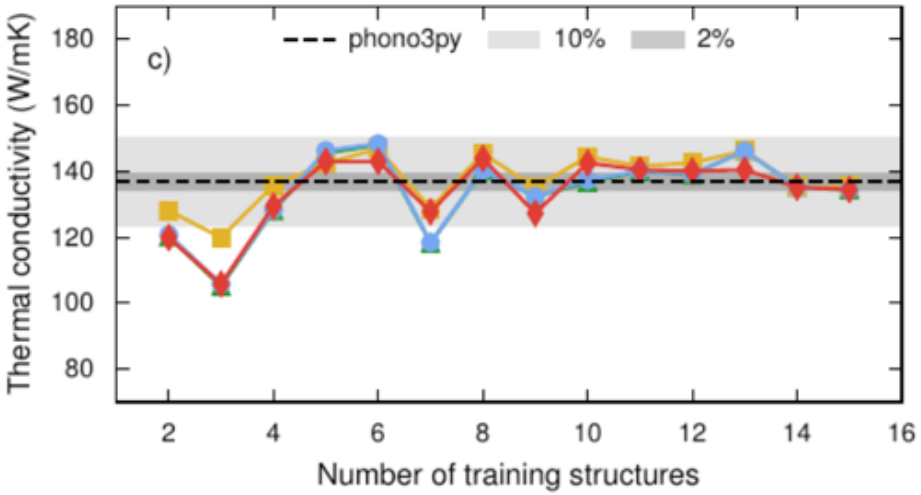


Fransson *et al.*, npg Comp. Mater. **6**, 135 (2020)

# Manual vs automated feature selection

Manual selection (via cutoff)



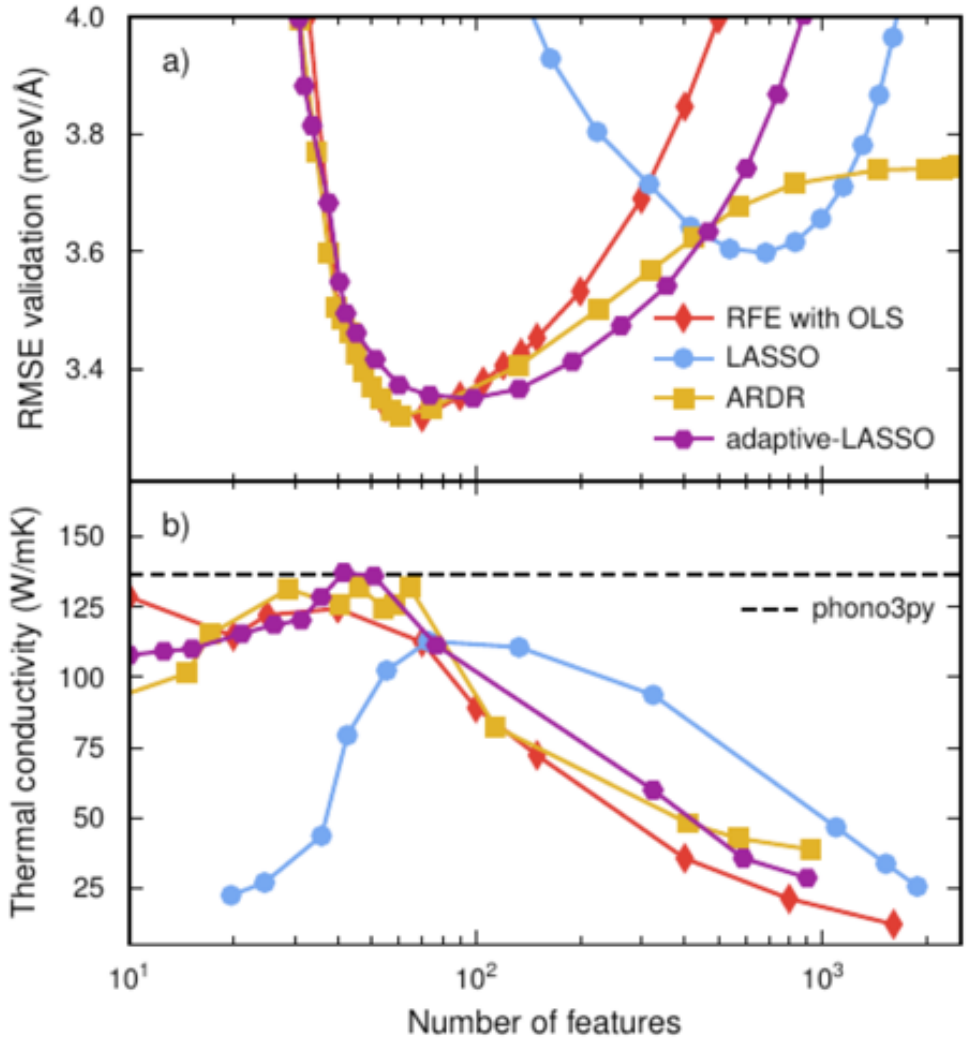

In the overdetermined limit all methods recover the reference value (which required about 10 times as many DFT calculations)



Fransson *et al.*, npg Comp. Mater. **6**, 135 (2020)

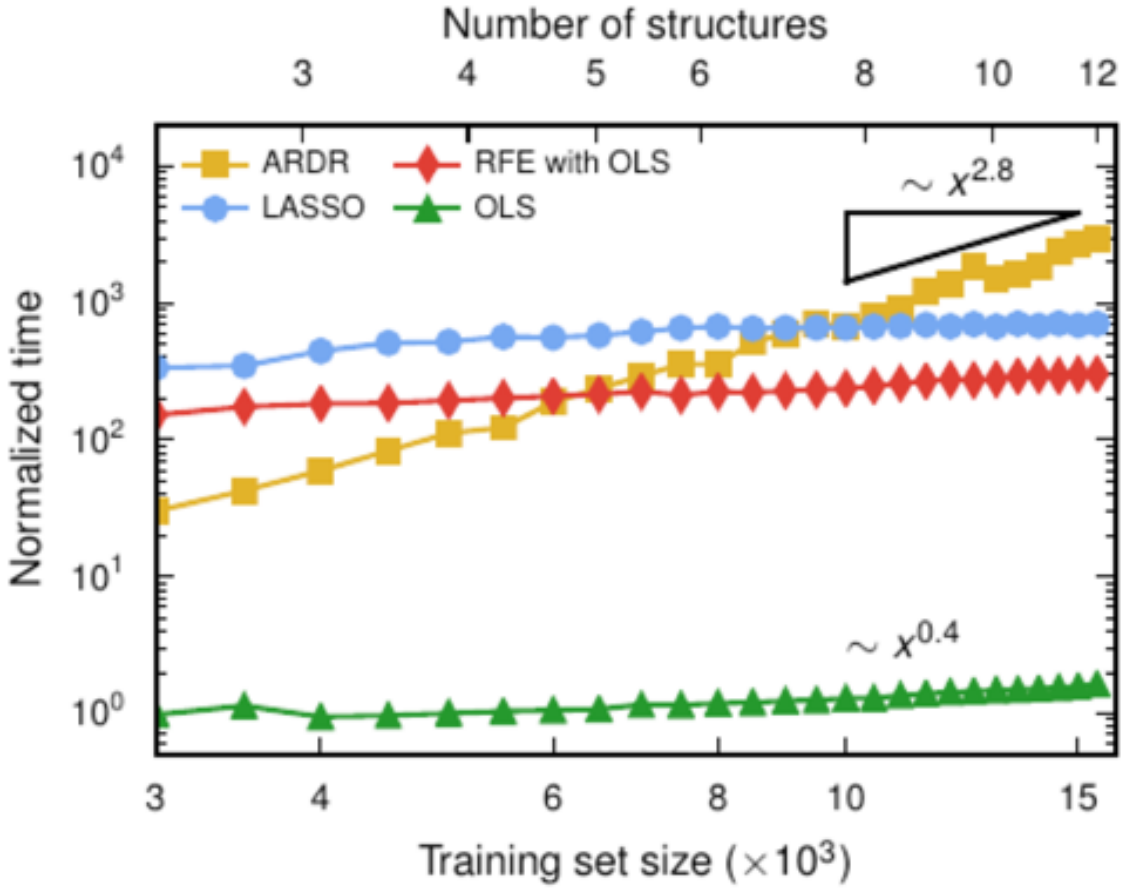
# Manual vs automated feature selection

Automated selection




- LASSO overselects (known) and fails at recovering derived property
- ARDR and RFE perform by far the best (reliably predicting derived properties)
- CV-RMSE poor predictor for quality of derived property

# Cost analysis



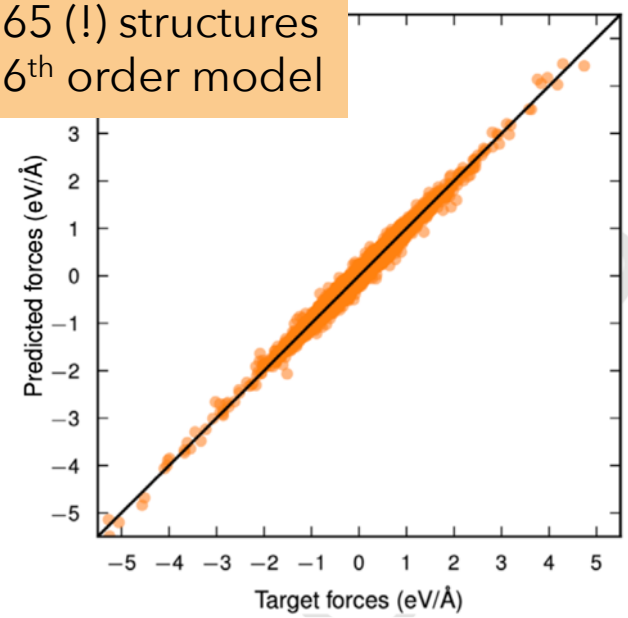

- Cost becomes significant for large problems (FC not CE)
- ARDR (Bayesian) becomes too costly due to scaling
- RFE or OLS with manual selection often the most performant (*bummer*)

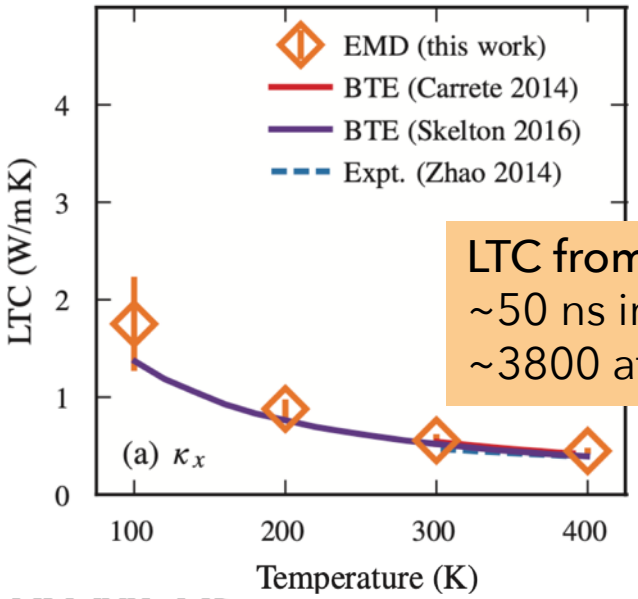


Fransson *et al.*, npg Comp. Mater. **6**, 135 (2020)

# GPU-accelerate MD with ab-initio accuracy

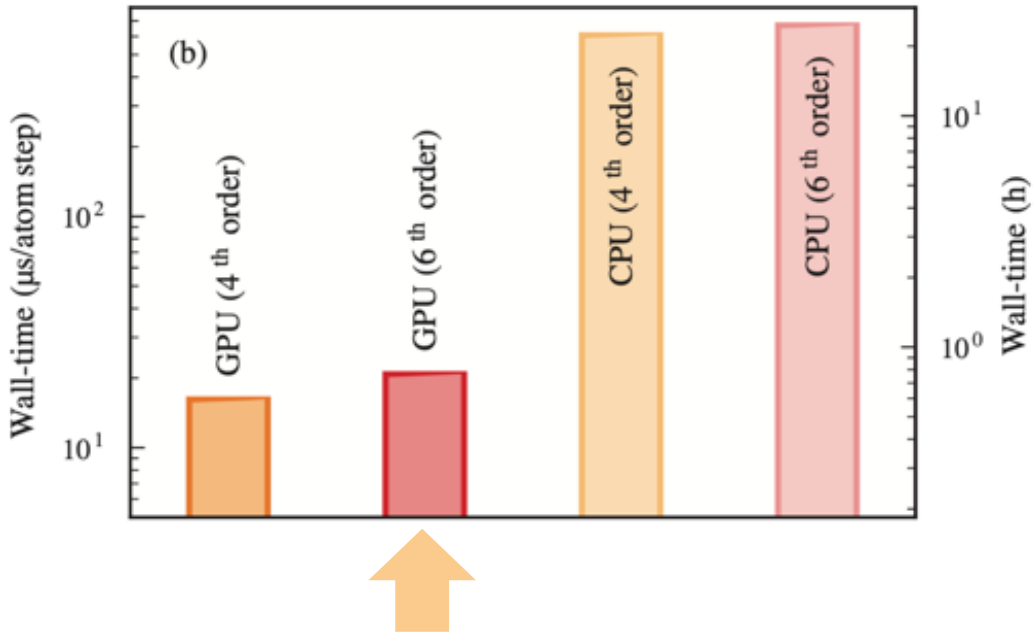
**SnSe**
65 (!) structures
6th order model




**LTC from Green-Kubo**
~50 ns in total
~3800 atoms

GPUMD




1000 steps with 1000 atoms
in 15 sec walltime (Tesla P100)



Brorsson *et al.*, Adv. Theo. Simul. (2021)

# Some semi-closing notes

- Sensing matrices have different structure for FCs and CEs
  → differences in performance of optimization methods

**Force constants**

- Information content *increases* with cell size (3N values per conf)
- Systems often close to (over)determined (formally)
- Recursive feature elimination works well

**Cluster expansions**

- information content *decreases* with cell size (1 value per conf)
- ARDR often yields best performance

Cross-validation and ensemble optimization are powerful

# Who did the work

icet

Mattias Ångqvist
Magnus Rahm
Erik Fransson
Pernilla Ekborg-Tanner
Joakim Brorsson
William Armando Muñoz
Thomas Holm Rod (ESS)
Celine Durniak (ESS)
Piotr Roczycko (ESS)



hiphive

Fredrik Eriksson
Erik Fransson

Zheyong Fan (Aalto)
Arsalan Hashemi (Aalto)
Joakim Brorsson

GPUMD

Knut and Alice Wallenberg Foundation

Vetenskapsrådet

Swedish Foundation for Strategic Research

Interreg Öresund-Kattegat-Skagerrak European Regional Development Fund

EUROPEAN UNION

CHALMERS UNIVERSITY OF TECHNOLOGY

https://materialsmodeling.org/software

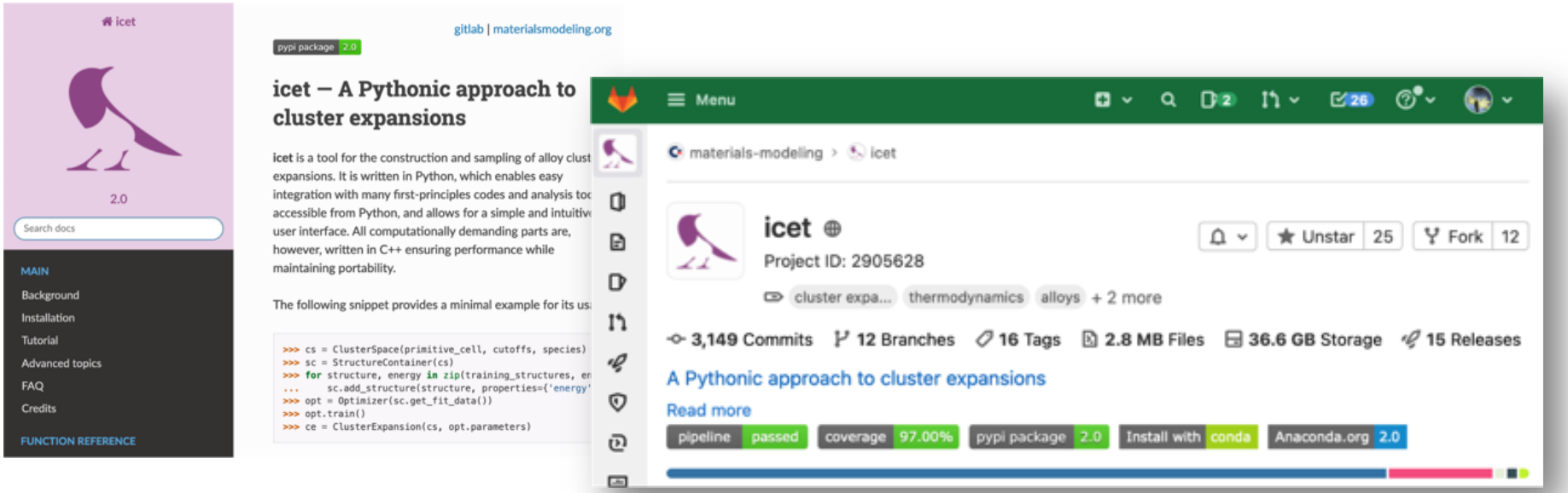
# Summary of updates

icet

Support for CEs with strain
Merging of orbits
Support for Bayesian regression
Improved MC performance
+ lots of small stuff

hiphive

Efficient regression
GPU-accelerated MD
Self-consistent phonons
+ lots of small stuff




https://materialsmodeling.org/software