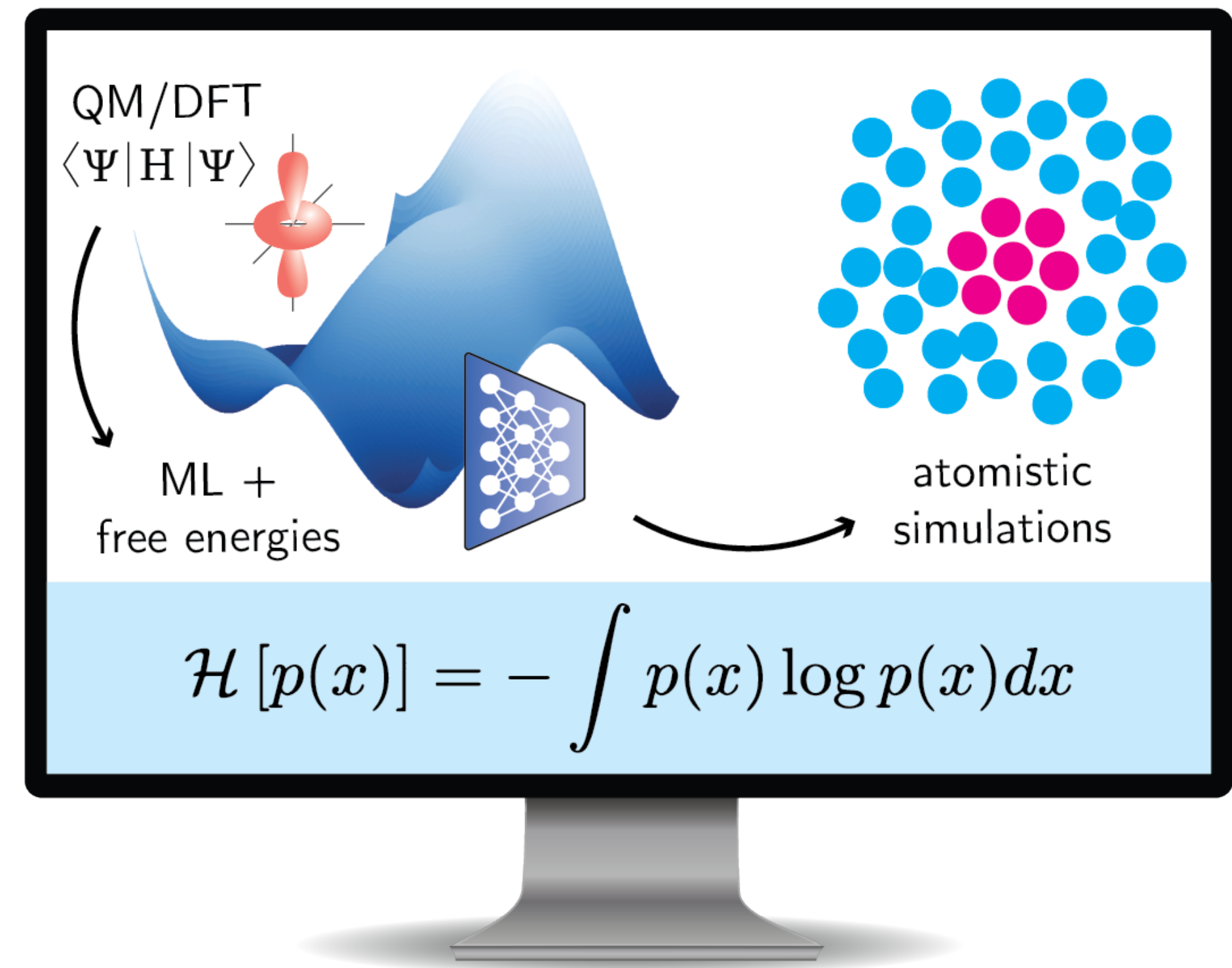


Bridging Scales in Materials Modeling with Atomistic Simulations, Information Theory, and Generative Models

Daniel Schwalbe-Koda

University of California, Los Angeles

dskoda@ucla.edu



IPAM Workshop, ECHWS2
October 9, 2025

UCLA

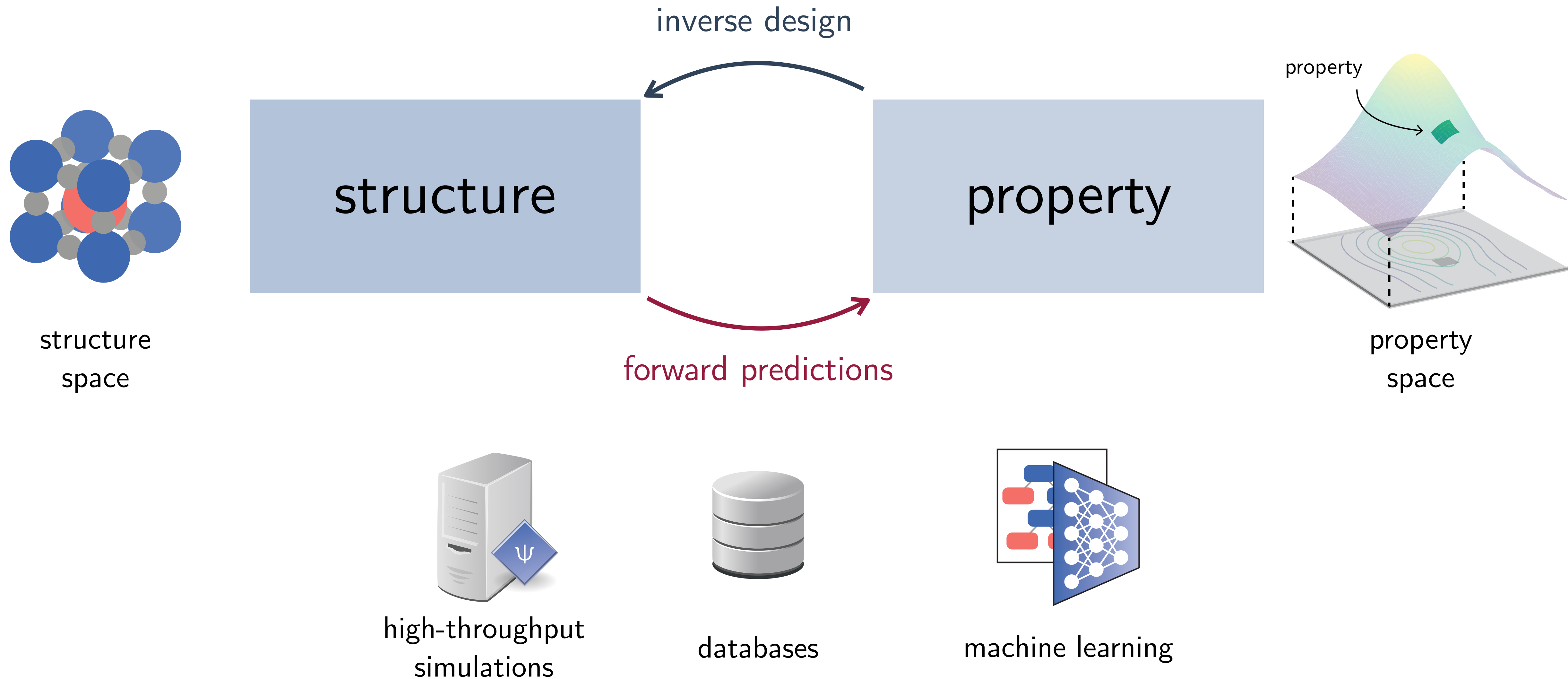
Samueli

Materials Science & Engineering

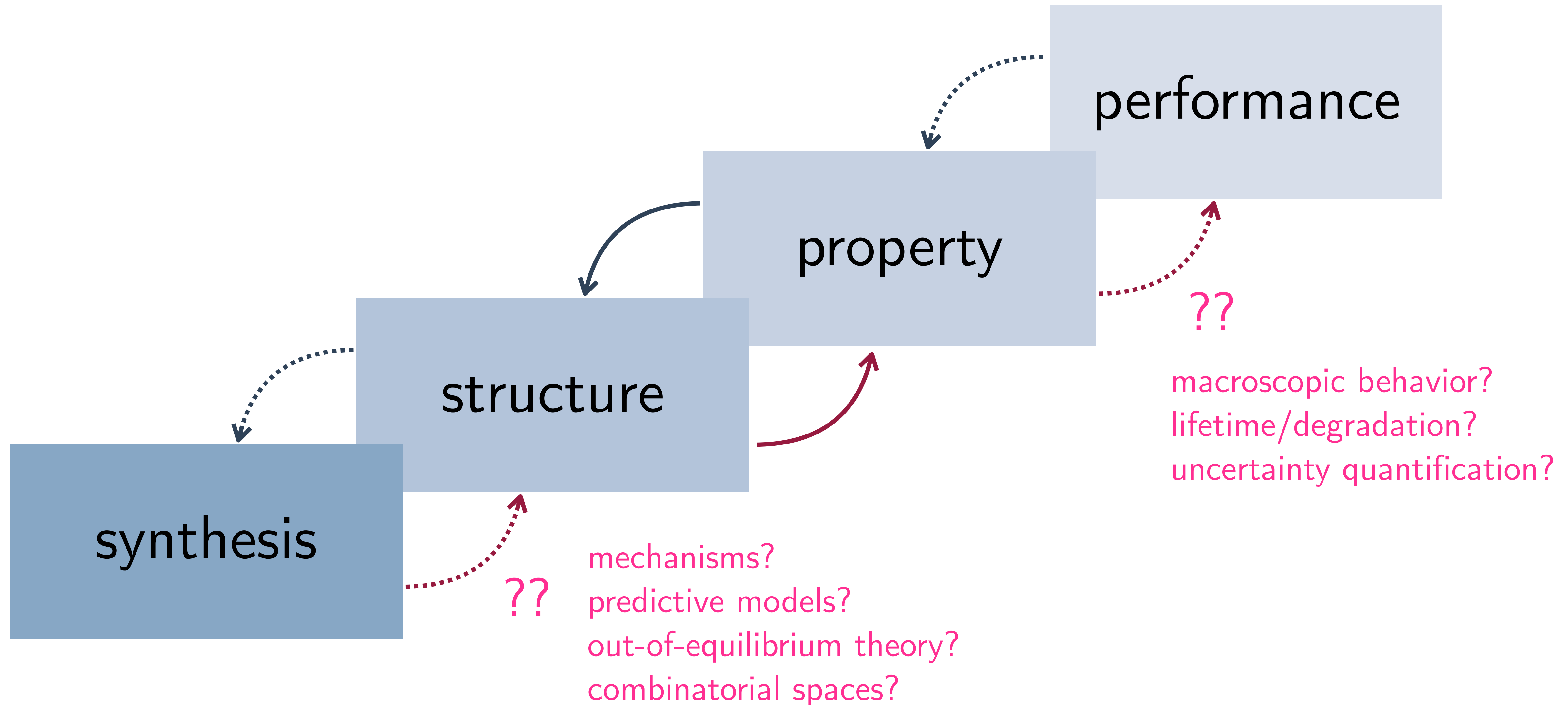
Bridging the Gap: Transitioning from Deterministic to Stochastic Interaction Modeling in Electrochemistry

Modern computational materials science does a good job in structure-property predictions

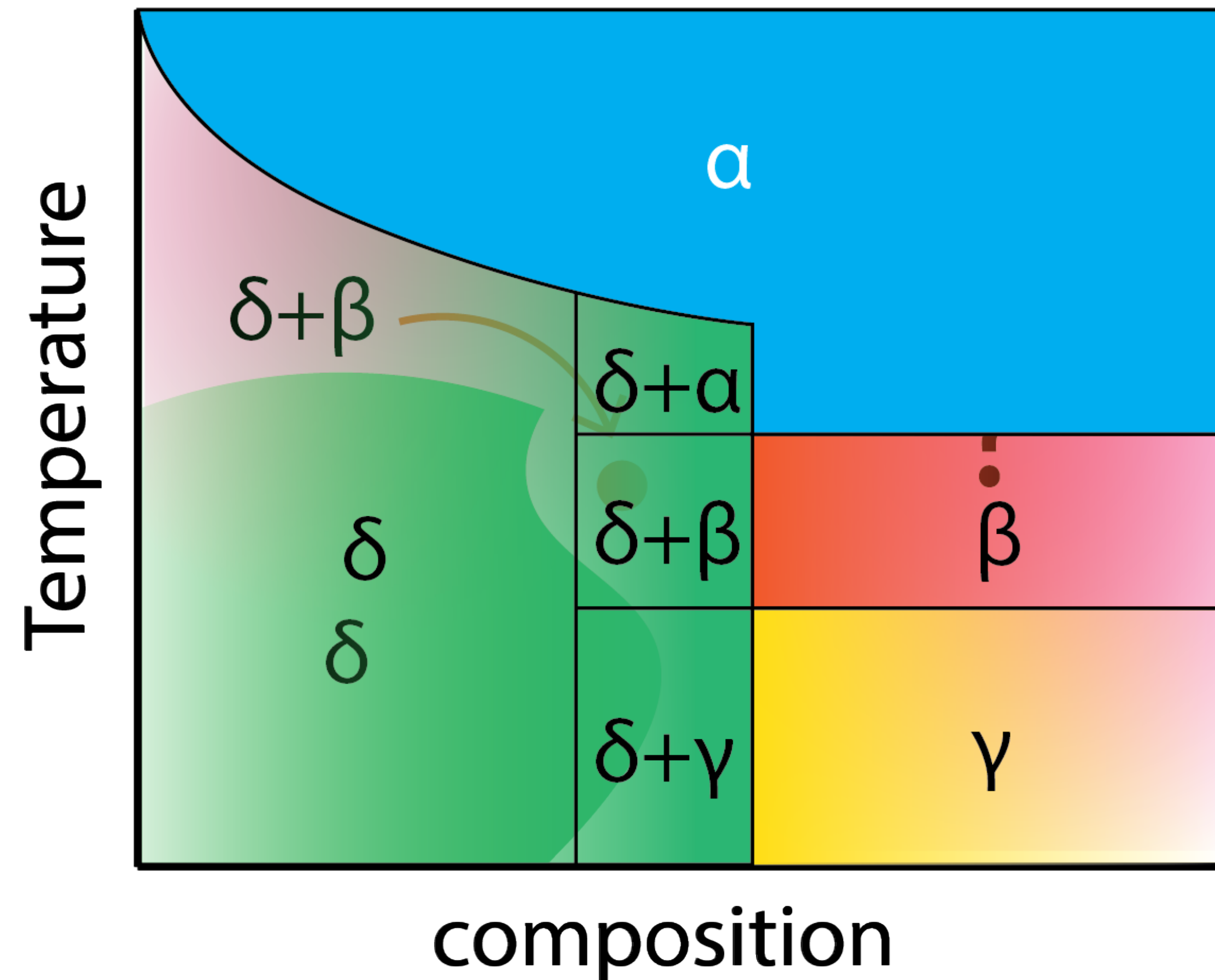
2



However, materials selection spans multiple scales, beyond structure-property models

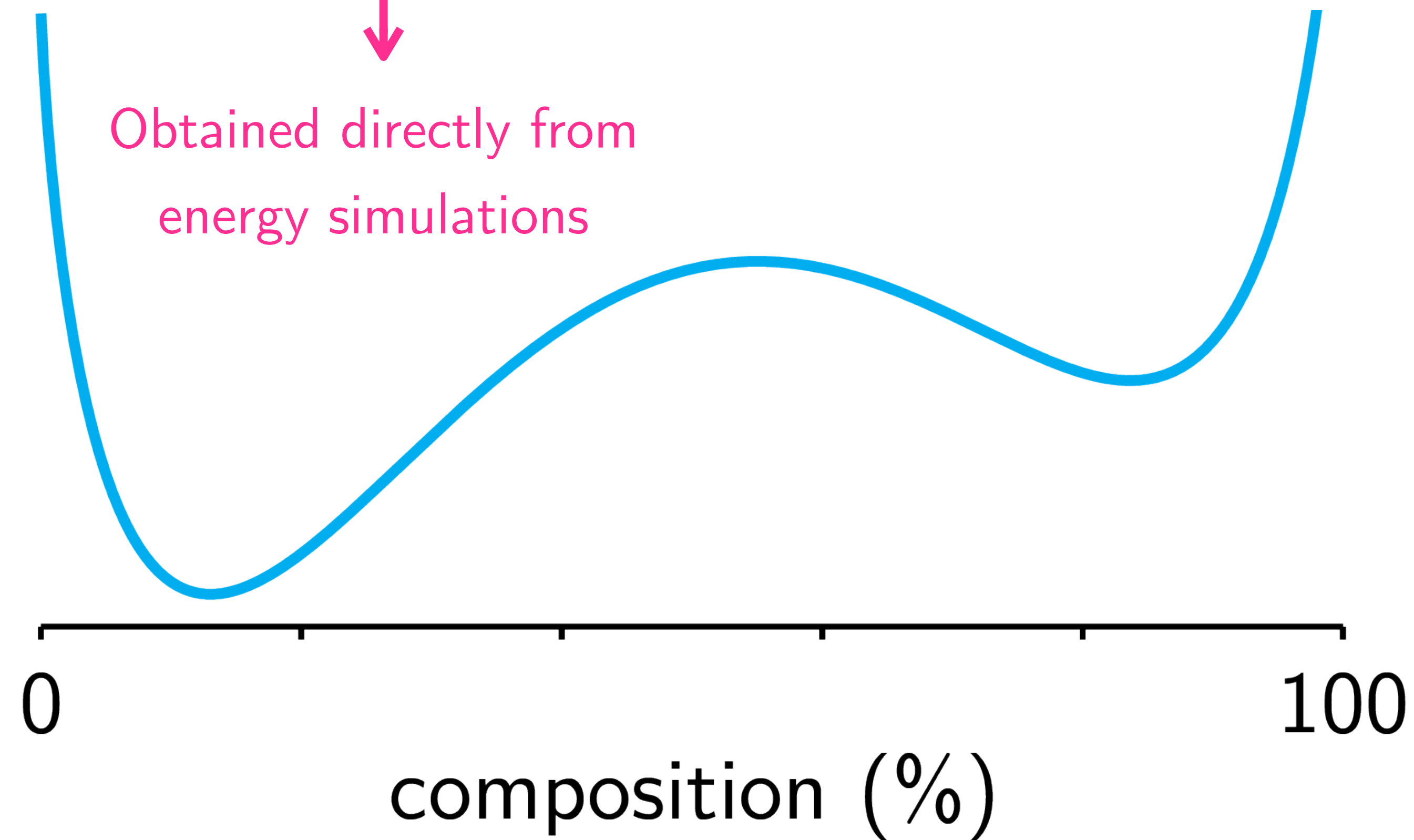


Target: α phase

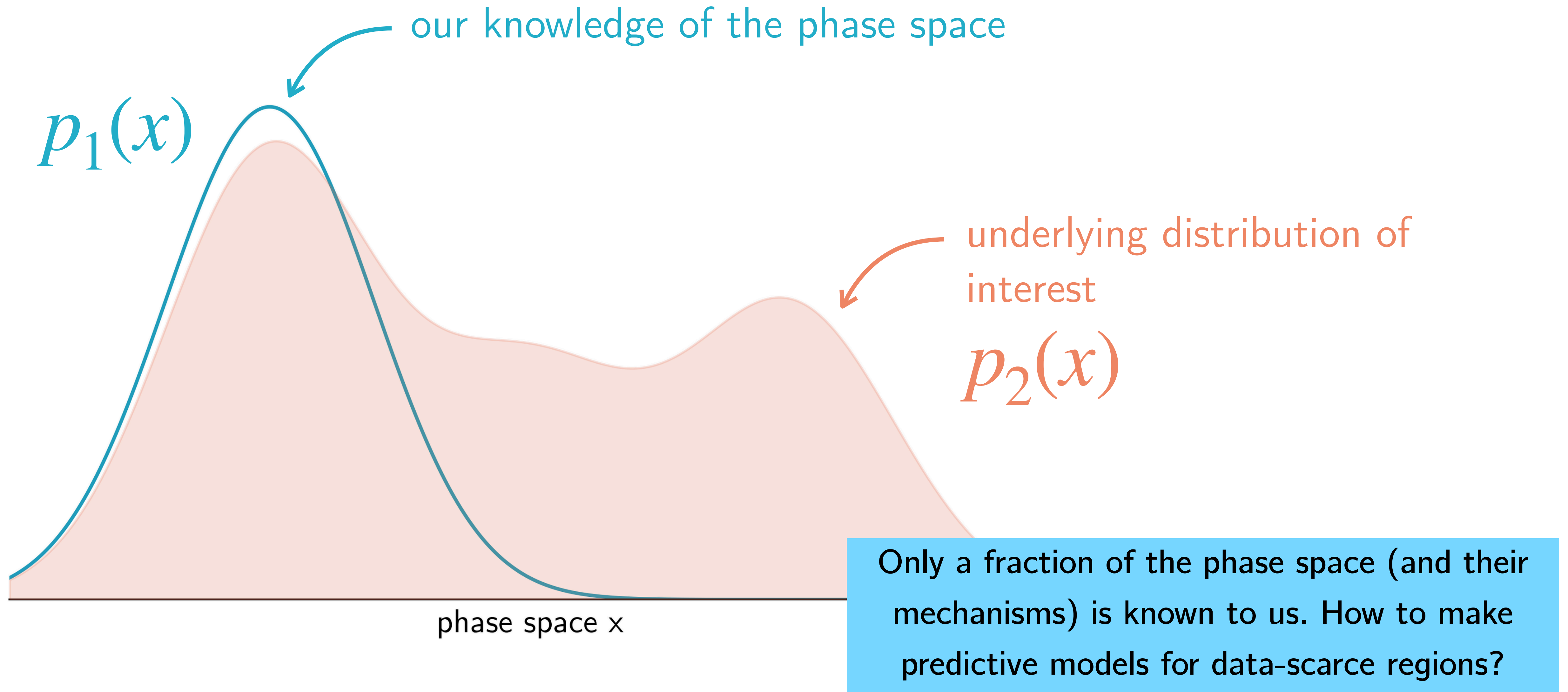


$$\Delta G = \Delta H - T\Delta S \rightarrow \text{Requires sampling}$$

Obtained directly from
energy simulations



Computing bottleneck: calculating, understanding,
and sampling the space of materials



Bridging Scales in Materials Modeling with Atomistic Simulations, Information Theory, and Generative Models

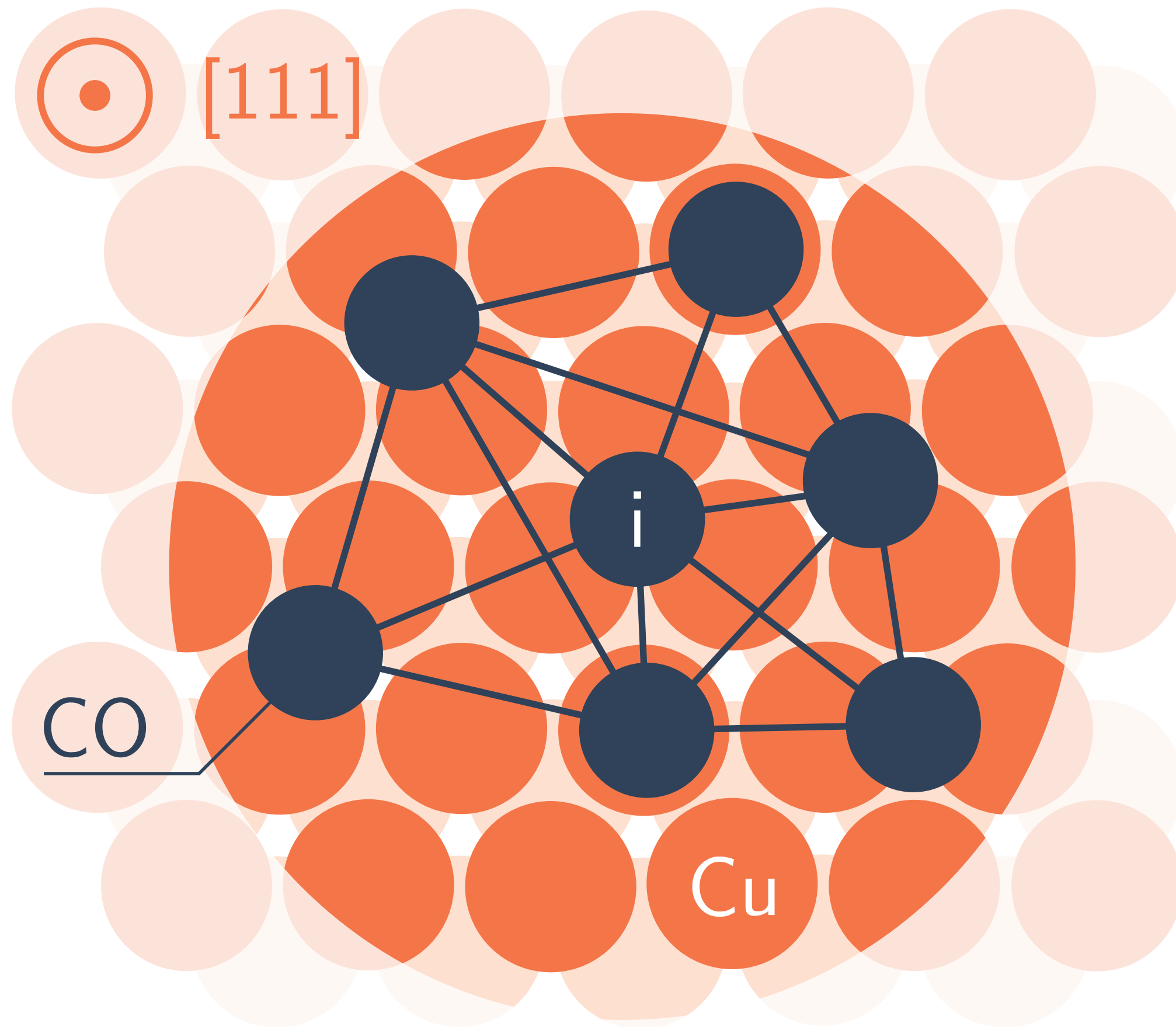
tl;dr:

Bridging scales in materials modeling - especially synthesis science - requires advances in:

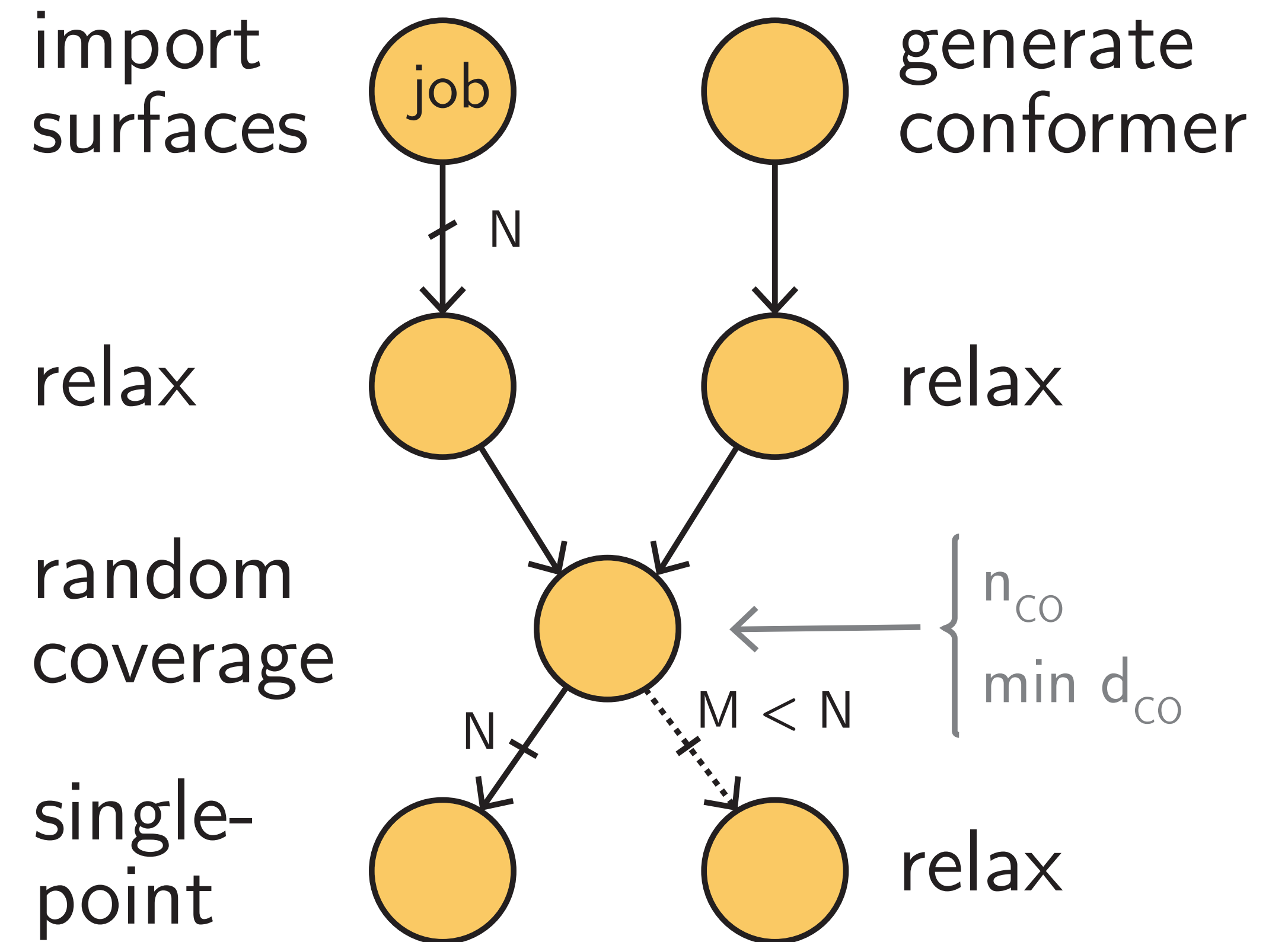
- (1) understanding model generalization
- (2) connecting atomistic simulations with macroscopic theories
- (3) sampling methods and generative models relevant to materials science

Example relevant to electrochemistry: lateral coverage of adsorbates

e.g., high coverage of CO on Cu facets



Pipeline: loop-free approach

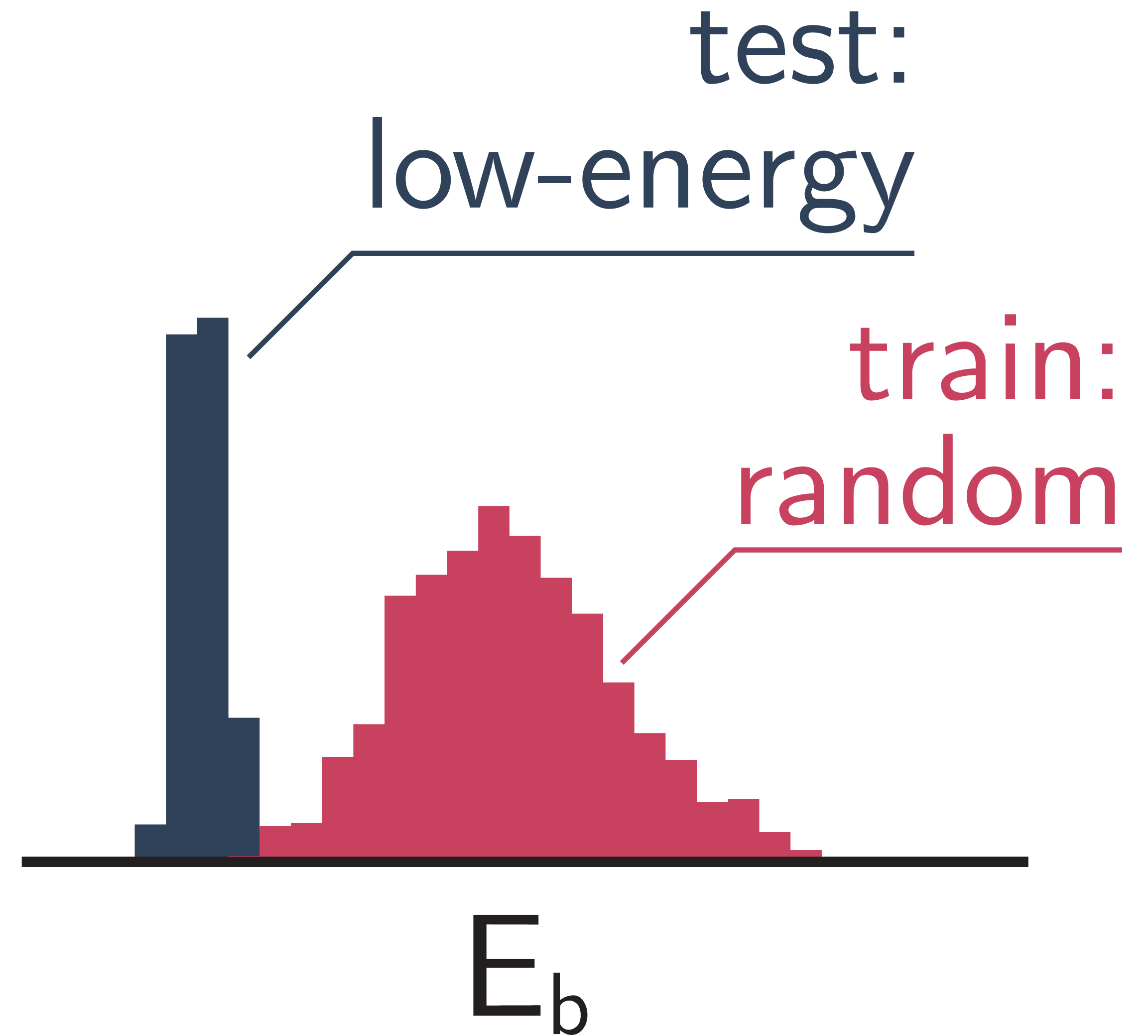


Data pipelines are accelerated if we do not worry about generalization

Step 1: randomly sample configurations +
generate training data

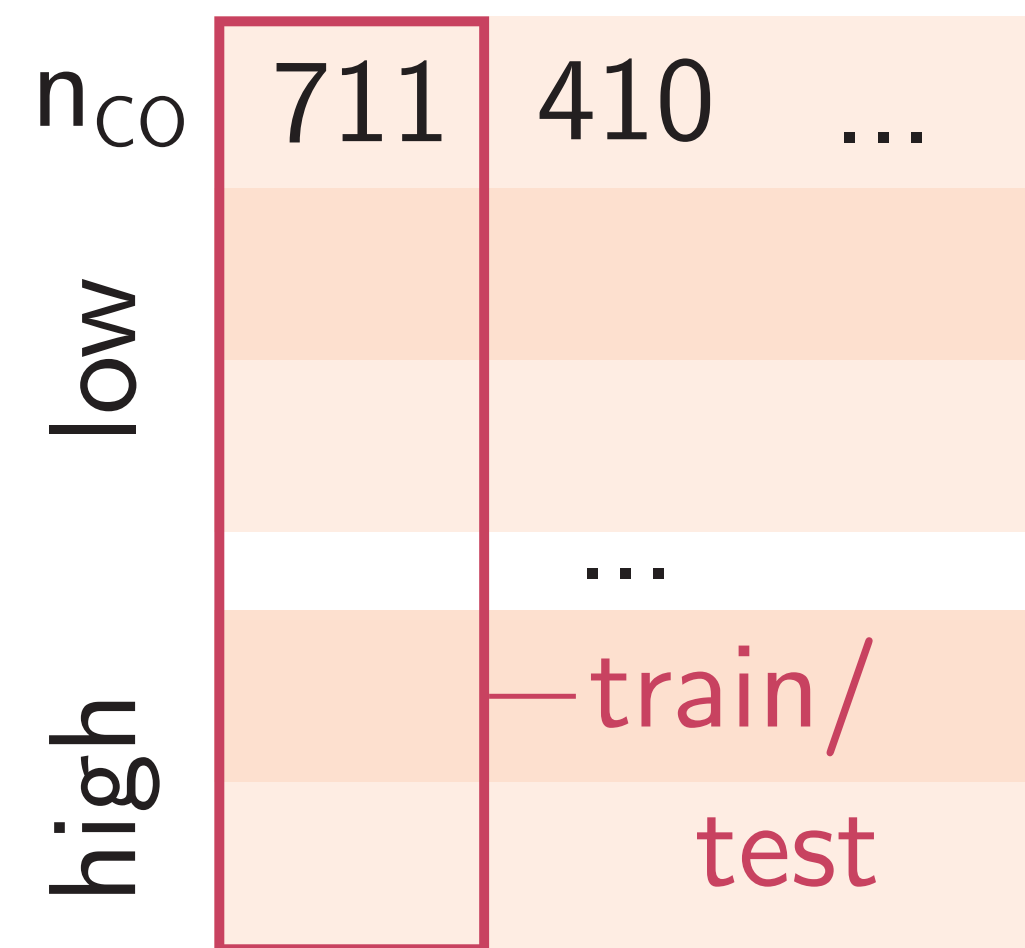
Step 2: control OOD performance
(more than just train/test)

Step 3: sample minima with confidence

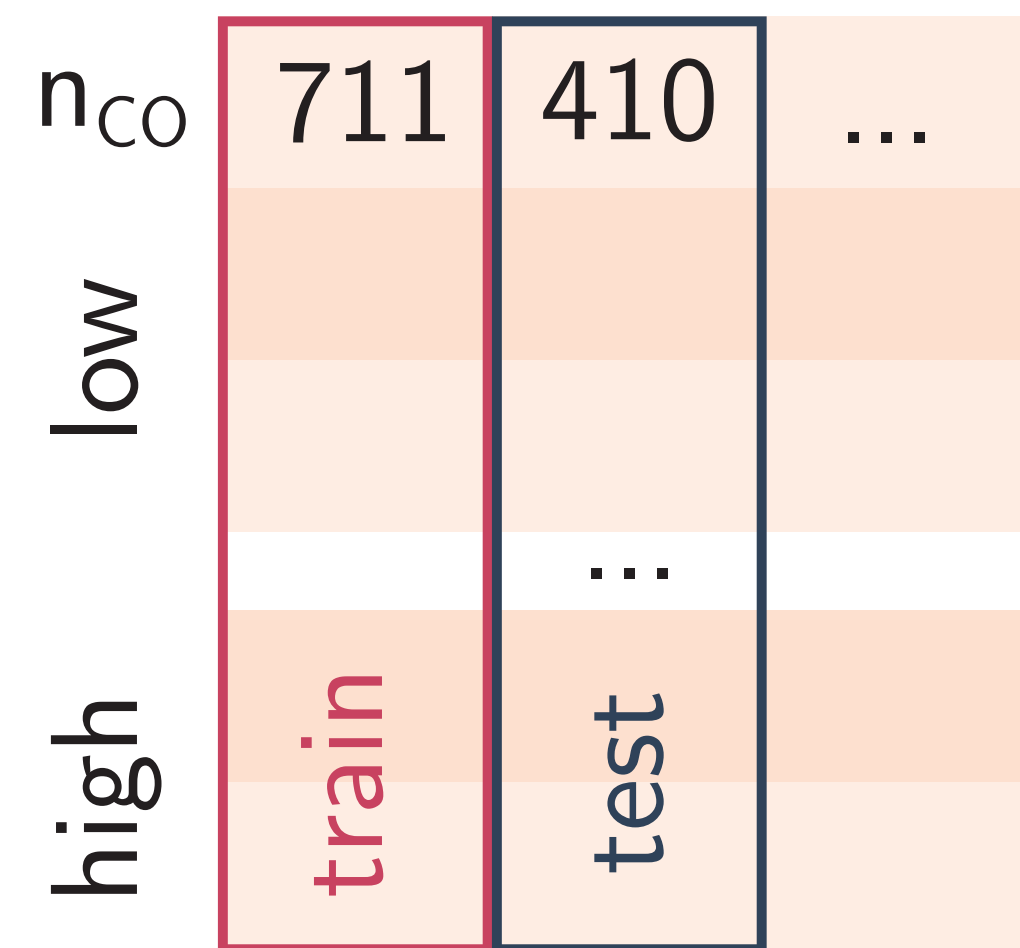


There are multiple forms of “out-of-distribution” to consider

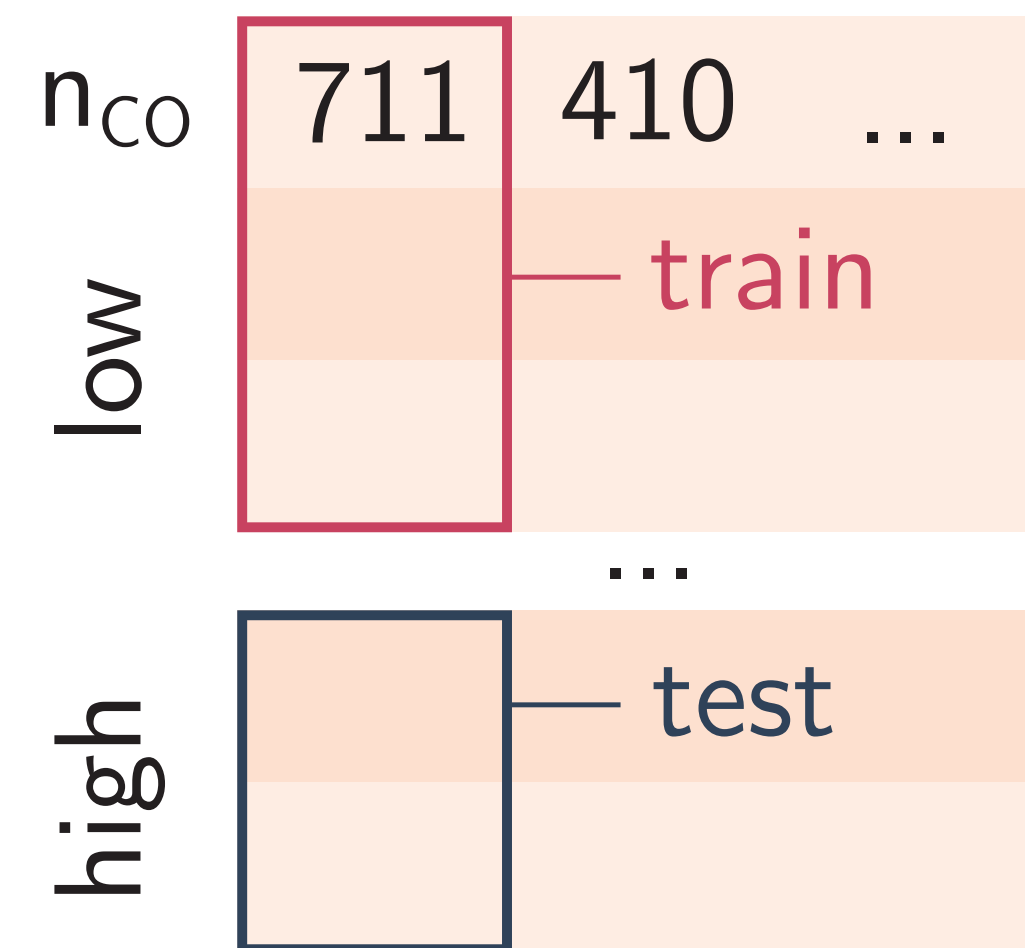
in-domain



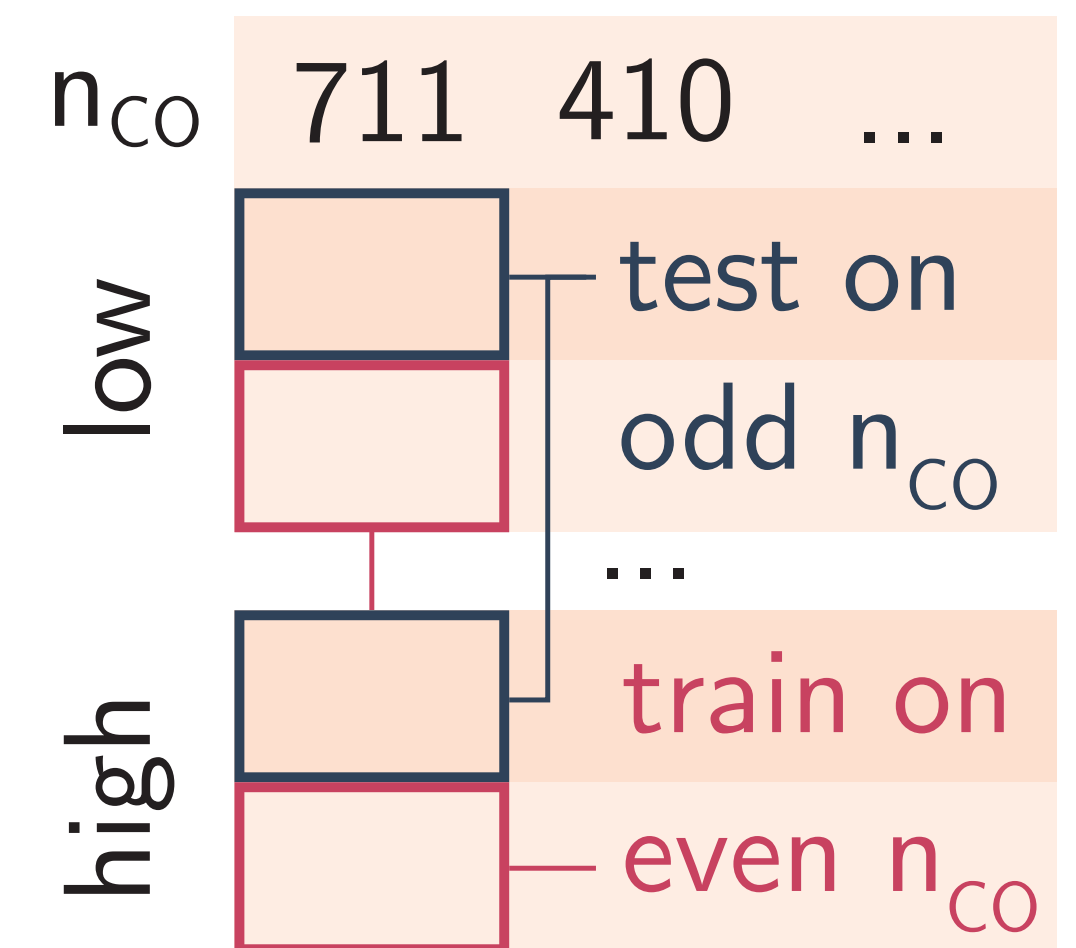
different facets

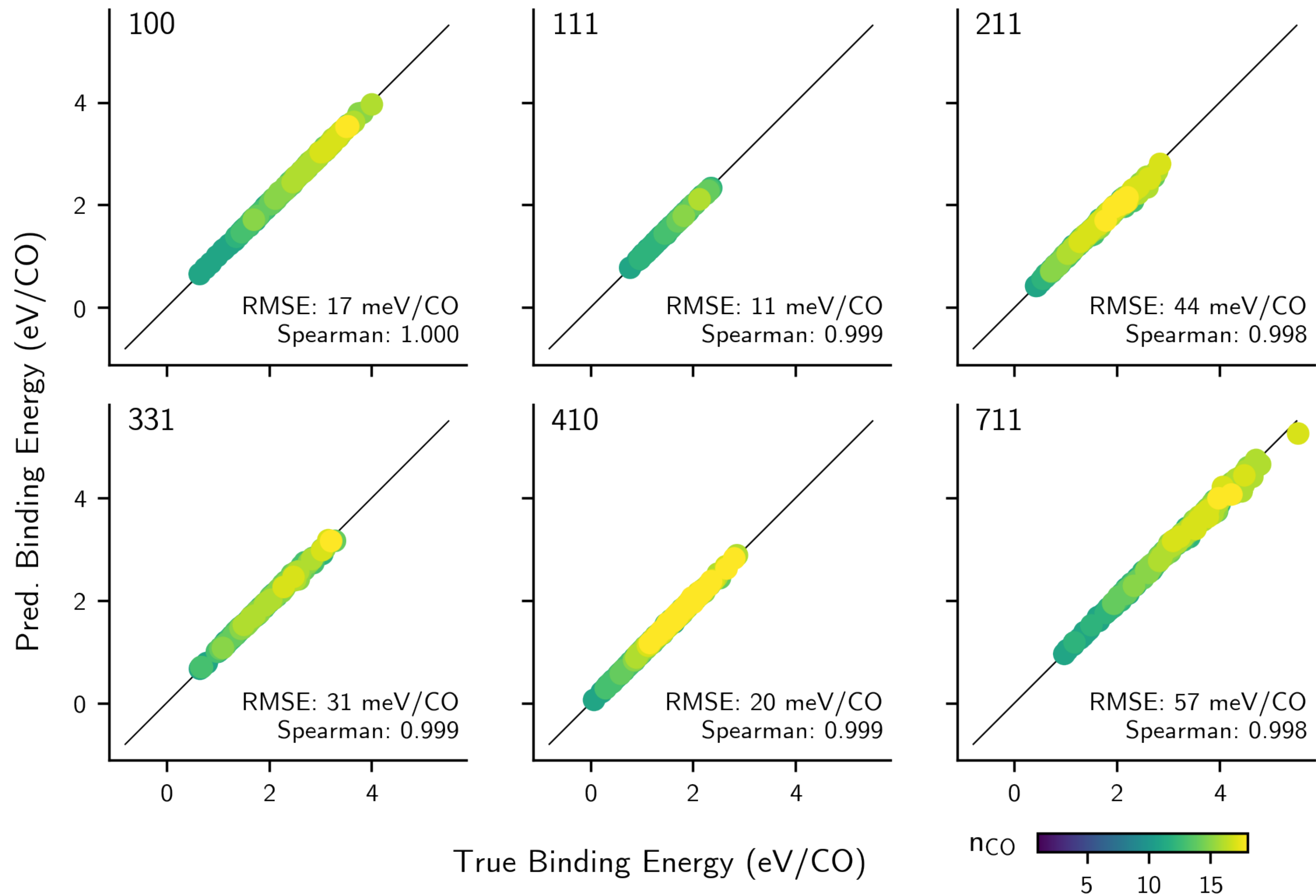


higher coverages

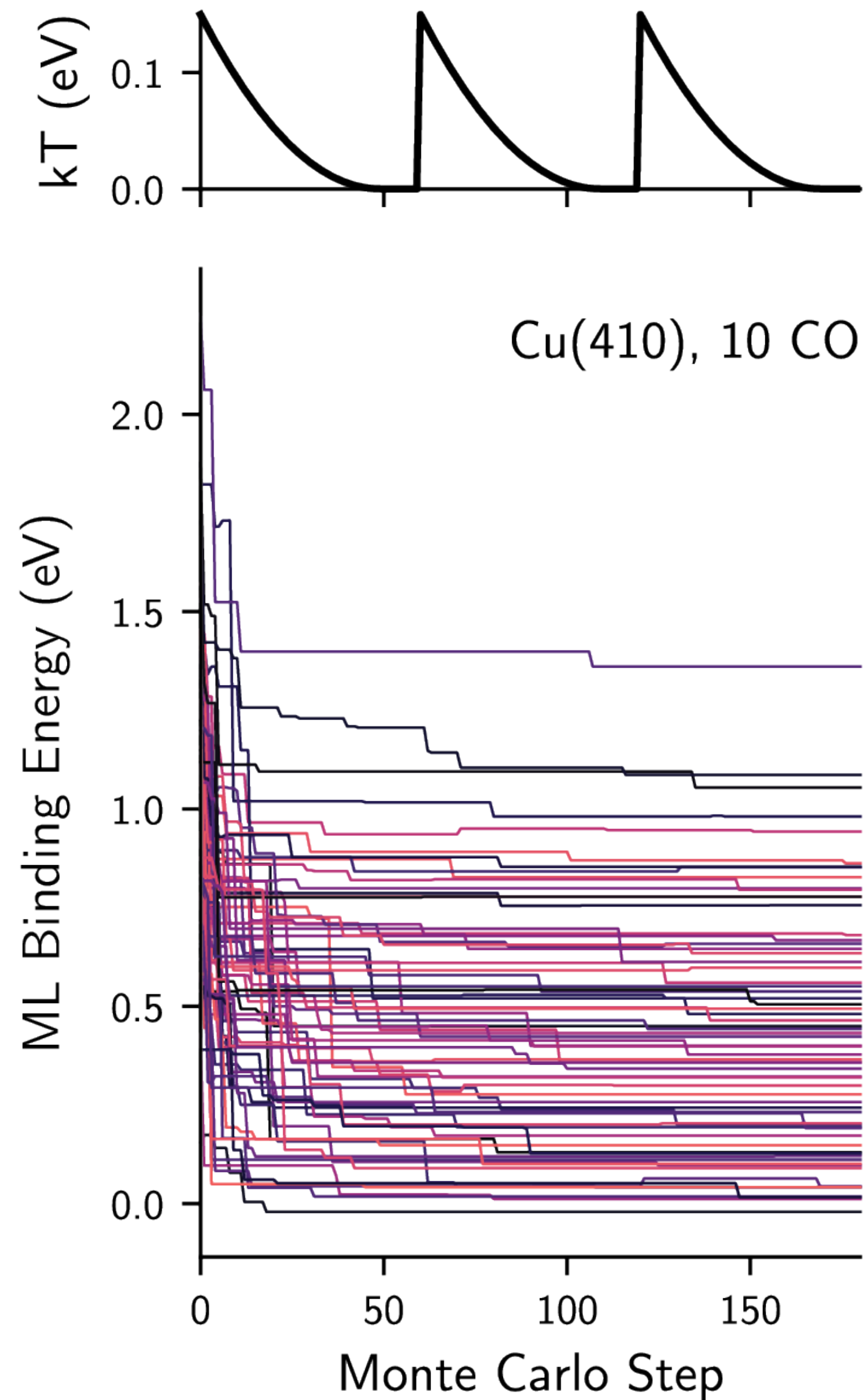


different coverages





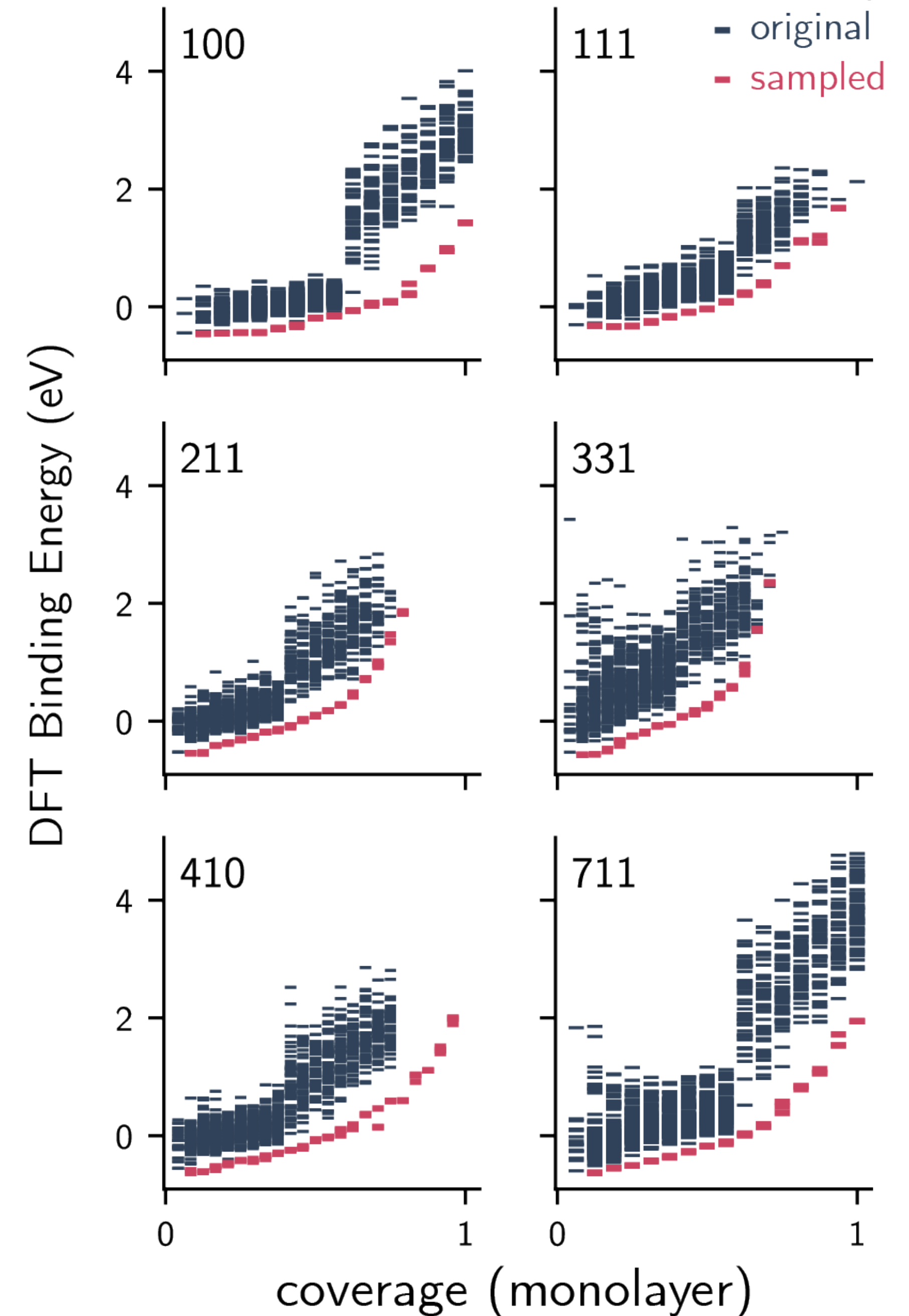
Using these models, we can sample for the lowest energy structures without active learning



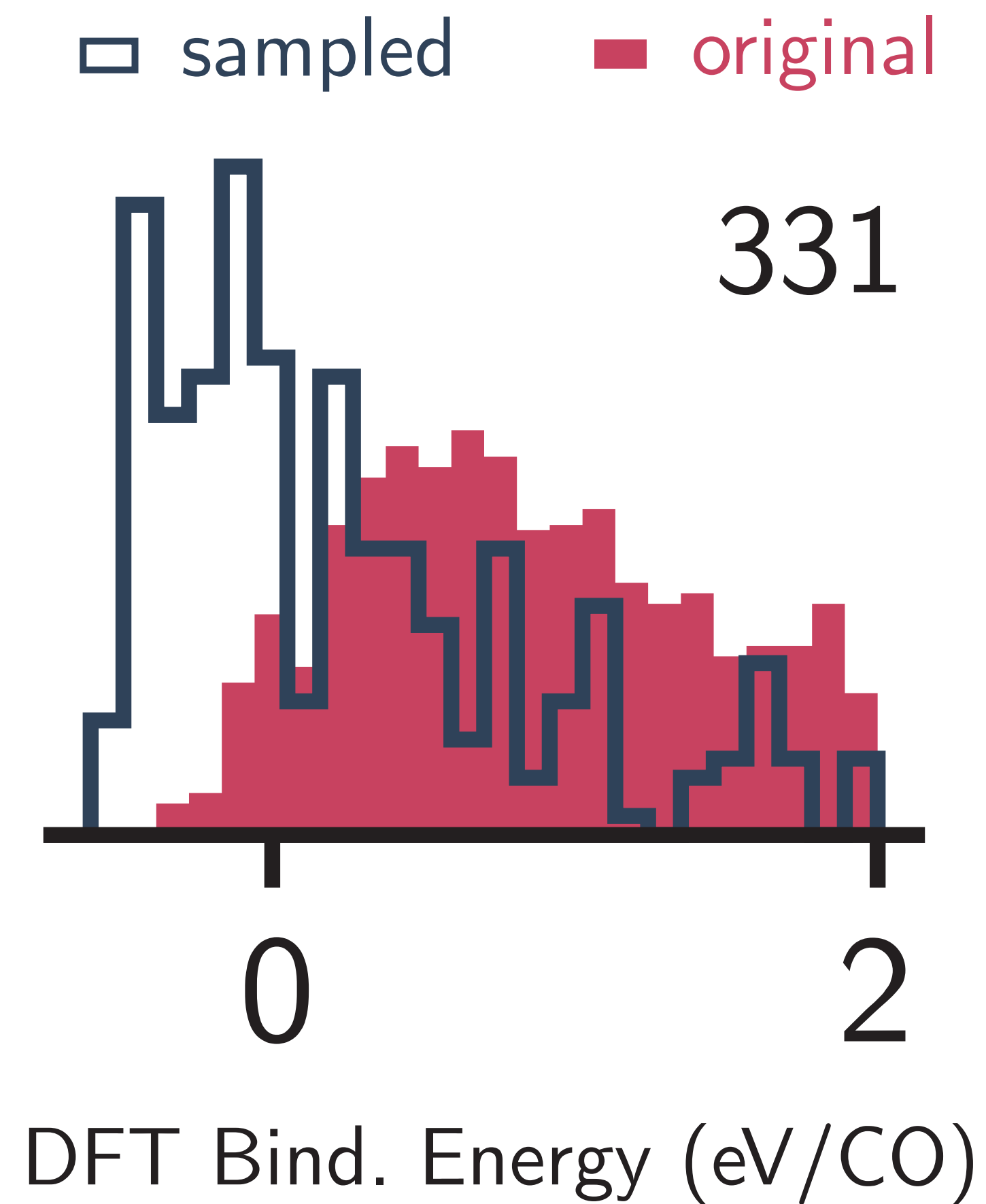
14M structures

DFT for low E_b

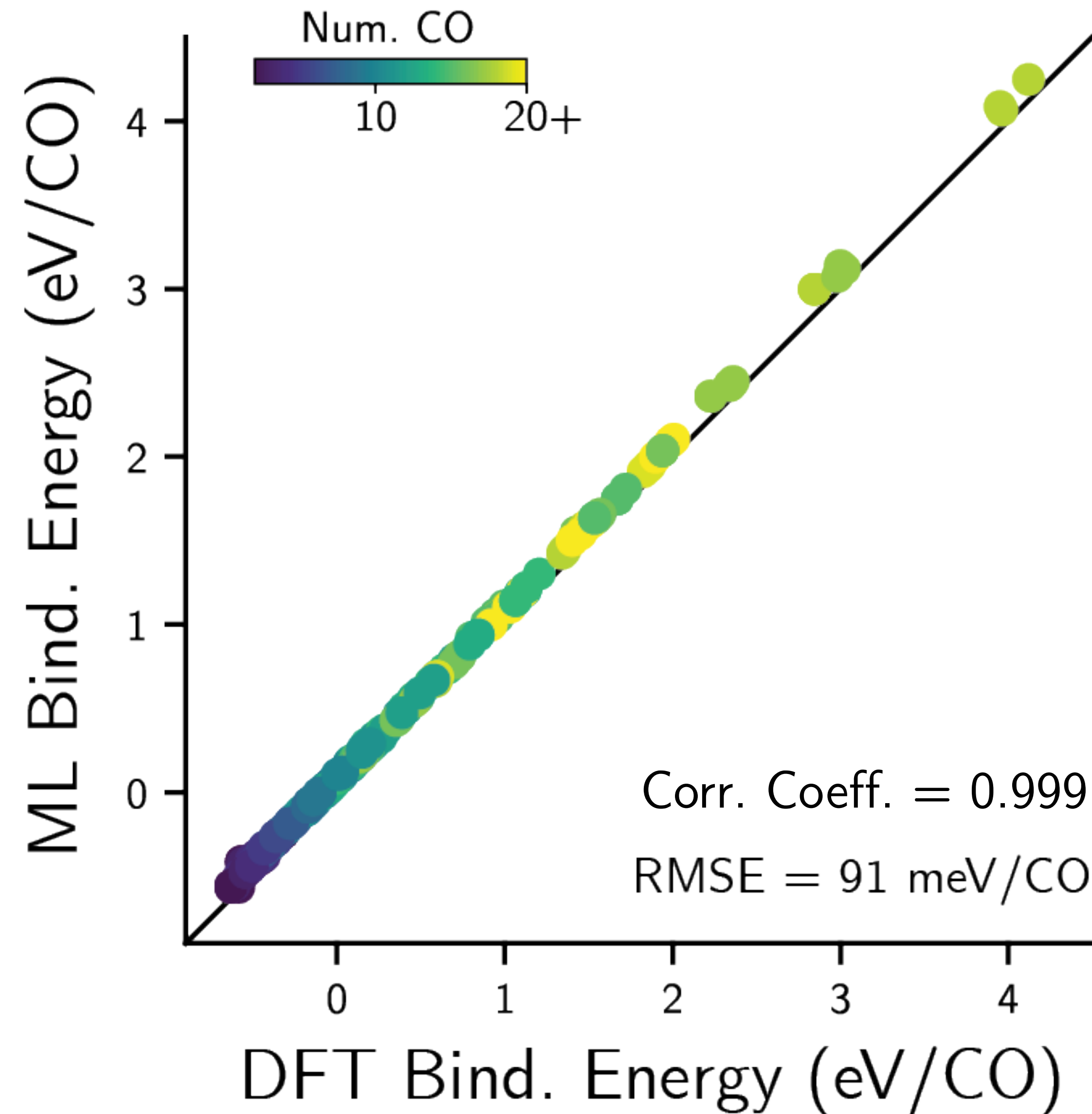
>180k energy
evaluations
per facet
per coverage

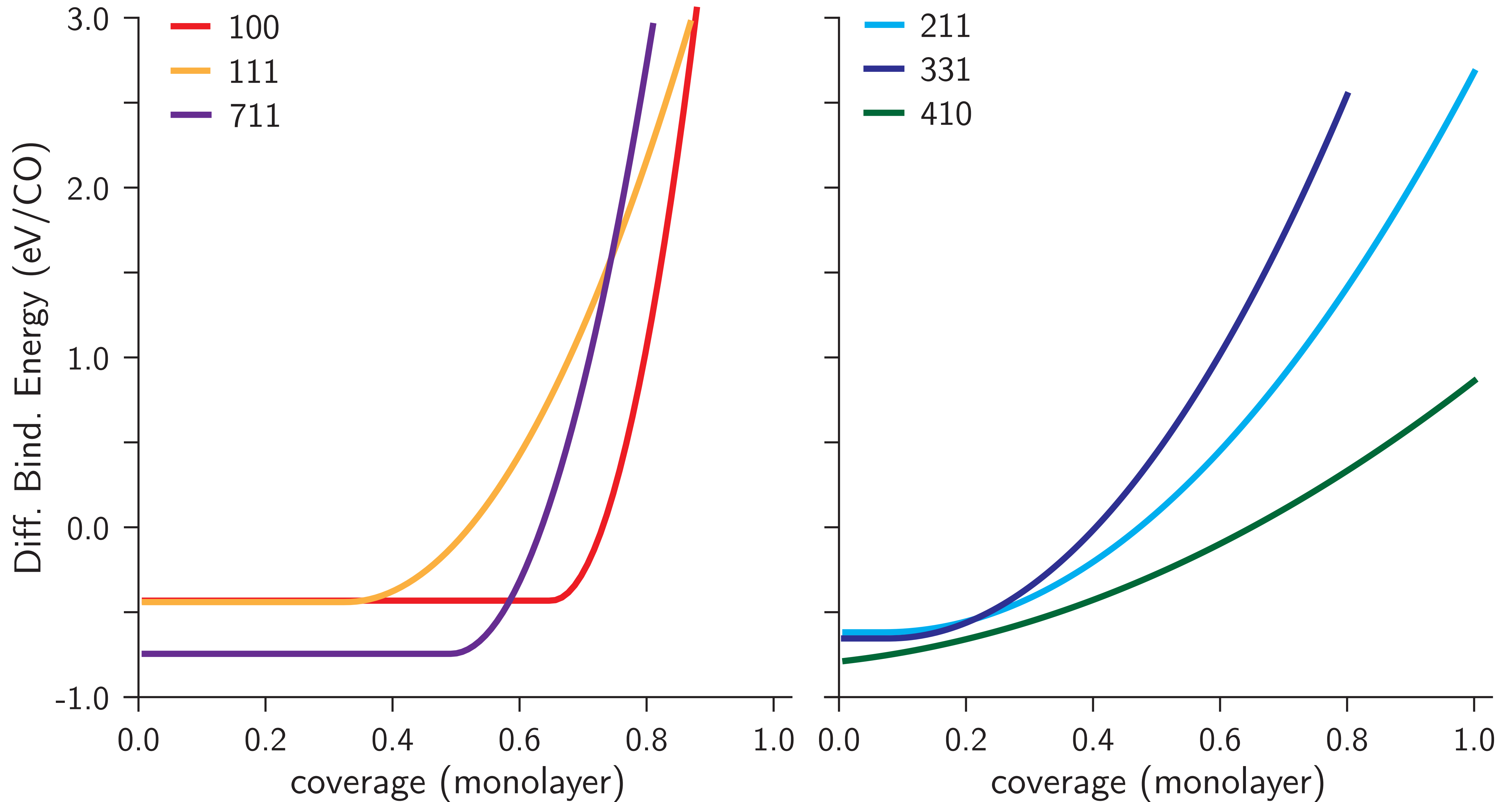


Revalidating: the model indeed performs really well for these low-energy structures



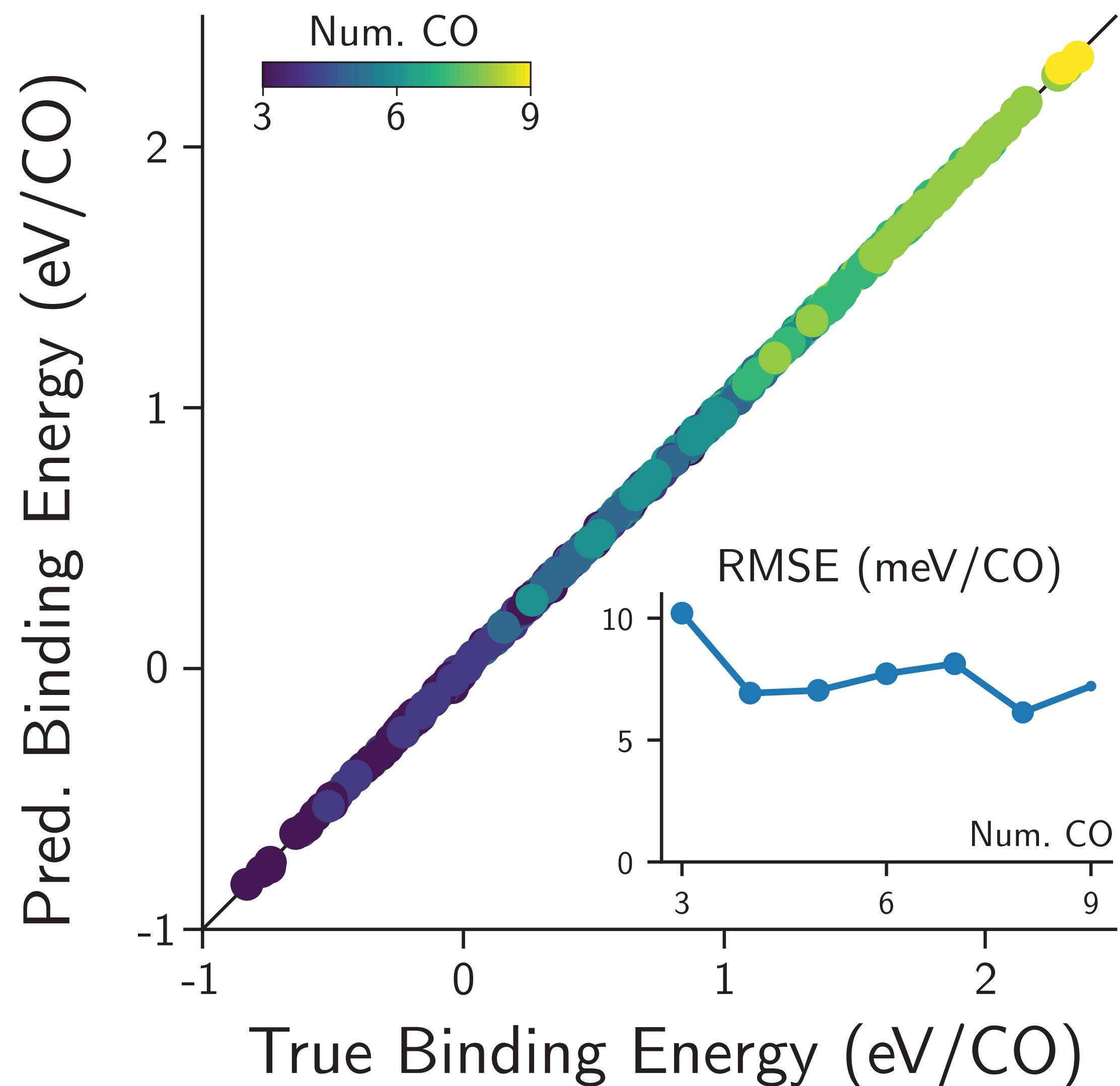
low E_b
↓



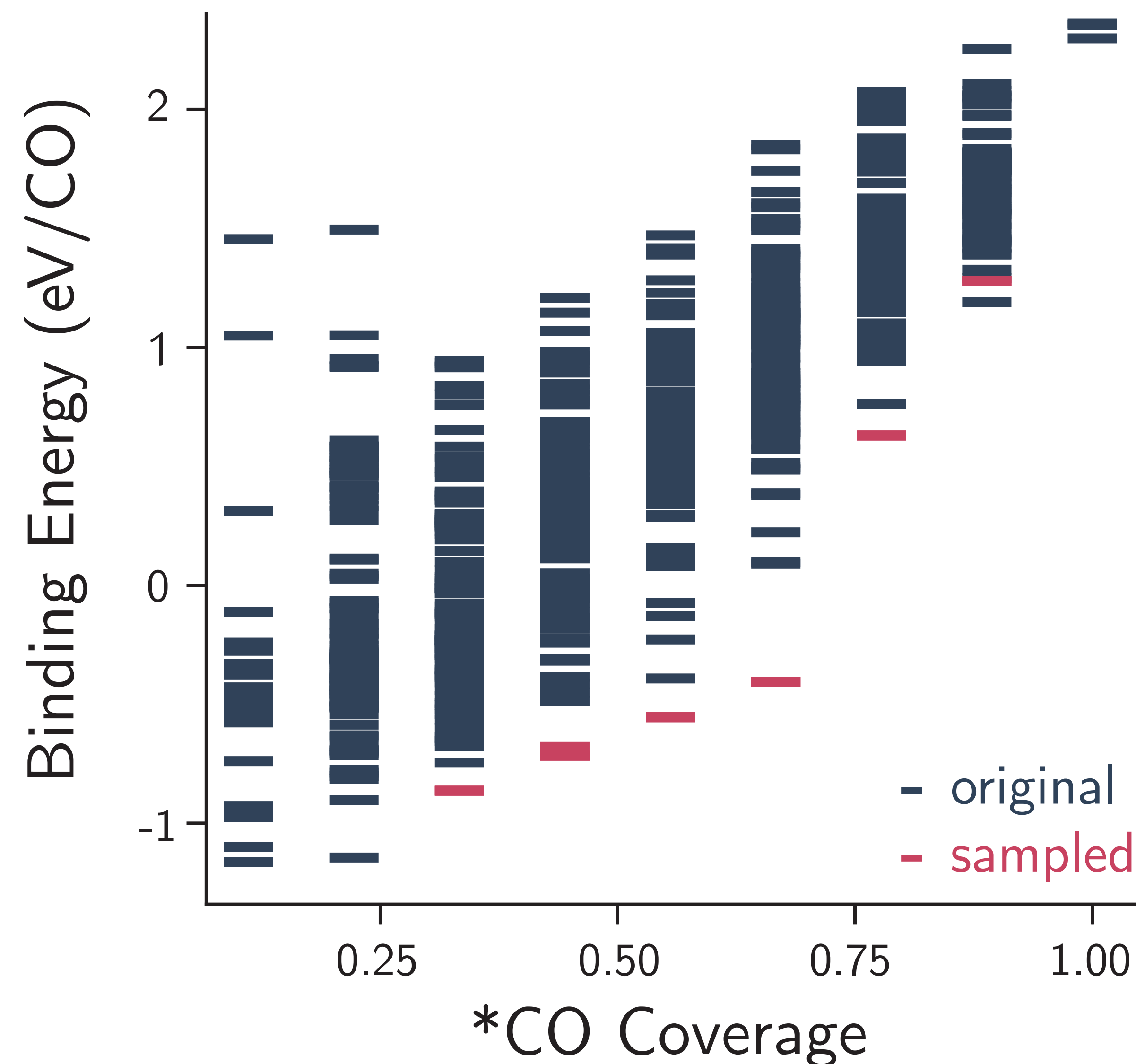


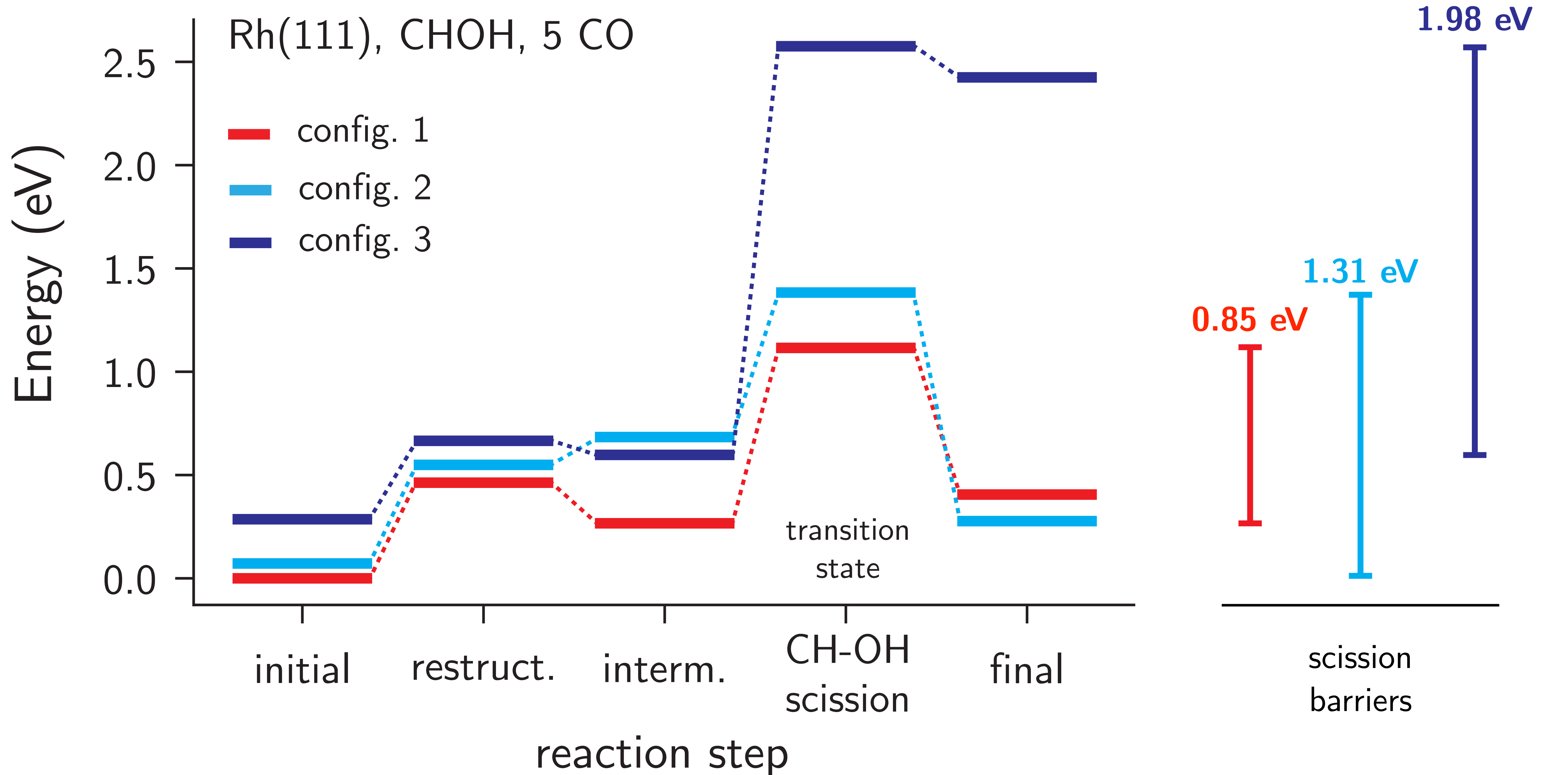
Harder example: co-adsorption of $\ast\text{CHOH}$ and $\ast\text{CO}$ on Rh(111)

Step 1: train model



Step 2: sample new data points

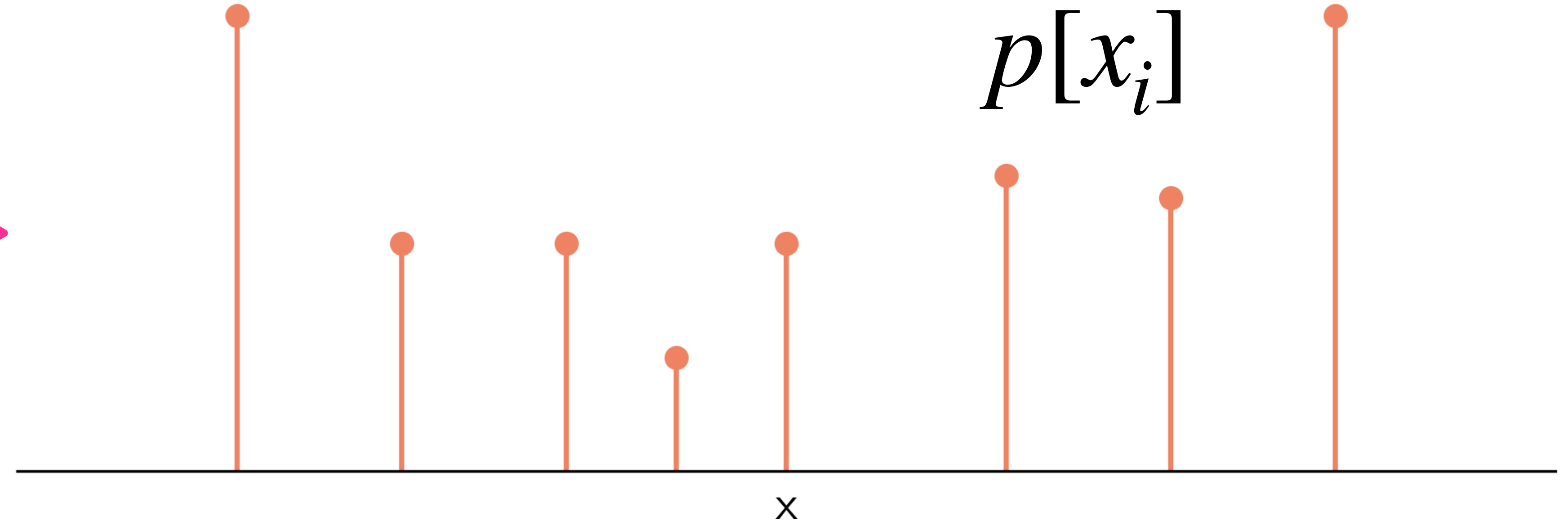




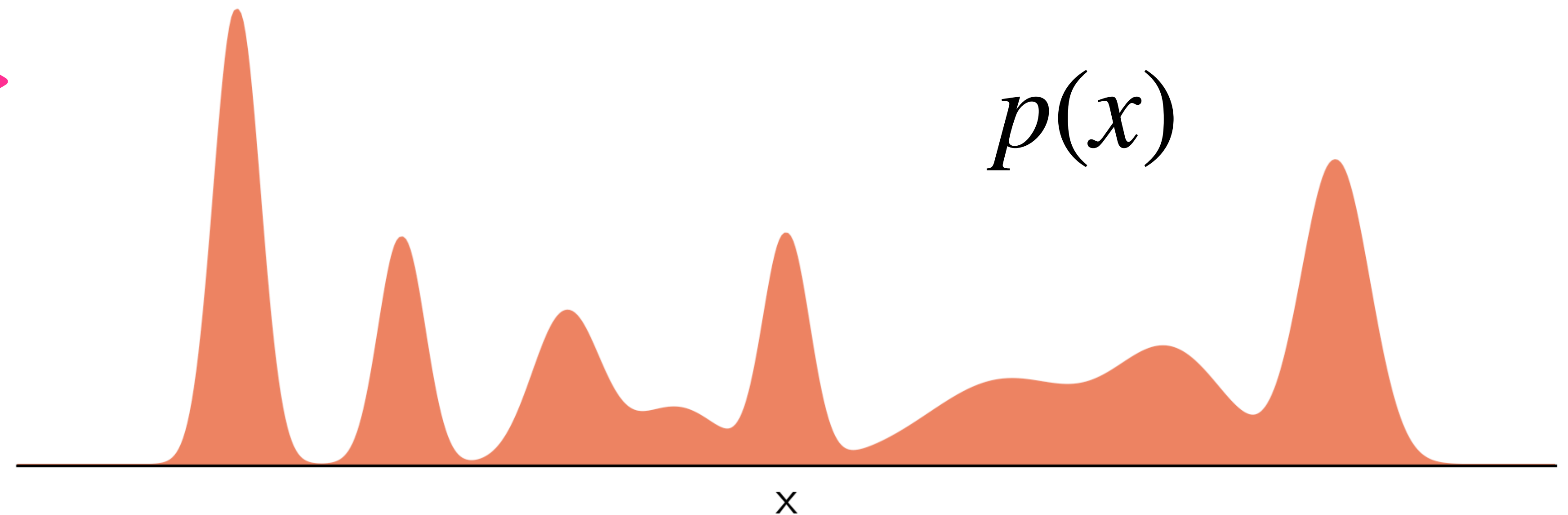
So far, however, we talked about discrete states, and not a continuous distributions

16

Final states:
equilibrium only

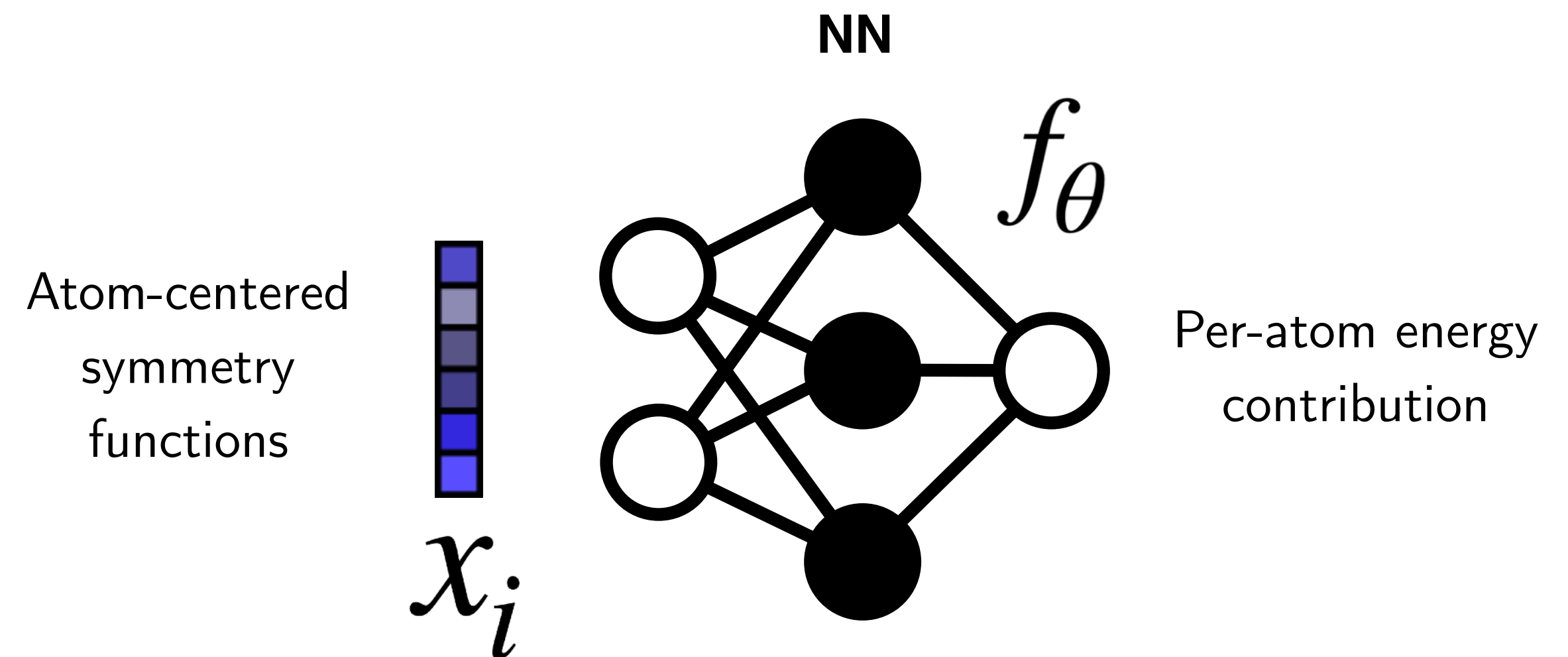
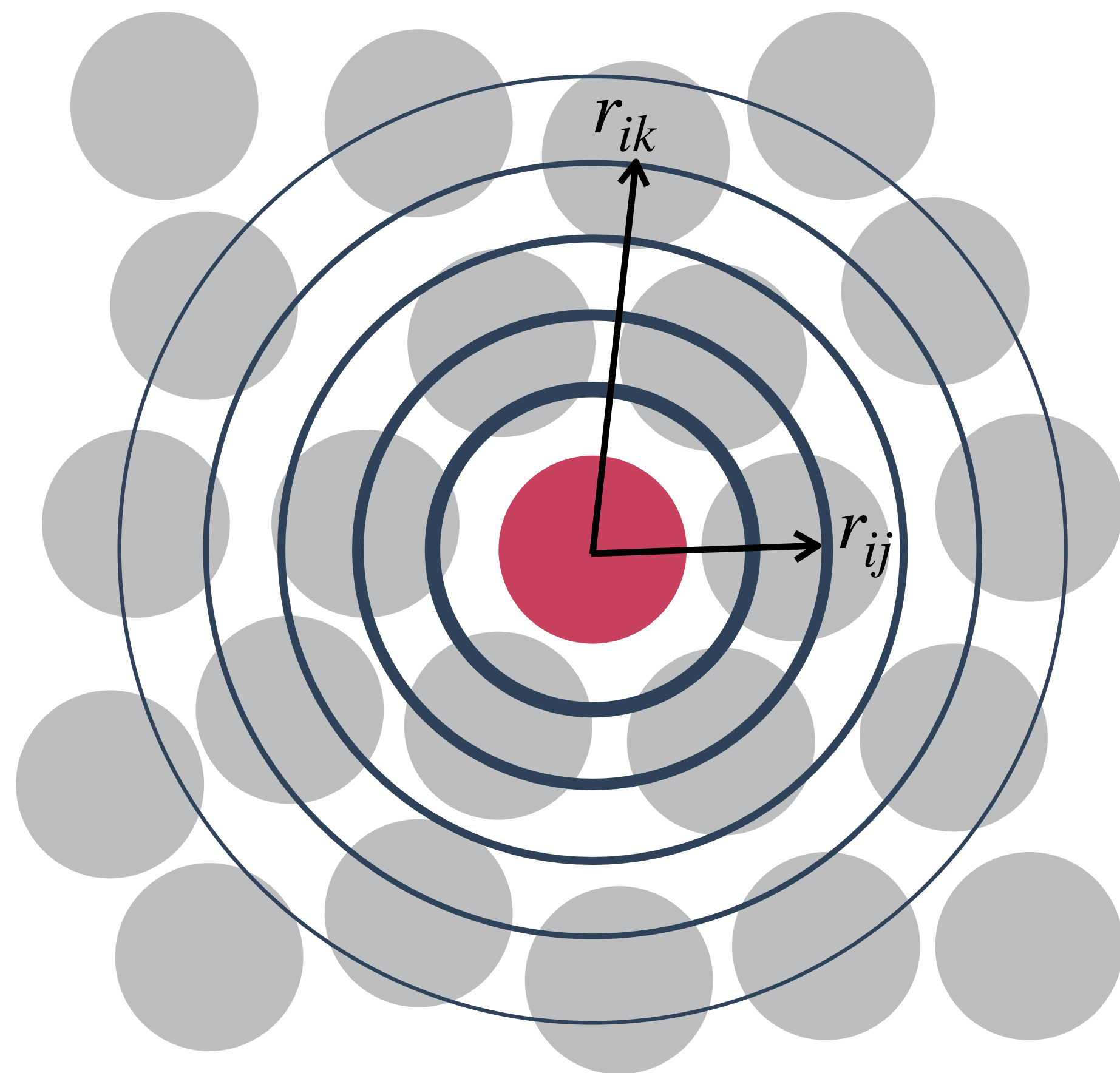


Intermediate states:
dynamics and pathways



How to sample these
Boltzmann-like
distributions for materials
dynamics?

From atomistic machine learning: the Behler-Parrinello scheme for potential energies



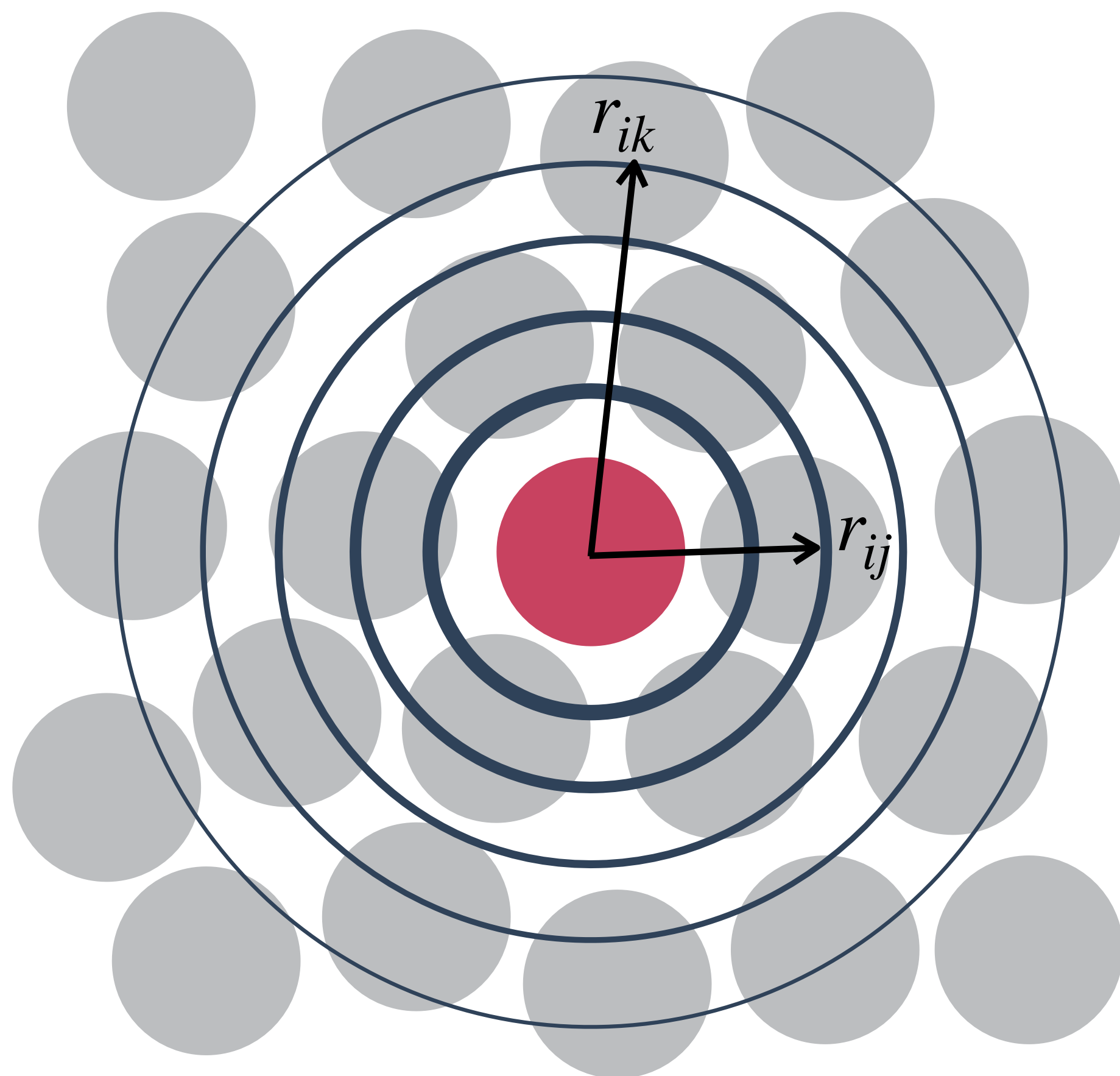
The final energy is computed as:

$$E = \sum_i f_\theta(x_i)$$

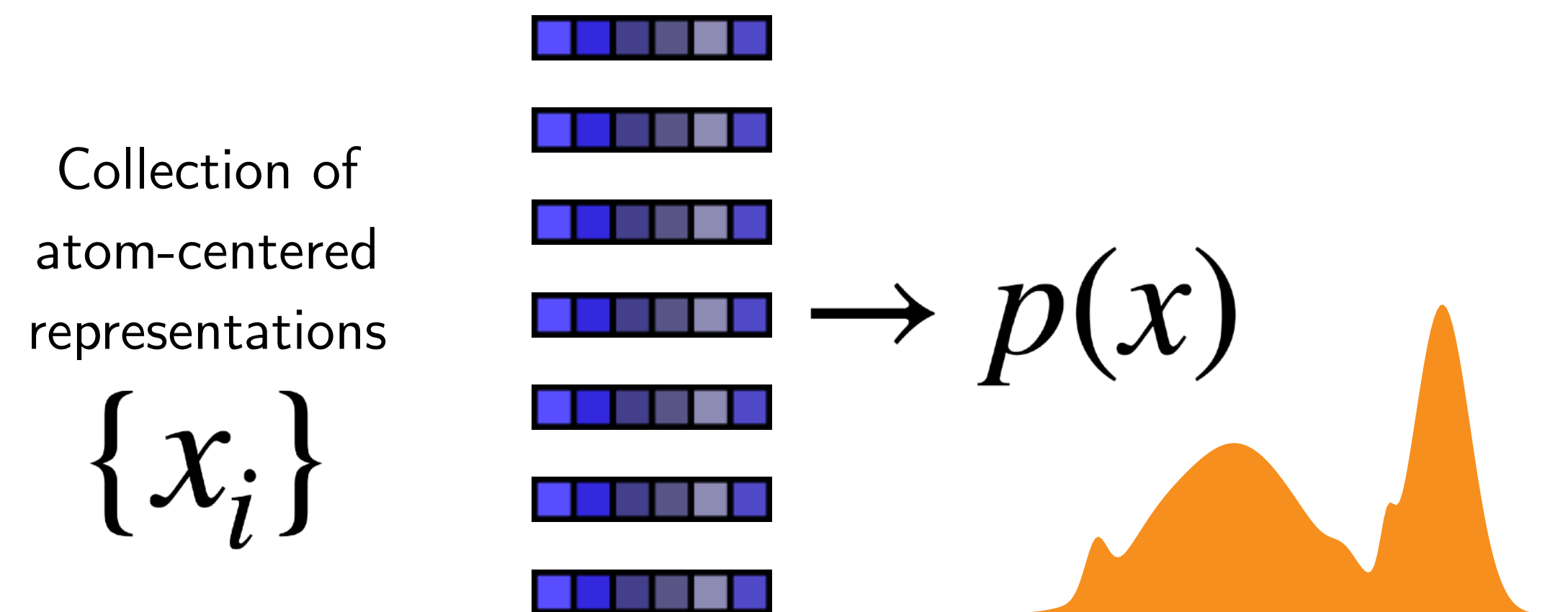
Atom-centered representations are computed for each atom of the structure up to a cutoff

Concept: atomistic structures $\rightarrow$ vector representations $\rightarrow$ model $\rightarrow$ energy

Similar approach for computing the (information) entropy



Atom-centered representations are computed for each atom of the structure up to a cutoff



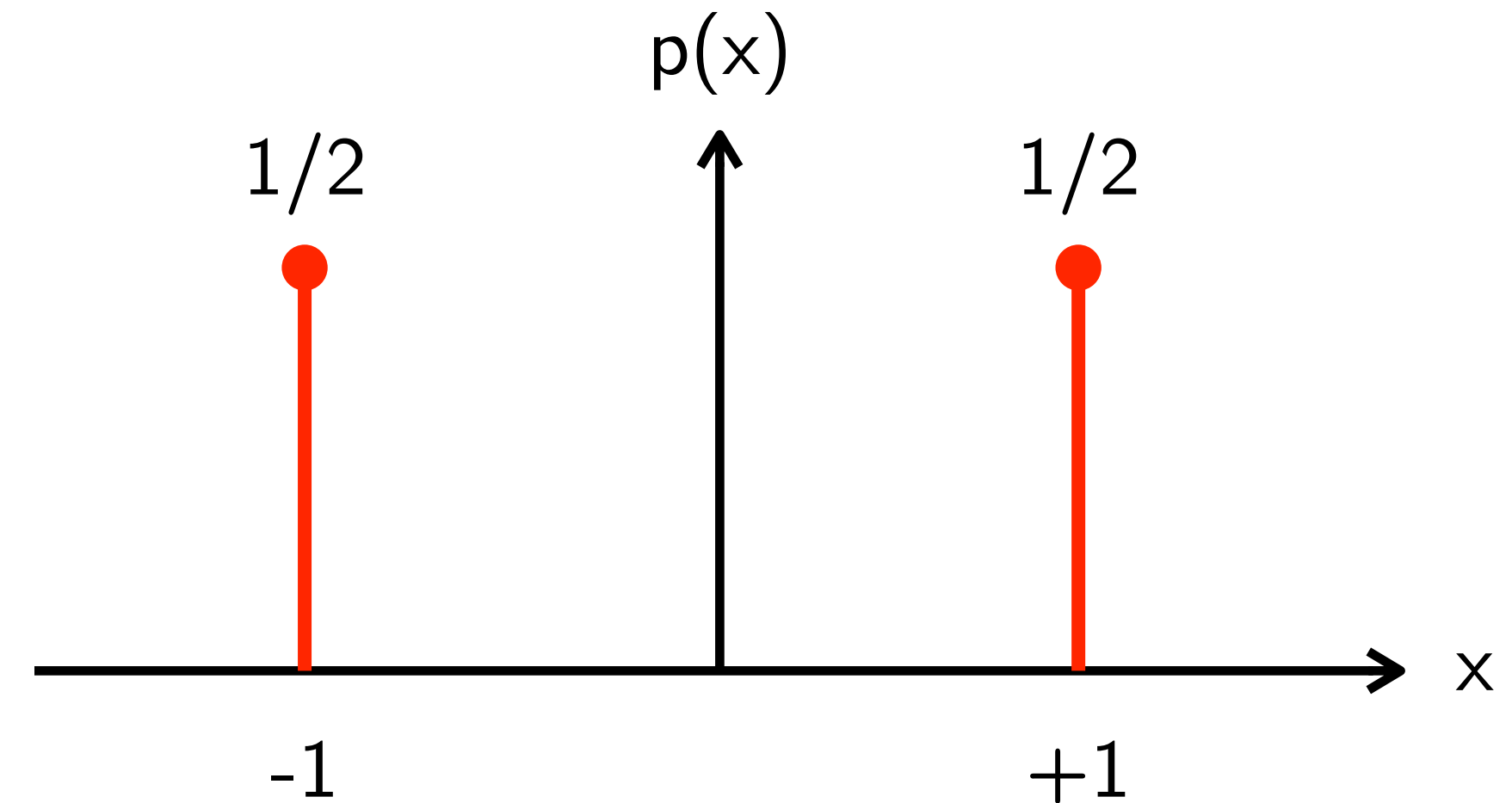
The information entropy is computed as:

$$\mathcal{H} [p(x)] = - \int p(x) \log p(x) dx,$$

Concept: atomistic structures $\rightarrow$ vector
representations $\rightarrow$ distribution $\rightarrow$ info. entropy



$$p(H) = 1/2 \quad p(T) = 1/2$$



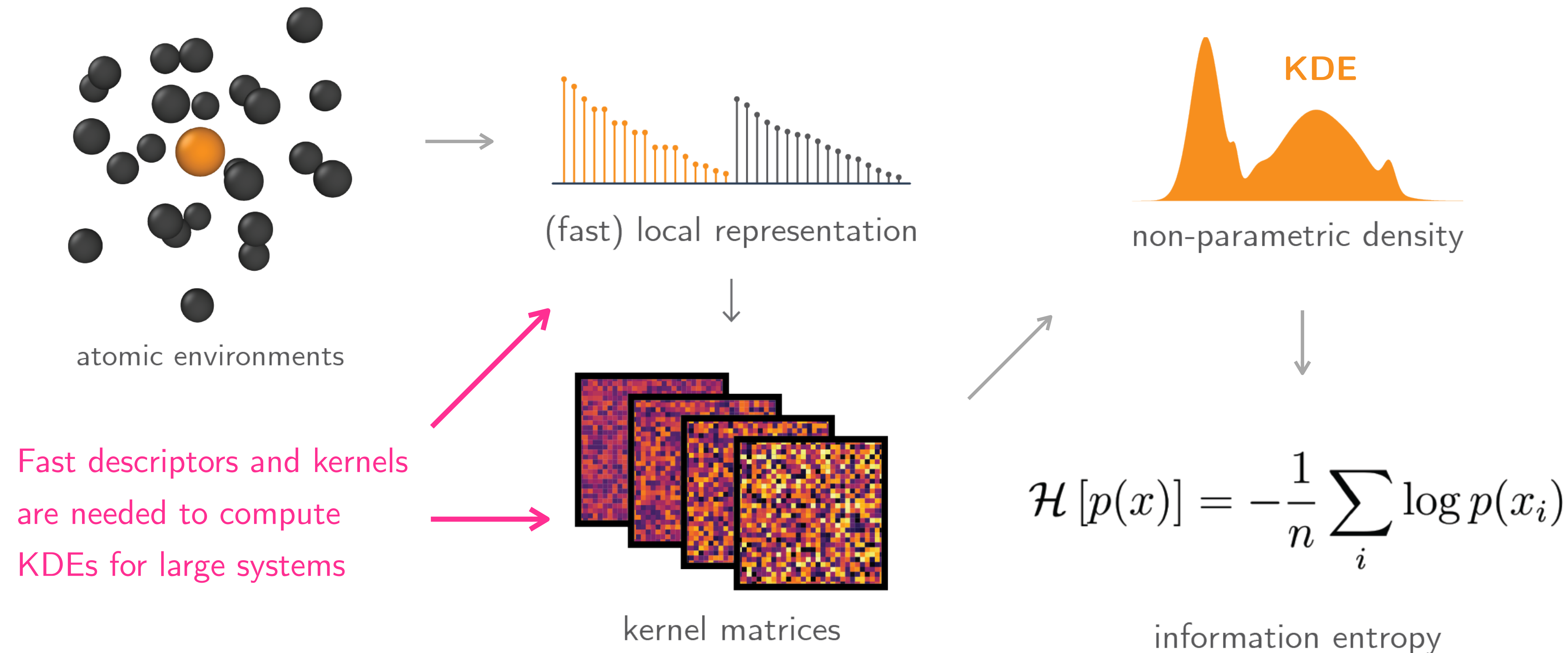
$$\mathcal{H} = - \left(\frac{1}{2} \log_2 \frac{1}{2} + \frac{1}{2} \log_2 \frac{1}{2} \right) = 1 \text{ bit} \rightarrow \text{Amount of information contained in a coin: info. entropy}$$

Useful
intuitions:

higher “spread” of a distribution
→ higher information entropy

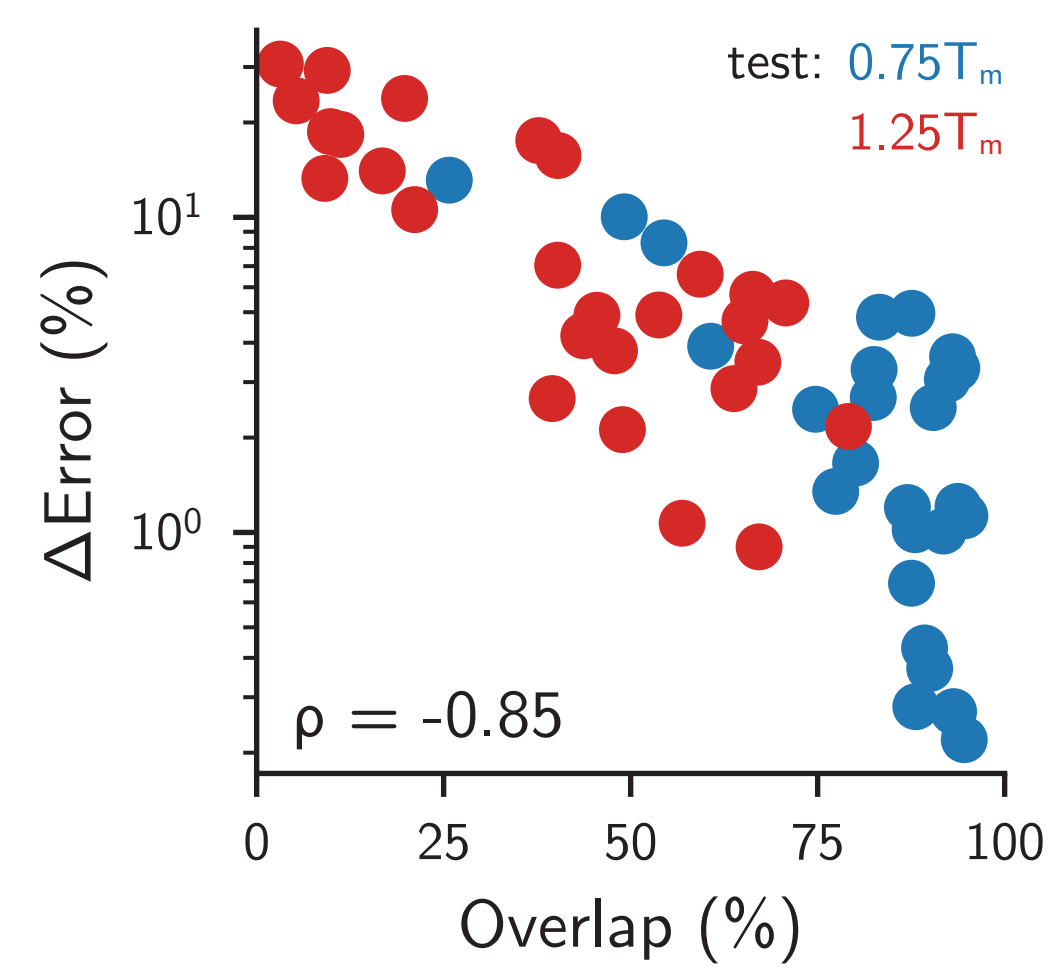
Min. Entropy = 0 (Dirac delta distribution)
Max. Entropy = $\log n$ (Uniform with n points)

Our Method: QUESTS = Quick Uncertainty and Entropy via STructural Similarity

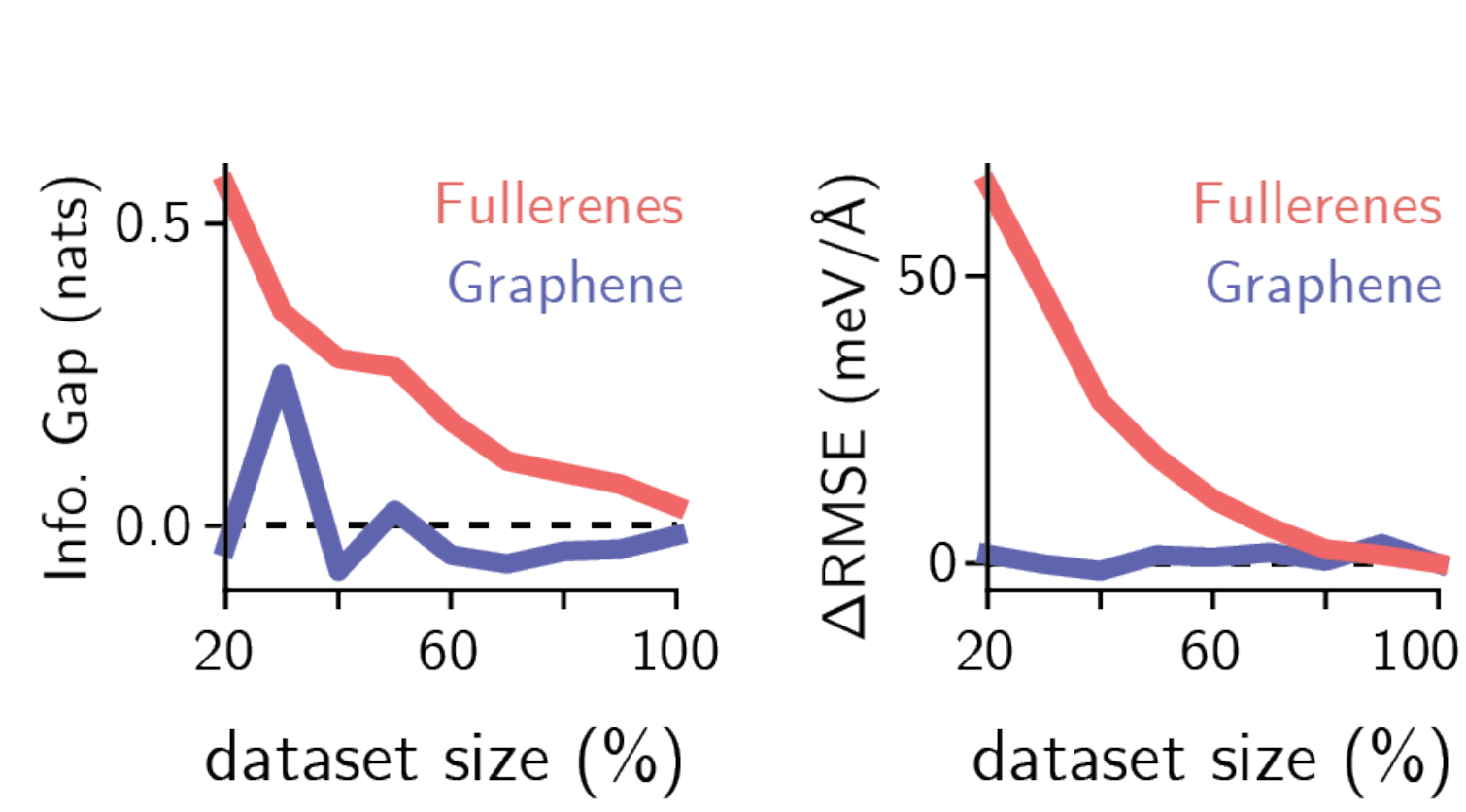


Information entropy can be used for numerous tasks in atomistic simulations

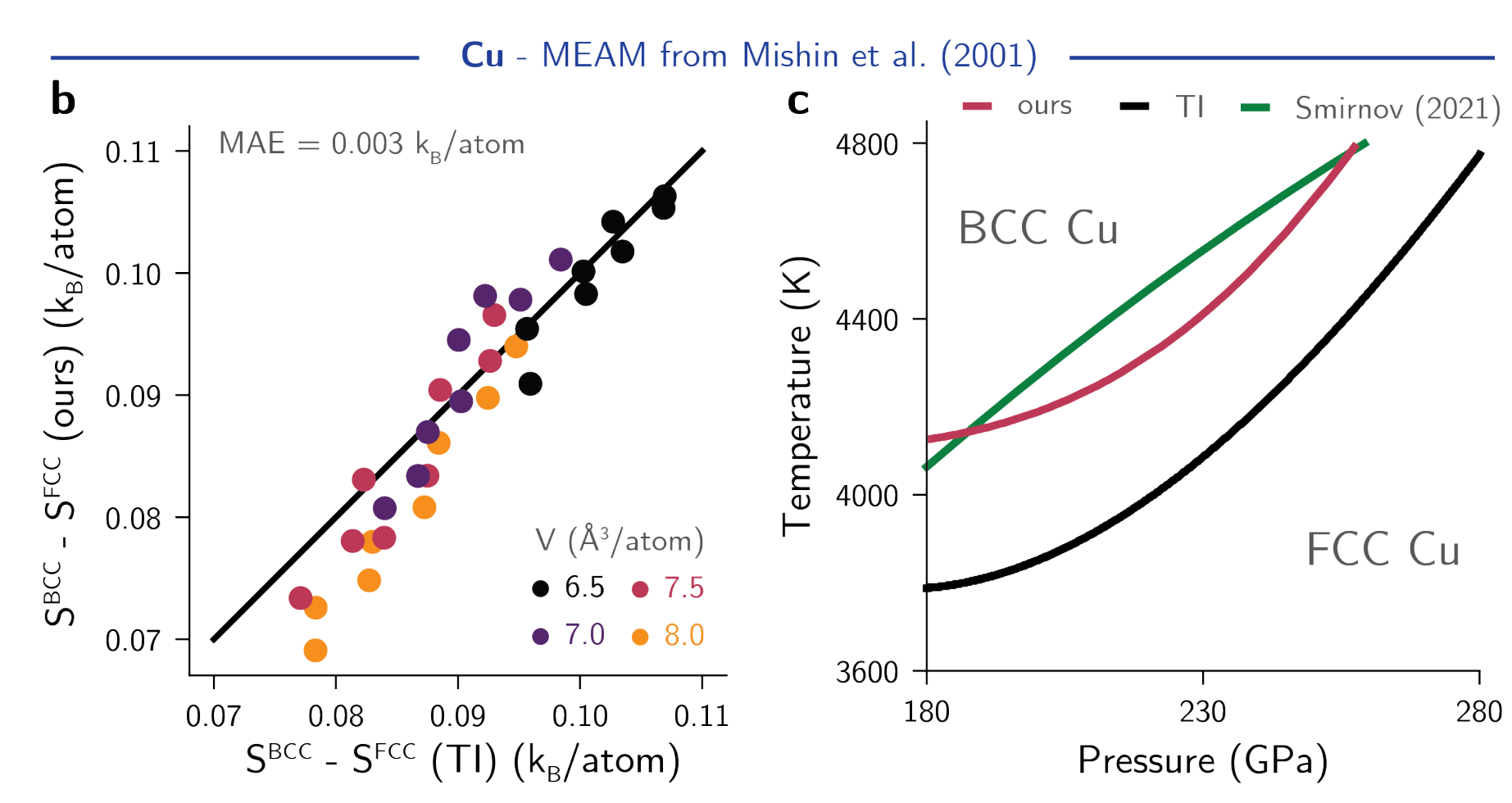
Rationalizing “extrapolation”



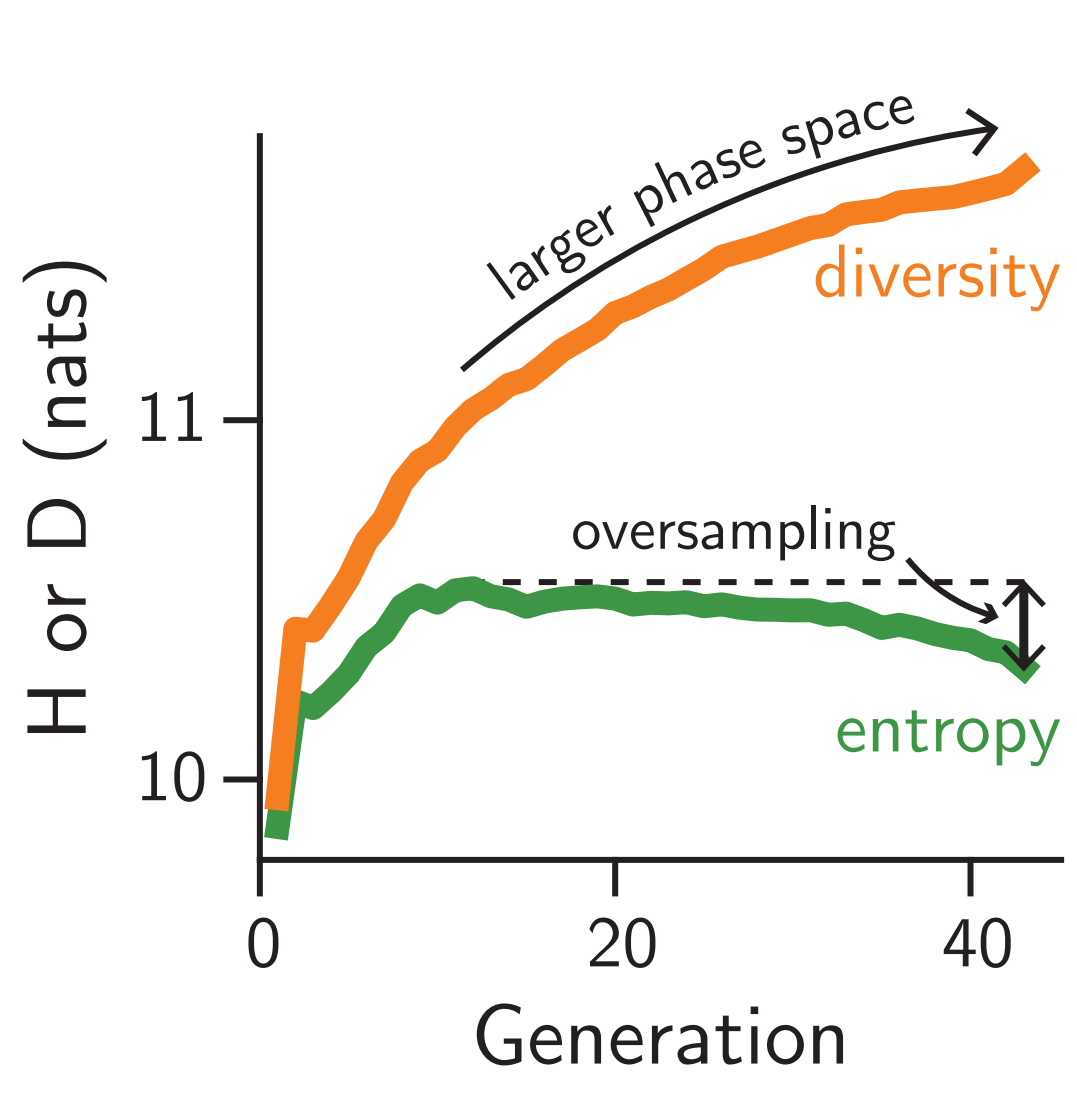
Dataset compression



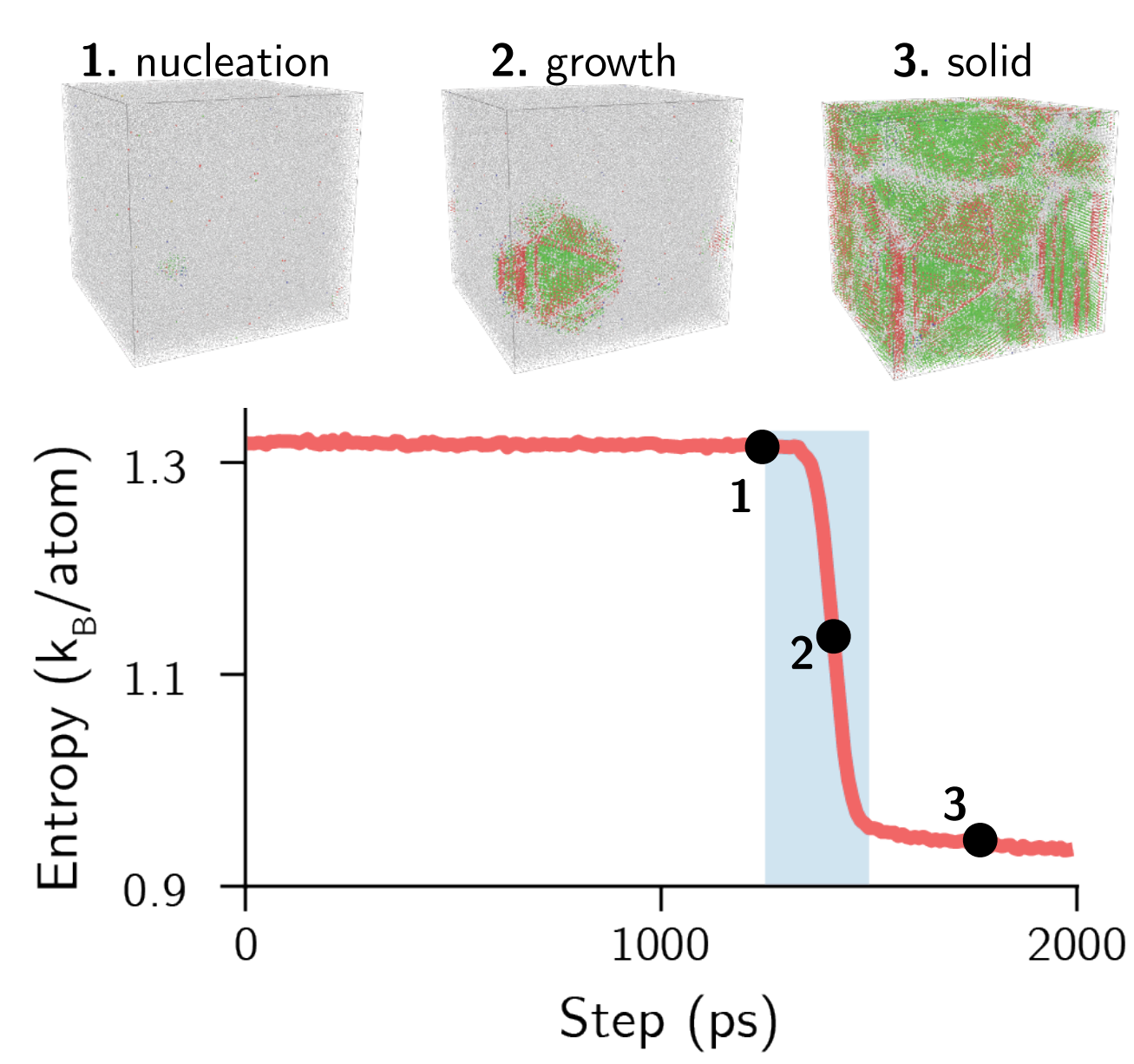
Phase diagrams and vibrational entropy



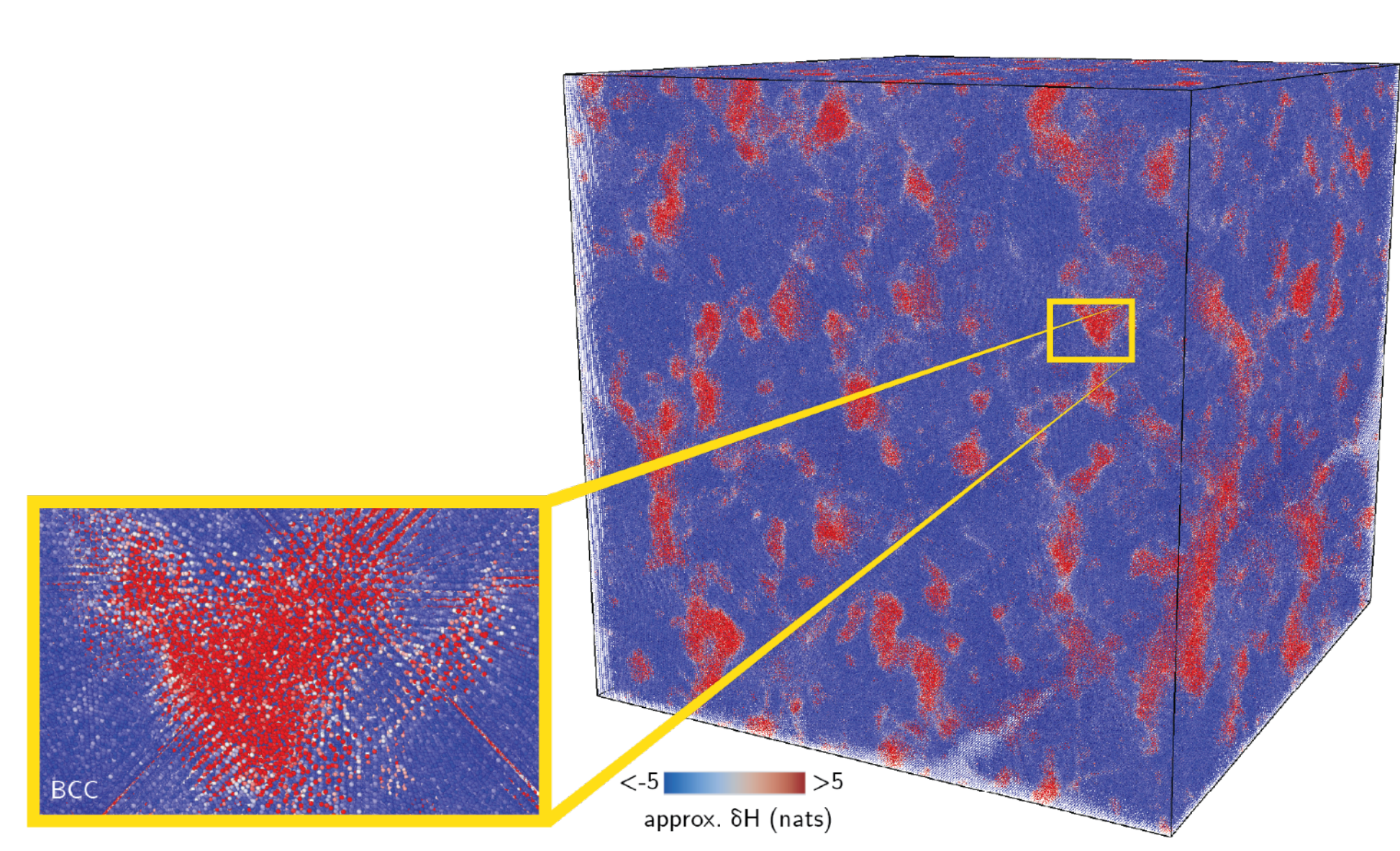
Measuring active learning performance



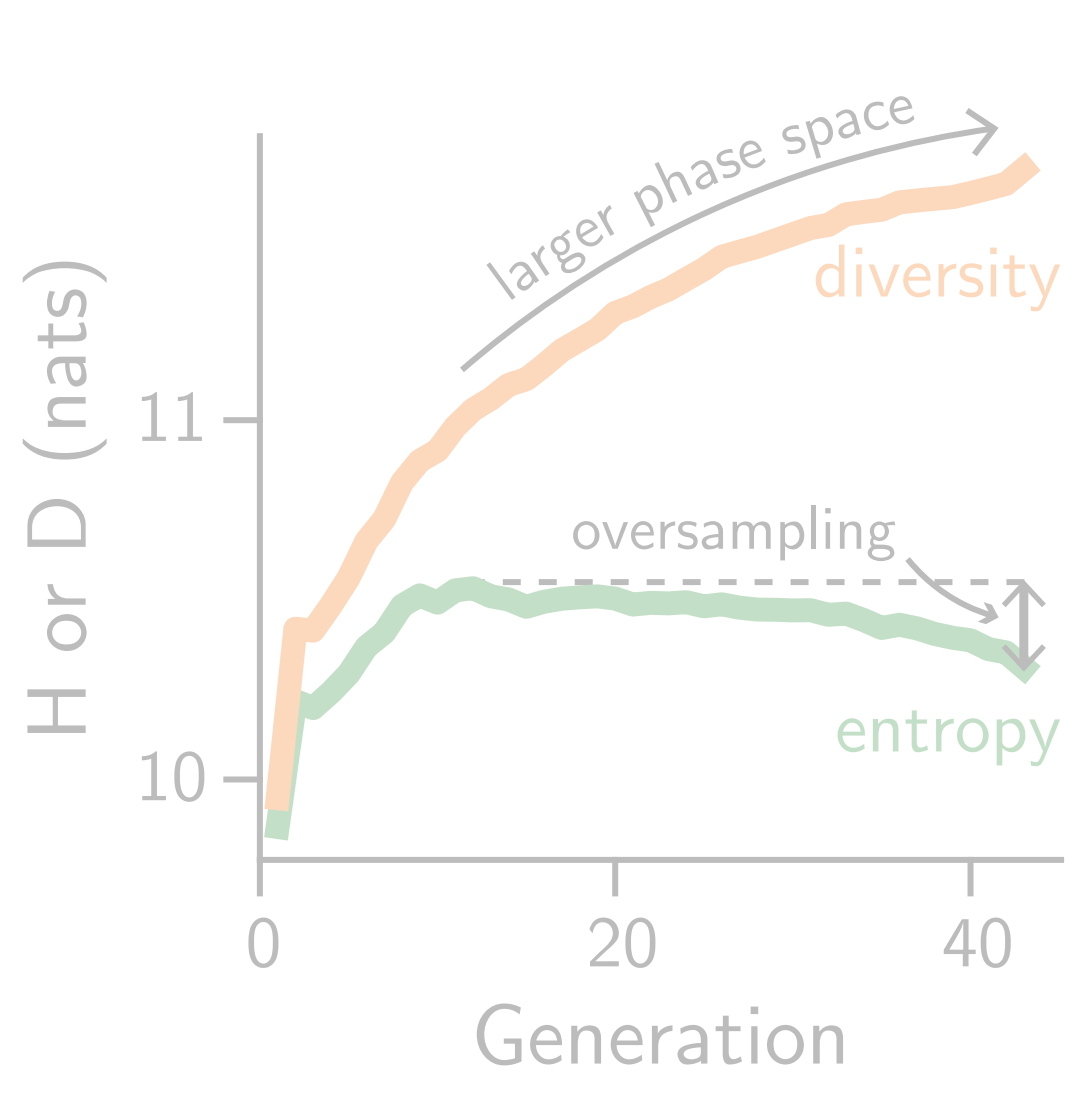
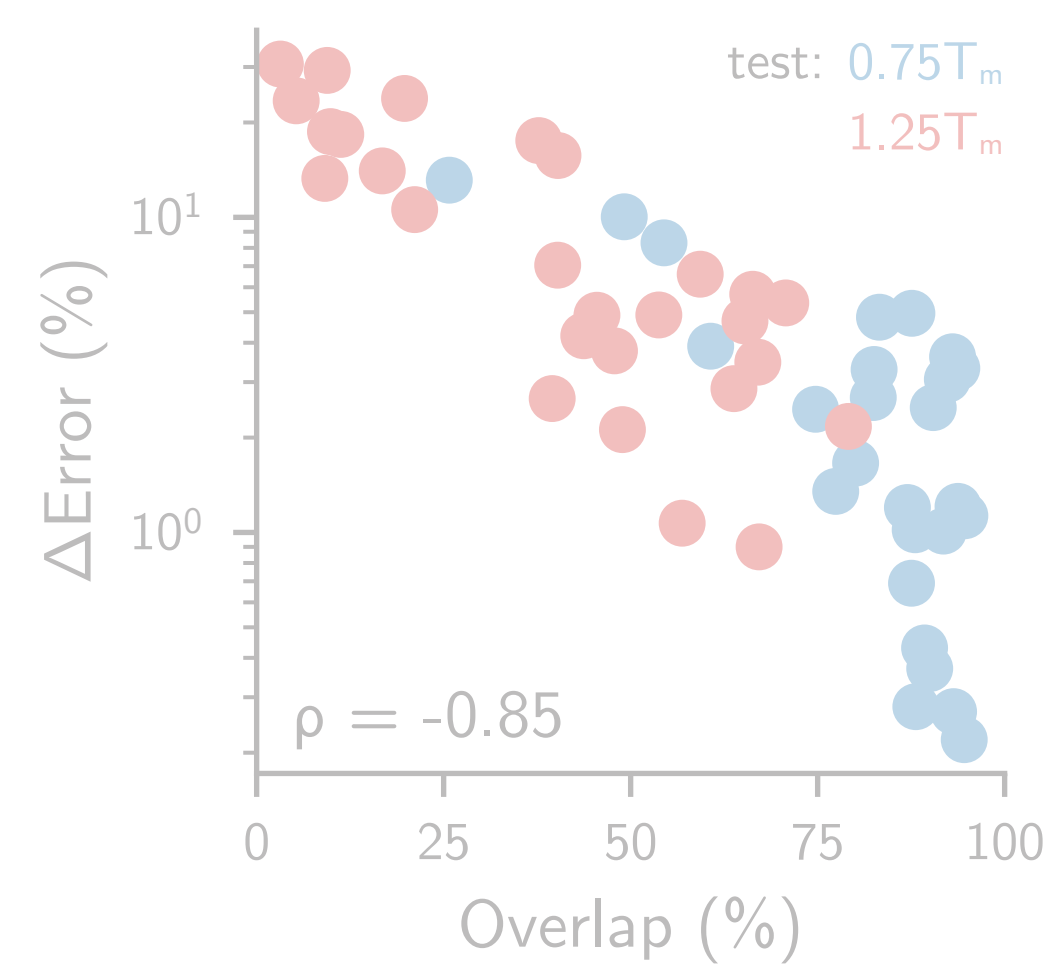
Modeling nucleation and growth



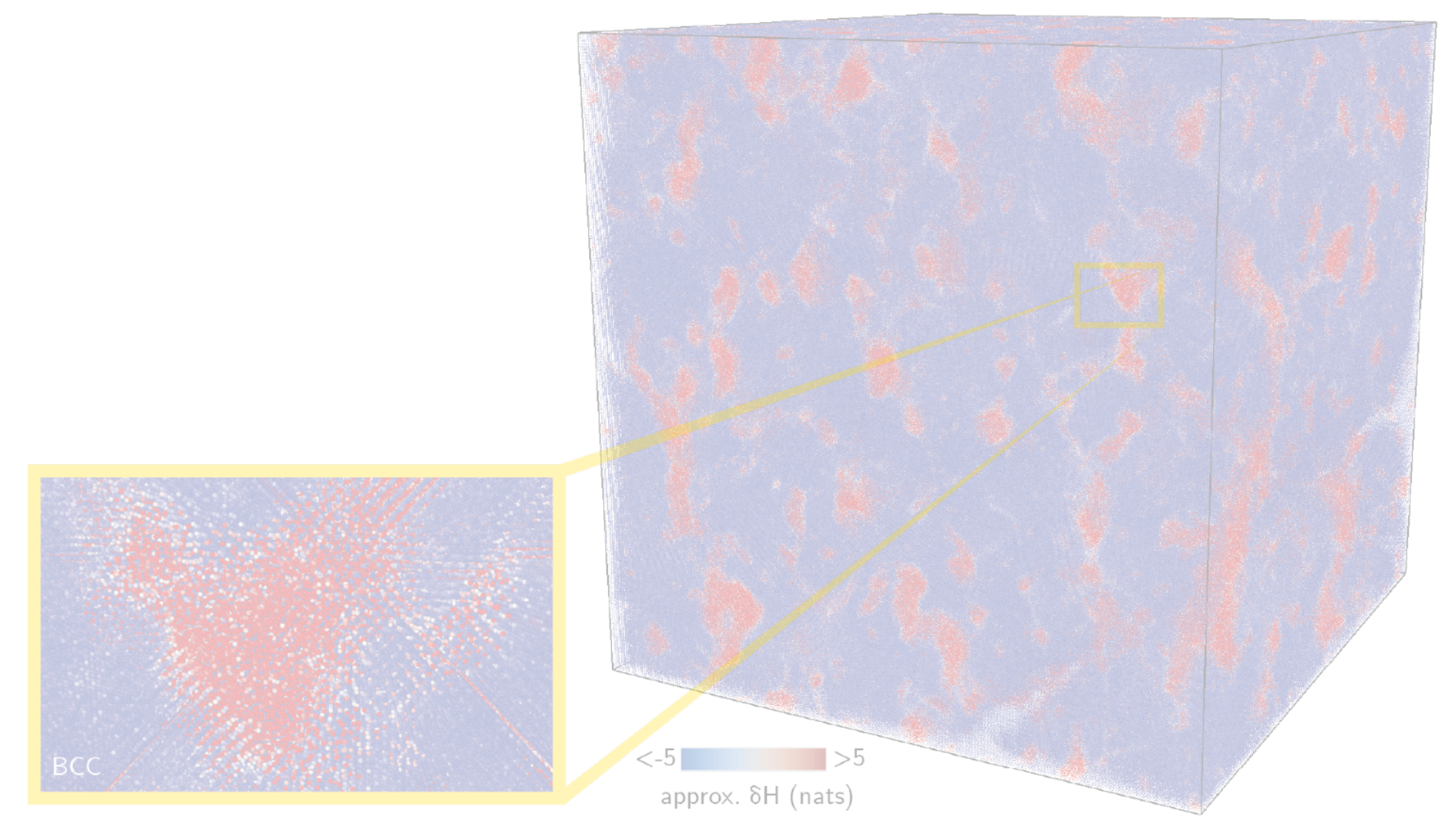
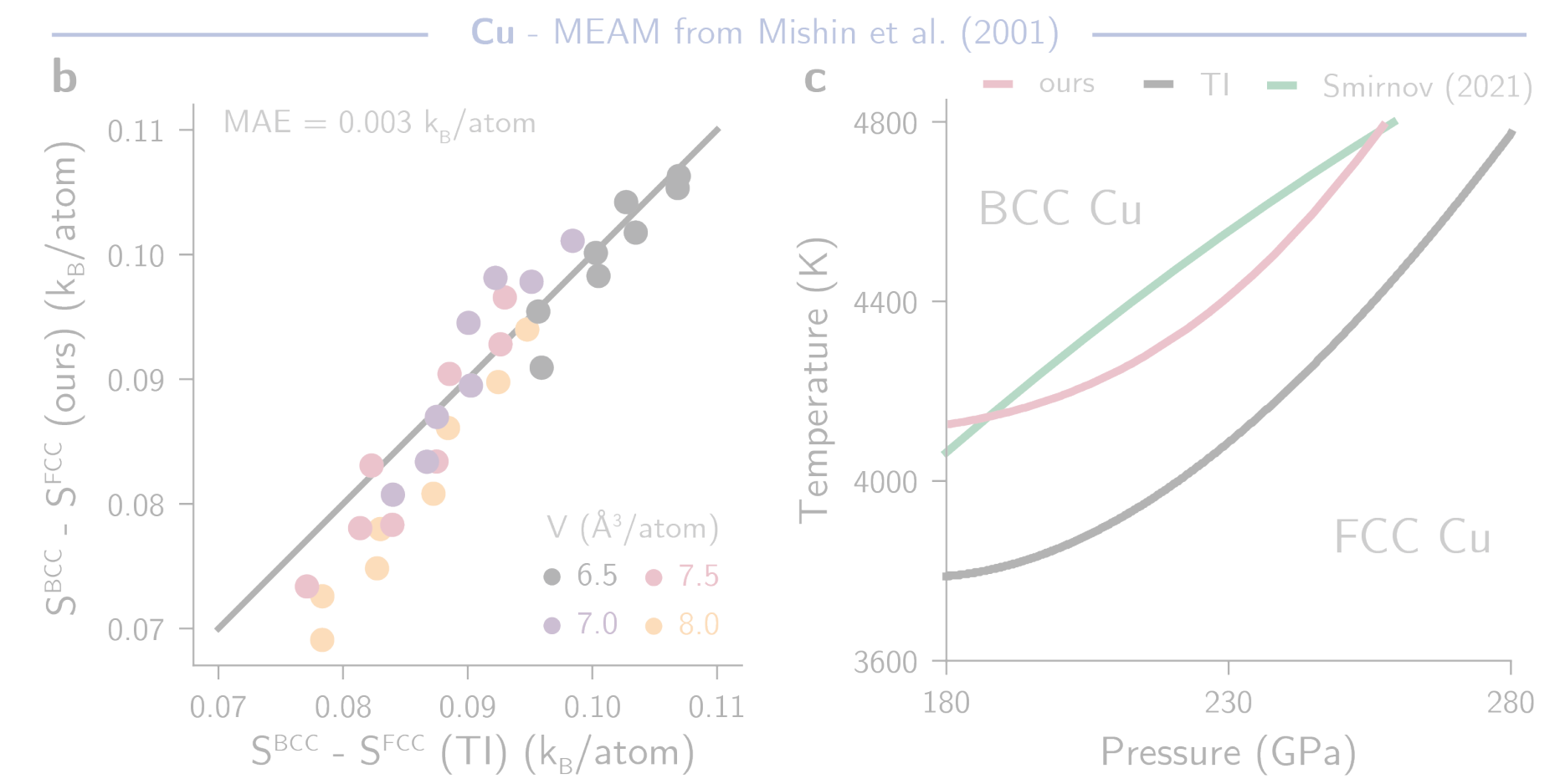
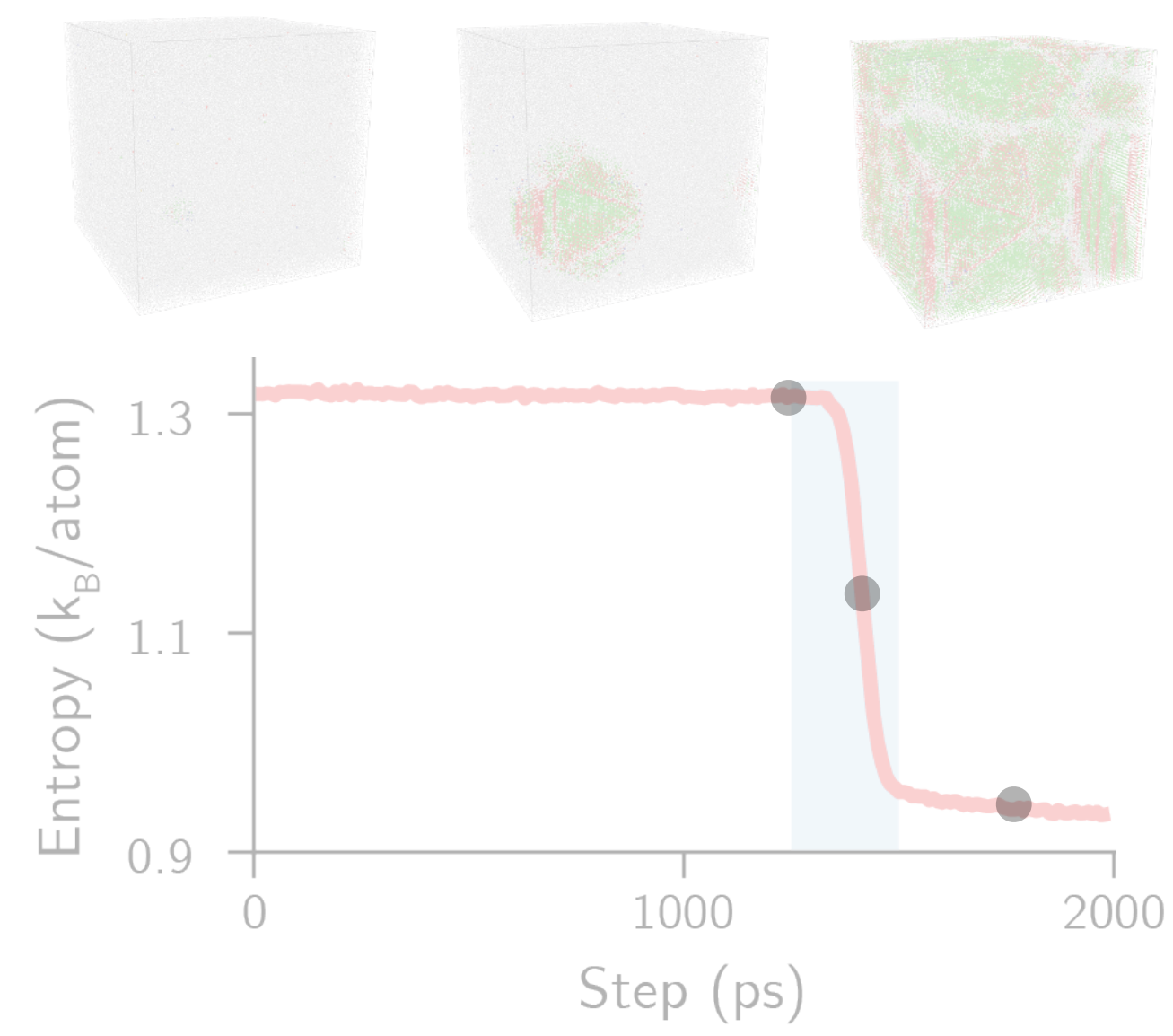
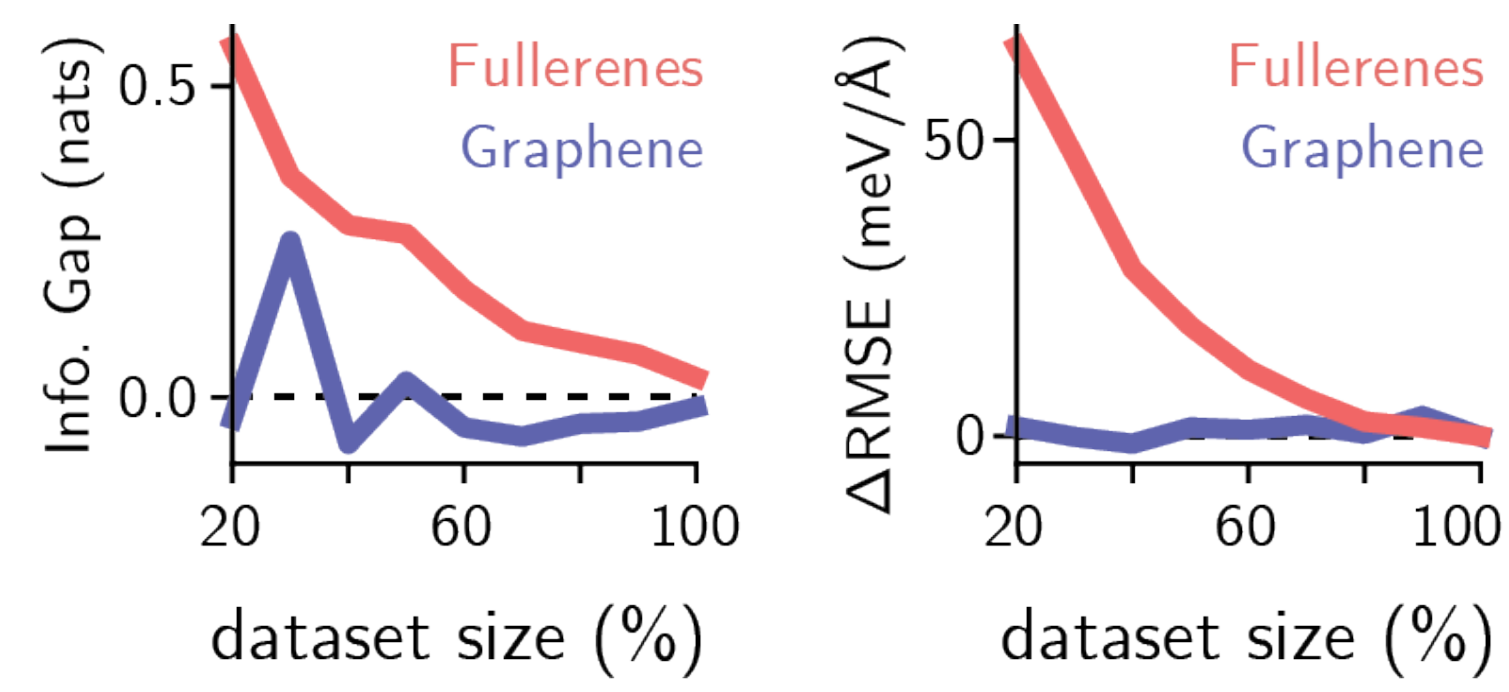
Detecting failed simulations

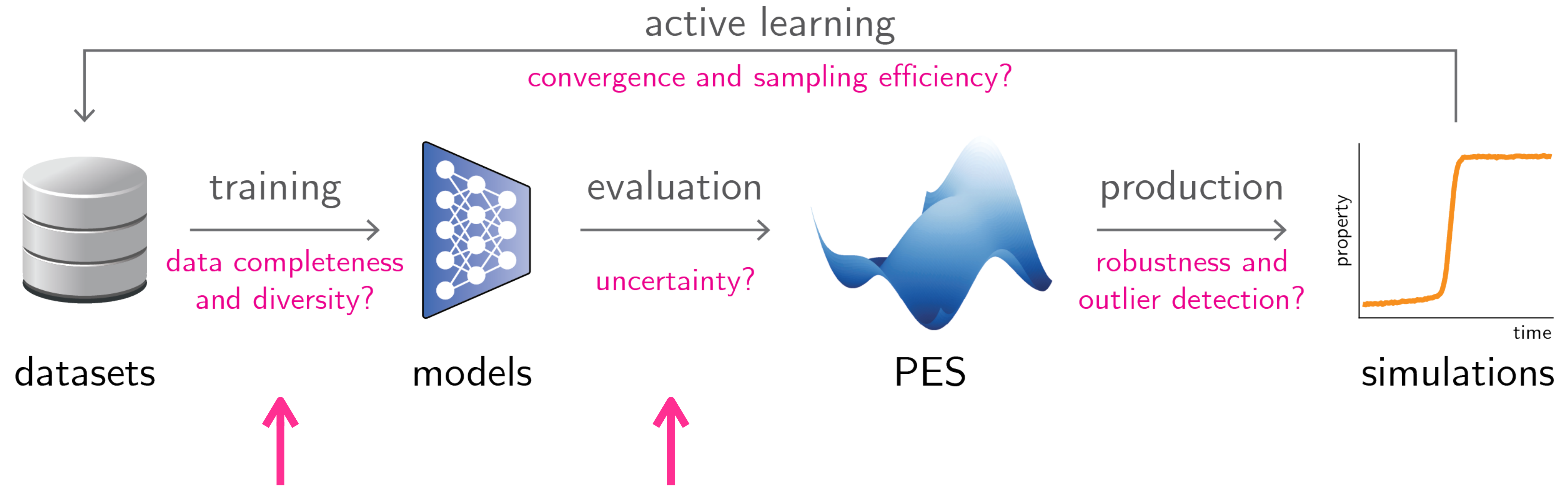


Information entropy can be used for numerous tasks in atomistic simulations



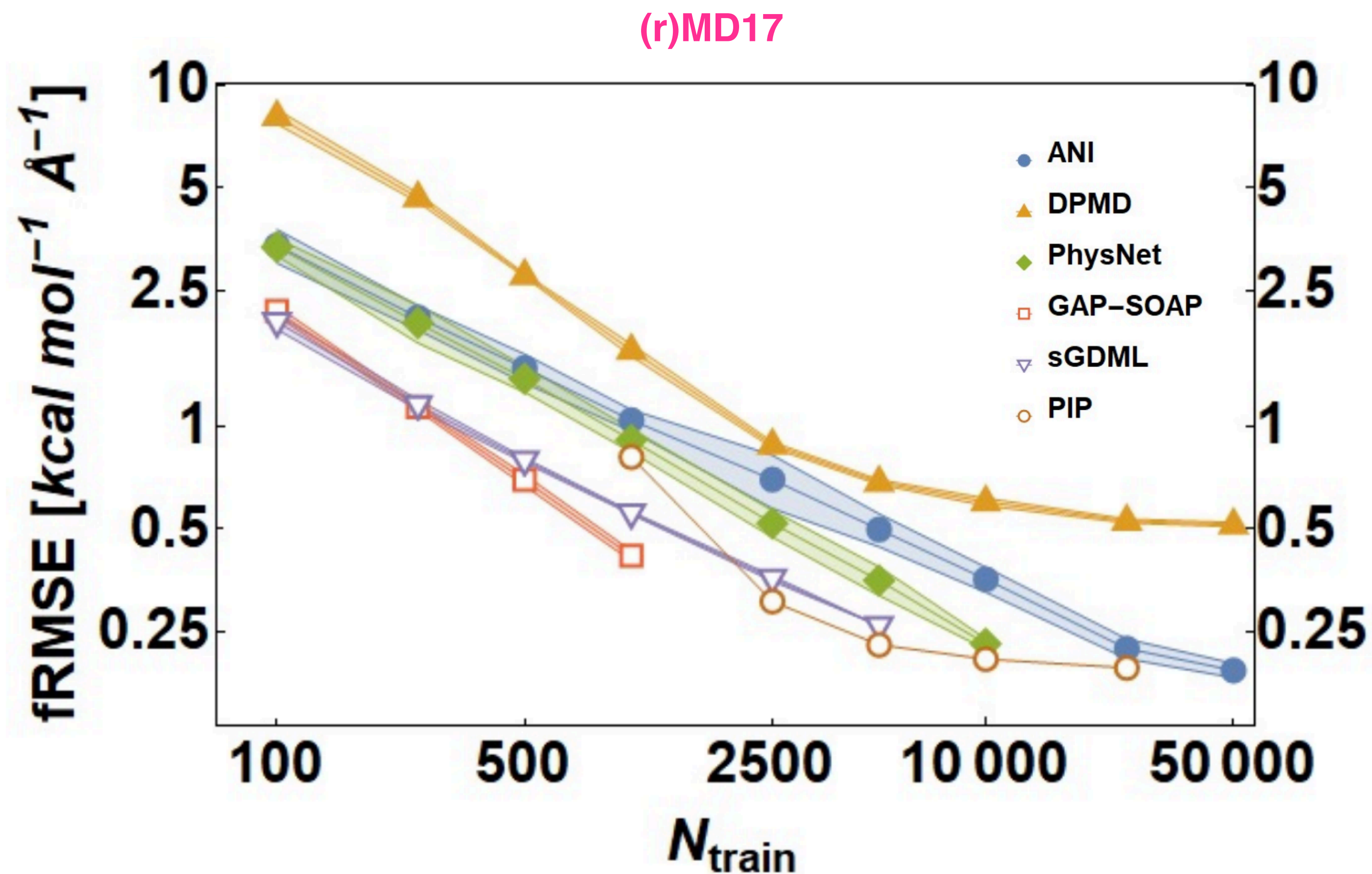
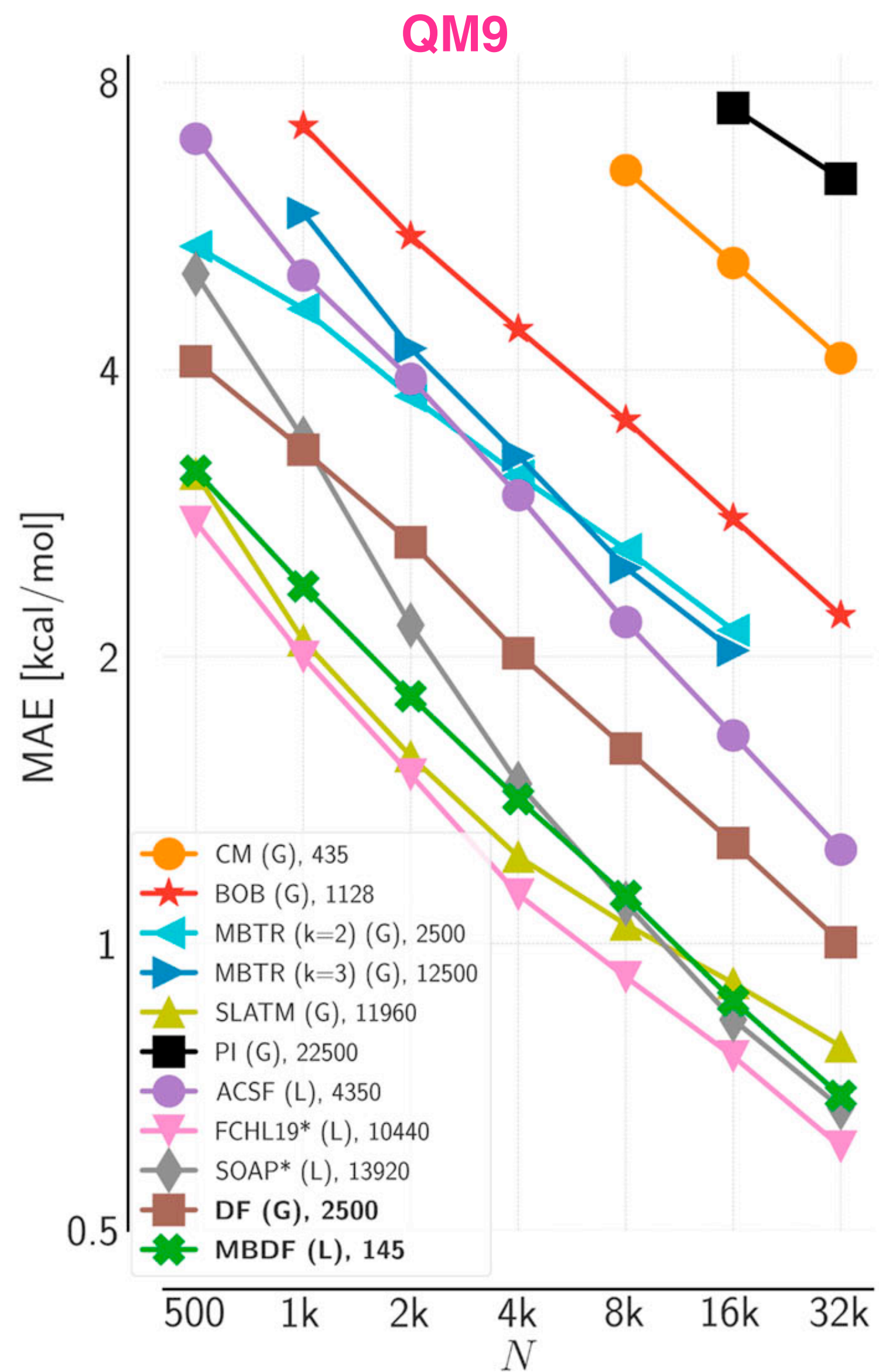
Dataset compression





How to quantify dataset completeness and model errors?

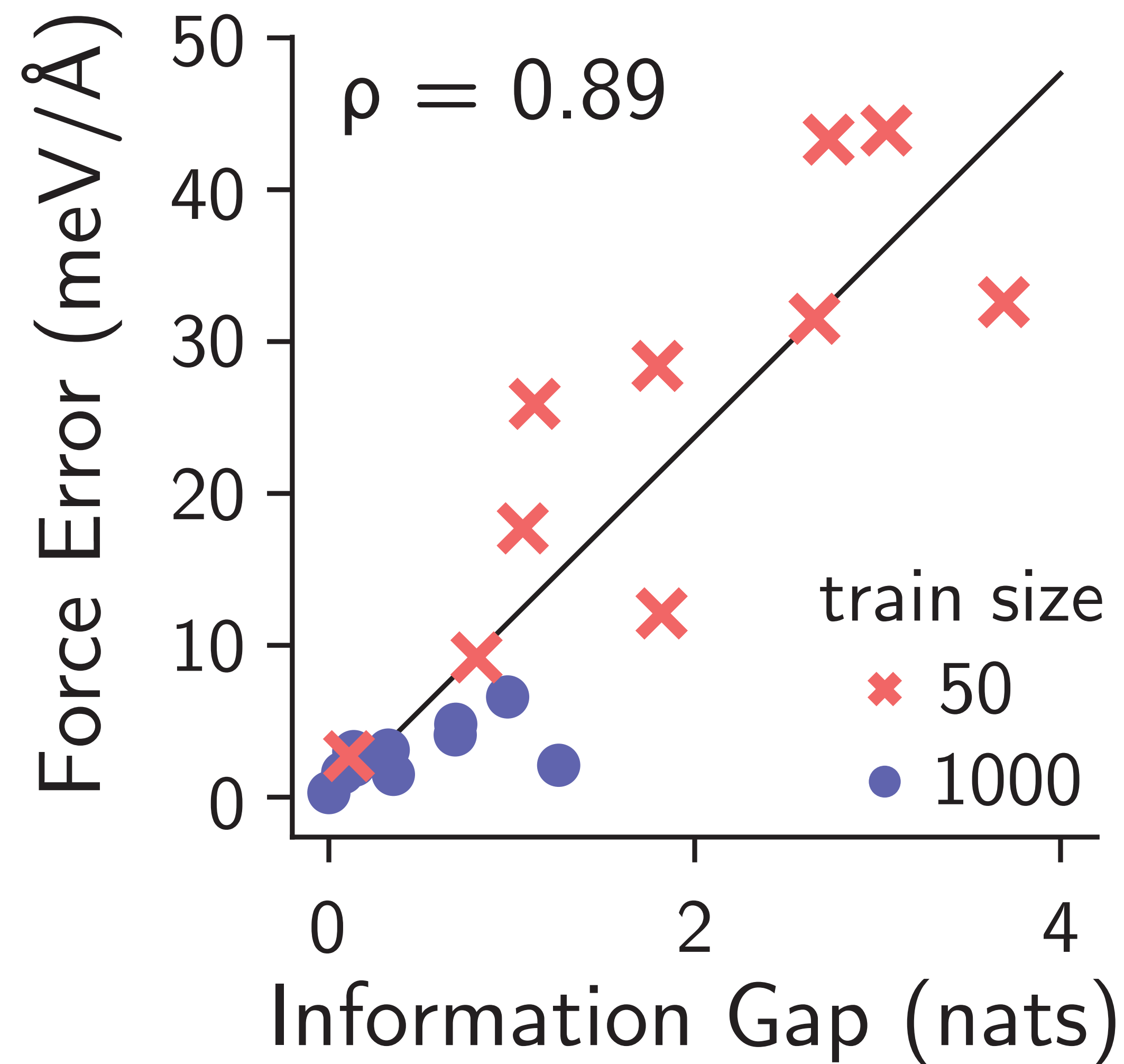
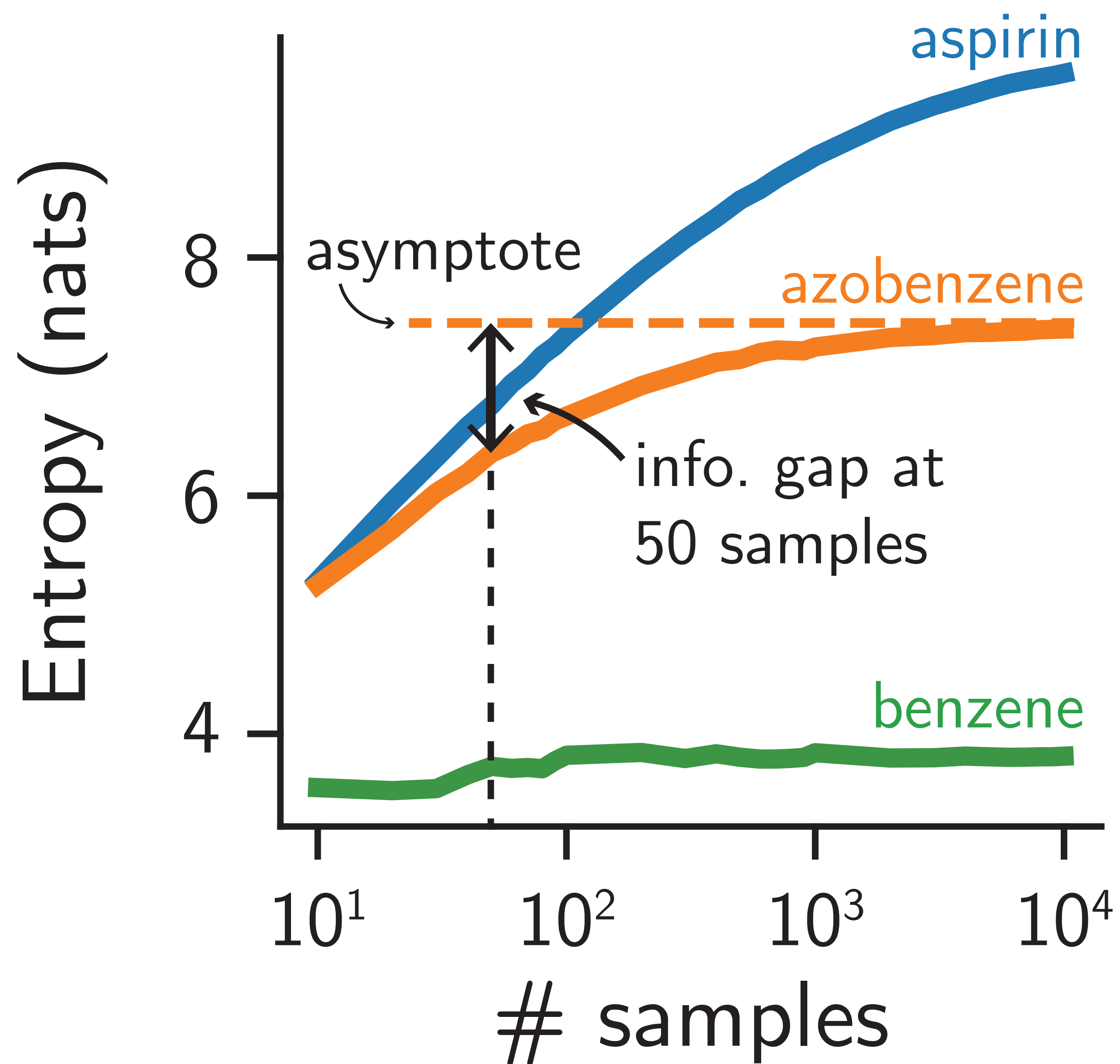
Explaining learning curves of datasets



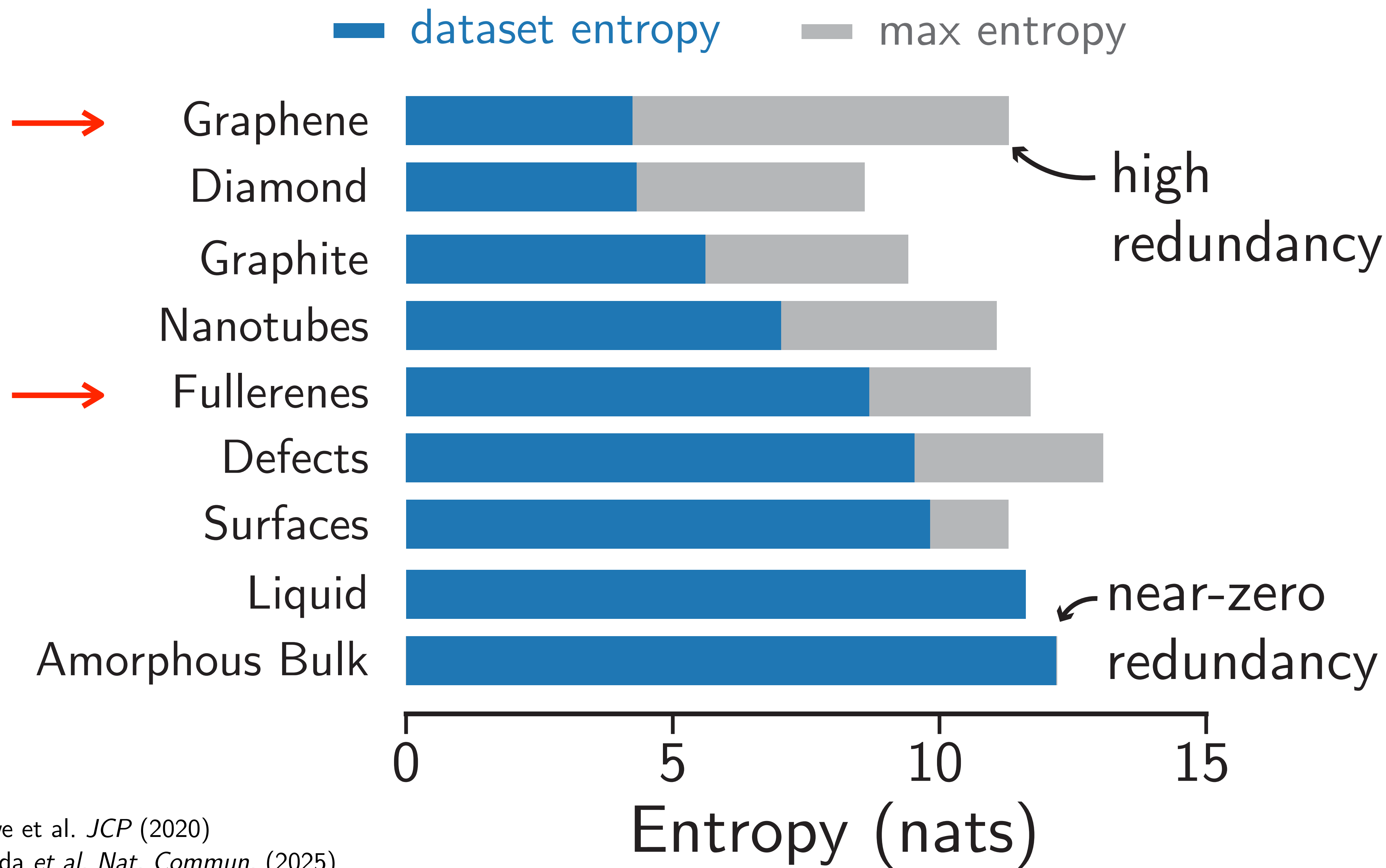
D. Khan et al. JCP (2023)

J. Bowman et al. JCP (2022)

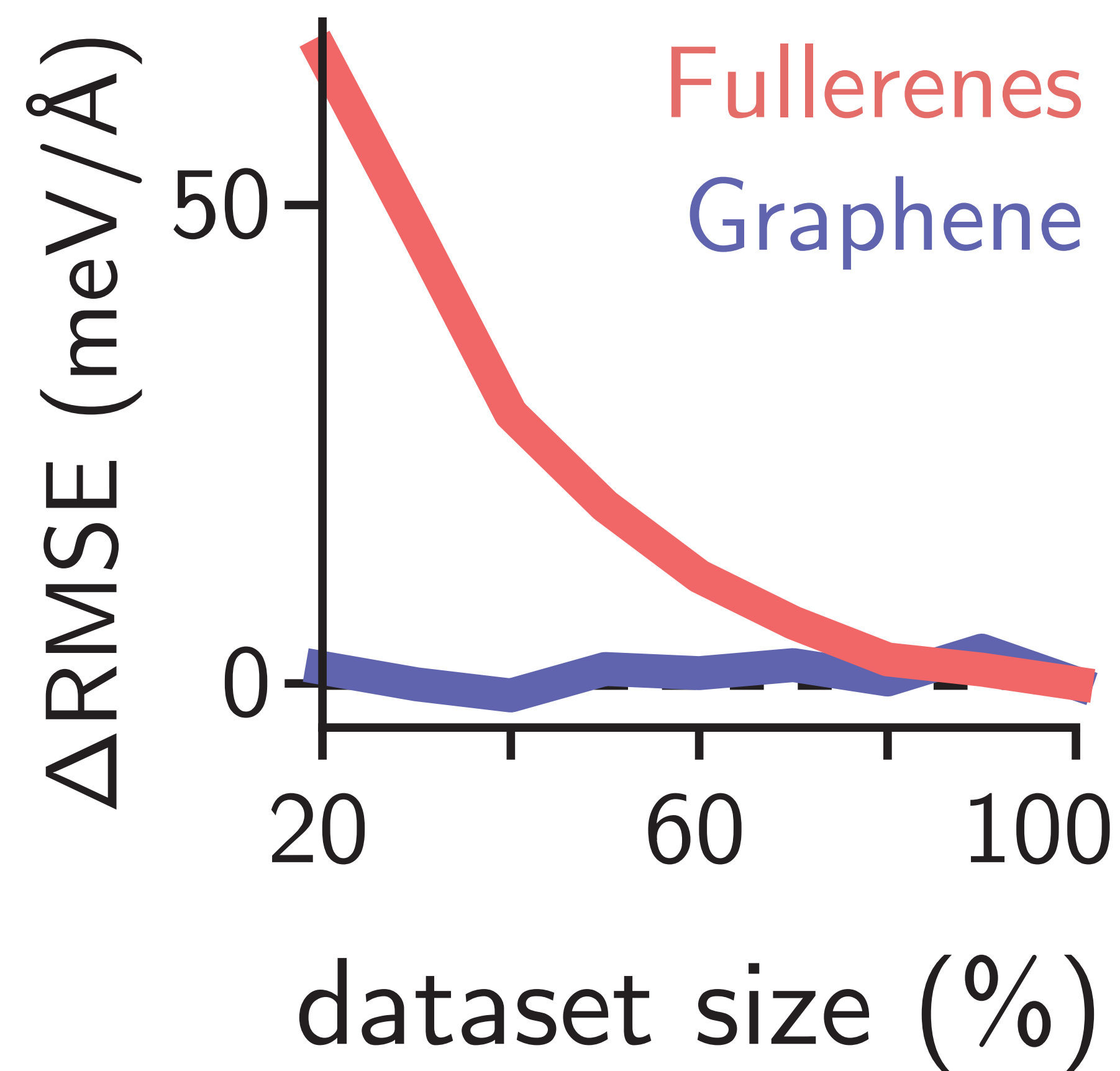
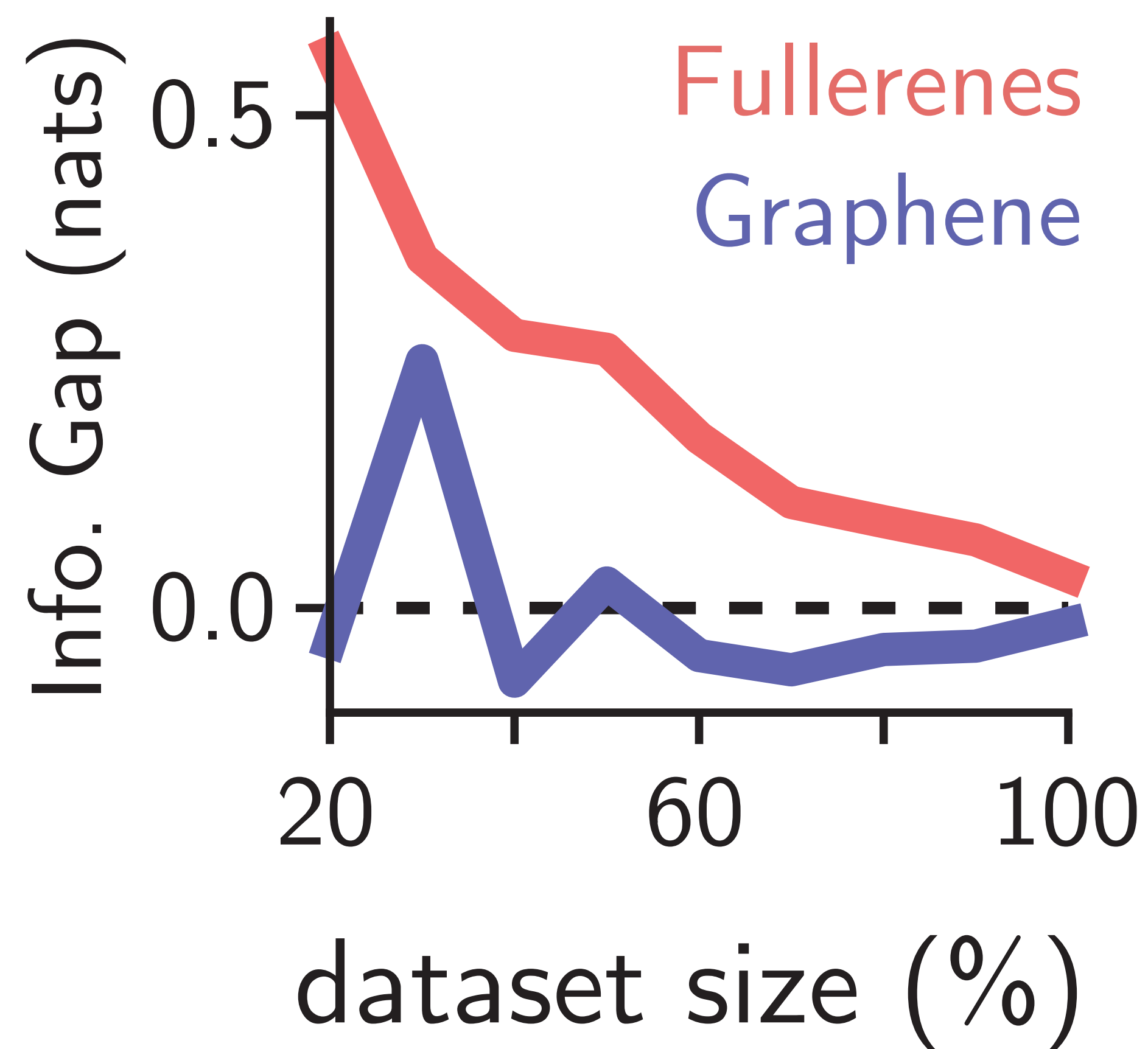
Information entropy explains trends in learning curves of datasets



Information entropy quantifies how diverse is an atomistic dataset

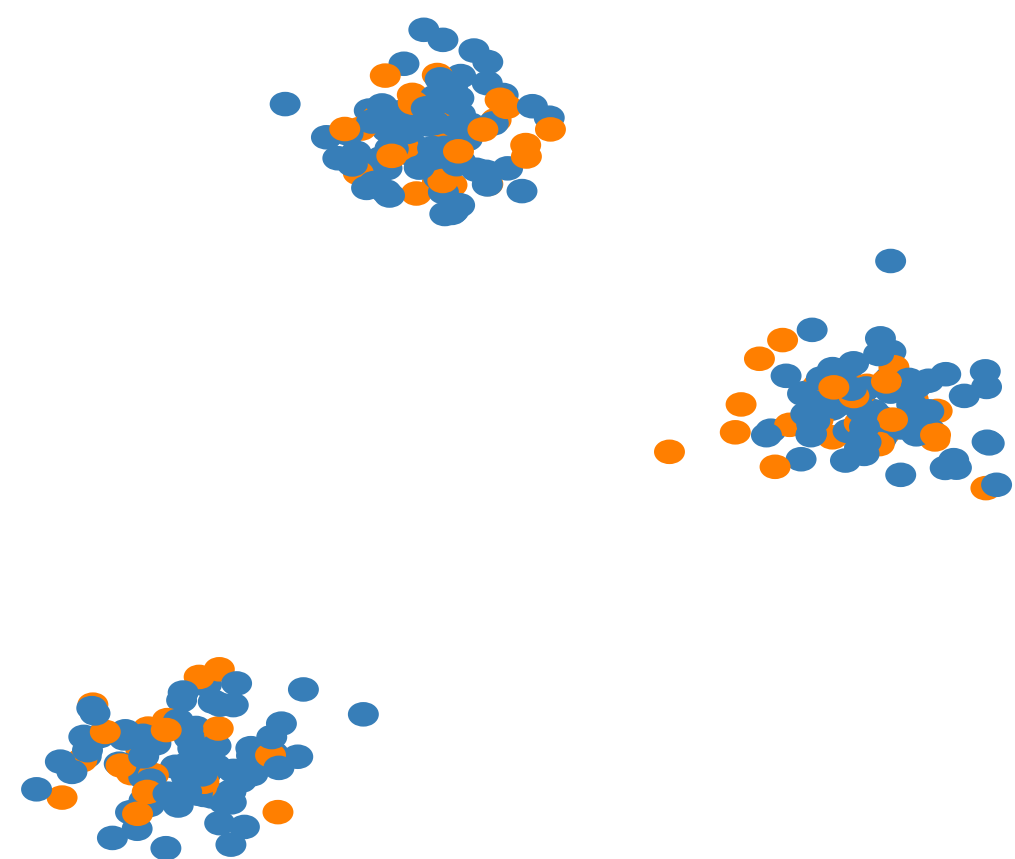


Randomly sampling datasets with different redundancy values



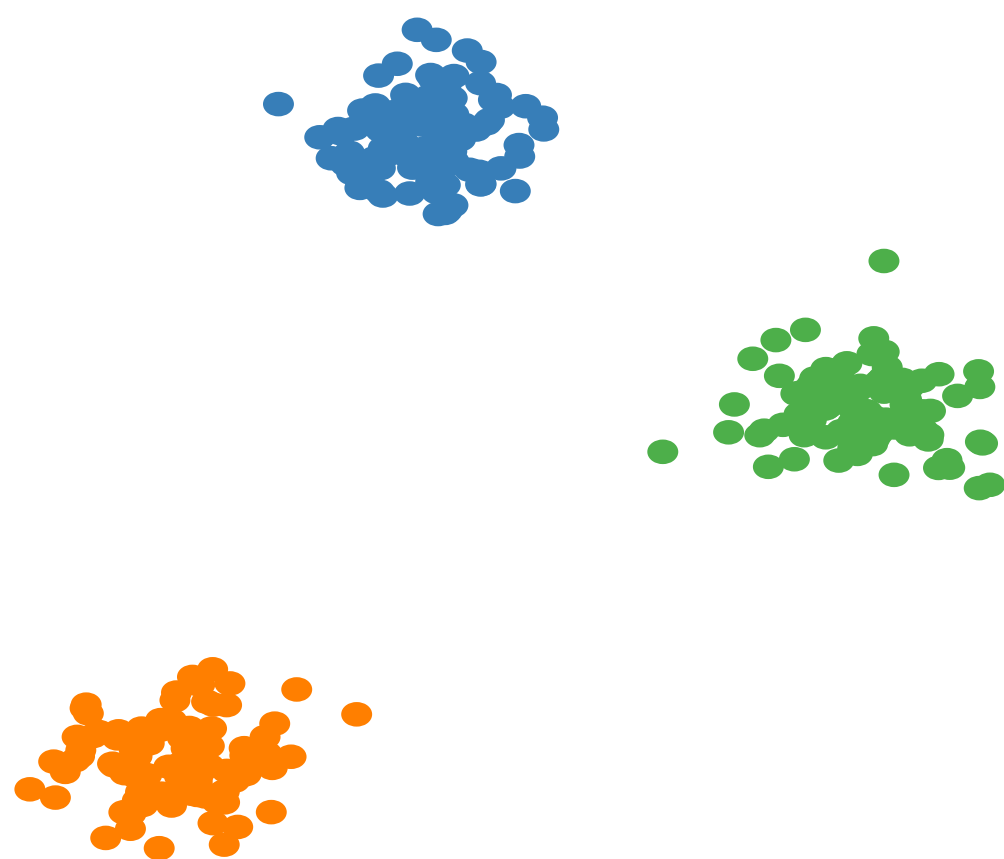
Information entropy is the metric that indicates data compression within information theory

a



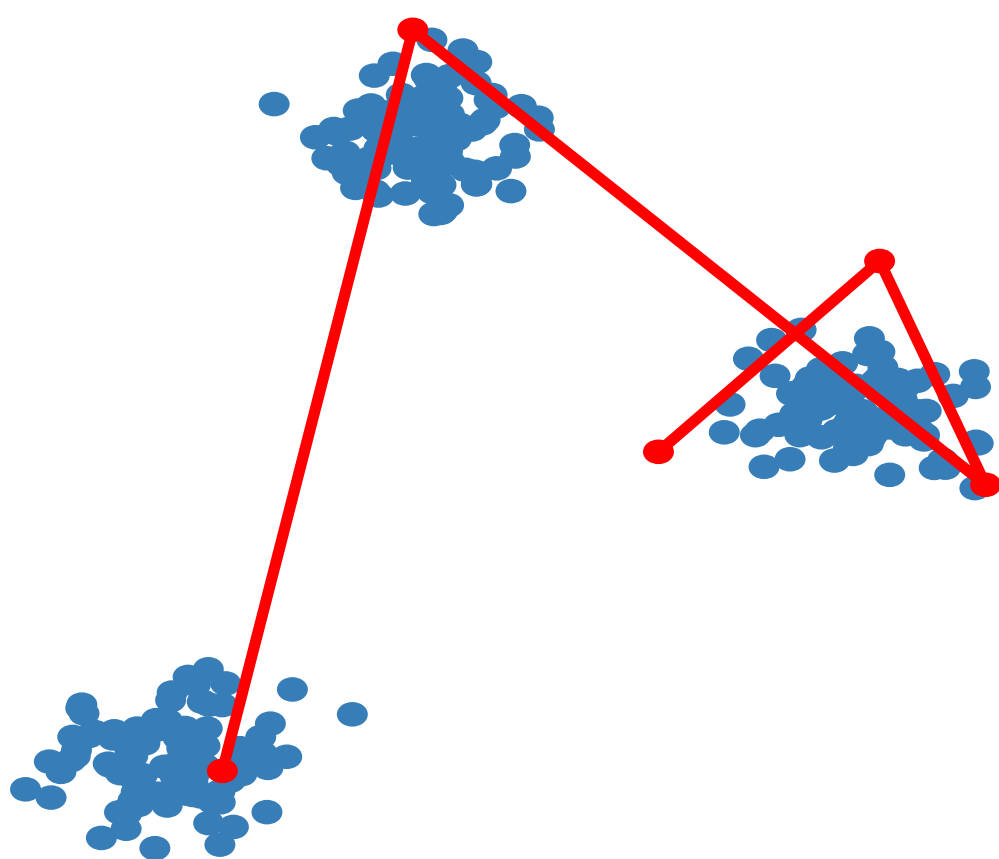
Random Sampling

b



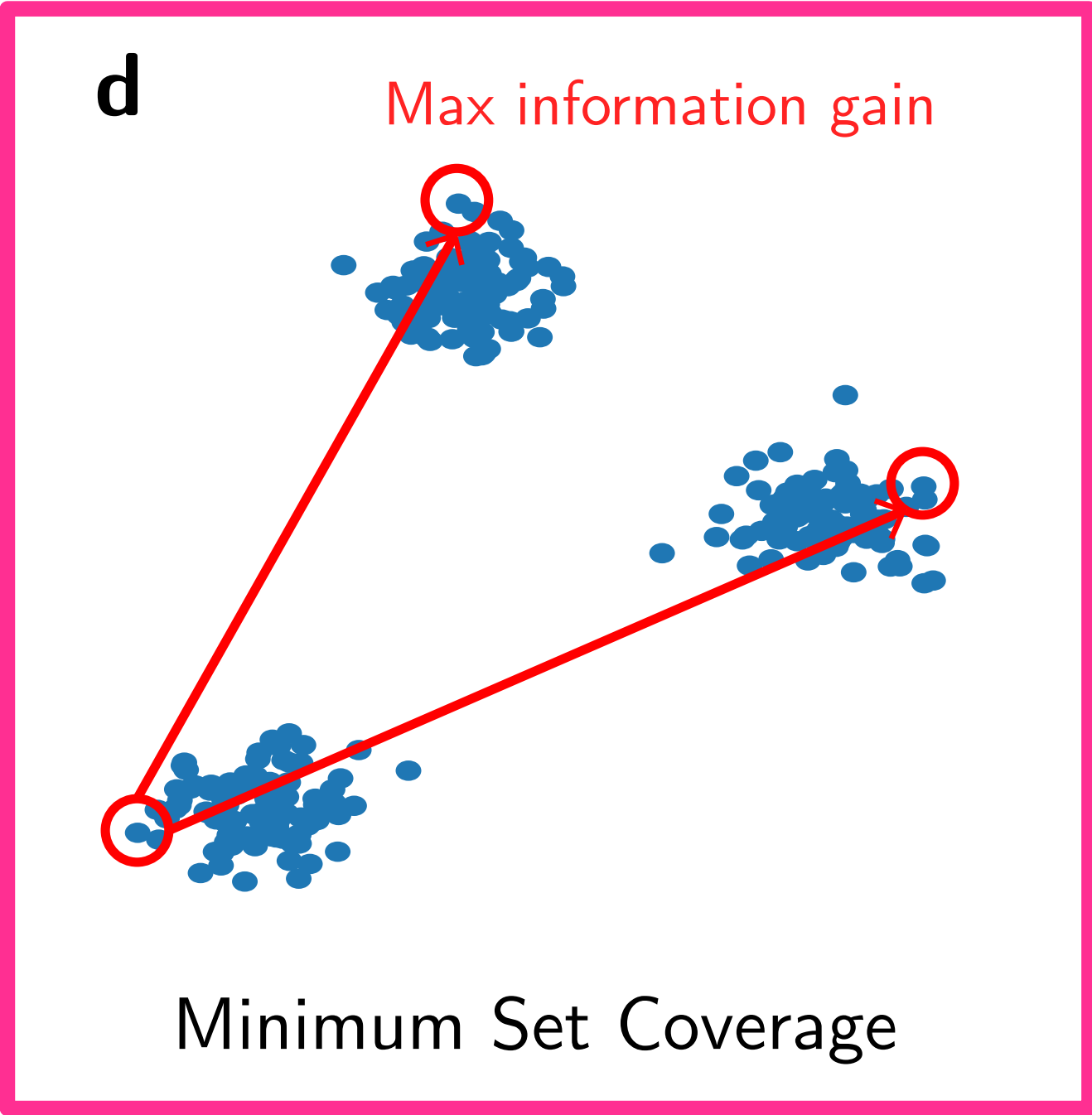
Mean K-means Sampling

c



Mean Farthest Point Sampling

d



Minimum Set Coverage



Ben Yu

Information theory is the field that studies data compression! Quantifying information gain in atomistic systems allows us to subsample datasets

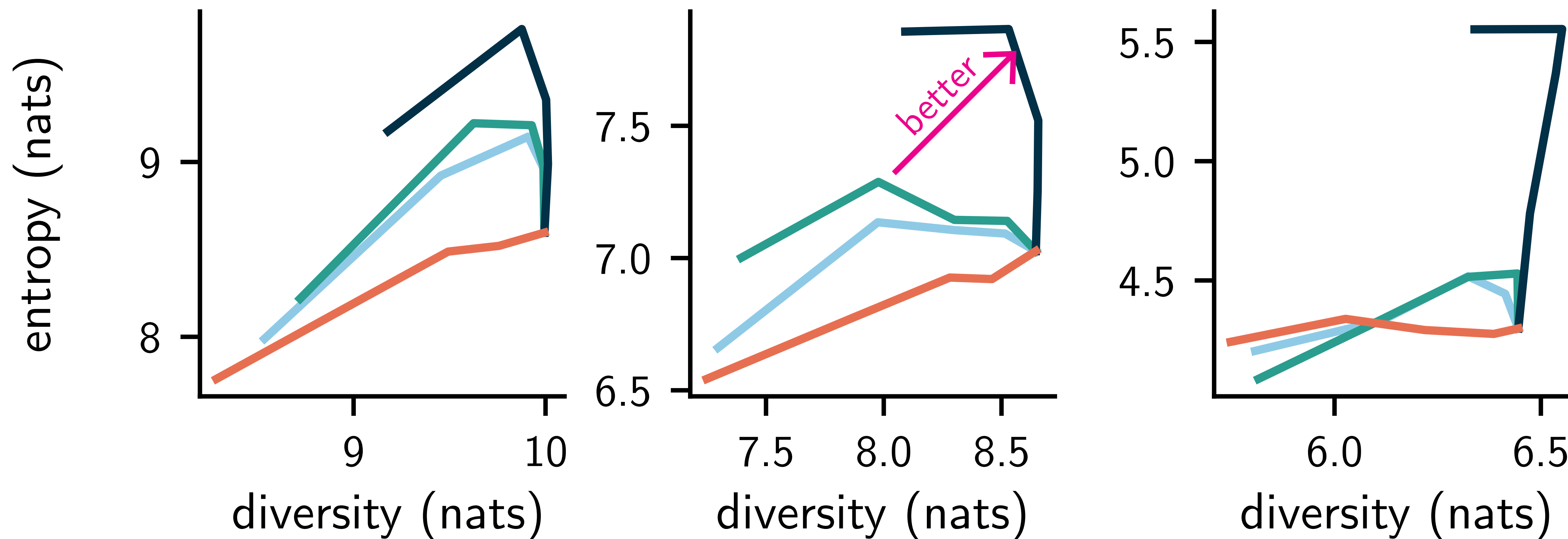
Our algorithm exhibits the best performance in dataset entropy and diversity

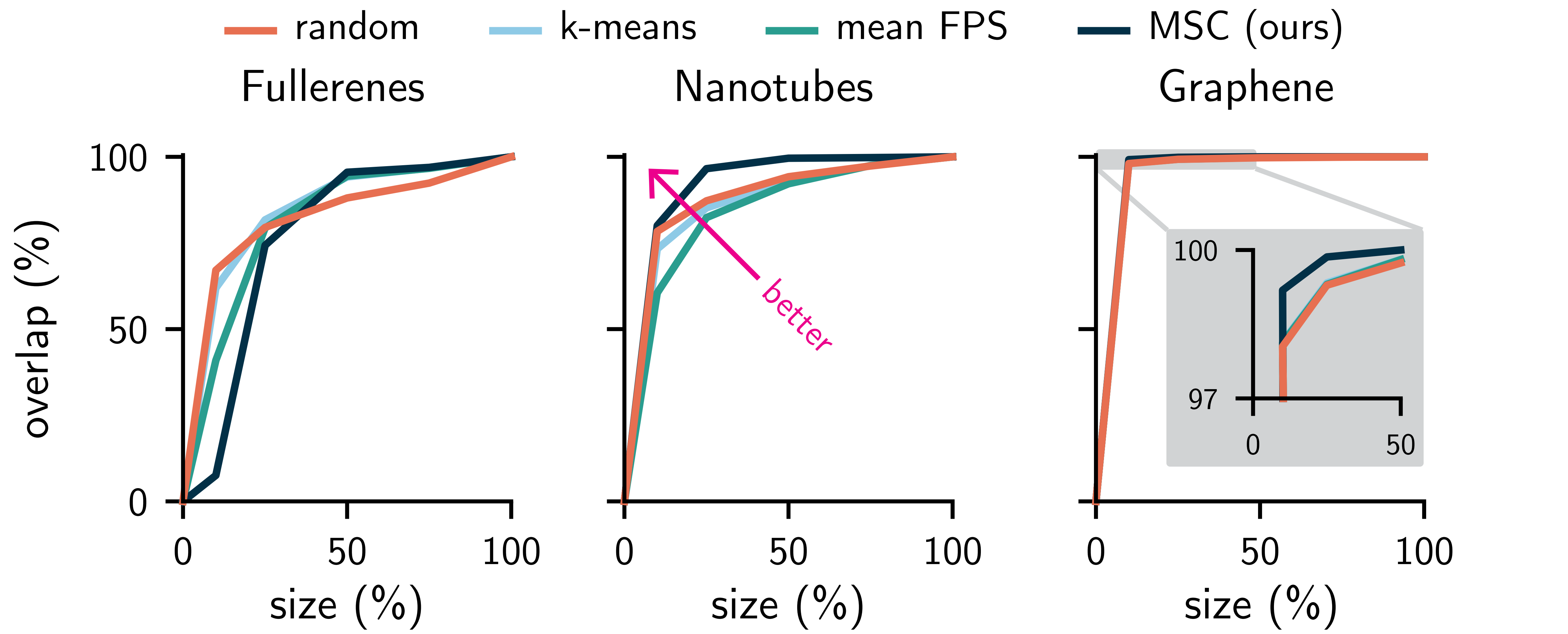
— random — k-means — mean FPS — MSC (ours)

Fullerenes

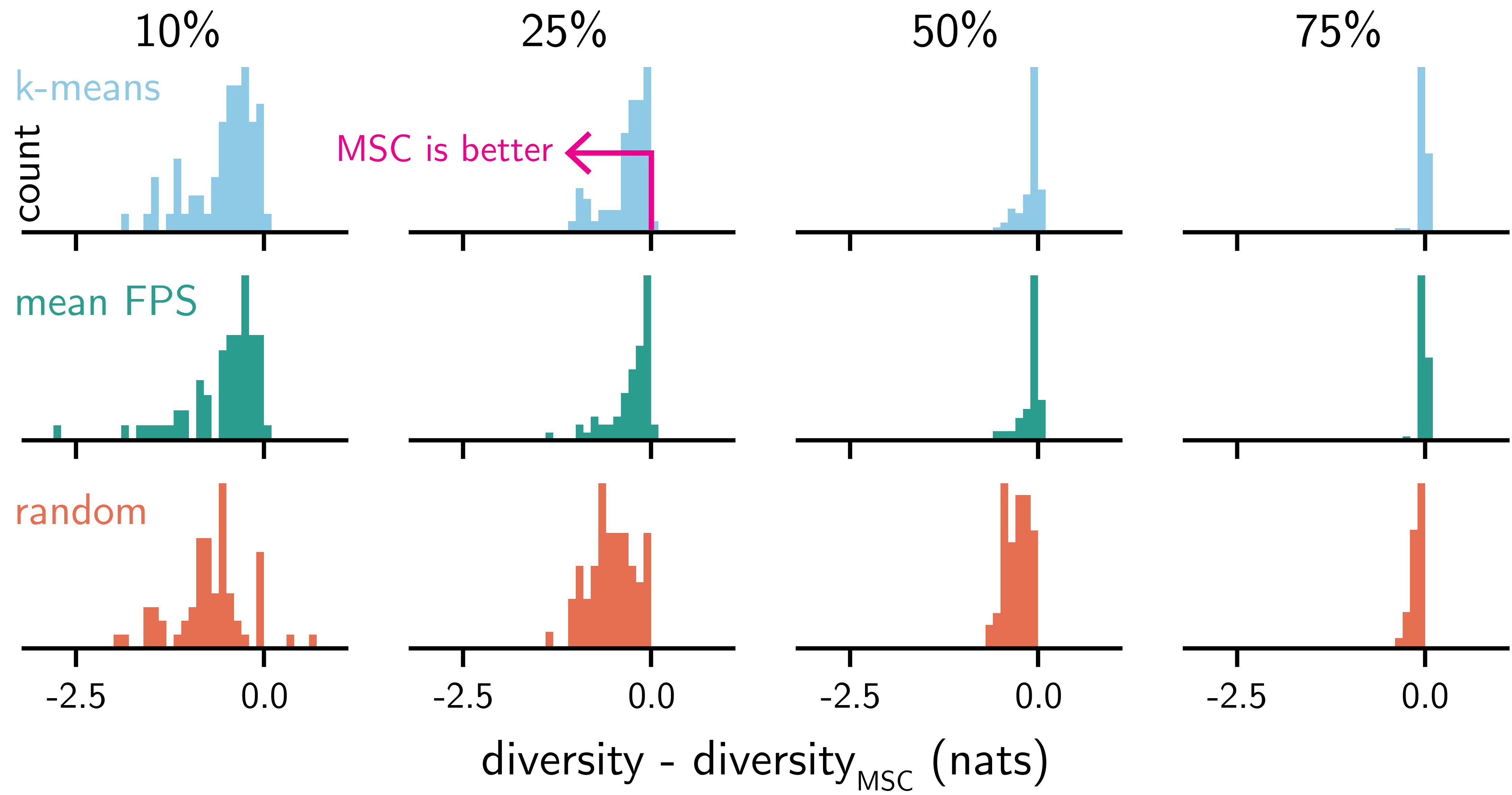
Nanotubes

Graphene

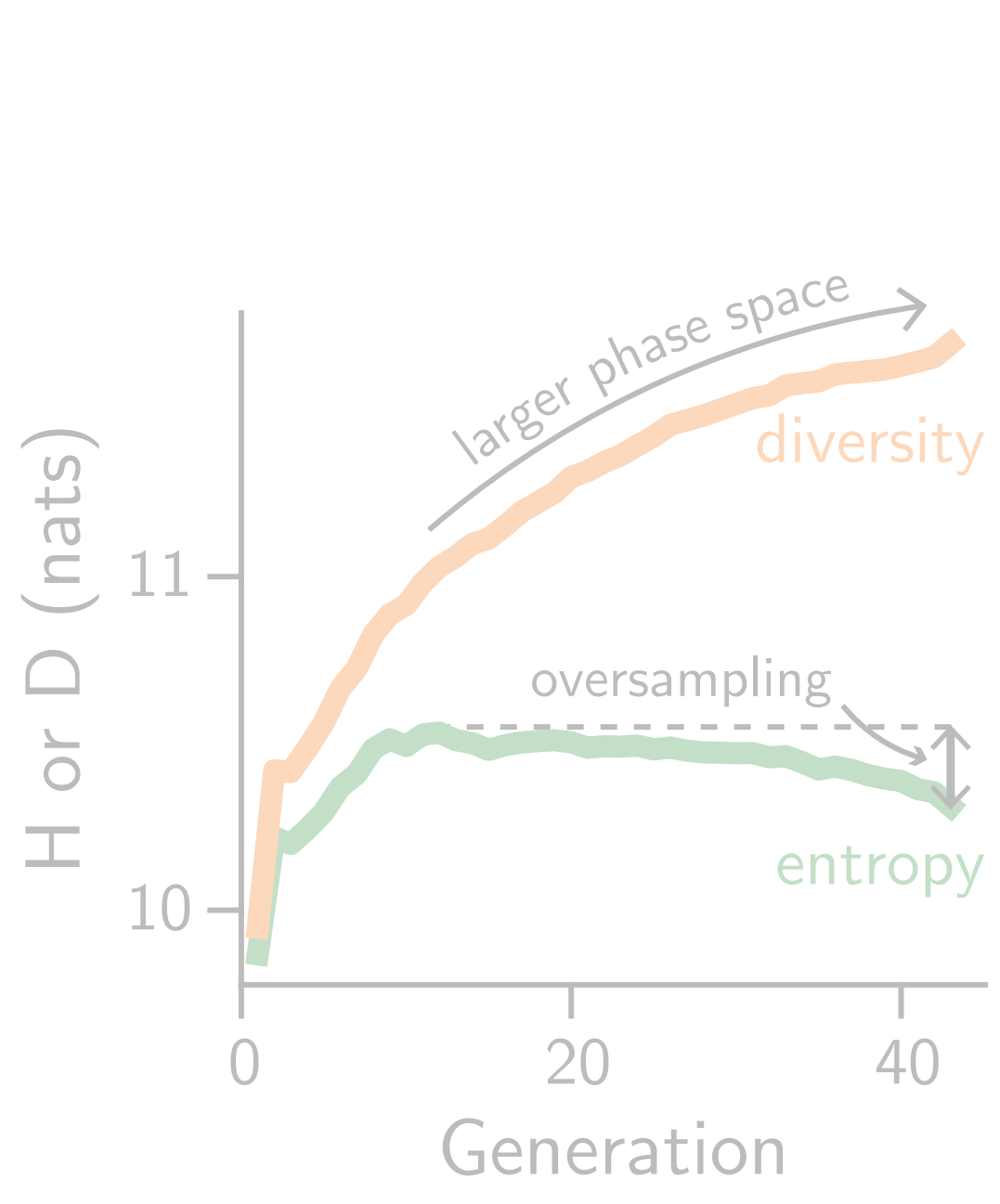
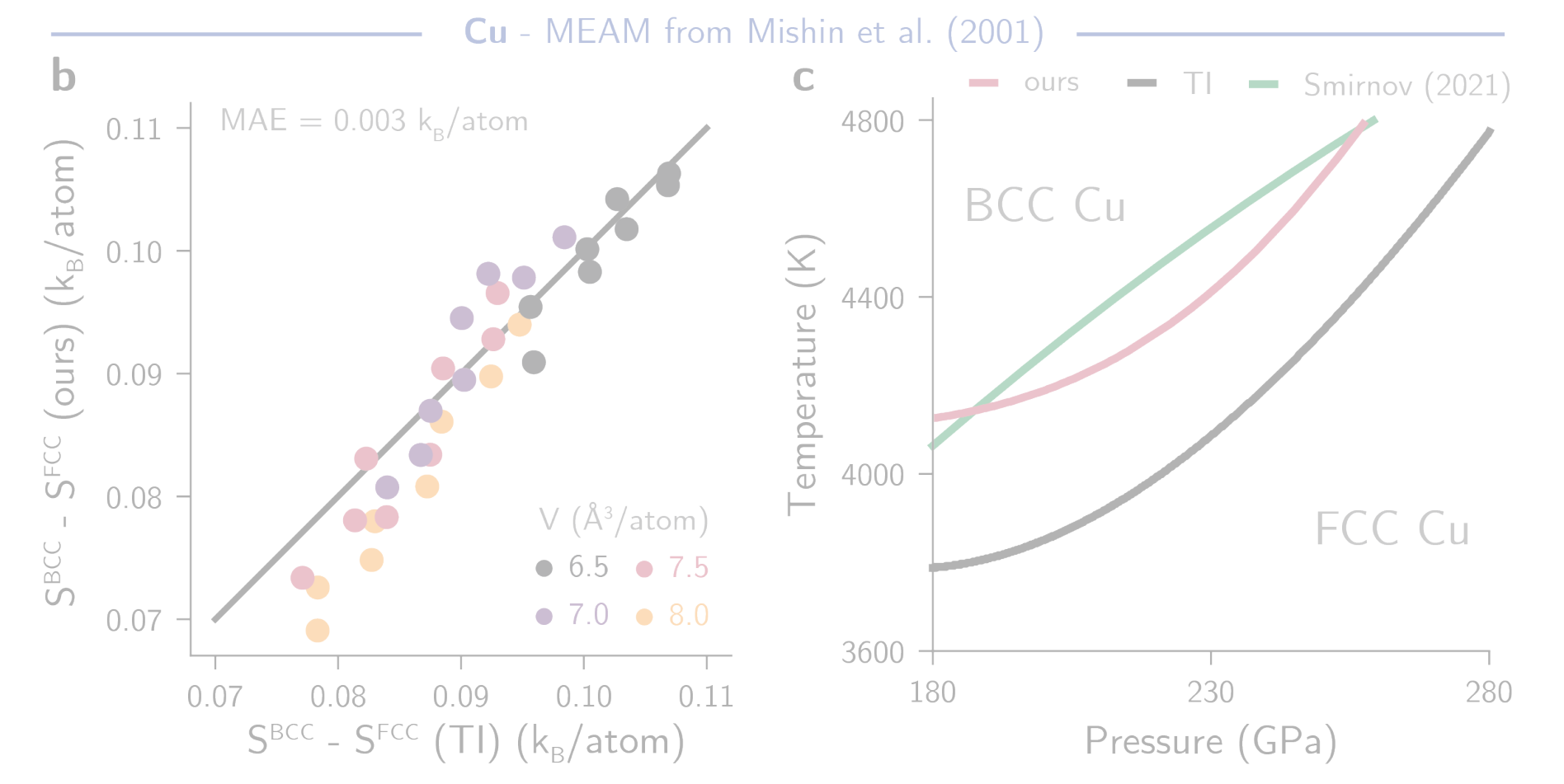
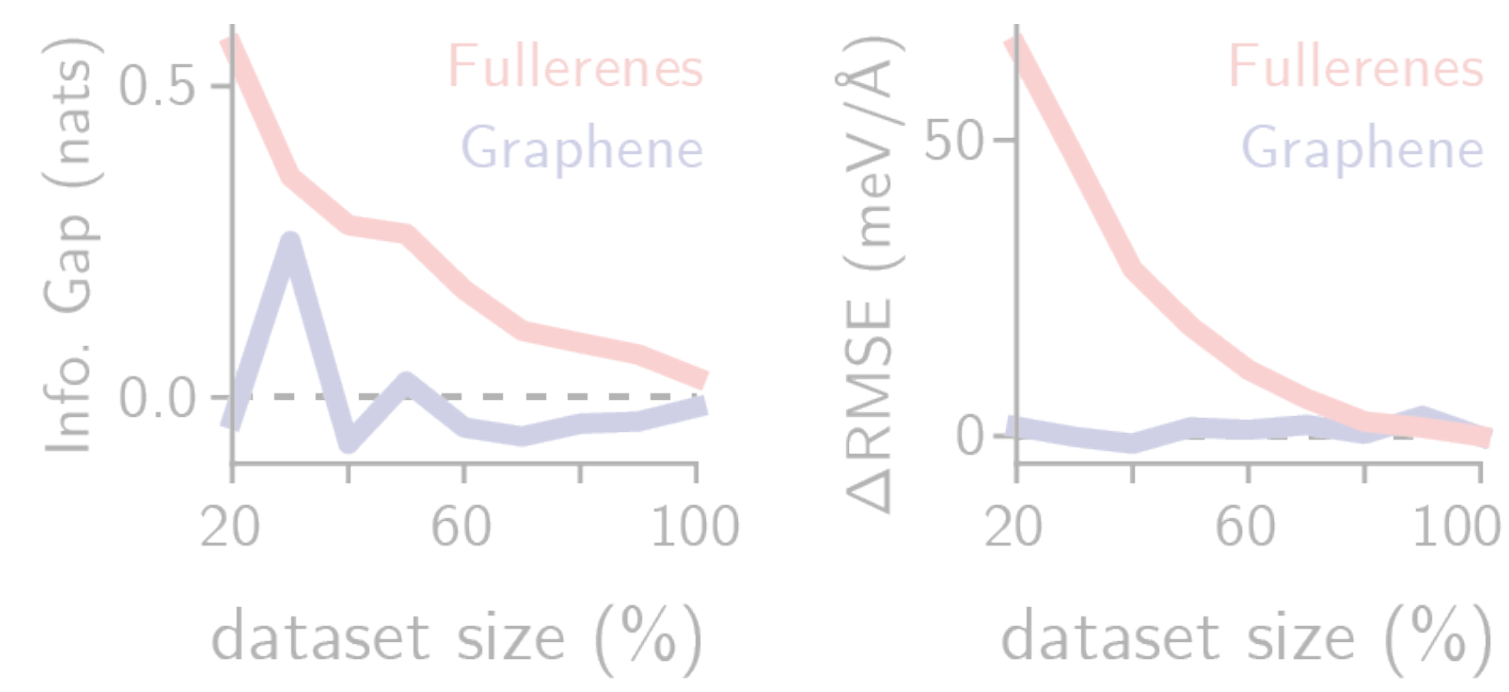
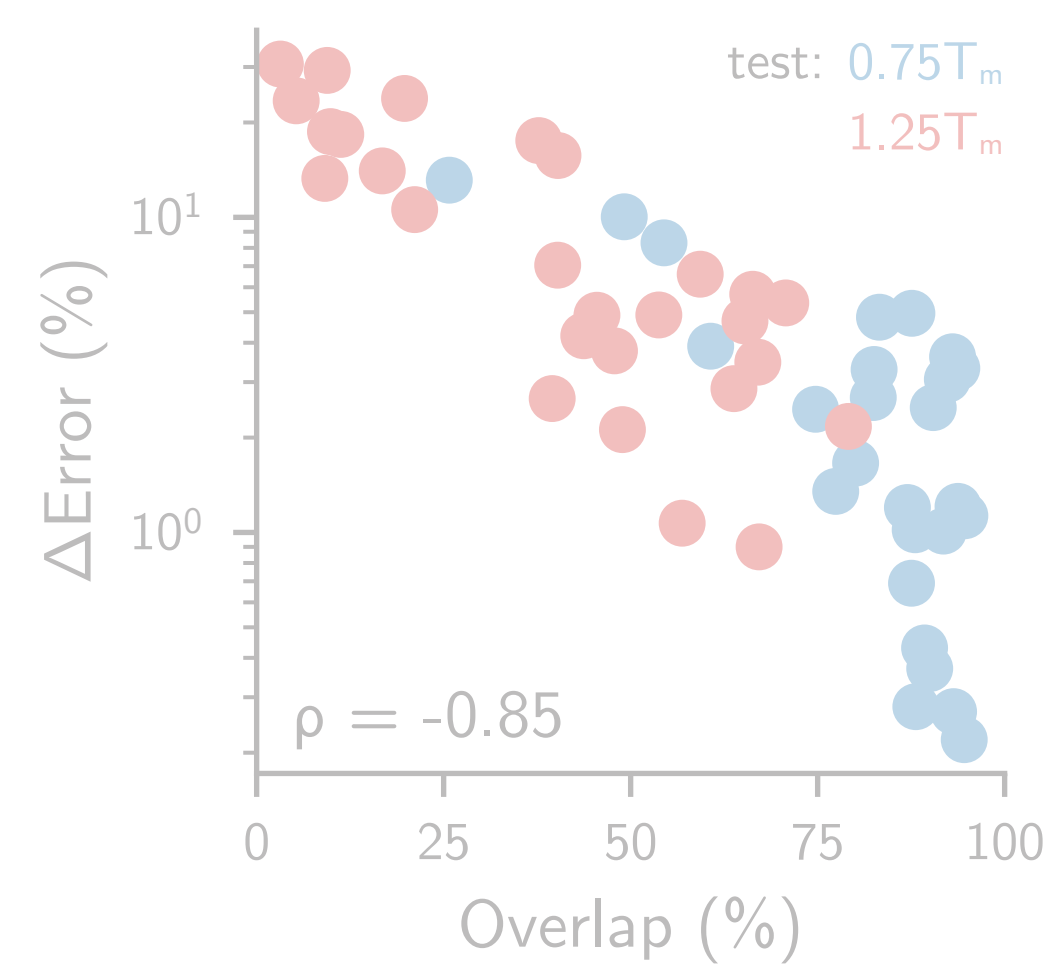




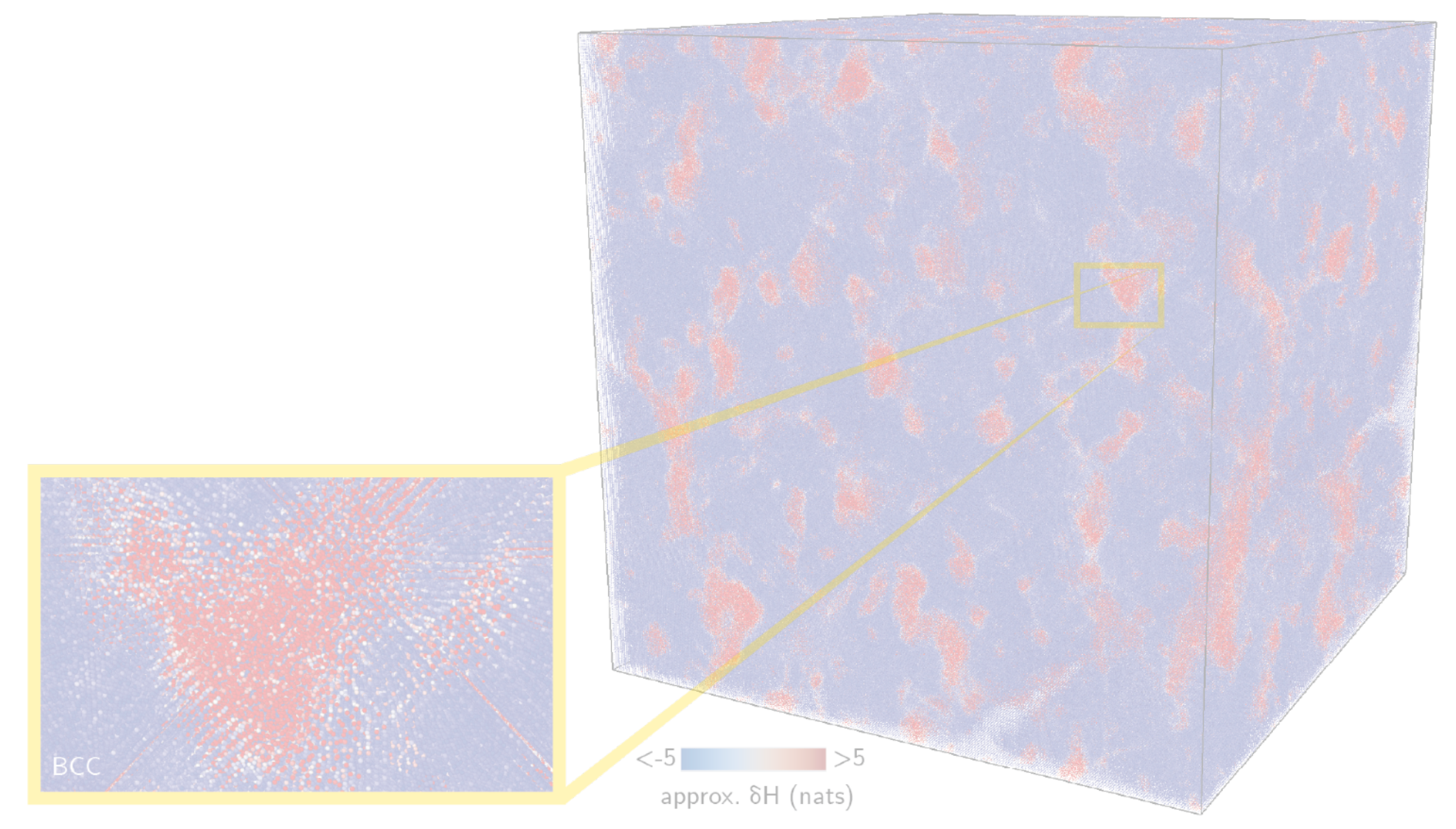
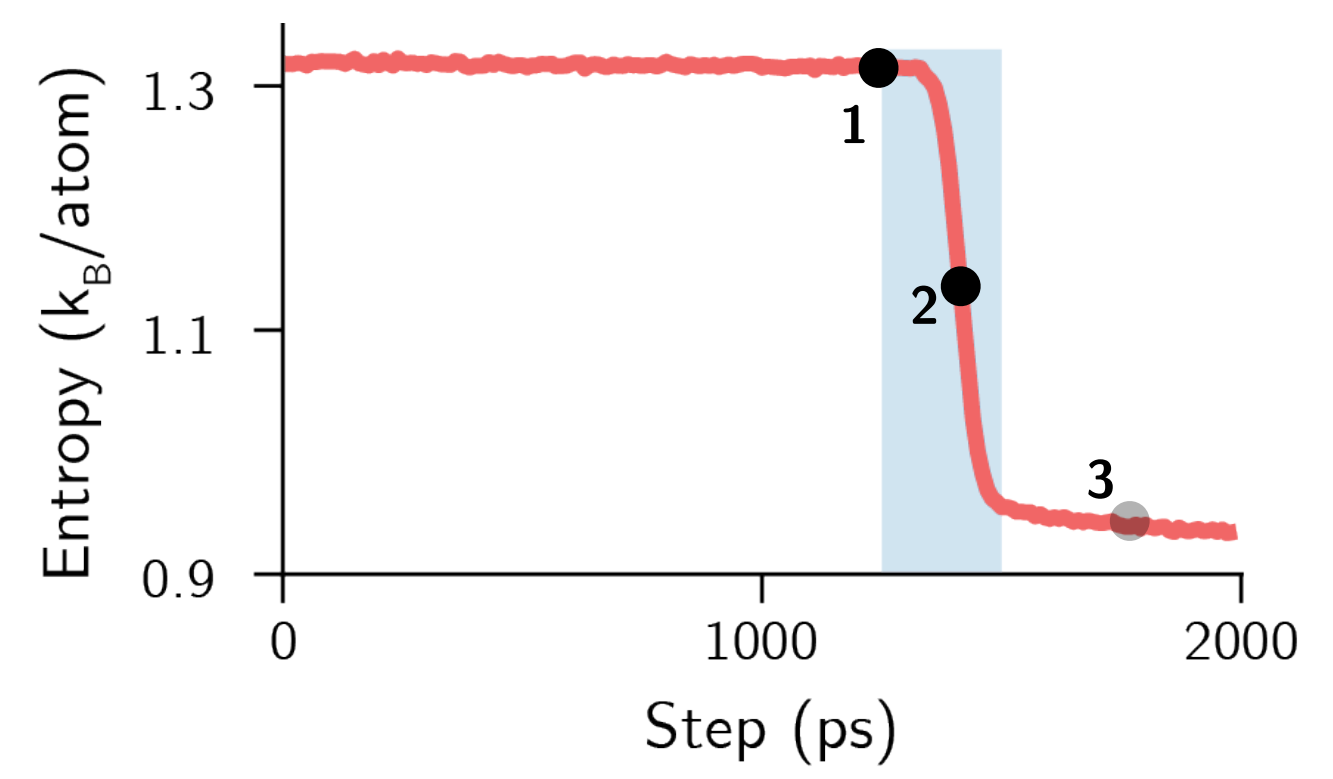
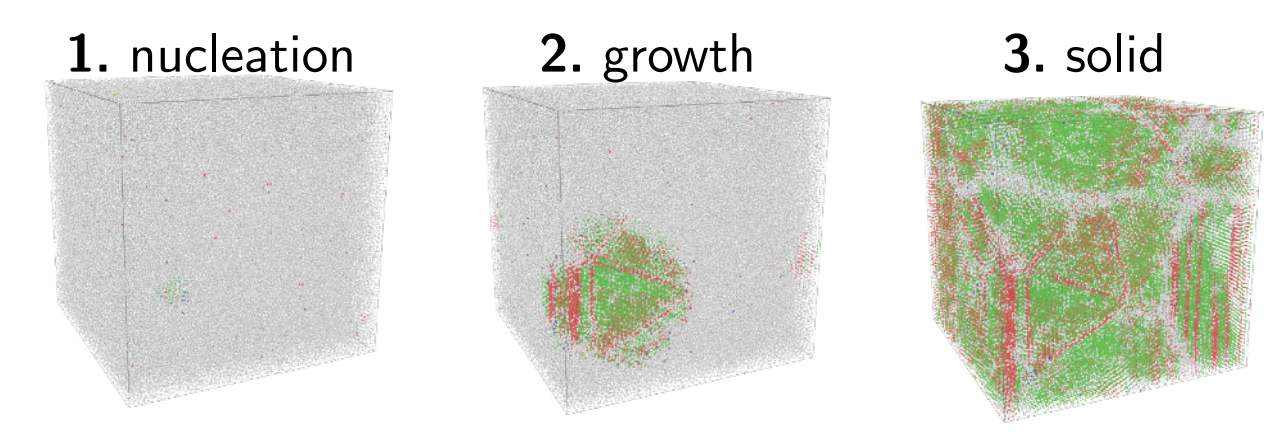
Compressing 64 datasets shows that MSC almost always leads to the highest diversity



Information entropy can be used for much more in atomistic simulations

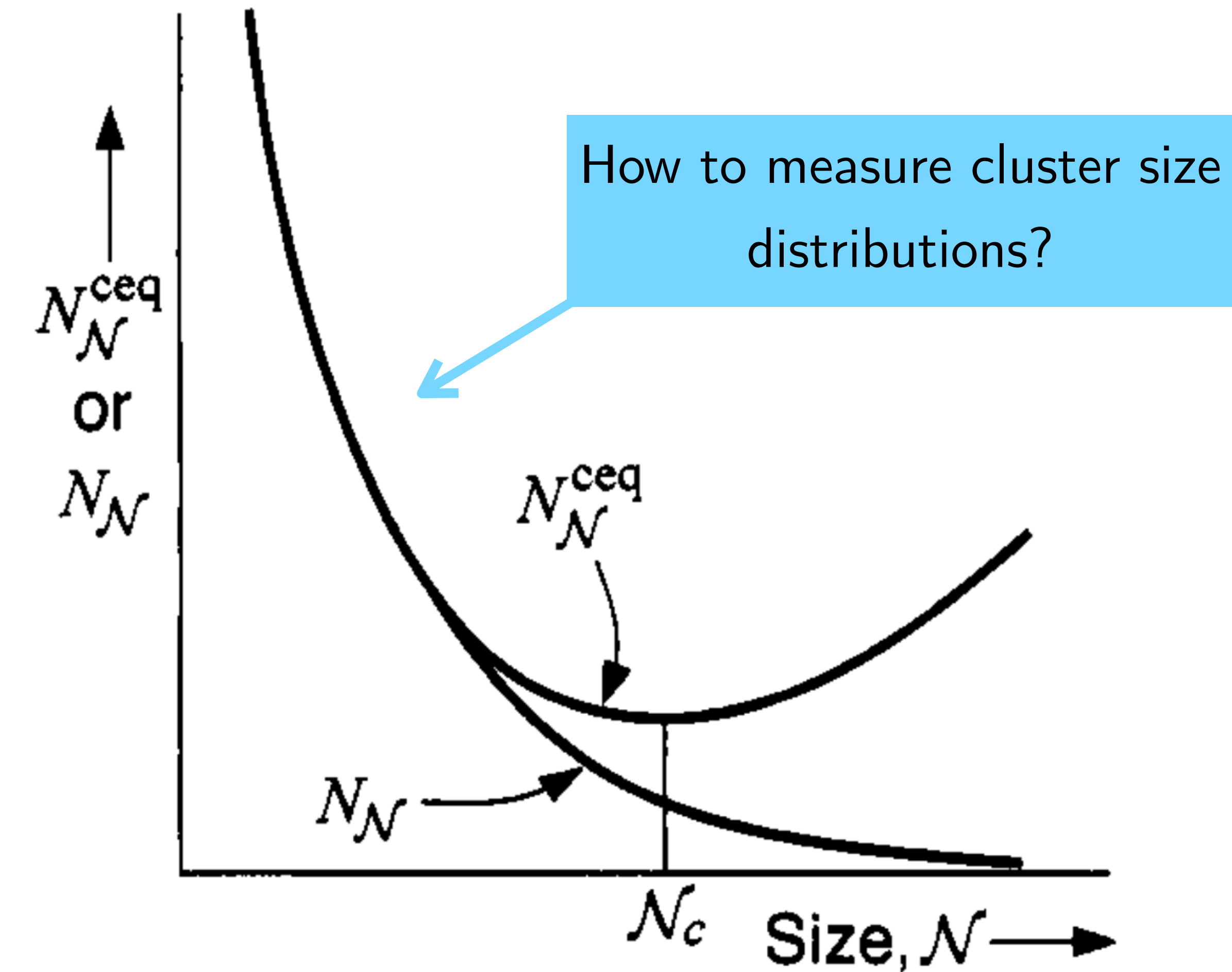


Modeling nucleation and growth

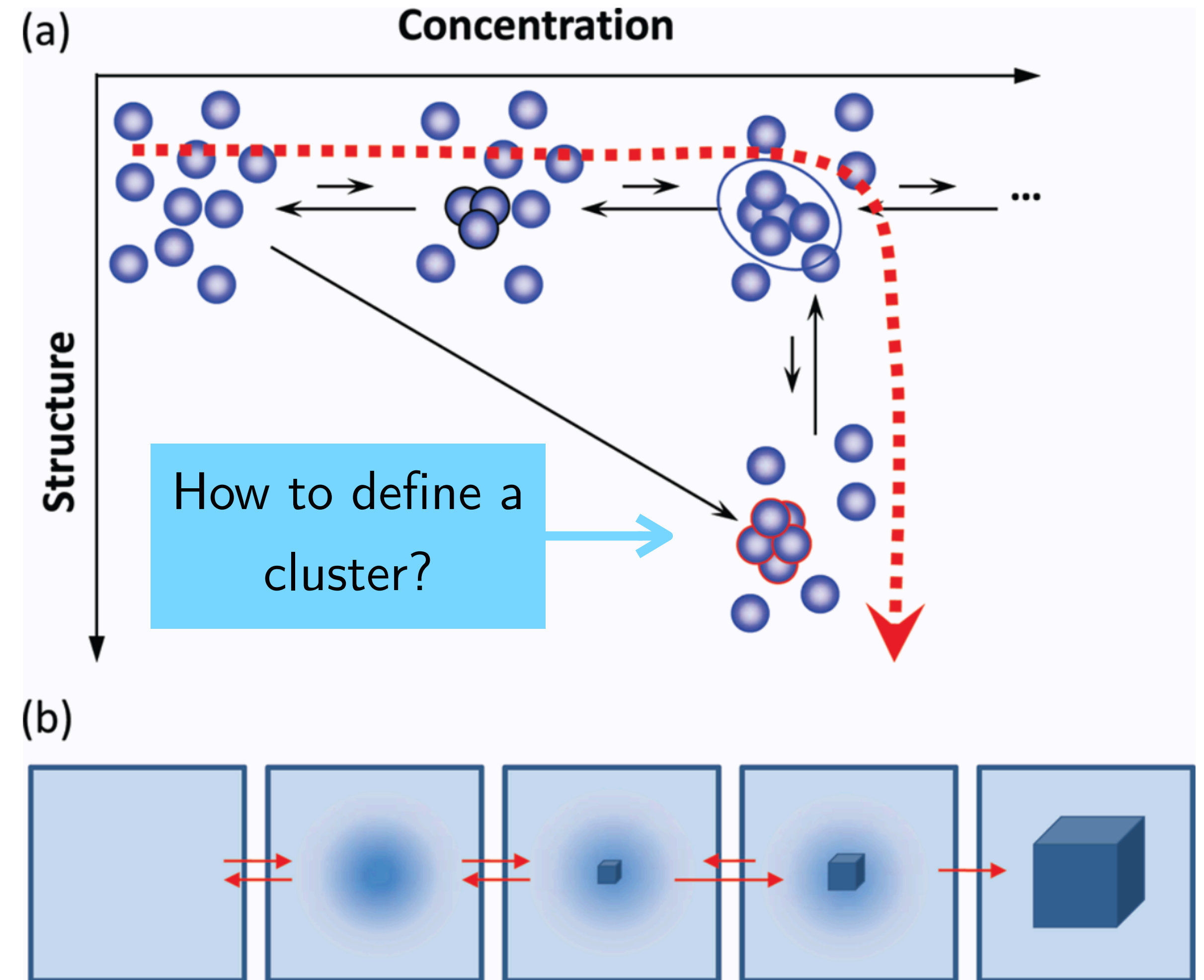


Examples in liquids and nucleation theory

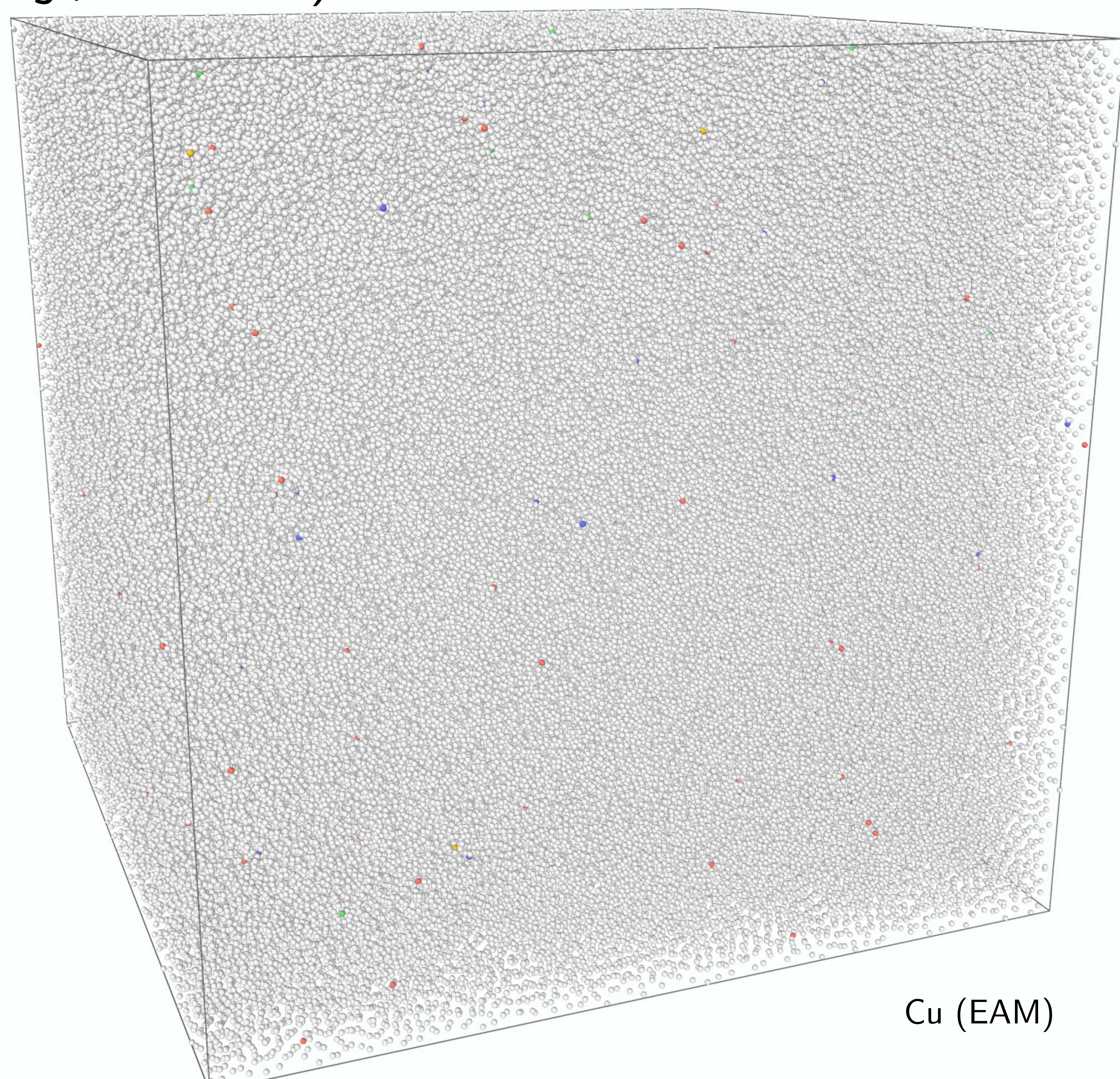
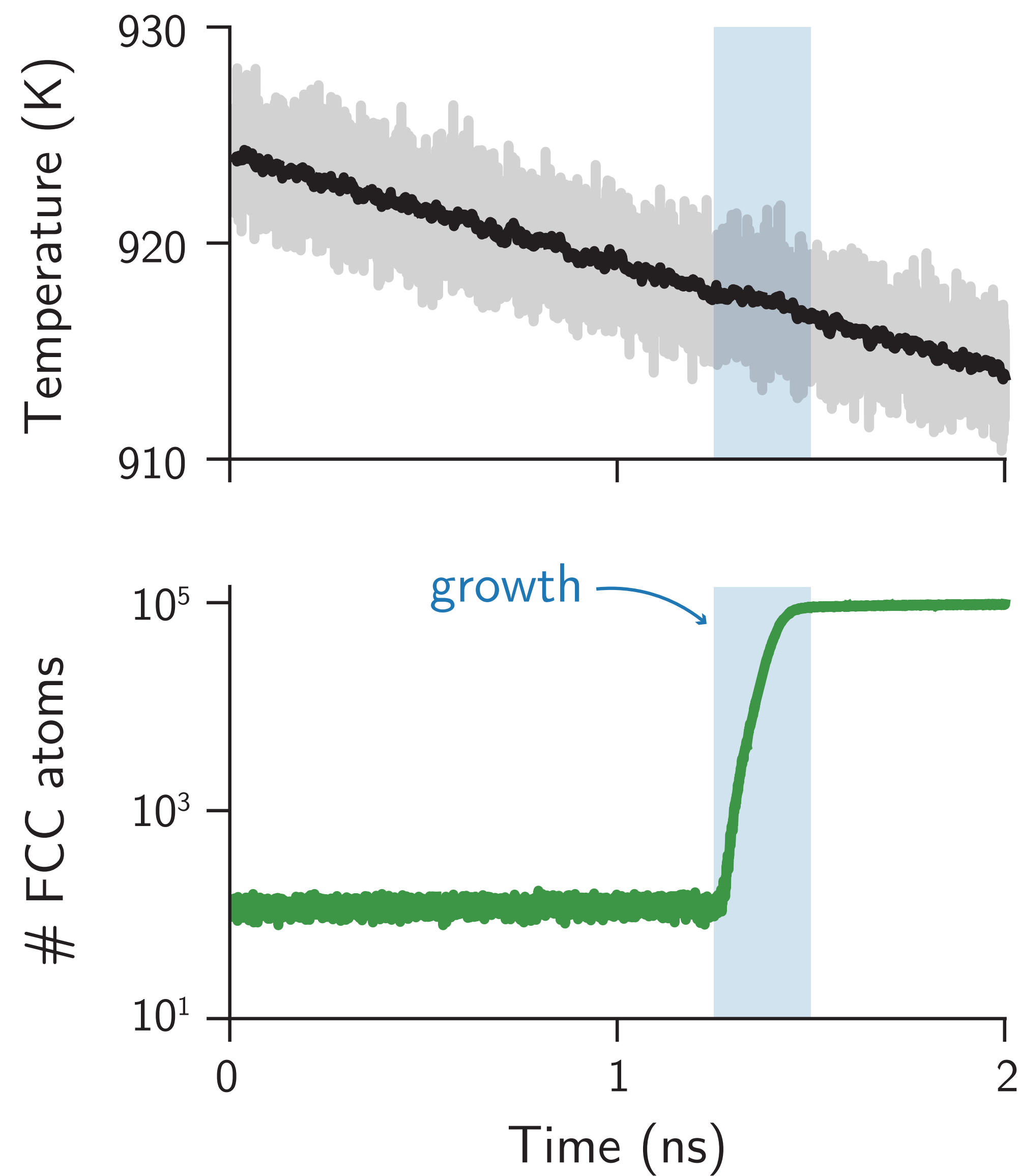
Cluster size distribution in CNT



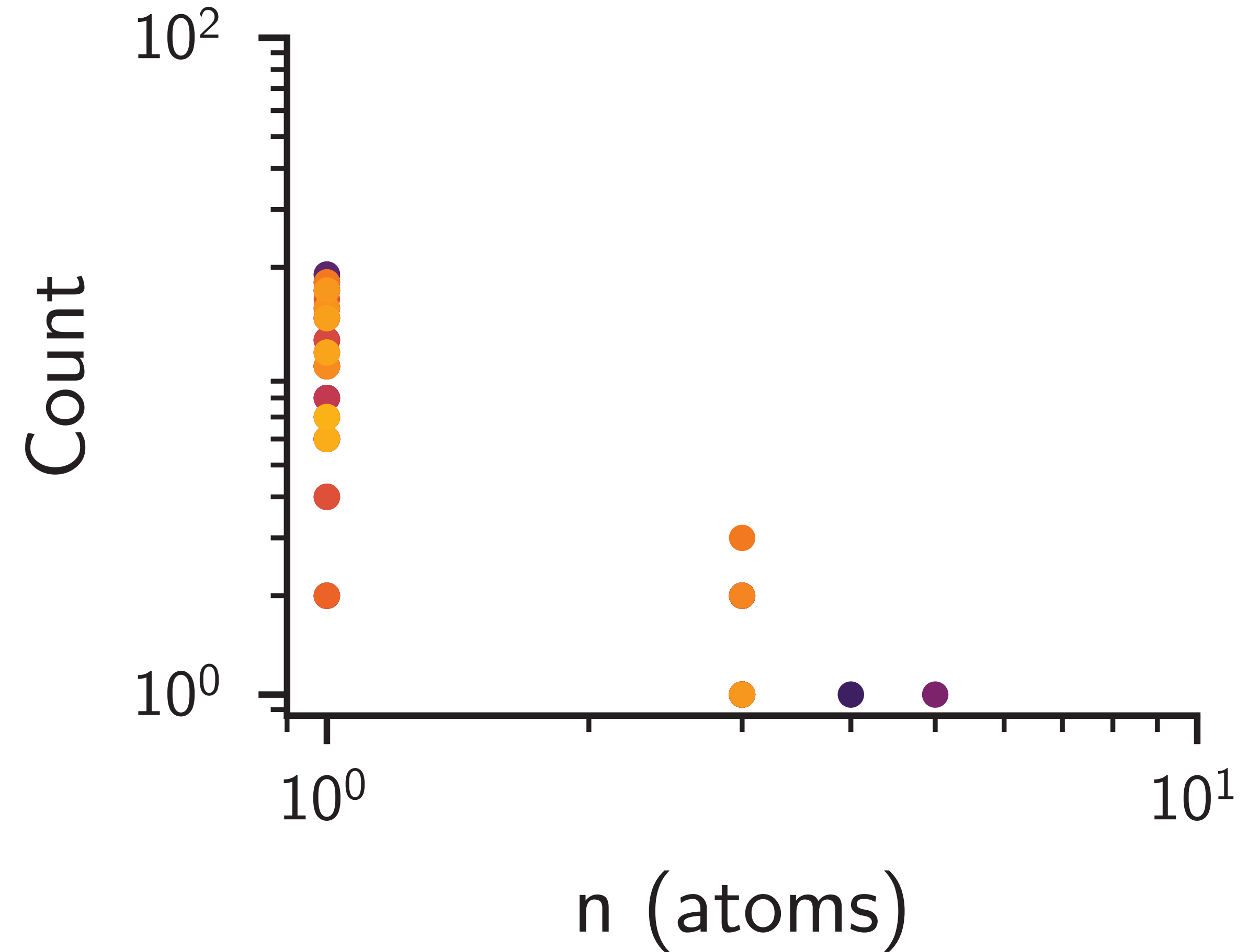
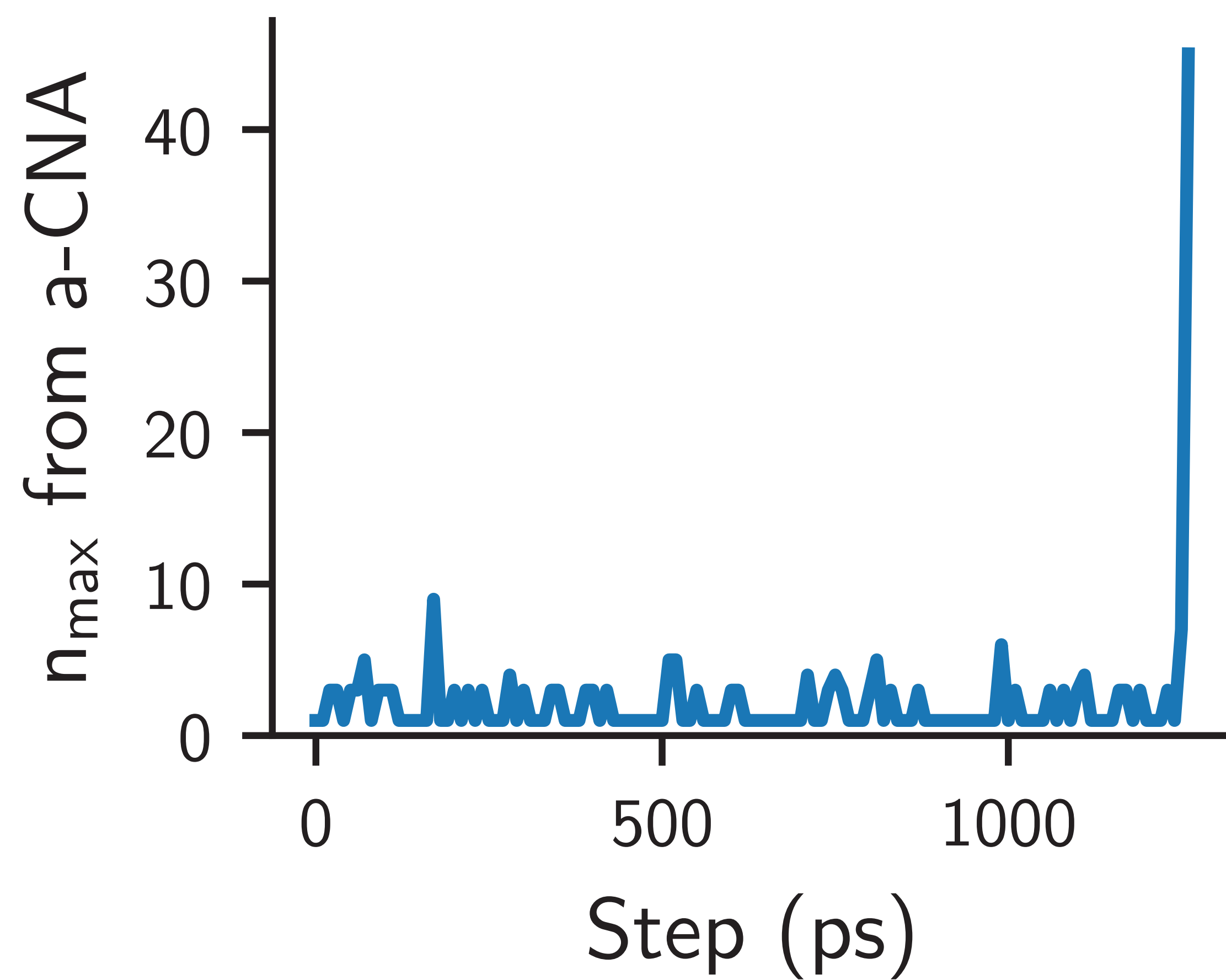
Two-step nucleation processes



Information theory can model rare events (e.g., nucleation)



Classical algorithms such as a-CNA cannot even detect clusters/nuclei

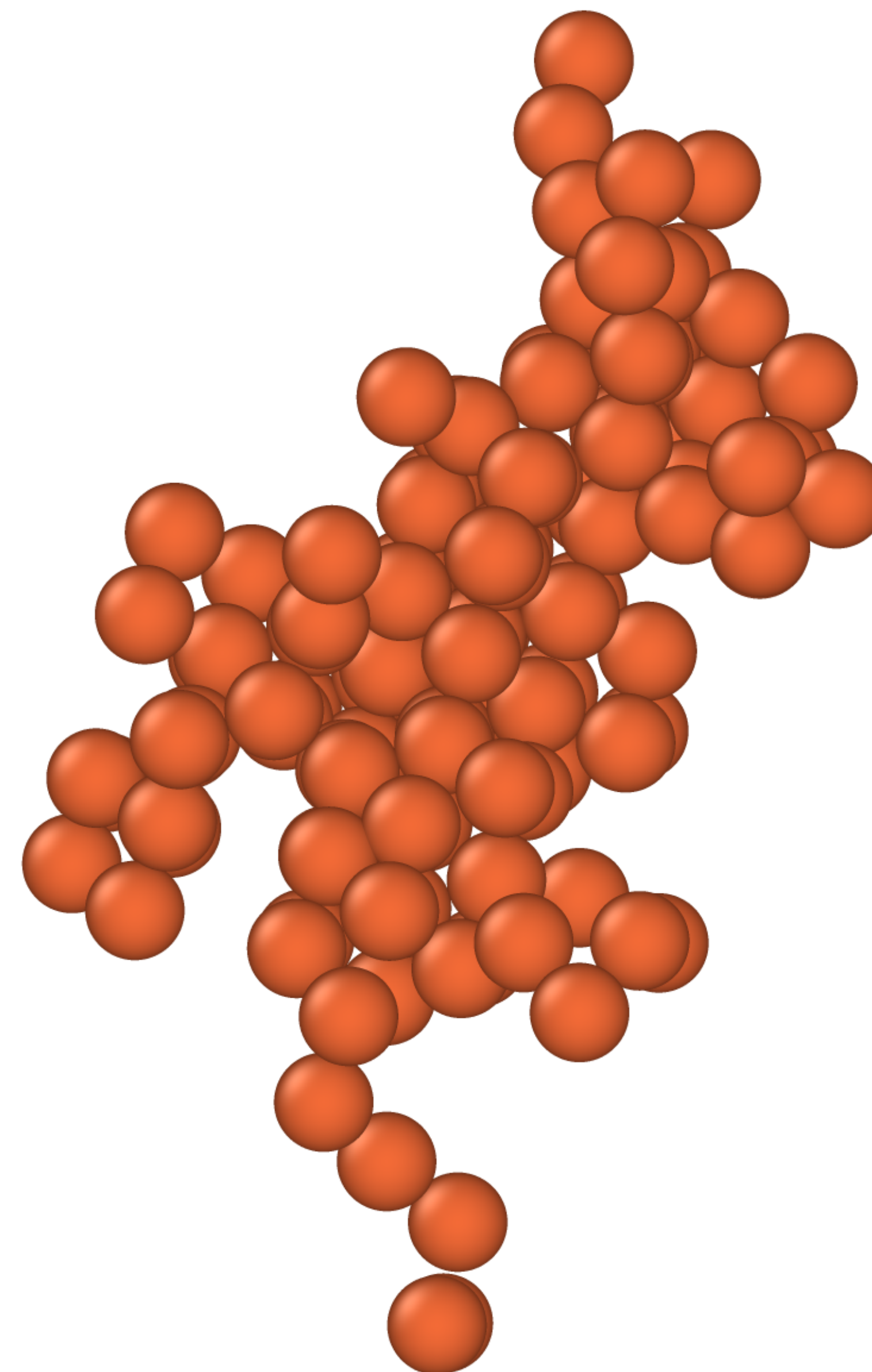
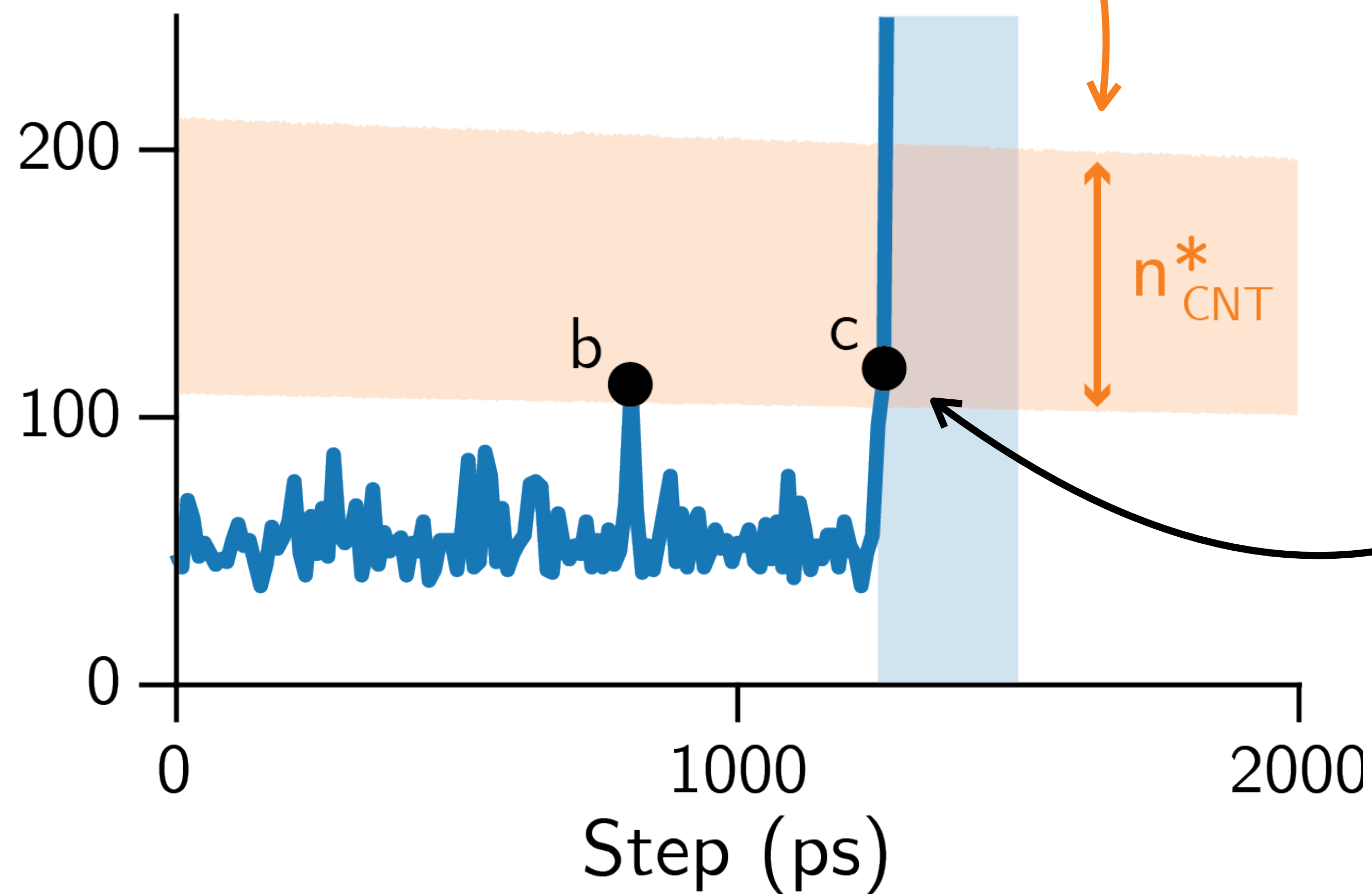


How to define a cluster in an information-theoretical perspective?

Information entropy allows us to extract structure from a melt

$$r^* = \frac{2\gamma_{\text{SL}}T_m}{\Delta H_m\Delta T}$$

a
largest cluster size



Information entropy allows us to verify predictions from the CNT

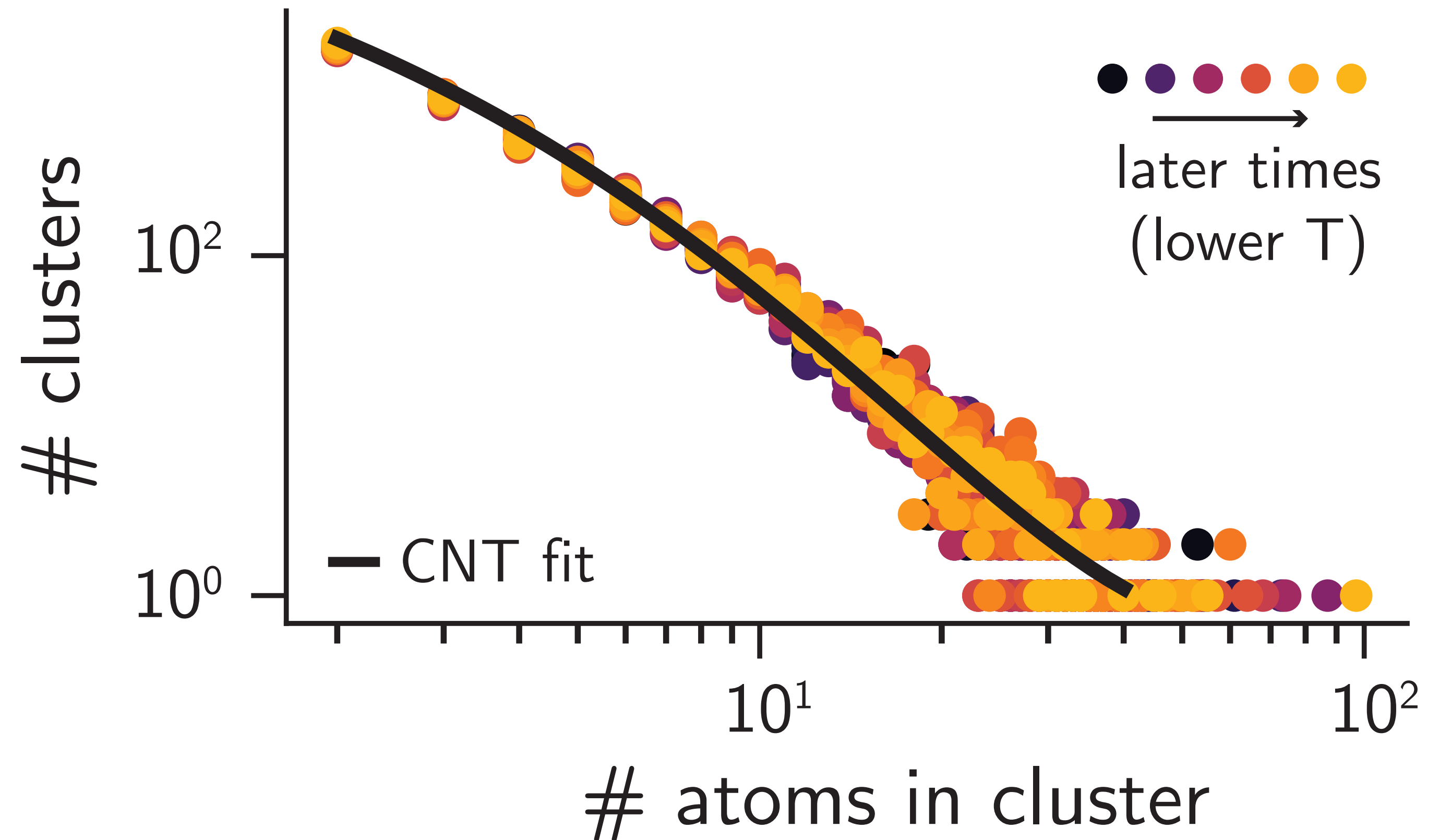
From the CNT:

$$N_r = N_0 \exp \left(-\frac{\Delta G_r}{k_B T} \right)$$

where

$$\Delta G_r = \frac{4}{3}\pi r^3 \Delta g_{\text{SL}} + 4\pi r^2 \gamma_{\text{SL}}$$

We can obtain the transformation and interfacial energies directly from the data →

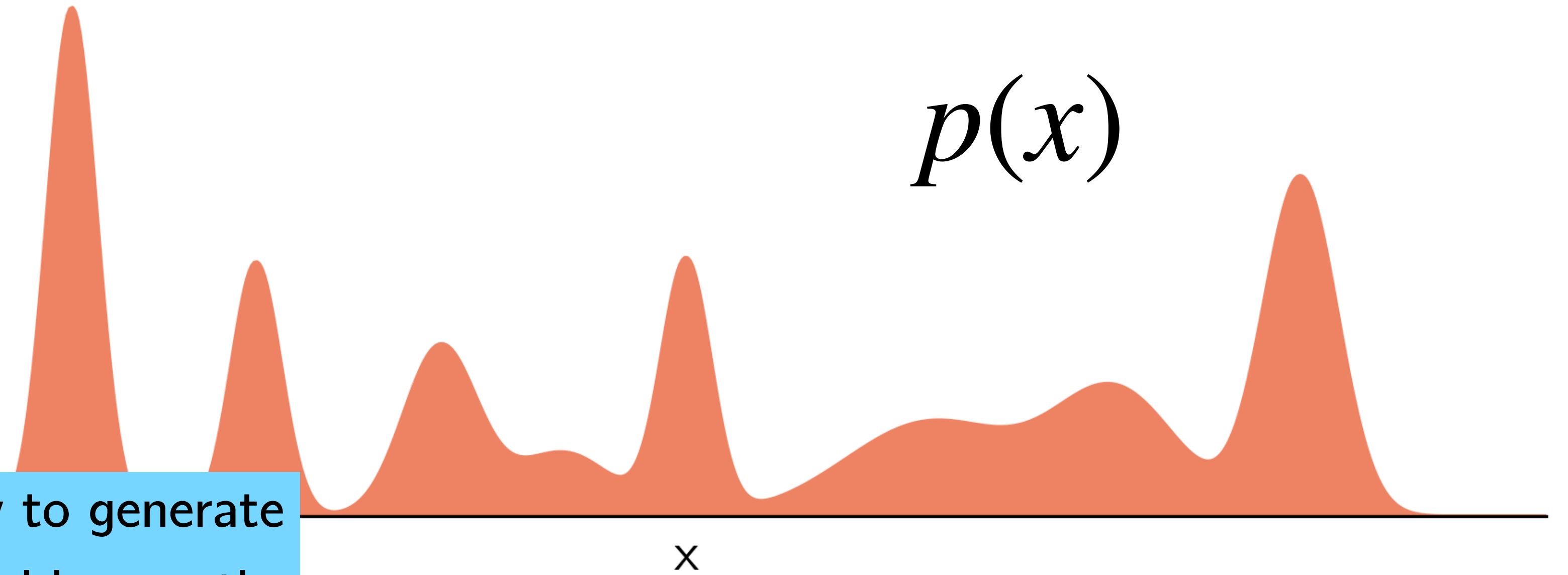


Information entropy allows atomistic understanding of nucleation and growth, but still recovers the continuum models

Now that we analyzed the distributions, we have one problem: how to generate them?

38

Entire distributions:
take a long time to
sample

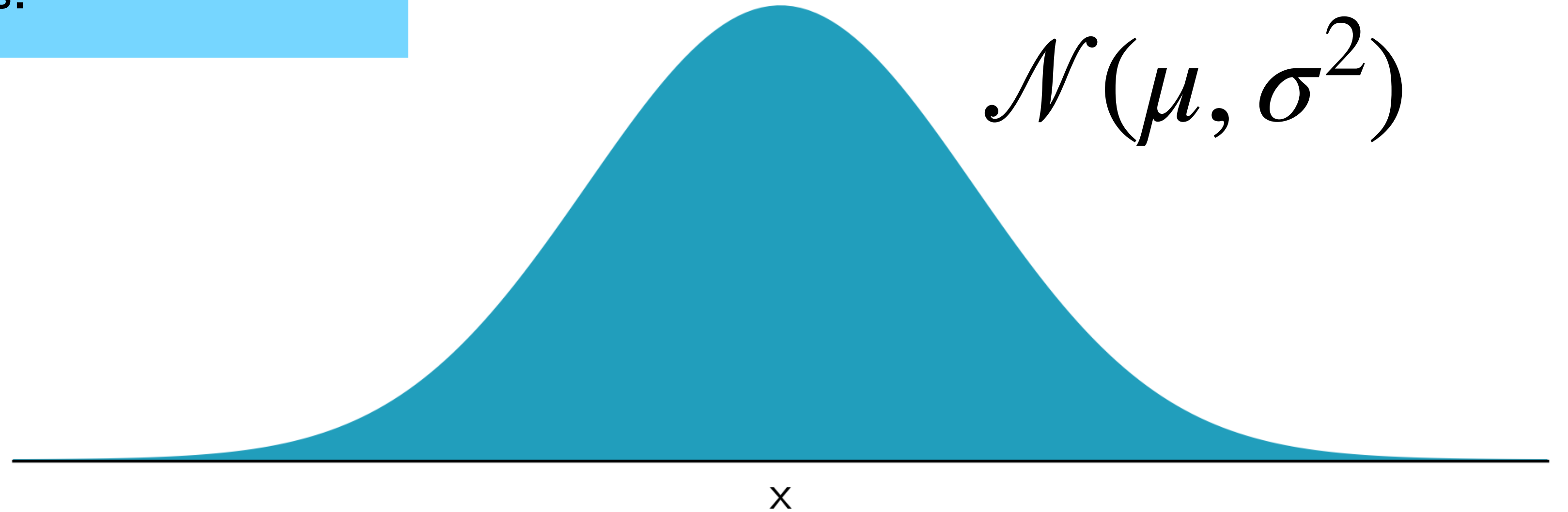


$p(x)$

x

It would be nice if we had an easy way to generate distributions of atomistic structures and bypass the sampling process!

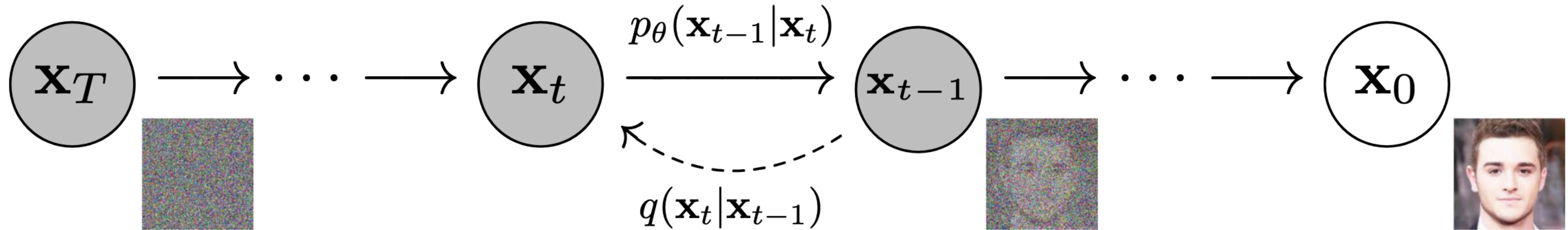
Simple distributions:
trivial to sample



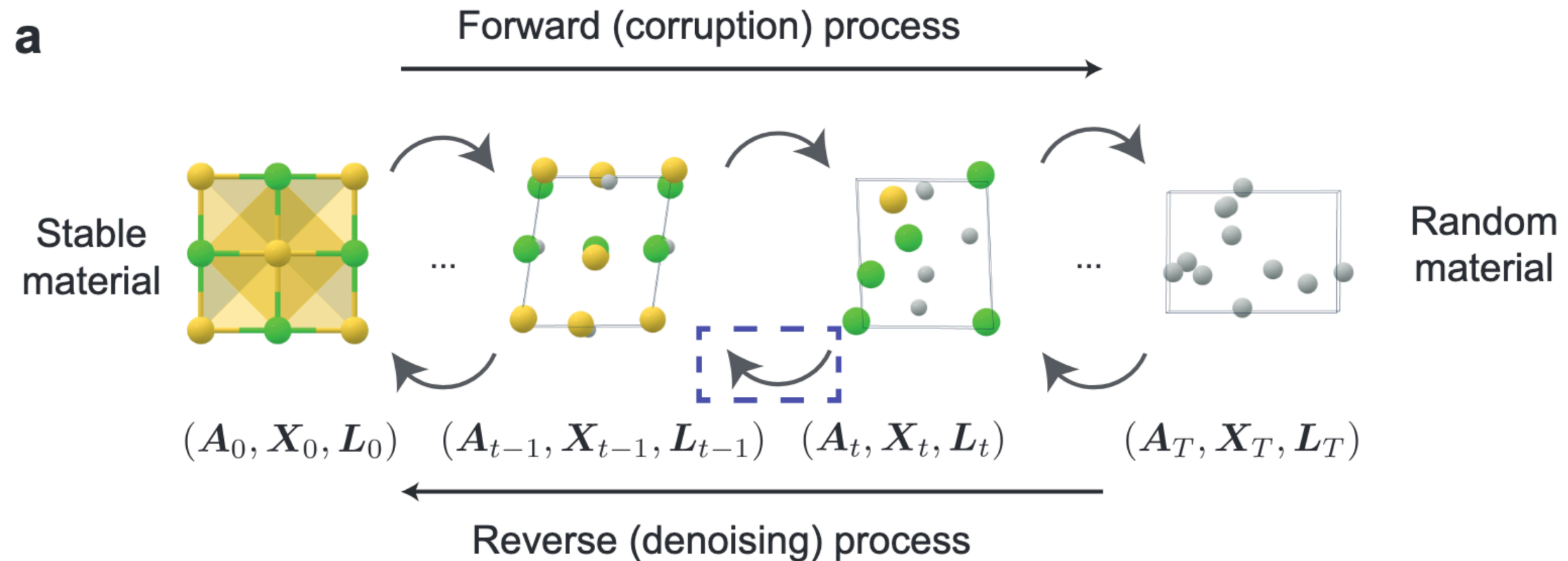
$\mathcal{N}(\mu, \sigma^2)$

x

Diffusion models have proven very effective in sampling/learning data distributions

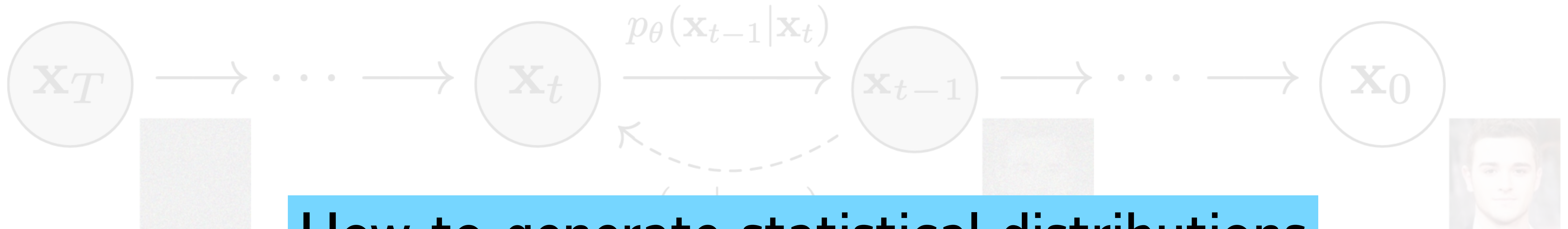


J. Ho et al. NeurIPS (2020). <https://arxiv.org/pdf/2006.11239>



C. Zeni et al. Nature (2025). <https://www.nature.com/articles/s41586-025-08628-5>

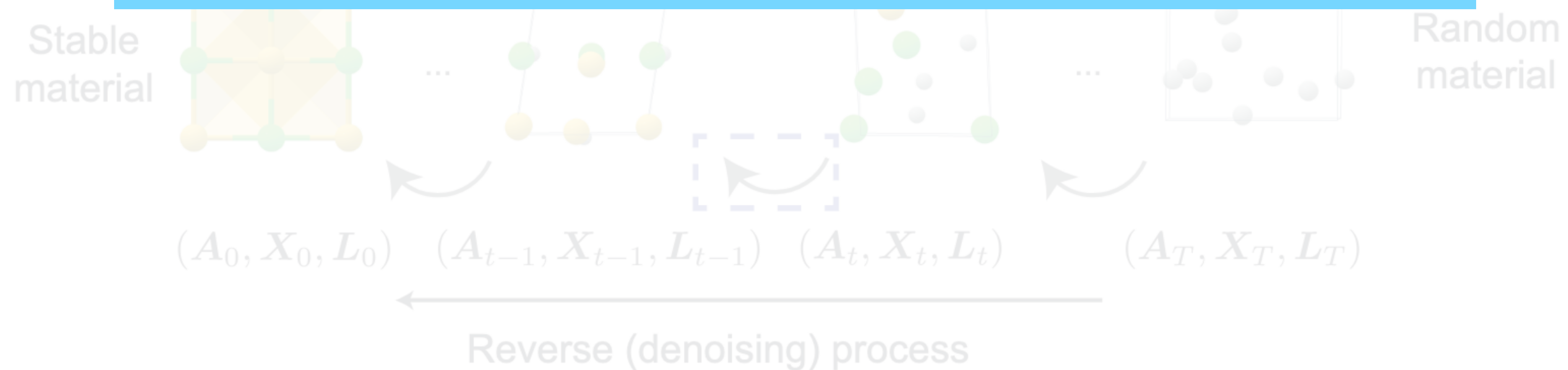
Diffusion models have proven very effective in sampling/learning data distributions



How to generate statistical distributions of complex atomistic systems such as amorphous materials and interfaces?

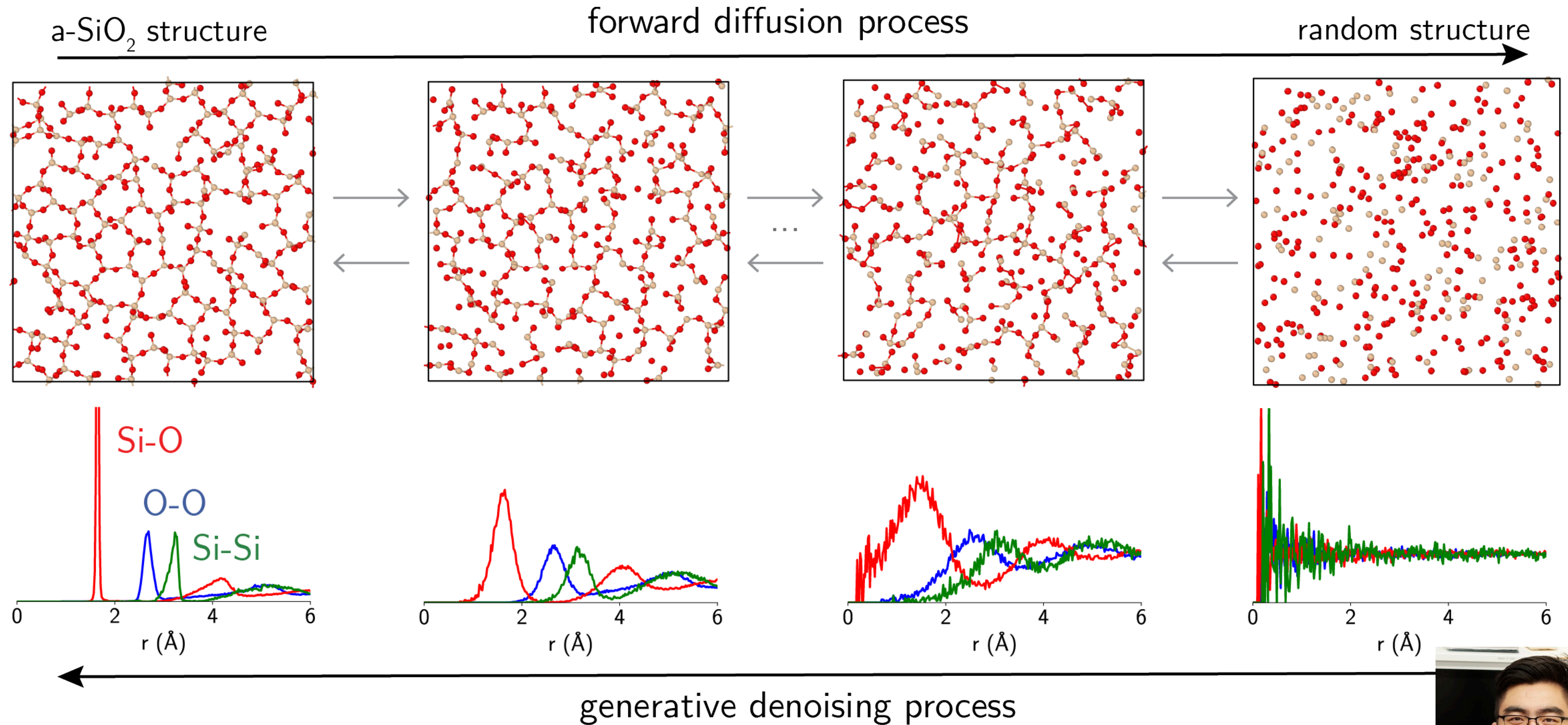
J. Ho et al. NeurIPS (2020). <https://arxiv.org/abs/2006.11238>

a

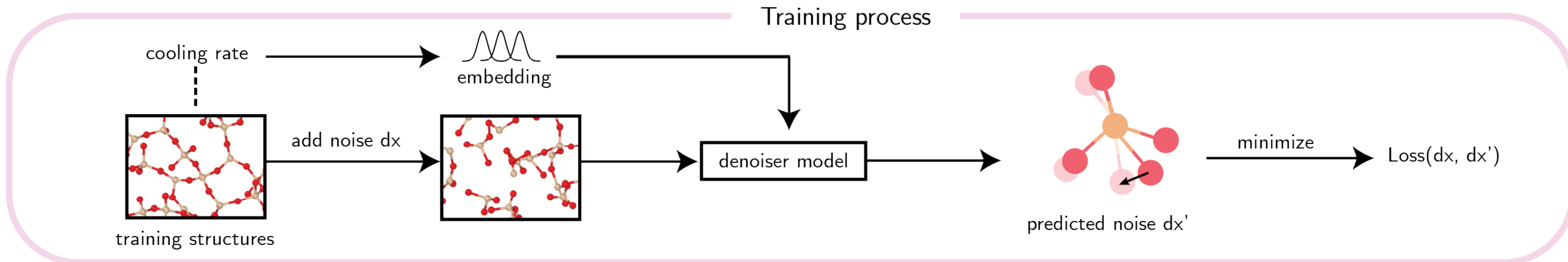


C. Zeni et al. Nature (2025). <https://www.nature.com/articles/s41586-025-08628-5>

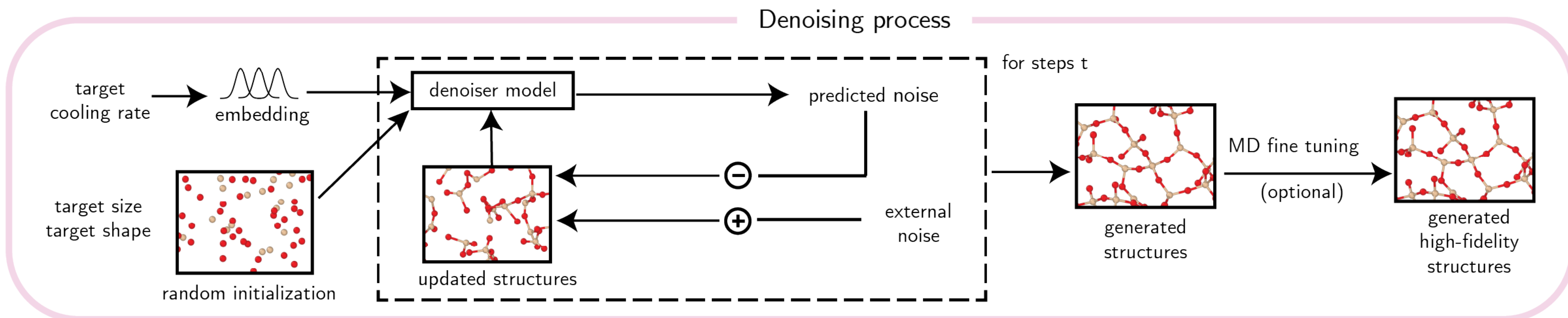
Proposing a diffusion model for amorphous materials



Training a diffusion model for amorphous materials

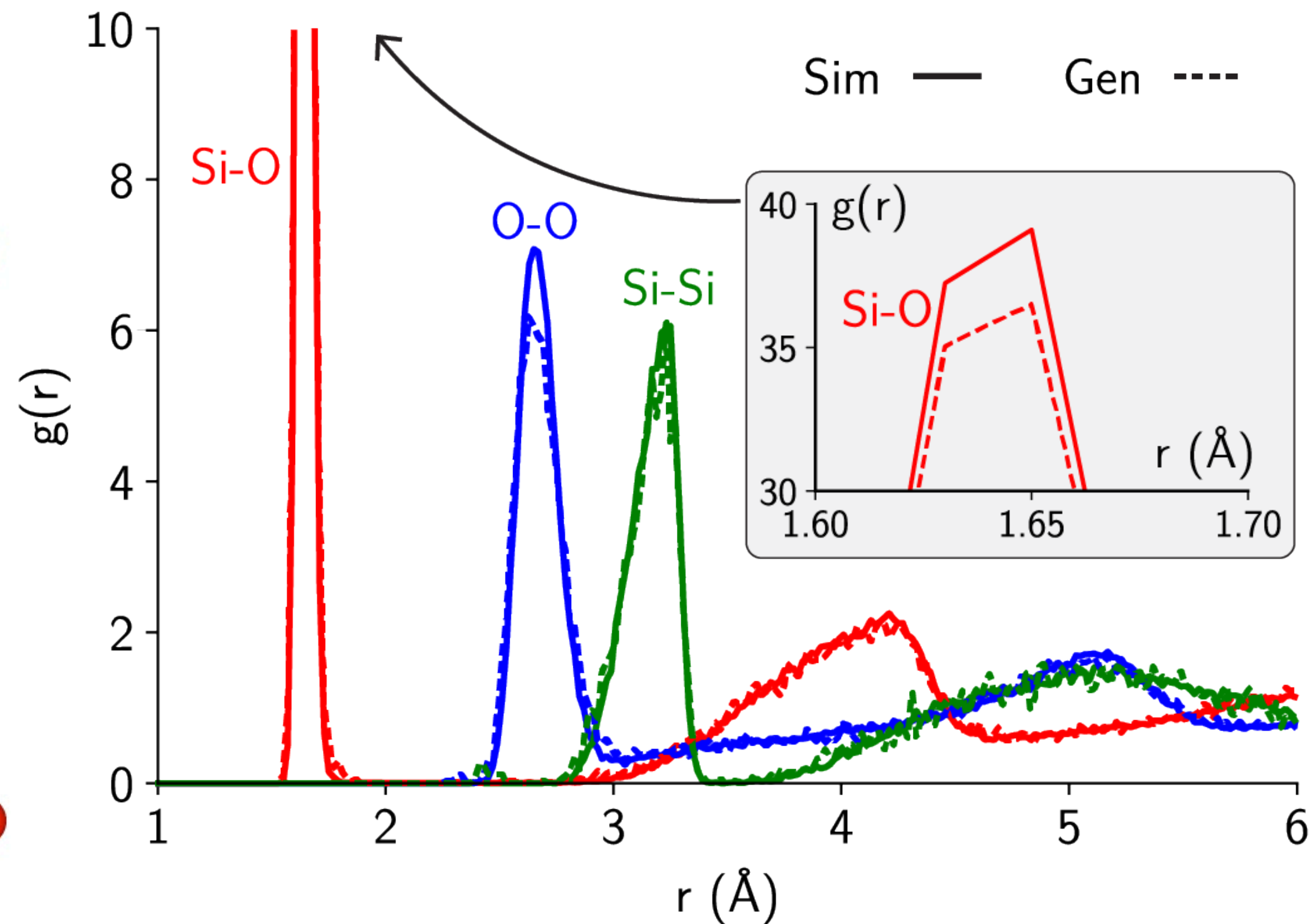
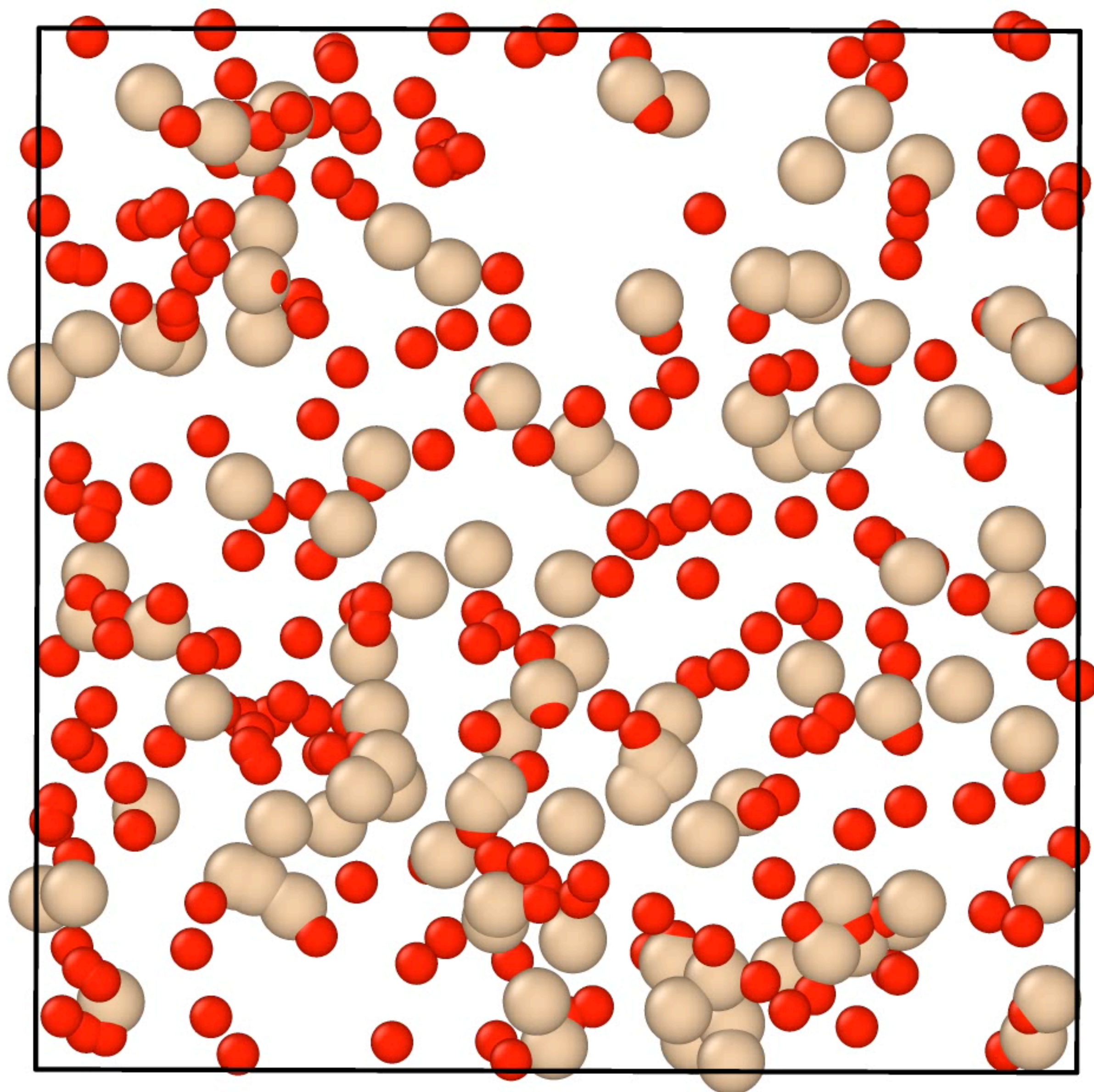


Training based on: T. Hsu et al. *npj Comp. Mater.* (2024)



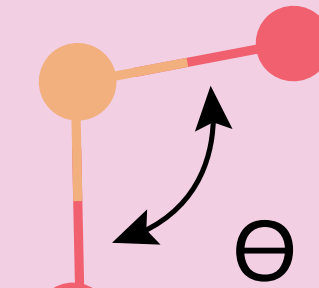
K. Yang and D. Schwalbe-Koda. *arXiv:2507:05024* (2025)

Diffusion model excels at sampling amorphous silica

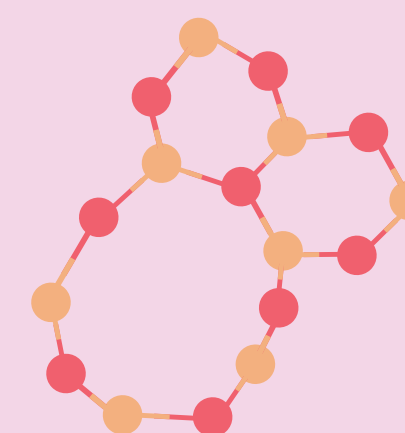


Gen. model
reproduces:

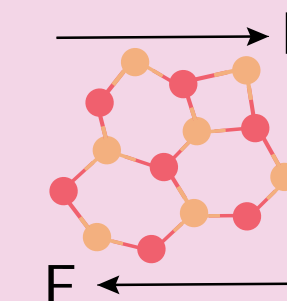
bond angle
distribution
(BAD)



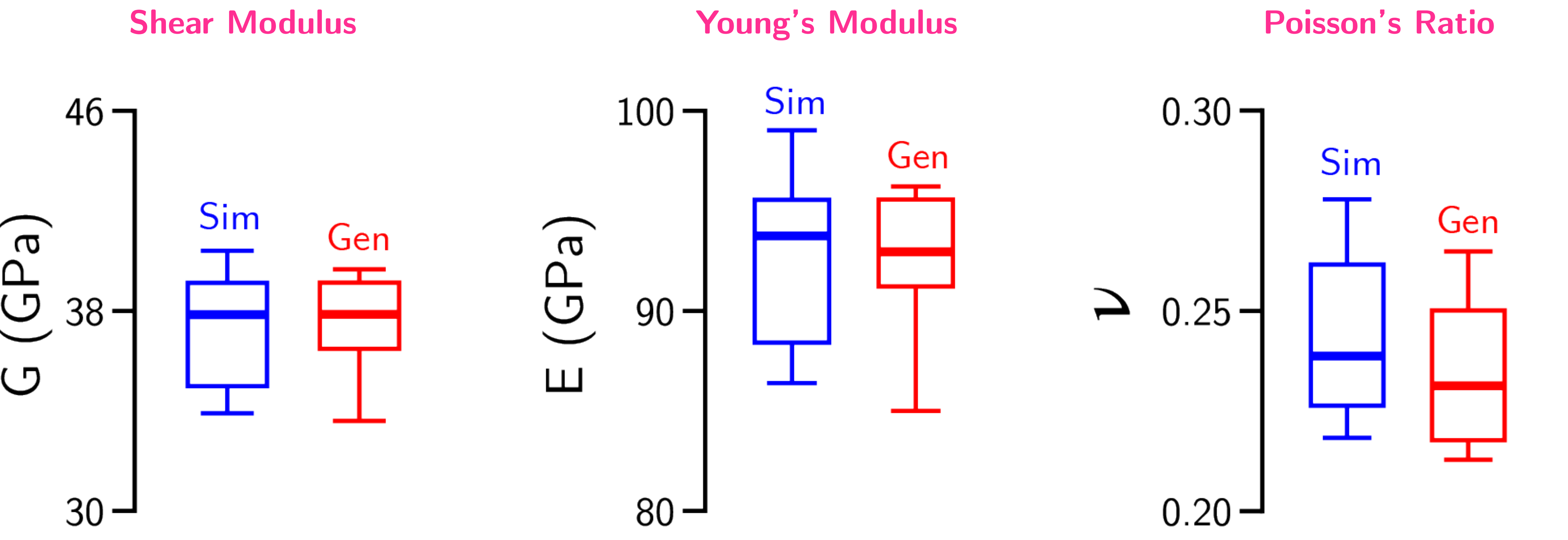
ring size
distribution

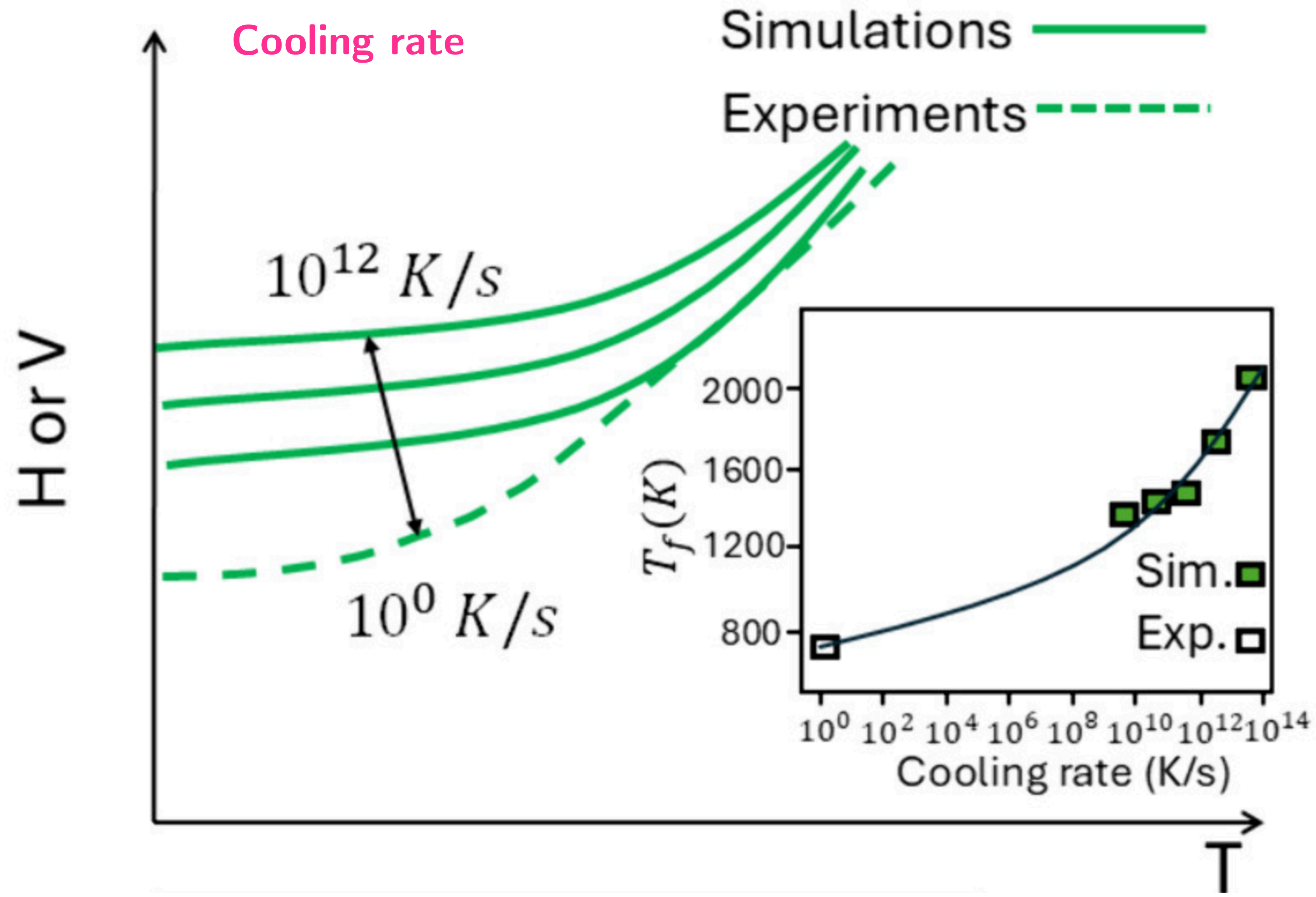


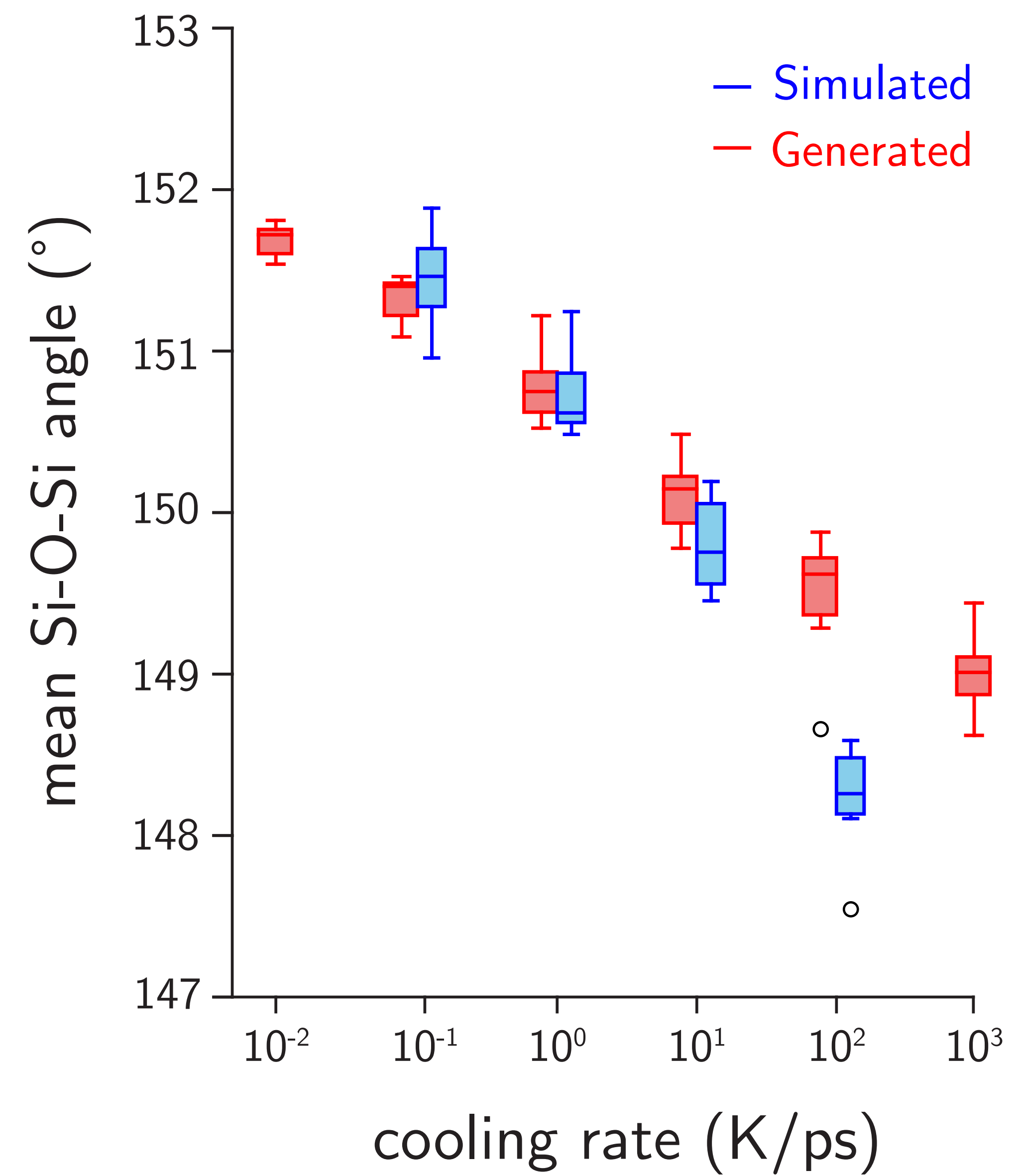
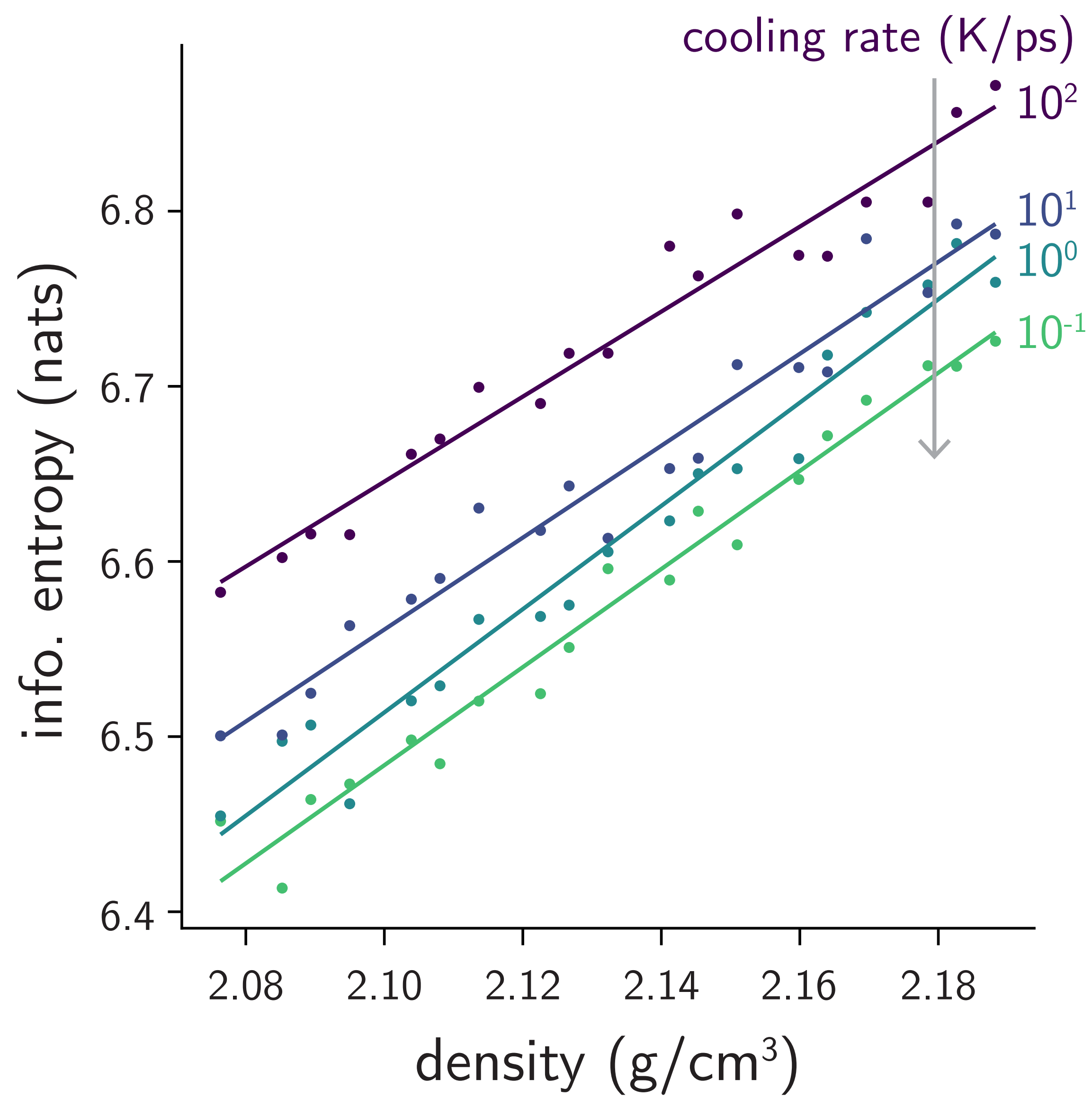
mechanical
properties

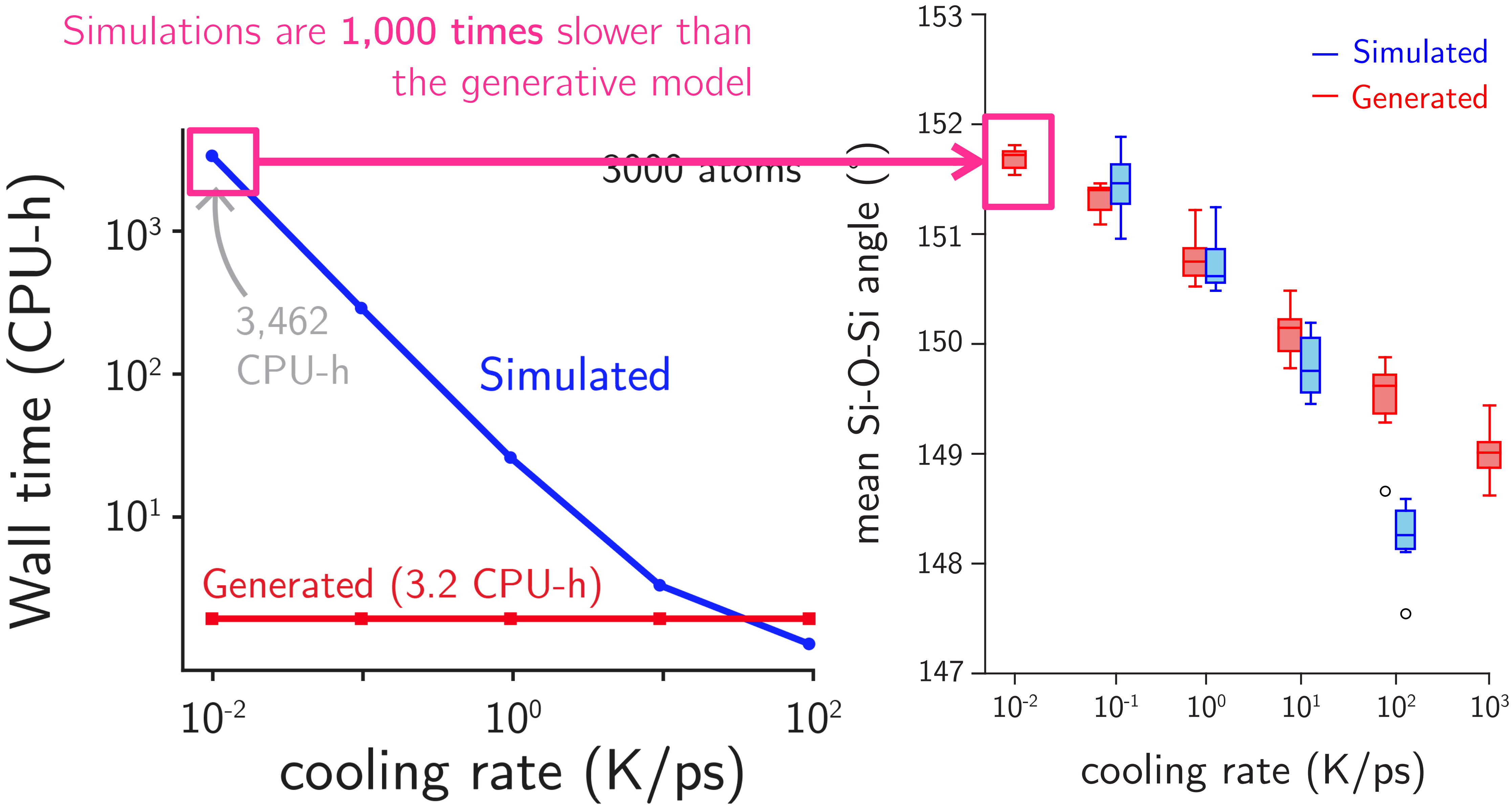


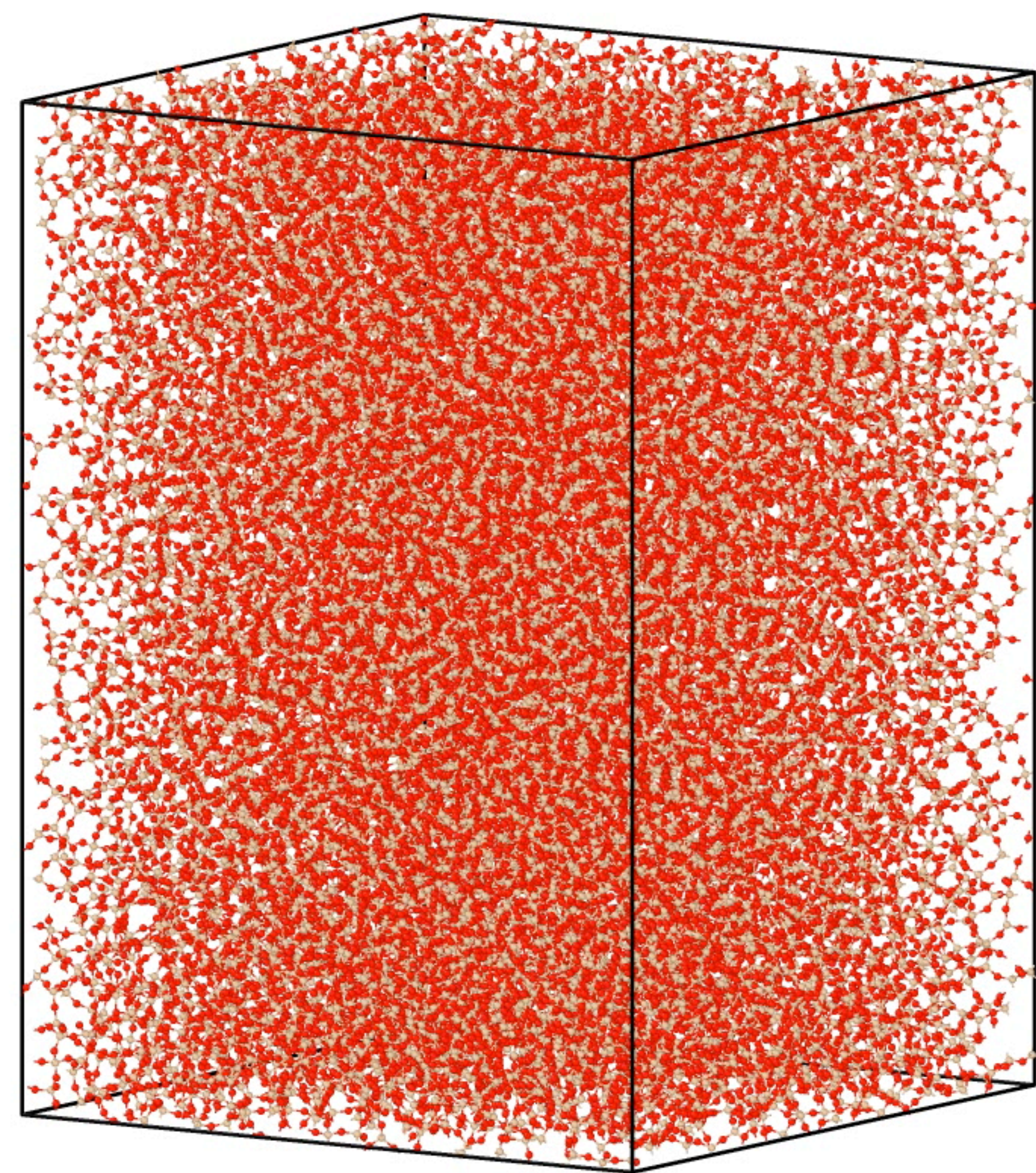
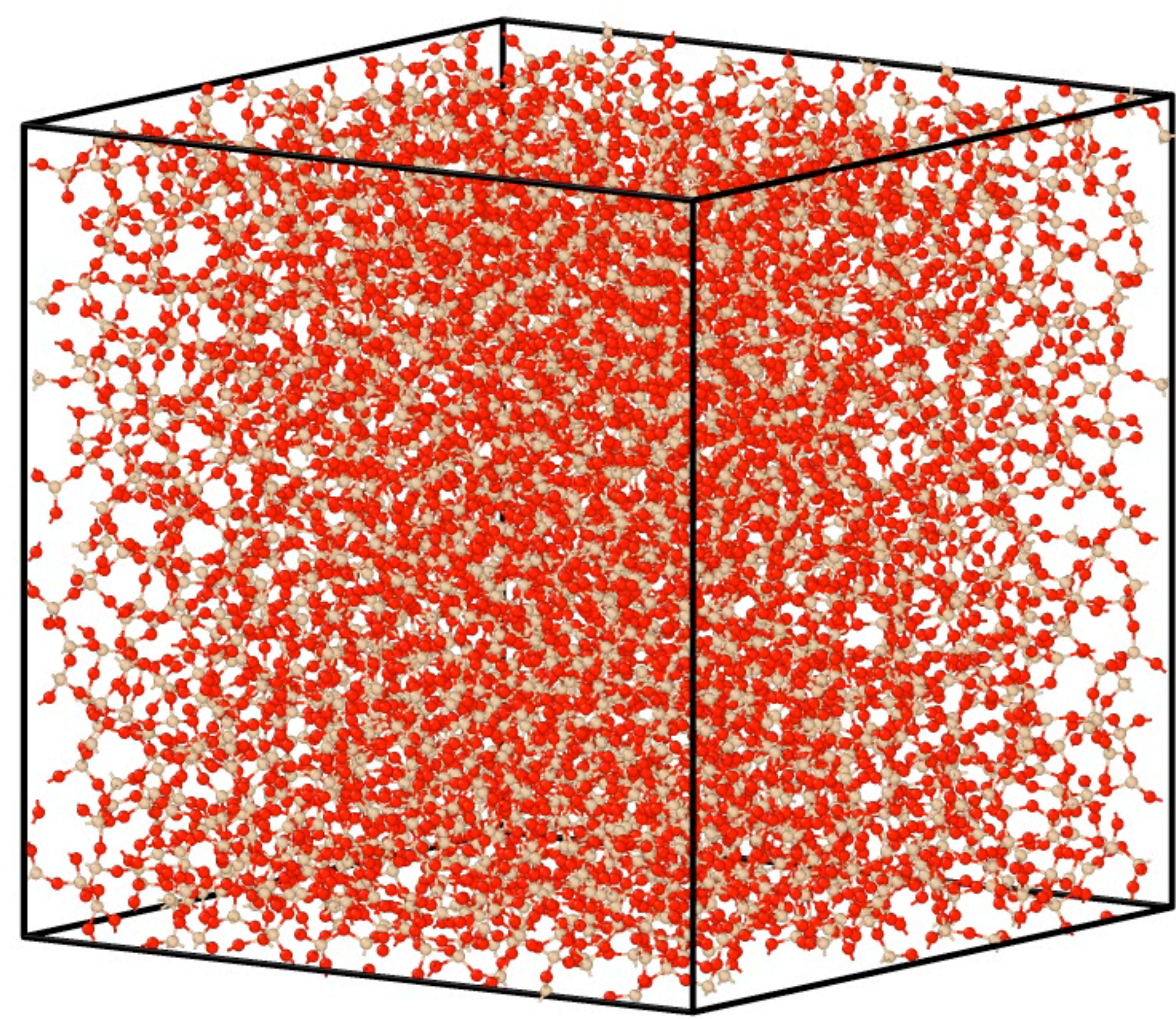
E, G, K, V



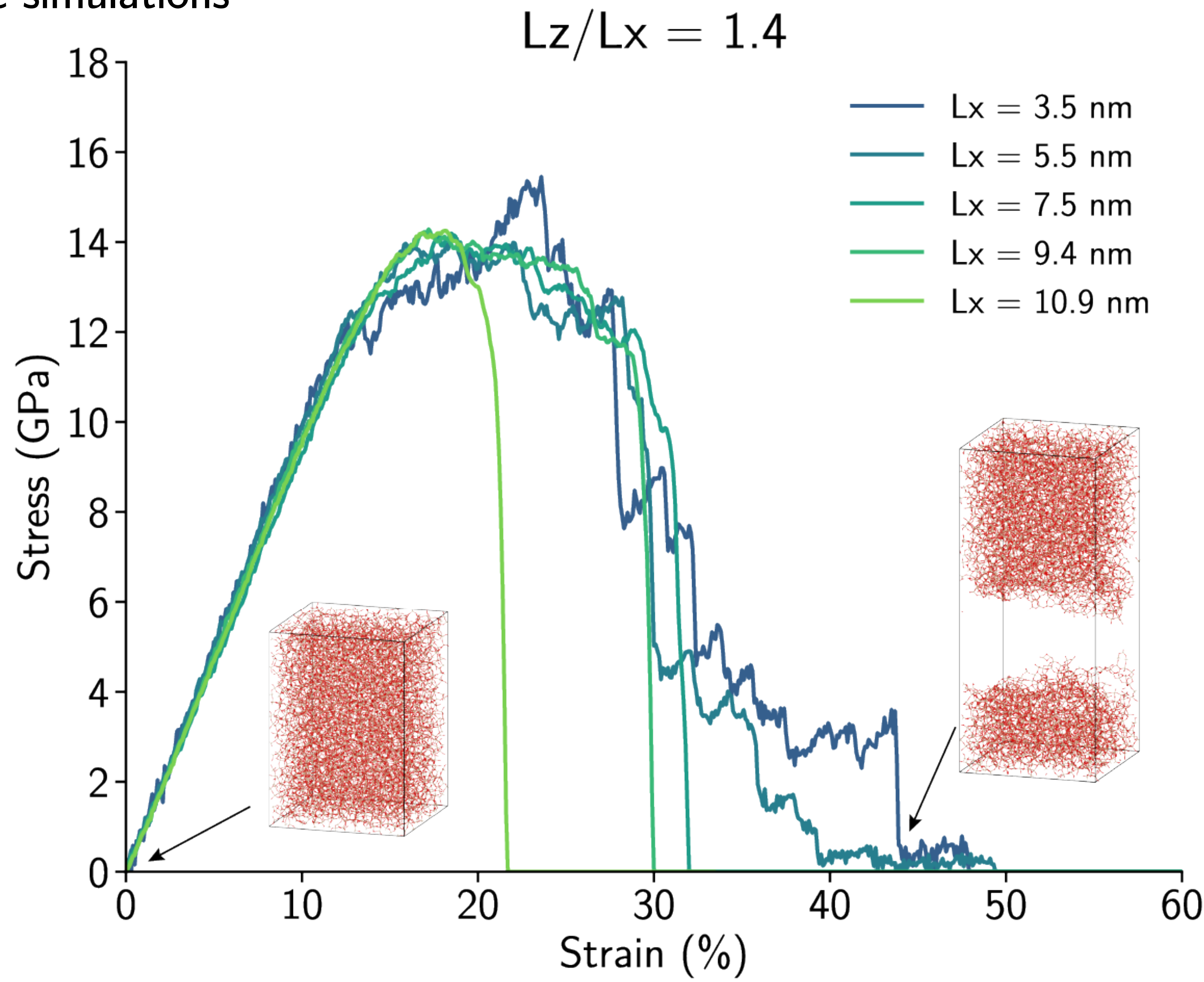


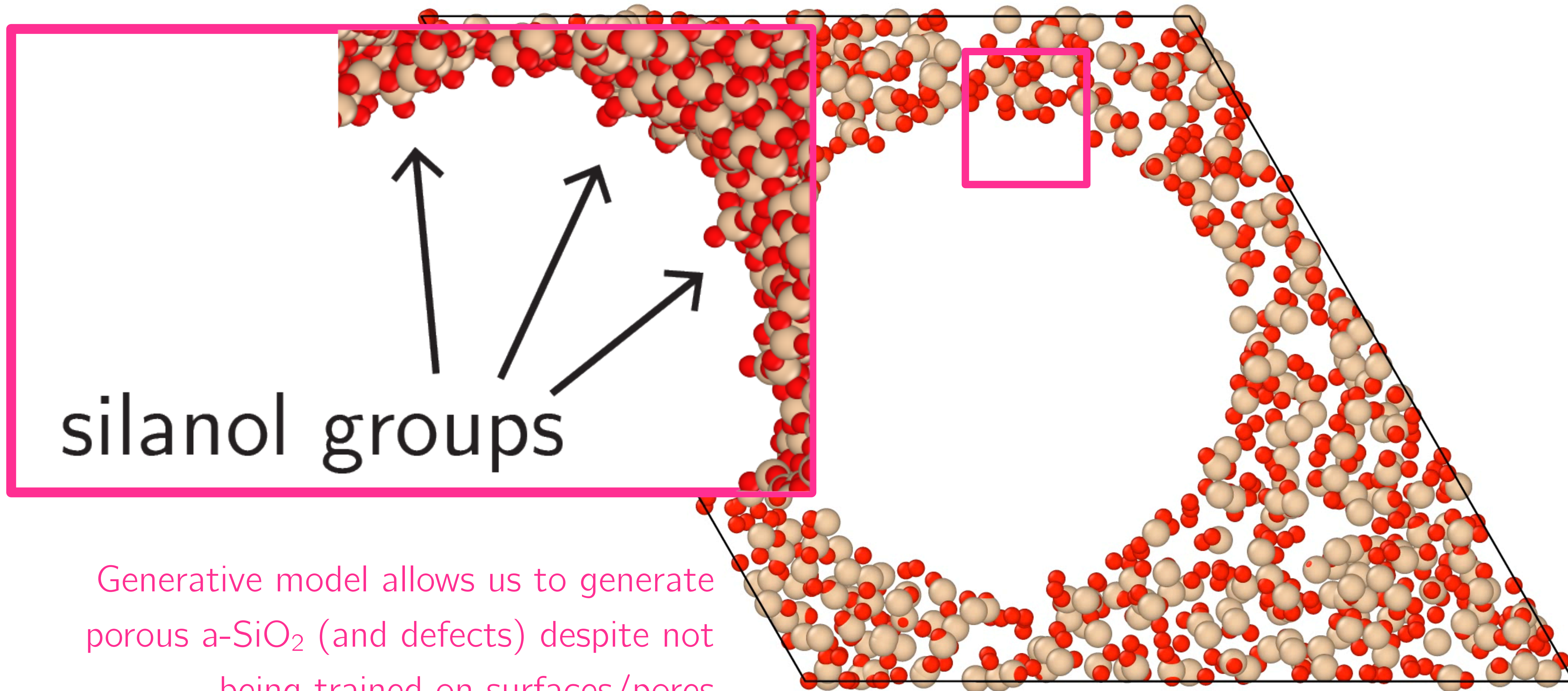


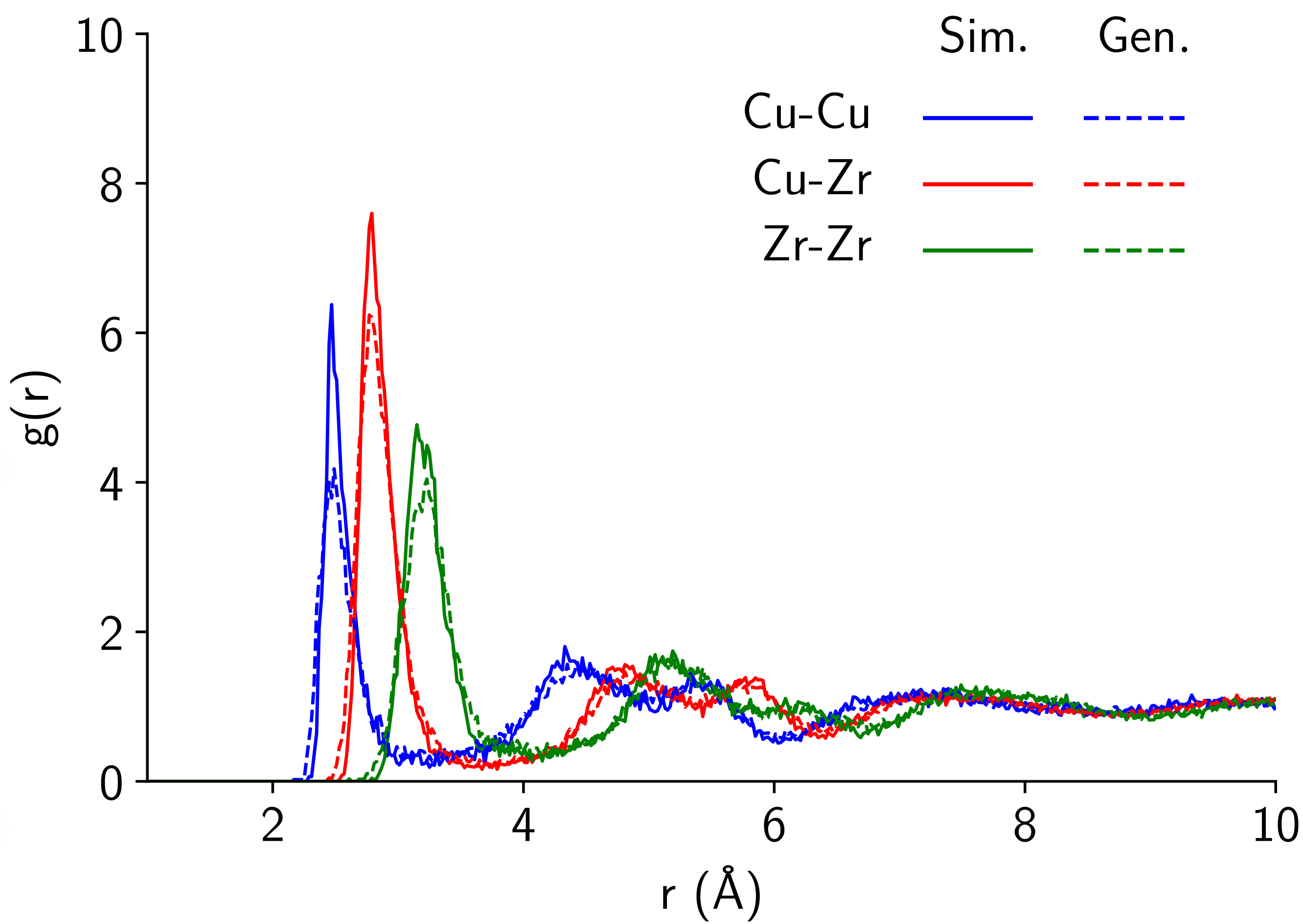
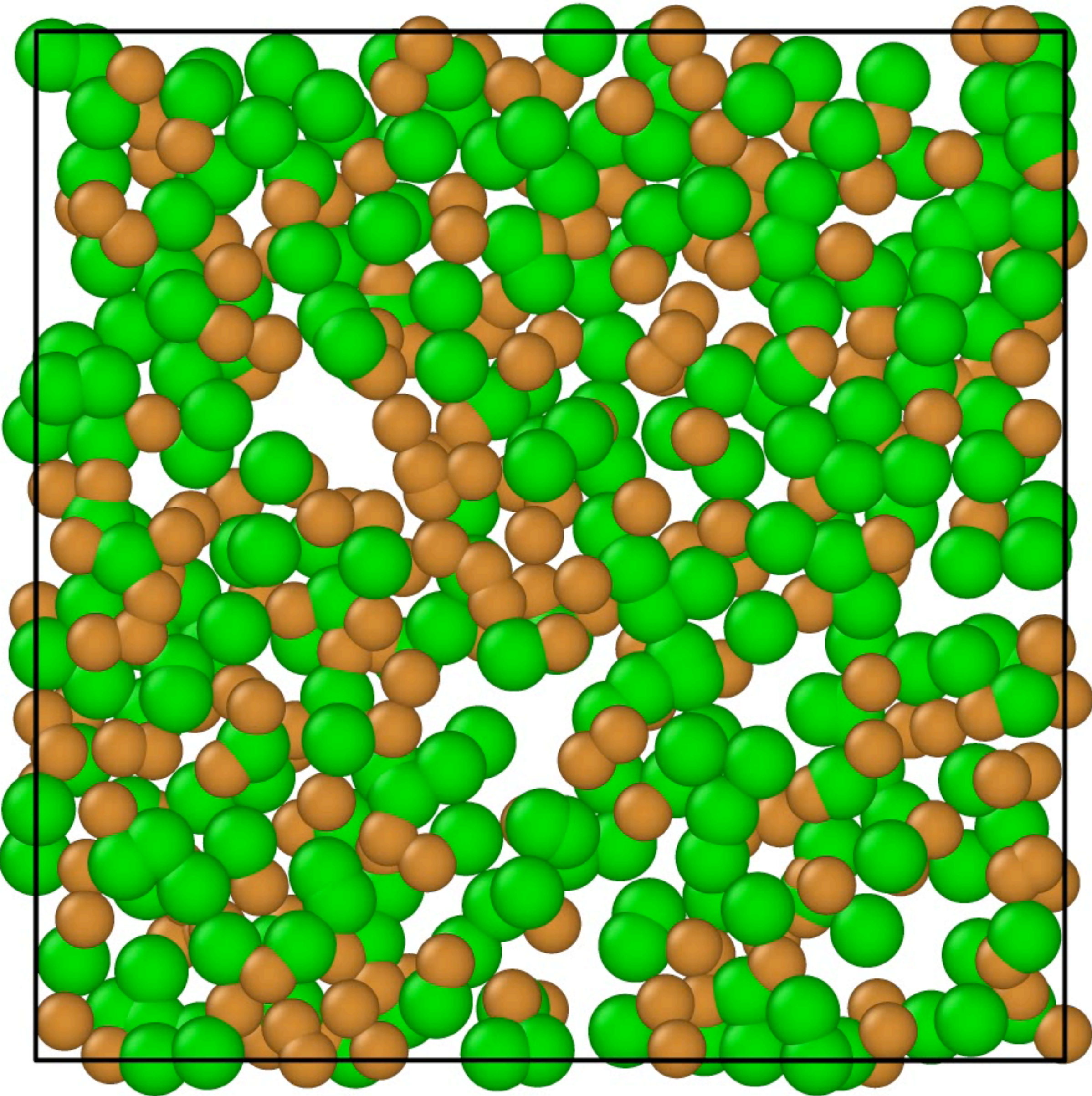




Generative model allows us
to capture the ductile-to-
brittle transition in a-SiO₂
due to length scales







Bridging Scales in Materials Modeling with Atomistic Simulations, Information Theory, and Generative Models

tl;dr:

Bridging scales in materials modeling - especially synthesis science - requires advances in:

- (1) understanding model generalization
- (2) connecting atomistic simulations with macroscopic theories
- (3) sampling methods and generative models relevant to materials science

Acknowledgements

LLNL

Sebastien Hamel

Babak Sadigh

Vasily Bulatov

Fei Zhou

Nitish Govindarajan (NTU)

Joel Varley

Vincenzo Lordi

Tuan Anh Pham

Brandon C. Wood

UCLA

Jiawei Guo

Kai Yang

Xiangrui Li

Yannah Melle

Jiayi Dong

Jiahao Chen

Ben Yu

Josh Lipman

Steven Wen



Funding



#DE-SC0025642, ECRP



Synthesis Advanced Research Challenge



Scialog AUT

