Uncertainty Quantification With Surrogate Models in Alloy Modeling

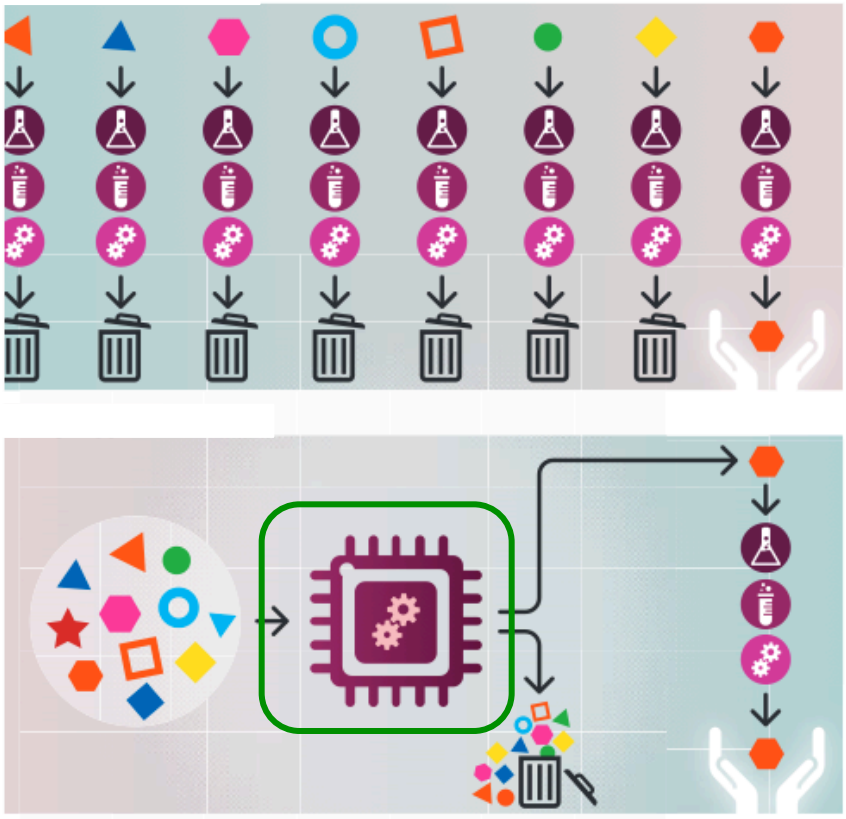
Jesper Kristensen,
(joint work with Prof. N. Zabaras*)

Applied and Engineering Physics &
Materials Process Design & Control Laboratory
Cornell University
271 Clark Hall, Ithaca, NY 14853-3501
and
***Warwick Centre for Predictive Modelling**
University of Warwick, Coventry, CV4 7AL, UK

Warwick Centre for Predictive Modelling Seminar Series, University of Warwick, Coventry, UK, January 8, 2015

FINDING THE BEST MATERIALS

Search for target material/property

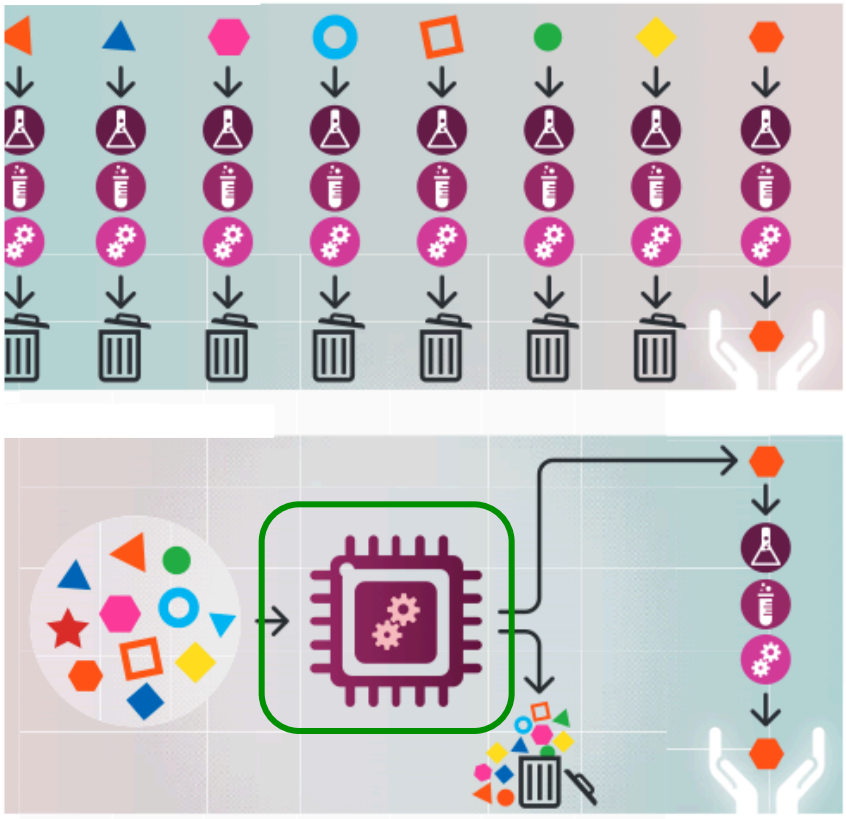


- Single candidate property (one column to the left)
  - Hours
- Million+ candidates
  - Infeasible!

Unmodified image in: G. Ceder and K. Persson. Scientific American (2013)

FINDING THE BEST MATERIALS

Search for target material/property

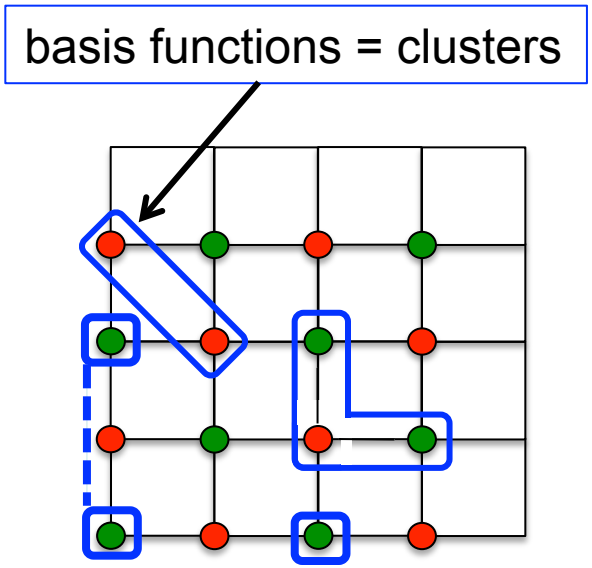


- Single candidate property (one column to the left)
  - Hours
- Million+ candidates
  - Infeasible!
- Need for surrogate models
  - Rapid configuration space exploration
  - Allows design of materials

Unmodified image in: G. Ceder and K. Persson. Scientific American (2013)

Using Surrogates in Alloy Modeling

THE CLUSTER EXPANSION



$$E(\boldsymbol{\sigma}) \approx \sum_i J_i \sigma_i + \sum_{i,j} J_{i,j} \sigma_i \sigma_j + \cdots + (M)$$

- Alloy surrogate model
  - The cluster expansion
- Cluster with *n* points: *n*-pt cluster
- Expansion coefficients J_k: ECI
  - ***Effective cluster interactions***
- Clusters similar under space group symmetries
  - ***Same ECI***
  - *"High symmetry: few unknowns"*

Cluster Expansions: J. Sanchez, F. Ducastelle, and D. Gratias, Physica A (1984)

Can Information Theory Improve Thermodynamic Alloy Modeling with Surrogates?

J. Kristensen, I. Bilionis, and N. Zabaras. Physical Review B 87.17 (2013)

Common Methodology

$$\mathcal{D} = \{\boldsymbol{\sigma}^{(i)}, E^{(i)}\}_{i=1}^{N}$$

- Expensive data set

$$\mathcal{L}(\boldsymbol{\gamma}) = ||\mathbf{E}^{(1:N)} - \Phi\boldsymbol{\gamma}||^2$$

- Much-used approach: Least squares

Design matrix

ECI

Can we do better if the objective is to obtain the ground states?

Fitting the Boltzmann Distribution

$$Z = \sum_{\boldsymbol{\sigma}} \exp(-\beta E(\boldsymbol{\sigma}))$$

- Partition function

σ: Configurational states of the system

Ab initio energy: expensive

$$p(\boldsymbol{\sigma}) = \frac{\exp(-\beta E(\boldsymbol{\sigma}))}{Z}$$

Replace Boltzmann with surrogate distribution

$$p(\boldsymbol{\sigma}|\boldsymbol{\gamma}) = \frac{\exp(-\beta E(\boldsymbol{\sigma};\boldsymbol{\gamma}))}{Z(\boldsymbol{\gamma})}$$

$$E(\boldsymbol{\sigma};\gamma) = \boldsymbol{\gamma}^T \phi(\boldsymbol{\sigma})$$

ECI

We aim to <u>match distributions</u> rather than energies!

Quantify Information Loss

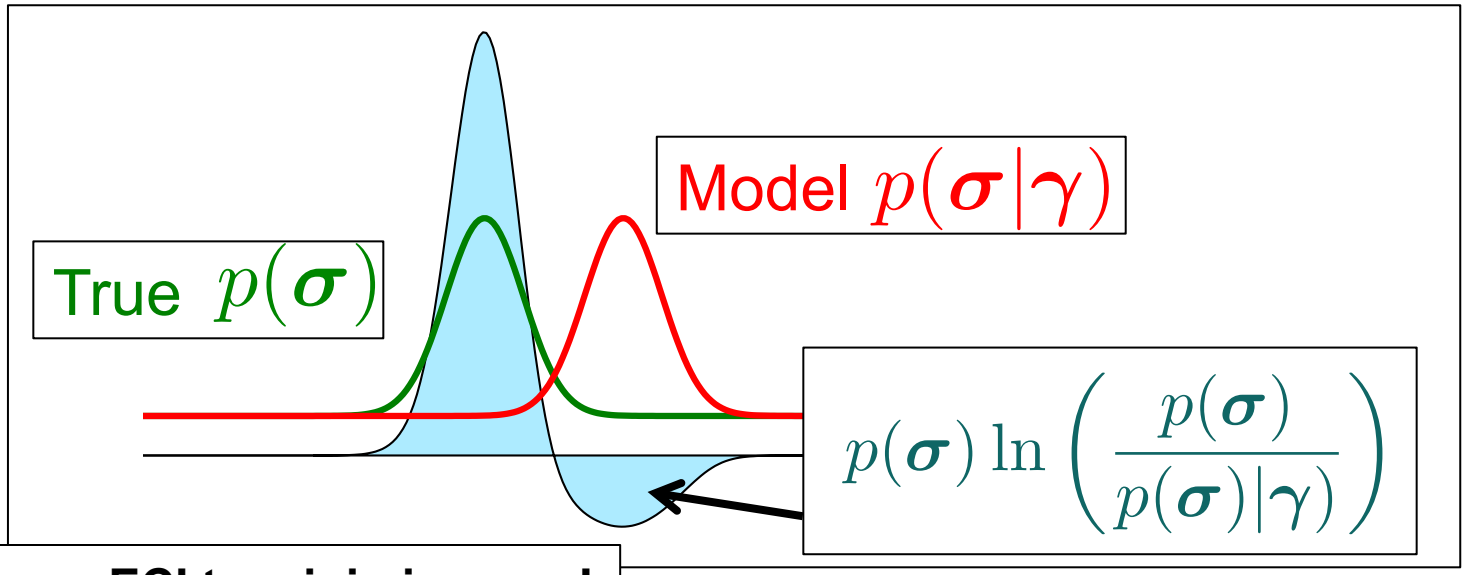
$$p(\boldsymbol{\sigma}) = \frac{\exp(-\beta E(\boldsymbol{\sigma}))}{Z}$$

$$p(\boldsymbol{\sigma}|\boldsymbol{\gamma}) = \frac{\exp(-\beta E(\boldsymbol{\sigma};\boldsymbol{\gamma}))}{Z(\boldsymbol{\gamma})}$$

"Distance" of choice:

$$S[\boldsymbol{\gamma}] = \int_{\mathcal{M}} p(\boldsymbol{\sigma}) \ln\left(\frac{p(\boldsymbol{\sigma})}{p(\boldsymbol{\sigma}|\boldsymbol{\gamma})}\right) \mathrm{d}\boldsymbol{\sigma} \geq 0$$

Relative Entropy (Rel Ent)



Choose ECI to minimize area!

Rel Ent Vs. Least Squares

Gaussian approximation of $S[\gamma] = \int_{\mathcal{M}} p(\boldsymbol{\sigma}) \ln\left(\frac{p(\boldsymbol{\sigma})}{p(\boldsymbol{\sigma}|\gamma)}\right) \mathrm{d}\boldsymbol{\sigma}$

We ideally minimize $$S[\gamma] \approx \frac{\beta^2}{2}\mathrm{Var}[E(\boldsymbol{\sigma}) - E(\boldsymbol{\sigma};\gamma)]$$

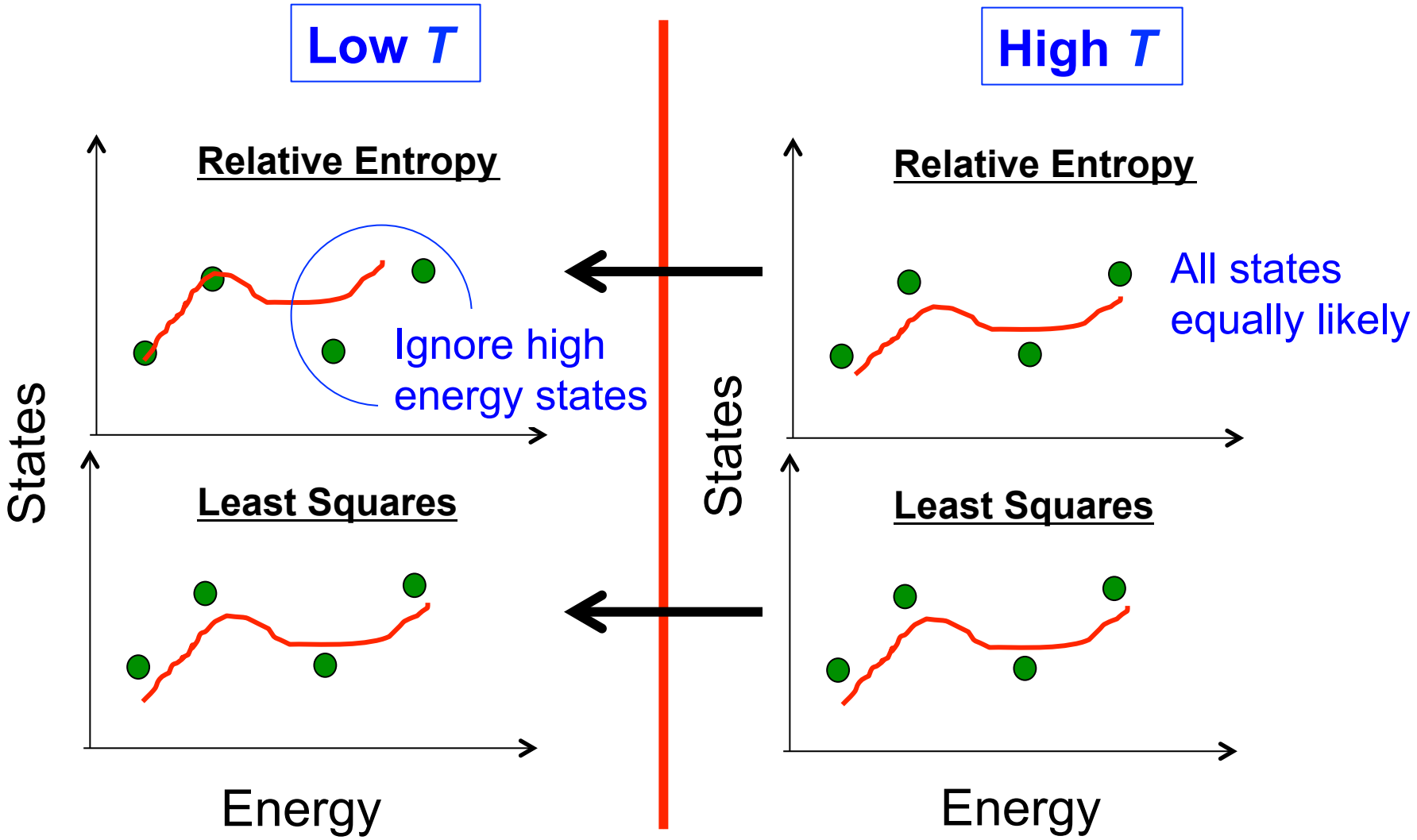
Least squares ideally minimizes $$\mathcal{L}[\gamma] = \sum_{\boldsymbol{\sigma}} (E(\boldsymbol{\sigma}) - E(\boldsymbol{\sigma};\gamma))^2$$

Matching distributions becomes a weighted least squares problem (from minimizing *S* above) with weights

$$(\mathbf{I}_N - \mathbf{p}_N \mathbf{1}_N)\mathrm{diag}(\mathbf{p}_N)(\mathbf{I}_N - \mathbf{p}_N \mathbf{1}_N)^t$$ *(t = transpose)*

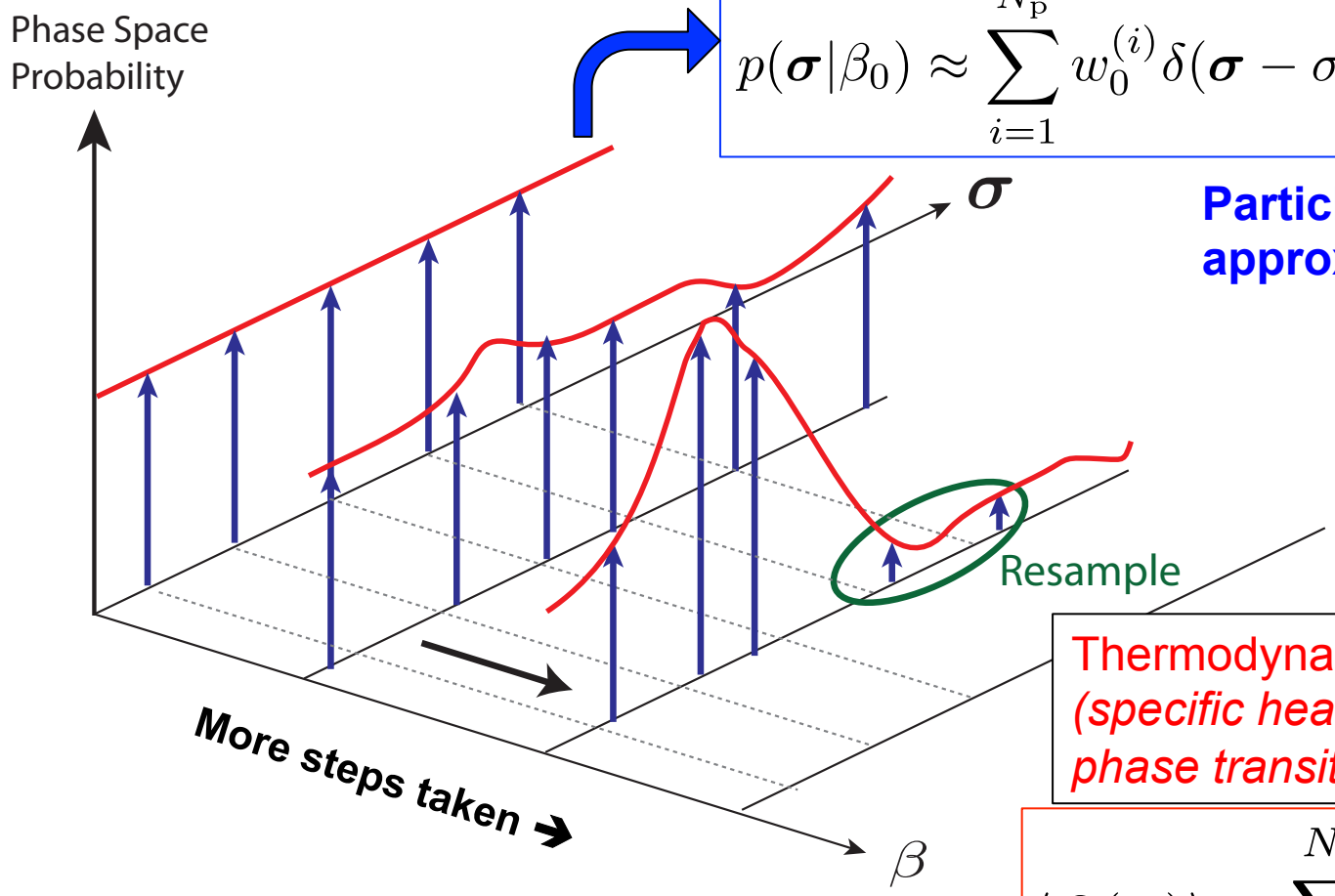
$$\mathbf{p}_N := \left(\frac{\exp(-\beta E(\boldsymbol{\sigma}^{(1)}))}{\mathbf{Z}_N}, \cdots, \frac{\exp(-\beta E(\boldsymbol{\sigma}^{(N)}))}{\mathbf{Z}_N}\right)$$

Rel Ent Behavior



How to Compute Phase Transitions

THERMODYNAMICS USING MCMC



Thermodynamic quantity Q *(specific heat in case of phase transition)*

$$\langle Q(T) \rangle = \sum_{j=1}^{N_\mathrm{p}} w_s^{(j)} Q(\boldsymbol{\sigma}_s^{(j)})$$

(step s)

Fully parallelizable:
Each particle on its own core

ASMC ALGORITHM: IMPLEMENTATION

1. Set $s = 0$ and $\gamma_s = 0$. Sample a particle approximation from $p_{\gamma_0}(\sigma)$, $\left\{w_0^{(i)}, \sigma_0^{(i)}\right\}_{i=1}^{N}$.

Adaptive step size according to how much distribution changes

2. Determine $\gamma_{s+1} \in [\gamma_s, 1]$ s.t:

$$ESS(\gamma_{s+1}) = \frac{1}{\sum_{i=1}^{N}\left(w_{s+1}^{(i)}\right)^2} = \zeta ESS(\gamma_s),$$

where $w_{s+1}^{(i)}$ is the **normalized** version of $W_{s+1}^{(i)} = w_s^{(i)}\, \widehat{w}_{s+1}^{(i)}$ with:

$$\widehat{w}_{s+1}^{(i)} = \frac{p_{\gamma_{s+1}}\left(\sigma_s^{(i)}\right)}{p_{\gamma_s}\left(\sigma_s^{(i)}\right)} = e^{-(\beta_{\gamma_{s+1}} - \beta_{\gamma_s})E_s^{(i)} + (\beta_{\gamma_{s+1}}\mu_{\gamma_{s+1}} - \beta_{\gamma_s}\mu_{\gamma_s})\sum_j \sigma_{s,j}^{(i)}}.$$

3. If $ESS(\gamma_{s+1}) < ESS_m$, then resample.

Threshold: Re-locate particles

4. Draw samples $\left\{\sigma_{s+1}^{i}\right\}_{i=1}^{N}$ from $p_{\gamma_{s+1}}(\sigma)$.

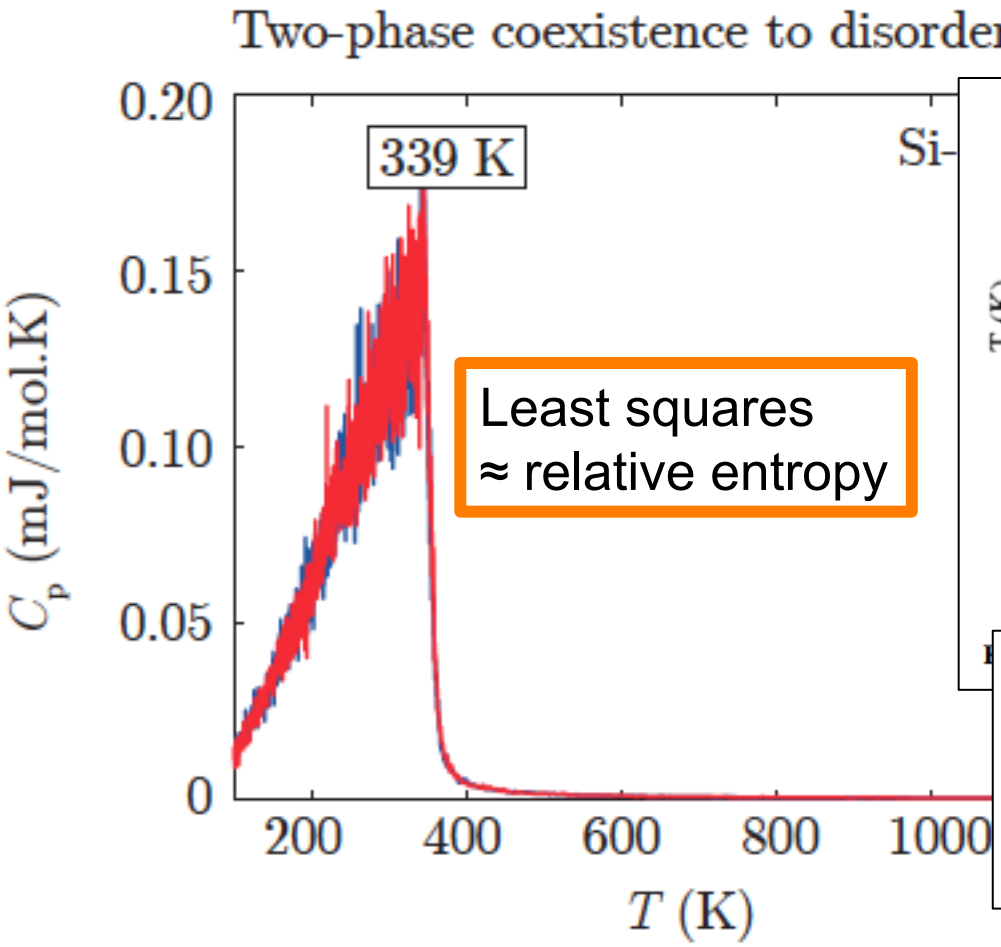
5. If $\gamma_s = 1$, STOP. Otherwise, set $s \leftarrow s + 1$ and go to 2.

Case Study: Silicon Germanium

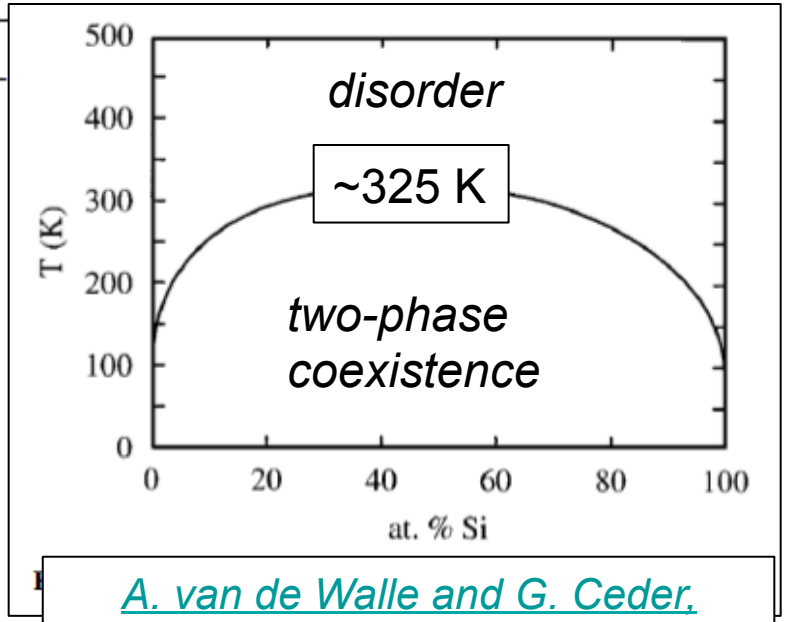
Predict two-phase coexistence to disorder phase transition at 50 % composition

50 % Si-Ge: Fit to Energies

- Use ASMC to obtain phase transitions



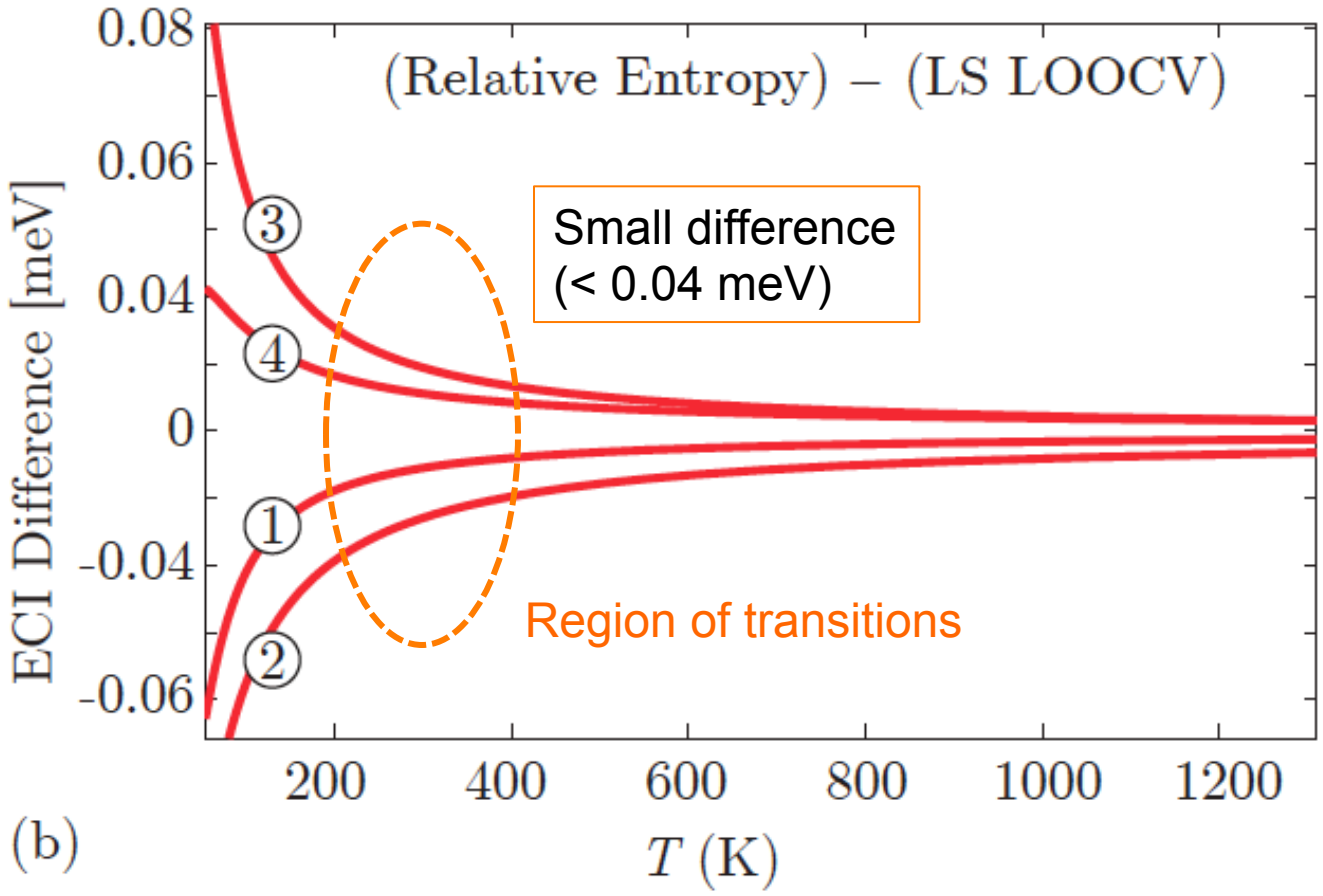
Existing literature



A. van de Walle and G. Ceder,
J. Phase Equilibria 23, 348 (2002)
Method:
Least Squares (ATAT)
+traditional MCMC

SILICON GERMANIUM

- Why the similarity?



Note: This is not probabilistic, but gives an idea of the behavior versus temperature

Case Study: Magnesium Lithium

Predict order/disorder phase transitions at 33 %, 50 %, and 66 % Mg

Magnesium Lithium

R. Taylor, S. Curtarolo, and G. Hart, Phys. Rev. B (2010)

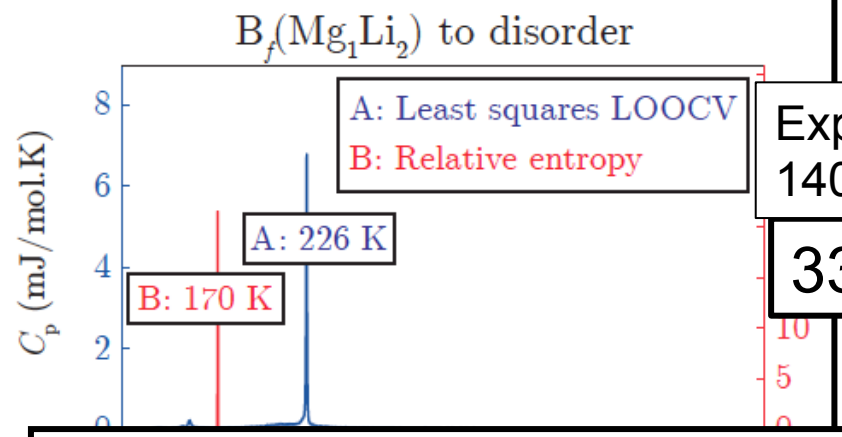
33 % Mg composition: ~190 K
50 % Mg composition: ~300-450 K
66 % Mg composition: ~210 K

Fitting observed energies (cluster expansion)

C. Barrett and O. Trautz. Trans. Am. Inst. 175 (1948)

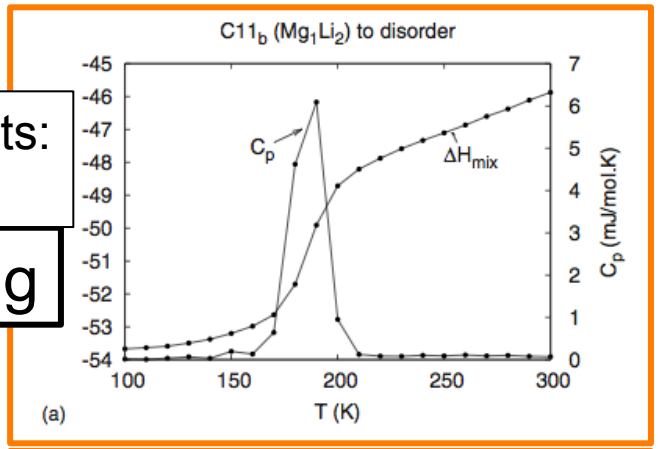
All compositions: ~140-200 K

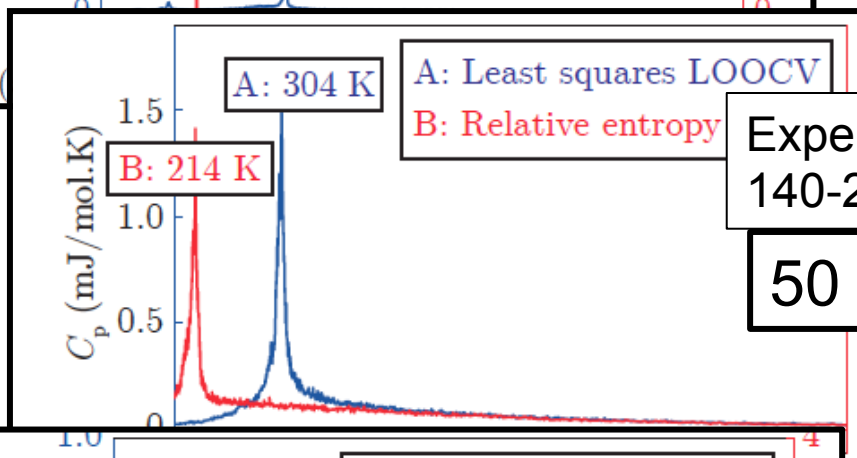
Experimental (non-conclusive)



Experiments:
140-200 K

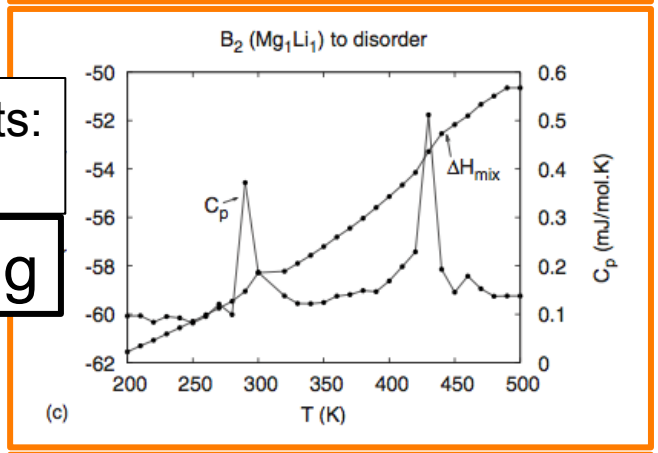
33 % Mg

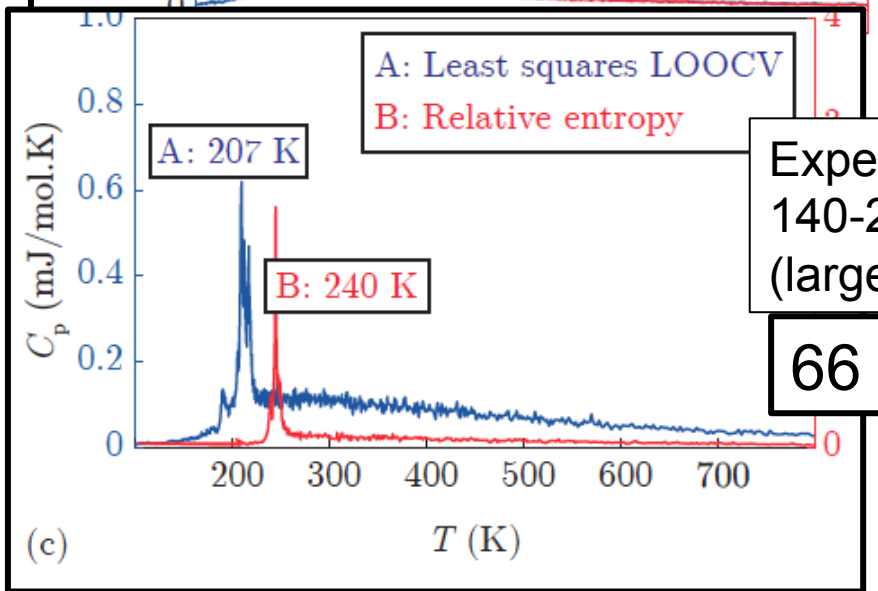




Experiments:
140-200 K

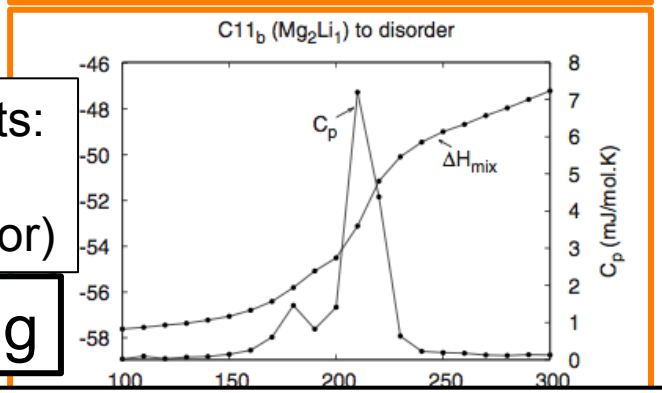
50 % Mg





Experiments:
140-200 K
(largest error)

66 % Mg



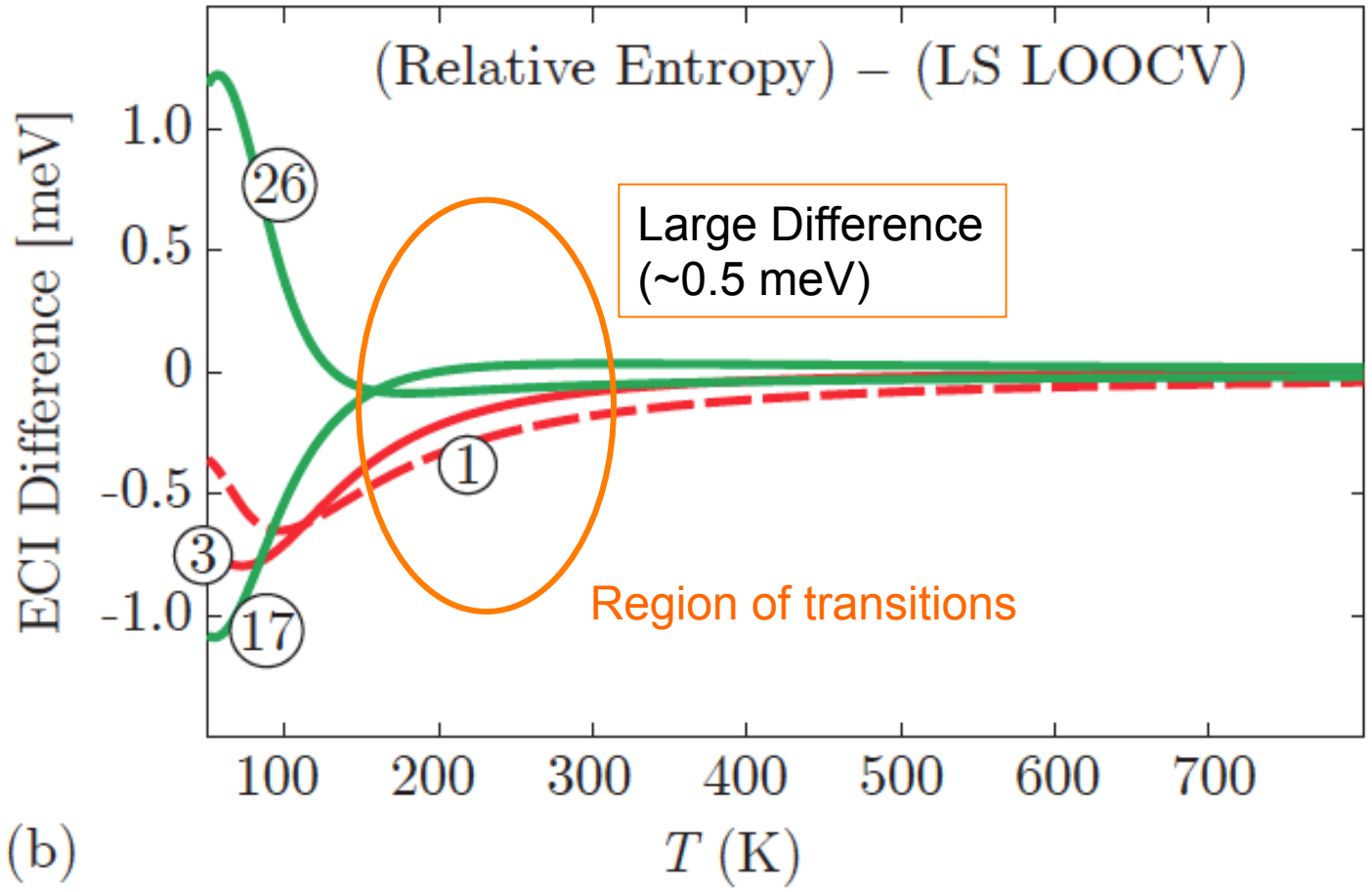
R. Taylor et al. Phys. Rev. B. 81 (2010)

Method:

Genetic Algorithm+MCMC

MAGNESIUM LITHIUM

- Why the difference?



Relative Entropy Conclusions

- Summary table

Alloy	x	Fit energies (our work)	Fit energies (literature)	Relative Entropy	Experiment
Si_xGe_{1-x}	50 %	~339 K	~325 K	**~339 K**	*N/A*
Mg_xLi_{1-x}	33 %	~226 K	~190 K	**~170 K**	~140-200 K
	50 %	~304 K	~300-450 K	**~214 K**	~140-200 K
	66 %	~207 K	~210 K	**~240 K**	~140-200 K (*largest error*)

Bayesian Approach to Predicting Materials Properties

Propagating Uncertainty From a Surrogate To, e.g., a Phase Transition

Fully Bayesian Approach

- The former part of this presentation does not *per se* offer ways of answering central questions such as:
  - **What is the uncertainty in a quantity of interest (e.g., a phase transition) given that we do not know the best cluster expansion and that we have limited data?**
- When computing a phase transition we want to know how uncertain we are about its value
  - Unknown whether what we predict is OK
- The Bayesian approach can provide answers to such questions
- We show how **surrogate models** can be used to accomplish this

Fully Bayesian Approach

- Probability means a reasonable degree of belief*
- Prior belief on clusters + ECI
- Likelihood function
  - Given a model (i.e., set of clusters + ECI) how likely is *D*
- Posterior belief on clusters + ECI
- Use Bayes theorem** to update degree of belief upon receiving new evidence *D* (what *D* is depends on the application)

Note:
D is limited, we can only see so many observations

$$p(\gamma|D) = \frac{p(D|\gamma)p(\gamma)}{p(D)}$$

*Laplace, Analytical Theory of Probability (1812)
**Bayes, Thomas. Philosophical Transactions (1763)

Propagating Uncertainty

Prior on property (quantity of interest) *"I"*

Prior on cluster and ECI *"θ"*

$$p(I) = \int \mathrm{d}\theta \delta(I[f(\cdot;\theta)] - I) p(\theta)$$

Notice how we integrate out the clusters and the ECI!
(In principle) all cluster and ECI choices (models) consistent with *D* are considered

- Then we observe an expensive data set *D* which helps us to learn more about the clusters and the ECI
  - In this work the data set was expensive energy computations

$$\mathcal{D} = \{x_i, f(x_i)\}$$

$$\mathcal{L}(\mathcal{D}|\theta, \cdot)$$

Likelihood: What information the data contains about the clusters and ECI

The Quantity of Interest

- Different quantities of interest *I* can require different data sets *D*
- This framework allows for very general quantities of interest *I*

- Some examples:

  - *I* = phase transition
    - The phase transition is found from the internal energy
    - The data set *D* consists of high-accuracy (expensive) energies

  - *I* = ground state line
    - The ground state line is found from the internal energy as well
    - The data set *D* consists of expensive energies

  - *I* = maximum band gap structure
    - The maximum band gap structure is found, e.g., from knowing the band gap of each structure or the entire band diagram
    - The data set *D* consists of expensive band gaps

Propagating Uncertainty In Surrogate

Posterior on truncation and ECI

$$p(\theta|\mathcal{D}) \propto \mathcal{L}(\mathcal{D}|\theta,\cdot)p(\theta)$$

How likely is the cluster expansion upon seeing *D* including what we knew before? (Bayes theorem!)

Posterior on property

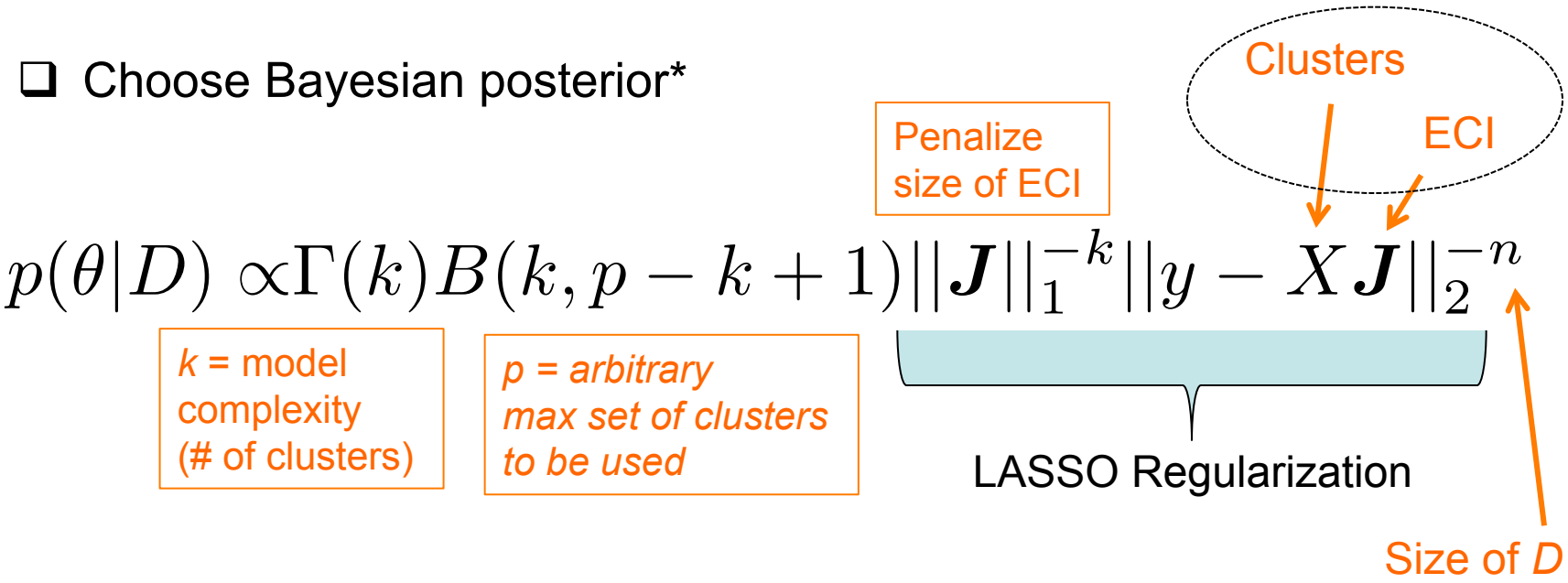
$$p(I|\mathcal{D}) = \int \mathrm{d}\theta \delta(I[f(\cdot;\theta)] - I) p(\theta|\mathcal{D})$$

- We stress here that we now have a probability distribution on the property—not a single estimate
  - From this, the uncertainty estimate follows
- We now have a way to **propagate uncertainty** from the cluster expansion to the quantity of interest

- Next, we select a Bayesian posterior

Bayesian Posterior using a Surrogate

- Choose Bayesian posterior*



- Based on LASSO-inspired priors
  - Models describing physics are typically sparse**
- Expectation values with this posterior not in closed form!
  - Resolution: MCMC Sampling

*C. Xiaohui, J. Wang, and M. McKeown (2011)

**L. Nelson et al. Physical Review B 87.3 (2013)

Motivating Model Selection

$$p(I|\mathcal{D}) = \int \mathrm{d}\theta \delta(I[f(\cdot;\theta)] - I) p(\theta|\mathcal{D})$$

- There is an infinite number of cluster expansions (each symbolized by its own θ in the integral above)
  - Which ones are ***most relevant*** to determining the value of the integral?
- **We now explore model selection as an option**

Can Model Selection be Used to Quantify Epistemic Uncertainties With Limited Data?

J. Kristensen and N. Zabaras. Computer Physics Communications 185.11 (2014)

Model Selection

Selection of both basis functions and expansion coefficients

Reversible Jump Markov Chain Monte Carlo

- We used reversible jump Markov chain Monte Carlo (RJMCMC)* to perform the model selection

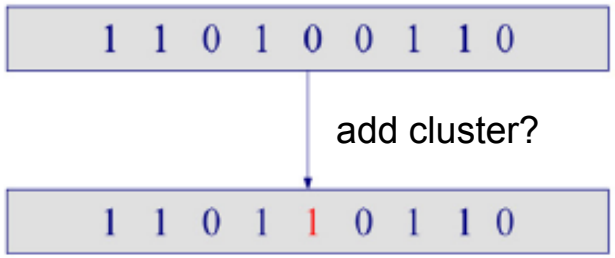
- Define 3 move types:

  Birth step (+1)

  Death step (-1)

  Update step (0)

Basis set (clusters) represented as a binary string (ECI not shown)



(Birth step)

- Practically speaking it behaves like a standard MCMC chain
  - Use 50 % burn-in
  - Use thinning if you want to (for memory reasons, e.g.)

*P. Green. Biometrika 82.4 (1995)

RJMCMC CHAIN: ALGORITHM

Input: The number of iterations T. Random walk step size ε.
Data: X and $\mathbf{y}$.
Output: $\{\boldsymbol{\theta}^{(t)} = (\boldsymbol{\beta}^{(t)}, \boldsymbol{\gamma}^{(t)}) | t \in \{0, \ldots, T\}\}$.

begin
  Initialization : set $\boldsymbol{\theta}^{(0)} = (\boldsymbol{\beta}^{(0)}, \boldsymbol{\gamma}^{(0)})$ and $t = 1$.
  repeat
    if $k^{(t-1)} = 1$ **then**
      $k^{(t)} \leftarrow k^{(t-1)} + U(\{0,1\})$.
    else if $k^{(t-1)} = p$ **then**
      $k^{(t)} \leftarrow k^{(t-1)} - U(\{0,1\})$.
    else
      $k^{(t)} \leftarrow k^{(t-1)} + U(\{-1,0,1\})$.
    end
    Sample $s \sim N(0, \varepsilon^2)$.
    $K \leftarrow \gamma^{(t-1)}$ and $K^c \leftarrow \{1, \ldots, p\} \backslash K$.
    if $k^{(t-1)} = k^{(t)}$ **then**
      Sample $j \sim U(K)$.
      Update $\beta_j^{(t)} \leftarrow \beta_j^{(t-1)} + s$ with an MH step, details in Section 3.2.
    else if $k^{(t)} = k^{(t-1)} + 1$ **then**
      Sample $j \sim U(K^c)$.
      Perform a " birth" move and update $\beta_j^{(t-1)}$, details in Section 3.3.
    else
      Sample $j \sim U(K)$.
      Perform a " death" move and update $\beta_j^{(t-1)}$, details in Section 3.3.
    end
    $t \leftarrow t + 1$.
  until $t = T$.
end

Initialization

Model selection of clusters

Model selection of the ECI

Standard Metropolis-Hastings*

$$\min\left\{\left(\frac{\|\boldsymbol{\beta}'\|_1}{\|\boldsymbol{\beta}\|_1}\right)^{-k} \times \left(\frac{\|\mathbf{y}-X\boldsymbol{\beta}'\|_2}{\|\mathbf{y}-X\boldsymbol{\beta}\|_2}\right)^{-(n-1)}, 1\right\}.$$

$$\min\left\{\frac{k^2}{p-k} \times \frac{\|\boldsymbol{\beta}'\|_1^{-(k+1)}\|\mathbf{y}-X\boldsymbol{\beta}'\|_2^{-(n-1)}}{\|\boldsymbol{\beta}\|_1^{-k}\|\mathbf{y}-X\boldsymbol{\beta}\|_2^{-(n-1)}} \times \frac{p(\boldsymbol{\gamma}' \to \boldsymbol{\gamma})}{p(\boldsymbol{\gamma} \to \boldsymbol{\gamma}')}, \times N(u; 0, \varepsilon^2)^{-1}, 1\right\}. \quad (10)$$

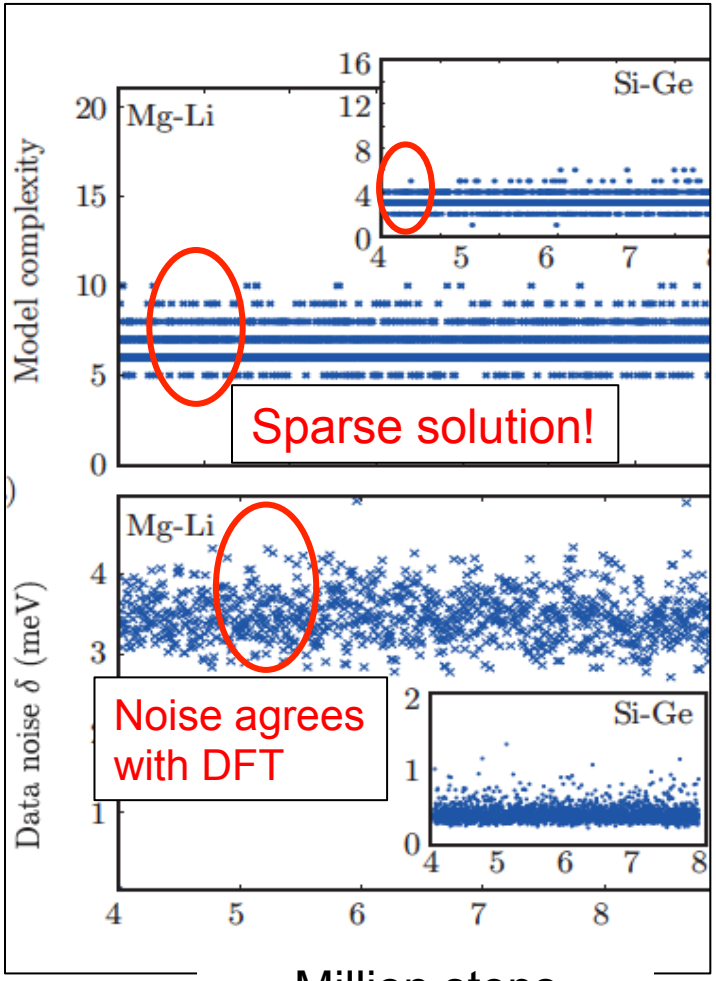
$$\min\left\{\frac{p-(k-1)}{(k-1)^2} \times \frac{\|\boldsymbol{\beta}'\|_1^{-(k-1)}\|\mathbf{y}-X\boldsymbol{\beta}'\|_2^{-(n-1)}}{\|\boldsymbol{\beta}\|_1^{-k}\|\mathbf{y}-X\boldsymbol{\beta}\|_2^{-(n-1)}} \times \frac{p(\boldsymbol{\gamma}' \to \boldsymbol{\gamma})}{p(\boldsymbol{\gamma} \to \boldsymbol{\gamma}')}, \times N(u; 0, \varepsilon^2), 1\right\}. \quad (11)$$

*N. Metropolis, et al. The journal of chemical physics 21.6 (1953)

Results on Real Alloys

Model Selection Results

$$p(\theta|D) \propto \Gamma(k) B(k, p-k+1) ||\boldsymbol{J}||_1^{-k} ||y - X\boldsymbol{J}||_2^{-n}$$ **+ RJMCMC**



Any particular blue point in upper plot represents a cluster expansion truncation:

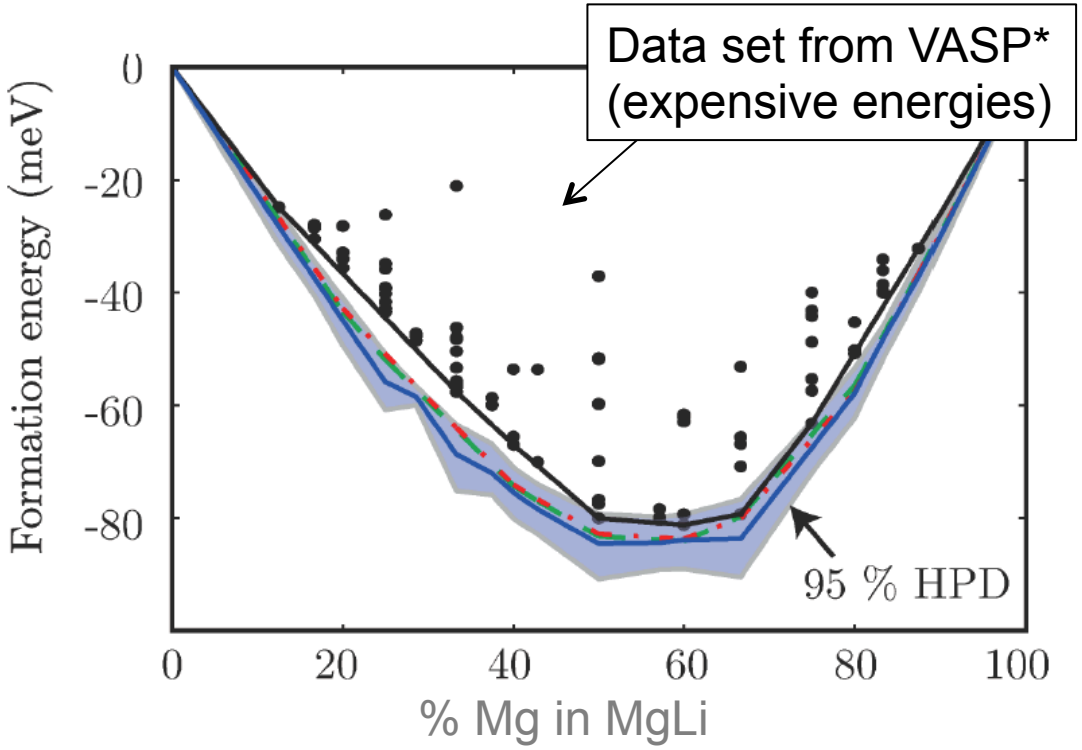
1) *y*-axis measures number of included clusters (but not which).
2) Actual values of ECI not shown

Ground State Line Uncertainty

- Bayesian uncertainty in ground state line with limited data

$$p(I|D,\cdot) = \int \mathrm{d}\theta \delta(I[f(\cdot;\theta)] - I) p(\theta|D,\cdot)$$

I = ground state line



Material: MgLi

Predictive variance of ground state line is around 12 %

We can implicitly conclude whether the data set is large enough!

- **This is the uncertainty induced in the quantity of interest from the uncertainty in the surrogate model**
  - If error bars too large: you need to increase/change your data set!

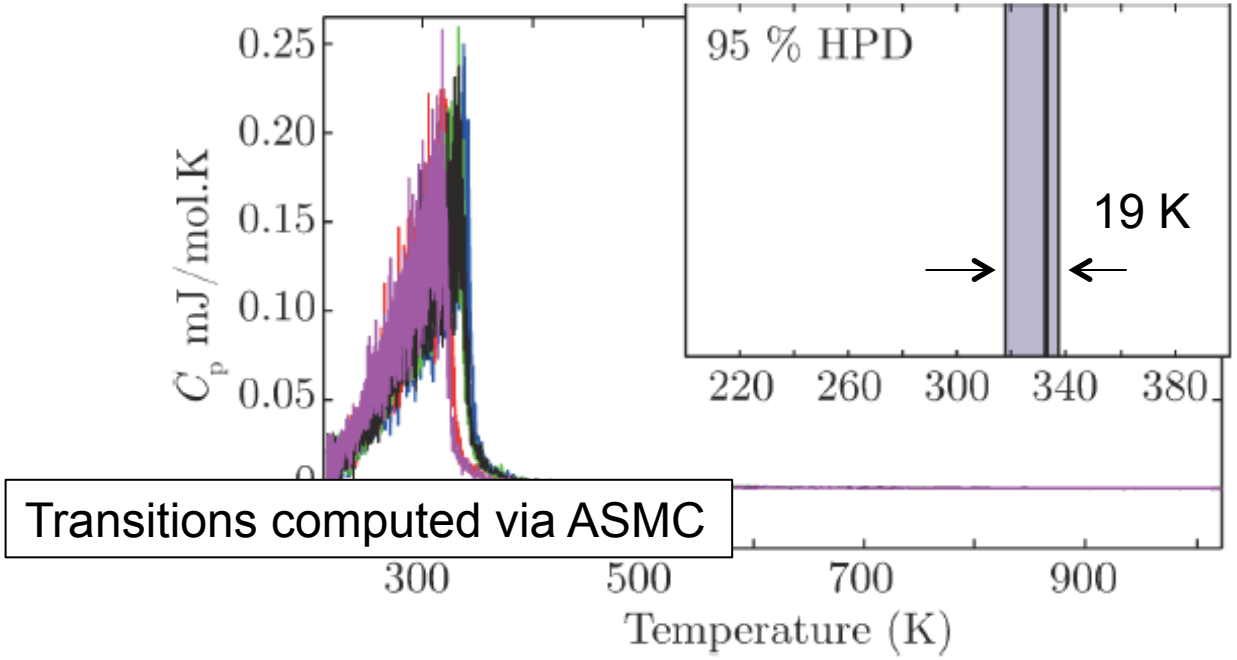
*G. Kresse and J. Hafner. Physical Review B 47.1 (1993)

PHASE TRANSITION UNCERTAINTY

- Bayesian uncertainty in phase transition from two-phase coexistence to disorder with limited data

$$p(I|D,\cdot) = \int \mathrm{d}\theta \delta(I[f(\cdot;\theta)] - I) p(\theta|D,\cdot)$$

I = two-phase coexistence to disorder phase transition



Transitions computed via ASMC

Material: SiGe at 50 %

Predictive variance of phase transition is around 6 %

- **This is the uncertainty induced in the quantity of interest from the uncertainty in the surrogate model**
  - If error bars too large: you need to increase/change your data set!

Using Surrogates for Designing Materials

J. Kristensen and N. Zabaras. In review (2014)

MATERIALS BY DESIGN

- We are now confident about the predictive capabilities of surrogate models
- Can we also use surrogates for designing new structures with specified properties?

  - **Materials by design**

- **Application:**
  - Optimize thermal conductivity in nanowires*
    - Heat dissipation in nanochips
    - Thermoelectric materials
      - Solar cells
      - Refrigeration

  - **But:** Nanowires require a different way of using the cluster expansion
    - by "the cluster expansion" we mean the standard bulk expansion implemented in, e.g., ATAT**
    - We show shortly how we addressed this issue

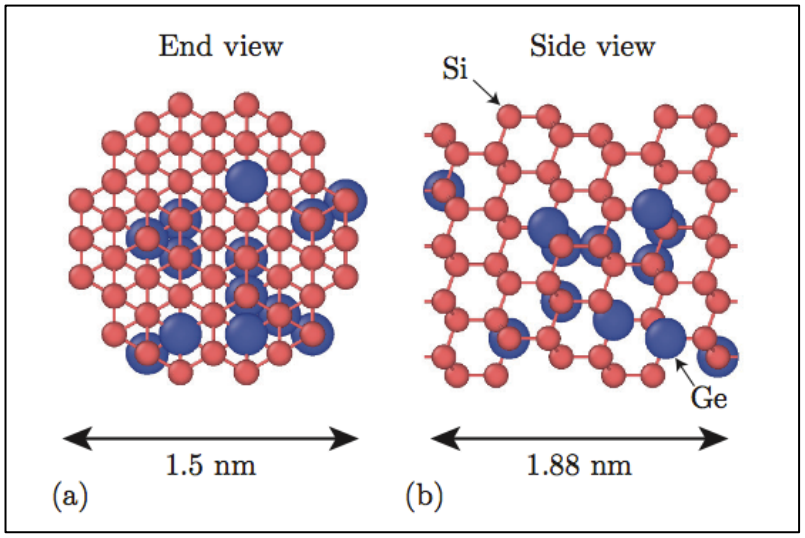
N. Mingo et al. Nano Letters 3.12 (2003)

**A. Walle, M. Asta, and G. Ceder. Calphad 26.4 (2002)*

Which Si-Ge Nanowire Configuration Minimizes the Thermal Conductivity?

DESIGN GOAL

- Find the configuration with **lowest thermal conductivity**



Green-Kubo method:*

$$\kappa = \frac{1}{V k_B T^2} \lim_{\tau_m \to \infty} \int_0^{\tau_m} \langle J_x(\tau) J_x(0) \rangle \mathrm{d}\tau$$

using microscopic heat current:

$$\mathbf{J} = \sum_{i=1}^{N} e_i \boldsymbol{v}_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \left(\frac{\partial V_{ij}}{\partial \boldsymbol{r}_j} \cdot \boldsymbol{v}_j \right) \boldsymbol{r}_{ij},$$

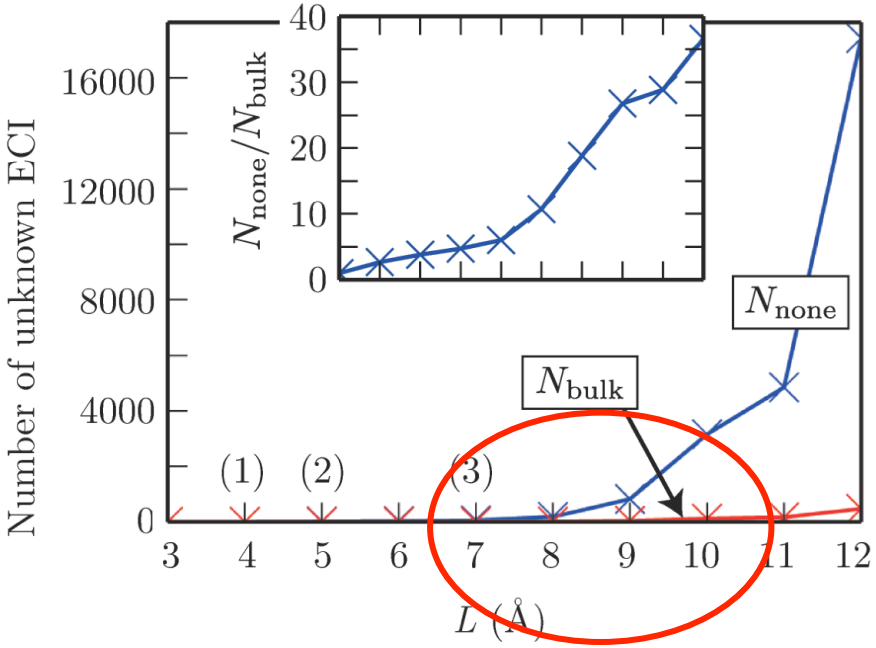
and a Tersoff** potential energy b/w bonds:

$$V_{ij} = f_C(\boldsymbol{r}_{ij}) \left[a_{ij} f_R(\boldsymbol{r}_{ij}) + b_{ij} f_A(\boldsymbol{r}_{ij}) \right]$$

*R. Kubo Journal of the Physical Society of Japan 12.6 (1957)
*M. Green The Journal of Chemical Physics 20.8 (1952)
**J. Tersoff Physical Review B 39.8 (1989)

Nanowire Challenge: Low Symmetry

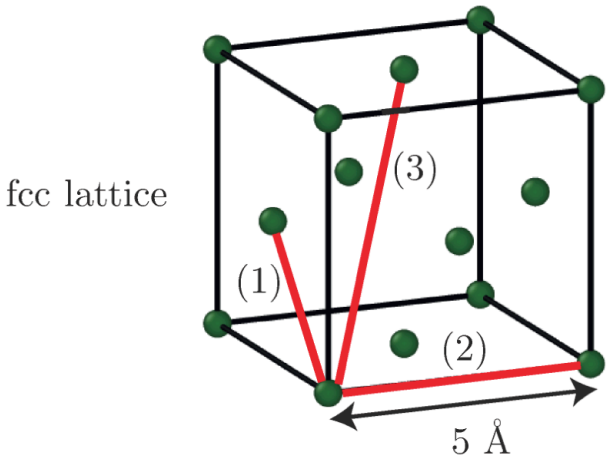
- Alloy optimization problem
  - Use cluster expansion surrogate
- **Problem for nanowires:**
  - Low-symmetry system
  - ECI become layer-dependent close to surfaces
  - Easily *thousands of unknowns*!



- Energy is additive so we can write*:

$$\Delta H_f^{\text{CE}} = \Delta H_f^{\text{vol}} + \Delta H_f^{\text{surf}}$$

Bulk part Surface part



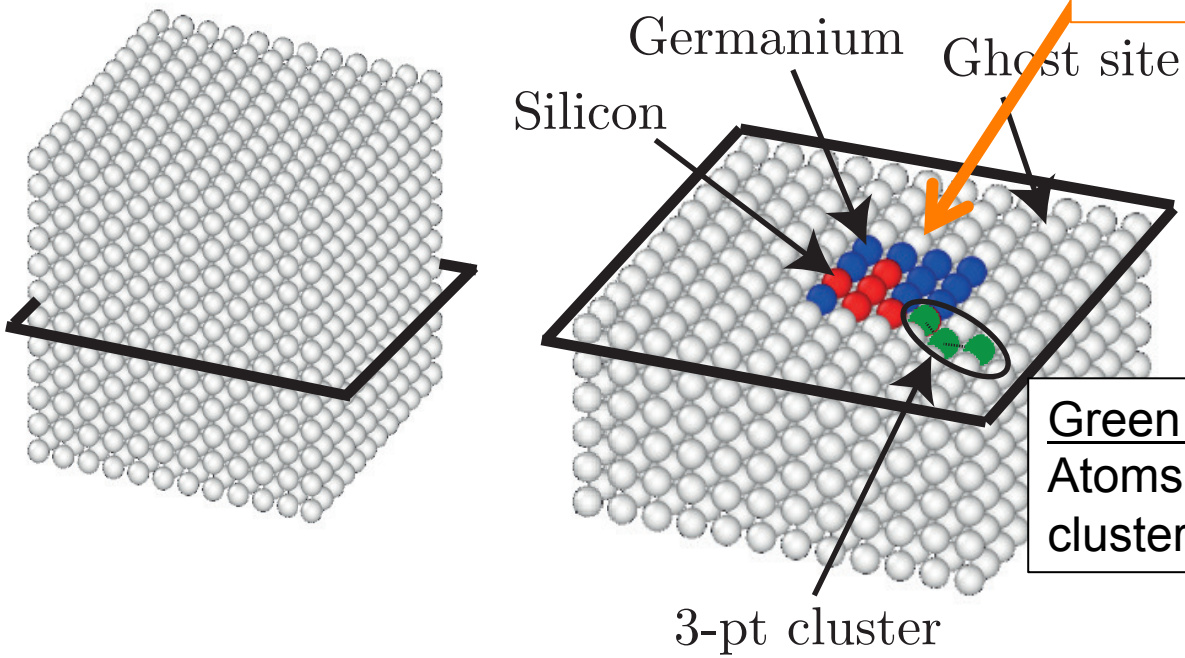
*D. Lerch et al. Modelling and Simulation in Materials Science and Engineering 17.5 (2009)

NEW CLUSTER EXPANSION APPROACH

- **Idea:** embed structure of *any* geometry in ***sea of ghost sites***
- Group clusters under ***bulk symmetries***
  - We get bulk contribution alone, but for ***any*** geometry!

$$\Delta H_f^{\mathrm{CE}} = \Delta H_f^{\mathrm{vol}} + \Delta H_f^{\mathrm{surf}}$$

Could be any shape on any lattice: bcc sphere, fcc nanowire, sc 2D sheets, etc.

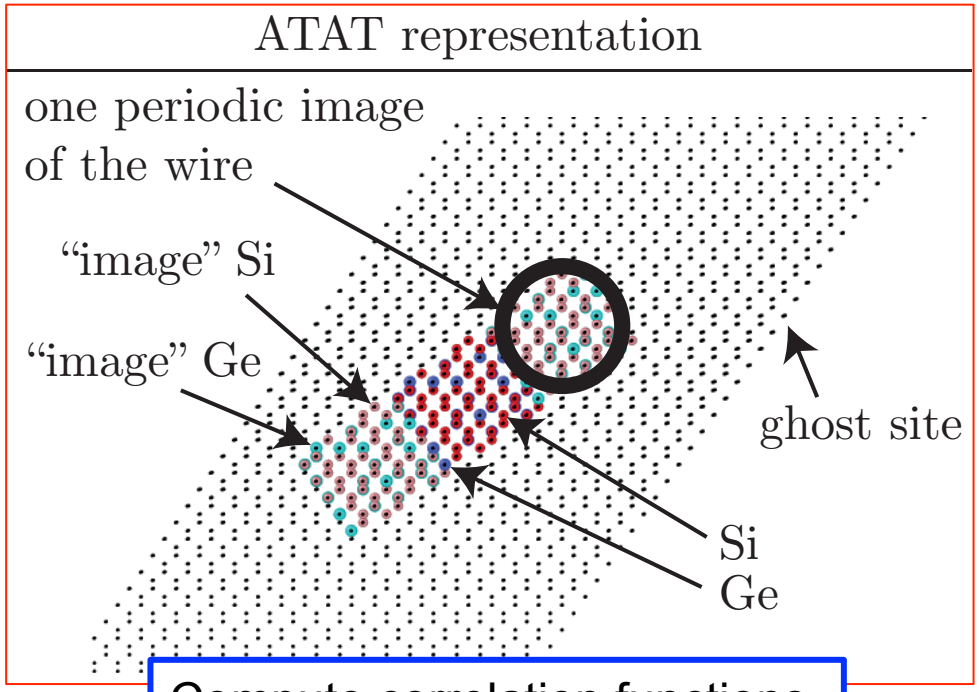


Green atoms: Atoms part of 3-pt cluster (visual aid)

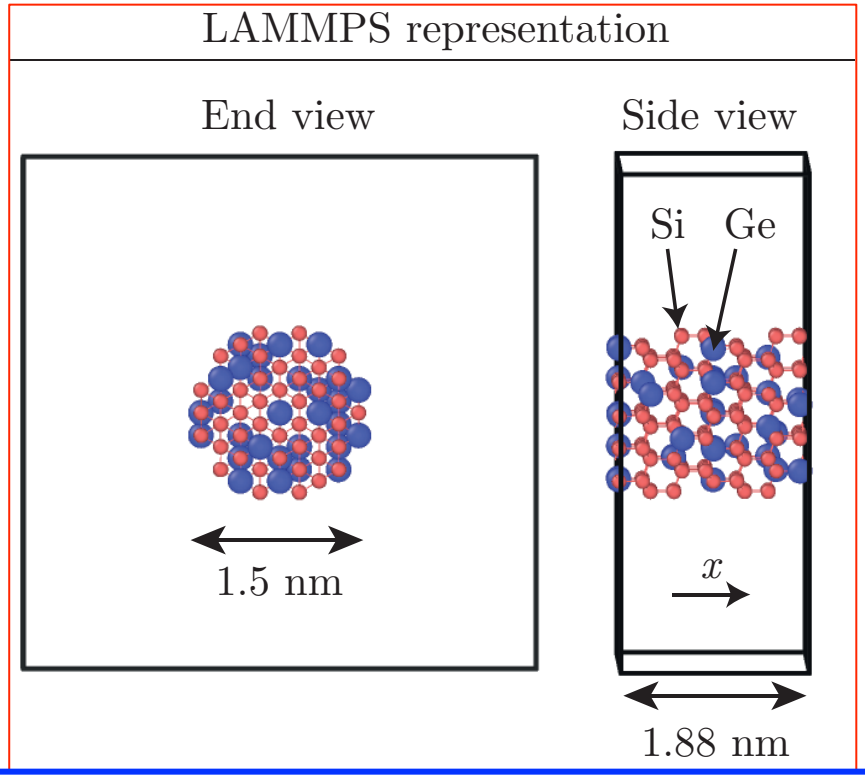
Ghost lattice method (GLM)

GLM ON NANOWIRE PROJECT

- Nanowire implementation with the GLM
  - Two different representations of the same wire (OVITO* used for visualization)



Compute correlation functions with ATAT modified for GLM (i.e., modified to parse ghosts)

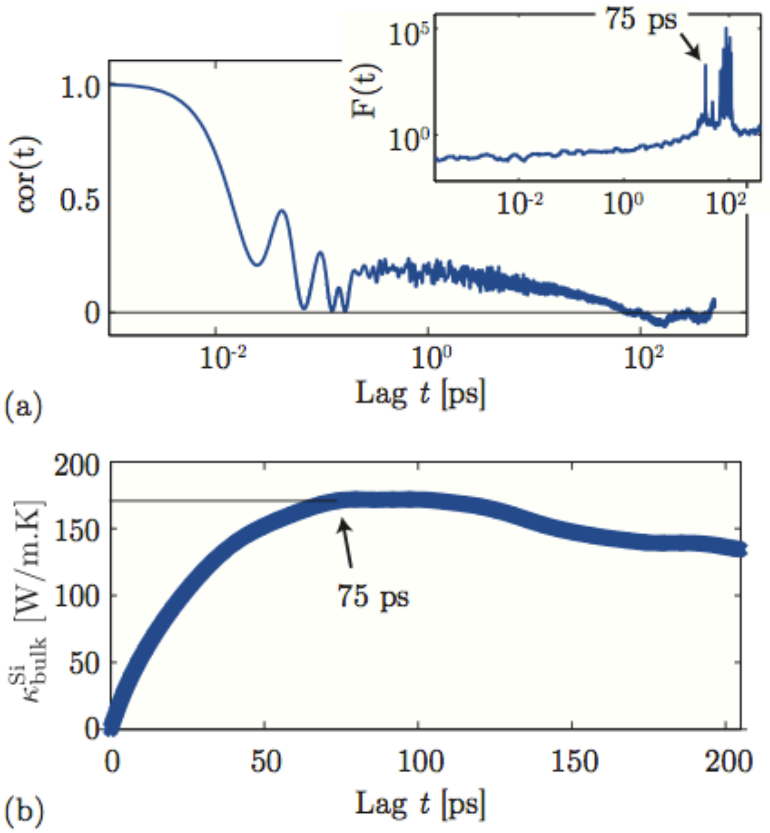


Compute thermal conductivity in LAMMPS

*A. Stukowski. Modelling and Simulation in Materials Science and Engineering 18.1 (2010)

VERIFY GENERAL LAMMPS IMPLEMENTATION

- **Bulk** Si and Ge (easy case): Use method in Ref. [*] in LAMMPS**



We predict 170 W/m.K for Silicon.
Experimental value = 150 W/m.K.

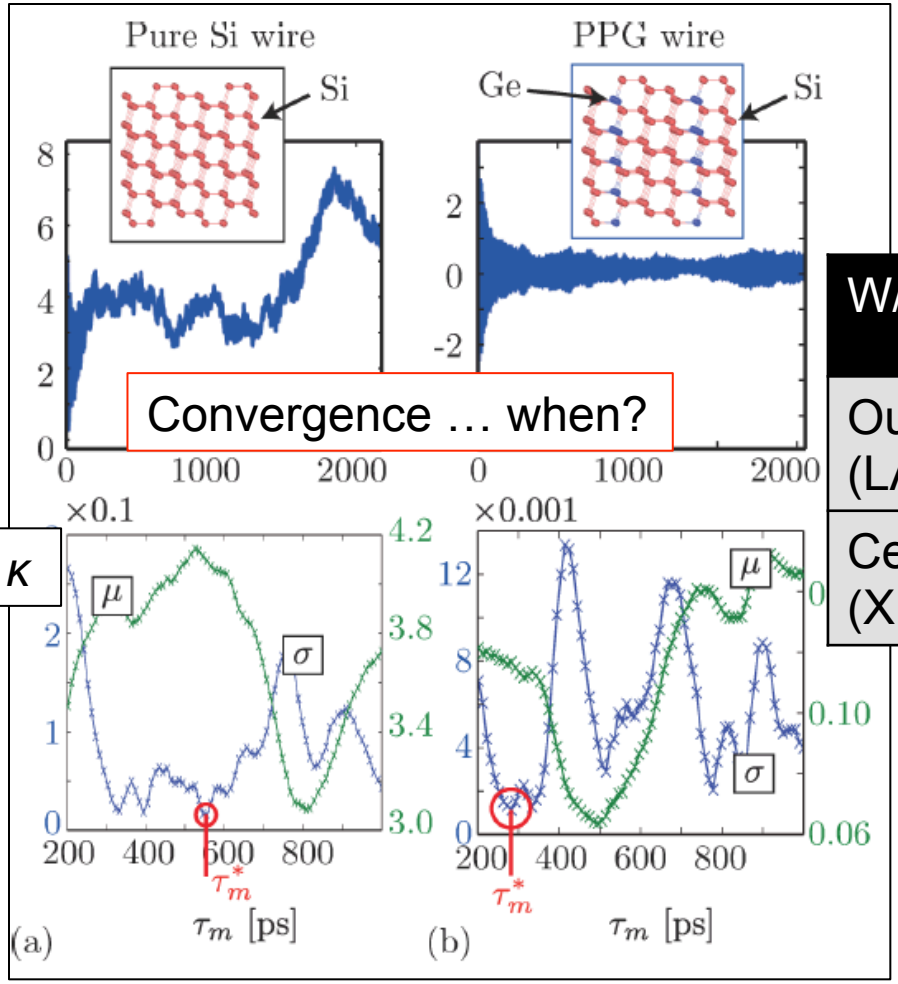
We predict 90 W/m.K for Germanium.
Experimental value is 60 W/m.K.

Tersoff is known to overshoot.
We obtain great agreement!

*J. Chen, G. Zhang, and B. Li. Physics Letters A 374.23 (2010)
**S. Plimpton. Journal of computational physics 117.1 (1995)

VERIFY NANOWIRE IMPLEMENTATION

- Compare data with Ceder's group at MIT*



Annealed heating of the (relatively large) surface area was necessary.

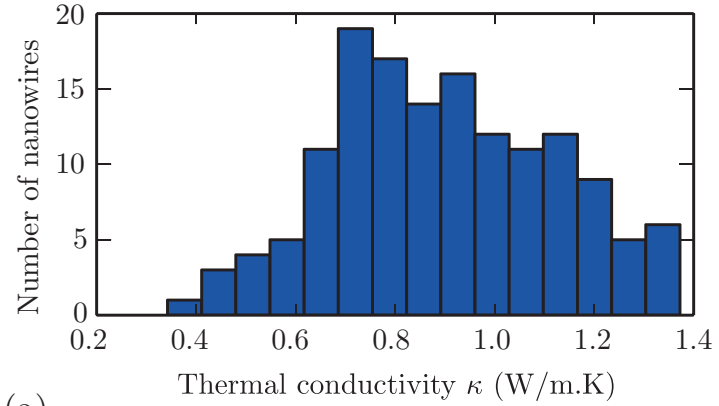
W/mK	Pure Si wire	PPG (defined later)
Our work (LAMMPS)	4.1 +/- 0.4	0.12 +/- 0.03
Ceder group (XMD)	4.1 +/- 0.3	0.23 +/- 0.05

Main sources of discrepancy

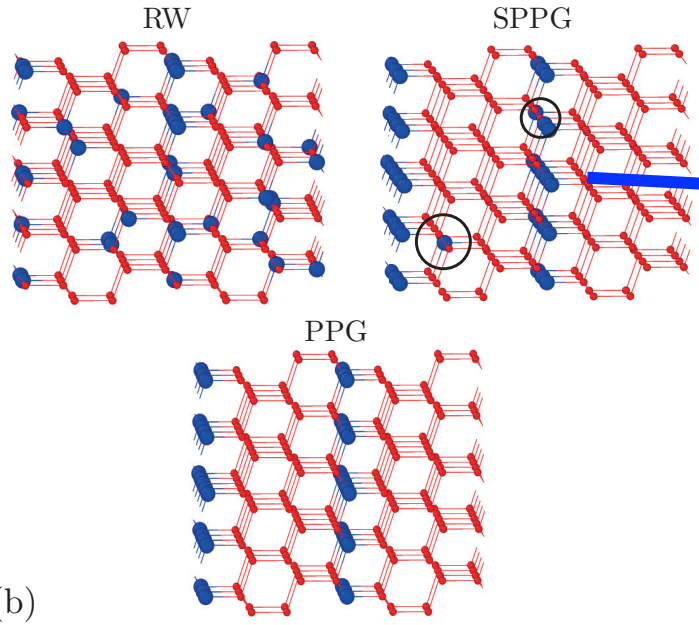
- thermalization techniques
- MD software
- thermalization times

*M. Chan et al. Physical Review B 81.17 (2010)

Nanowire Data set



(a)

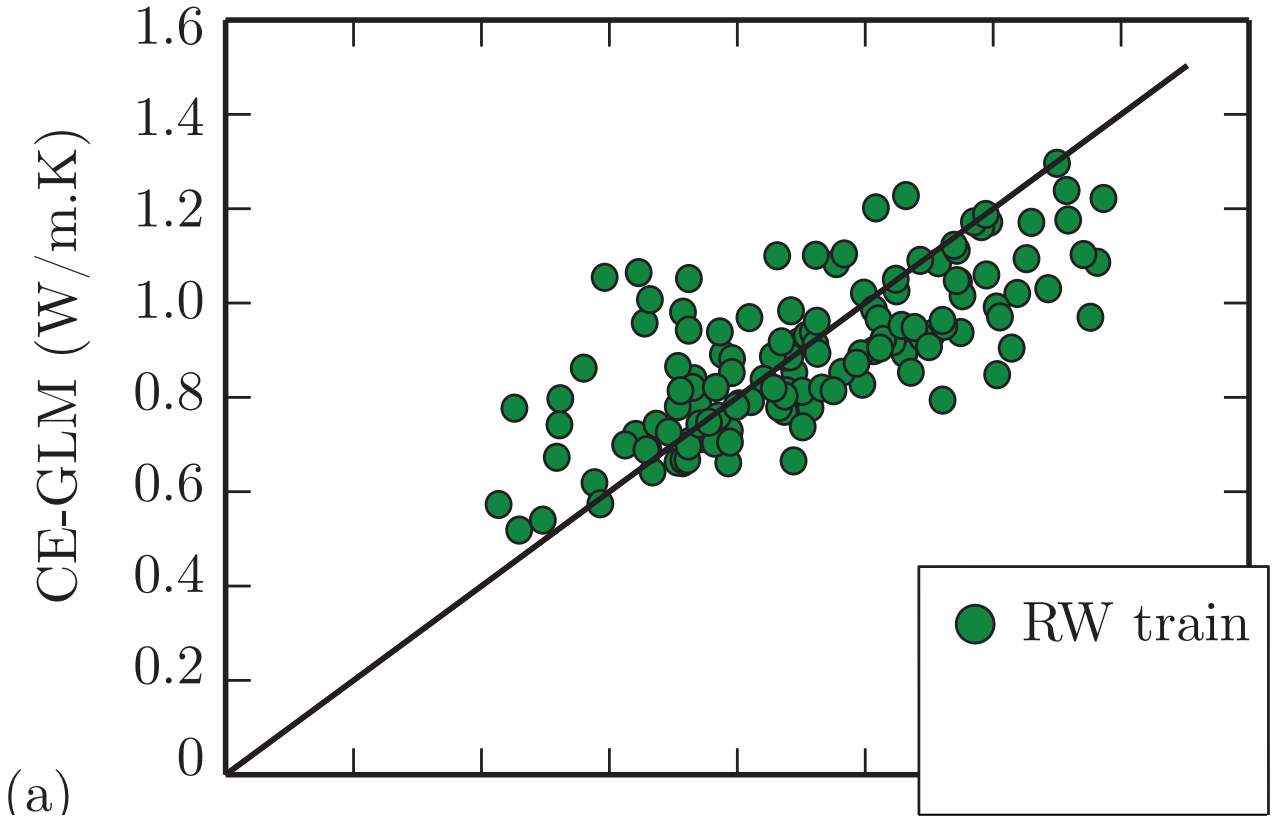


(b)

- 140 wires each with random Si/Ge configuration
  - This is the random nanowire dataset (RW)
- Split RW into train and test sets
  - Train CE with GLM on train
- Additional data sets:
  - Planes of pure Ge (PPG)
  - Similar to PPG (SPPG)
    - Perturbed: atom(s) from plane swapped with atom (s) from non-plane region

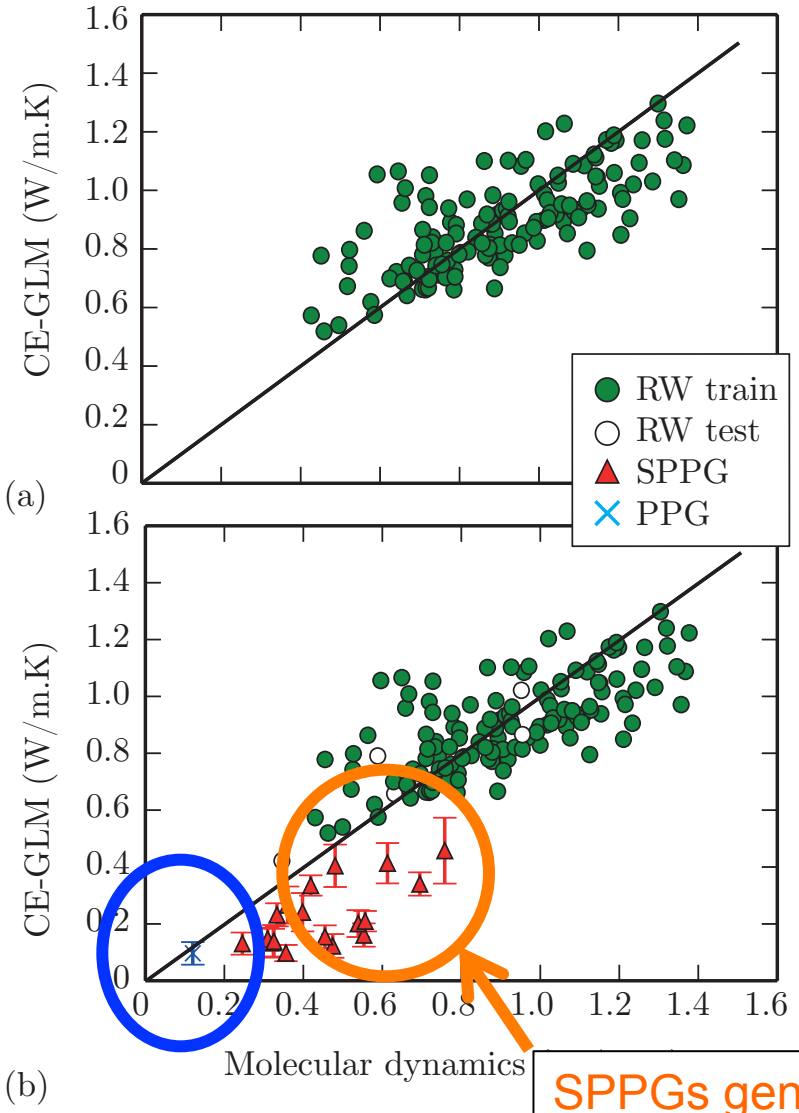
CE-GLM Fit on Random Nanowires

Using the new cluster expansion surrogate approach to fit the nanowire data set



Now that we have surrogate; find global minimum

Lowest Thermal-Conductivity Structure?



- We find the PPG to have lowest thermal conductivity
- Very strong case for the GLM
  - Evidence that thermal conductivity of nanowires is well captured by first term

$$\Delta H_f^{\mathrm{CE}} = \Delta H_f^{\mathrm{vol}} + \Delta H_f^{\mathrm{surf}}$$

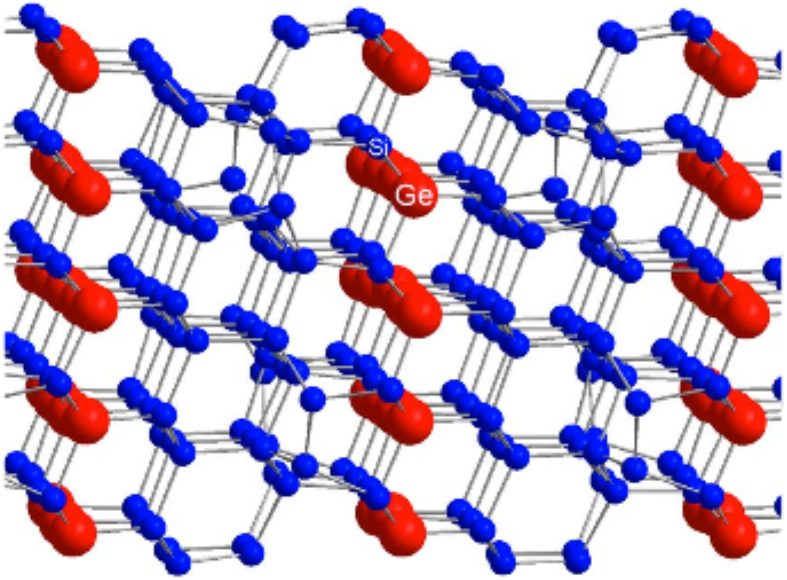
$\Delta H_f^{\mathrm{surf}} \approx 0$ (in our case)

SPPGs generally lower than RW train and test sets as expected

Compare with Literature

From Ref. *[*]* on the same problem
(but using a slightly different surrogate model)

They found as well that the PPG wire has lowest κ



(this image of the PPG wire is from Ref. [*])

*M. Chan et al. Physical Review B 81.17 (2010)

CONCLUDING REMARKS

Work in Progress

- Quantify uncertainties in
  - Band gaps
  - Energies
  - Phase Diagrams
  - Thermal Conductivities
  - *Any material property*
- Use information theory to design materials
- Help improve how data is collected (and the resources spent in doing so) in general
  - Choosing the limited data set in most informed way
- What happens to uncertainty quantification across length and time scales?
  - How do uncertainties in microscopic properties affect macroscopic properties?