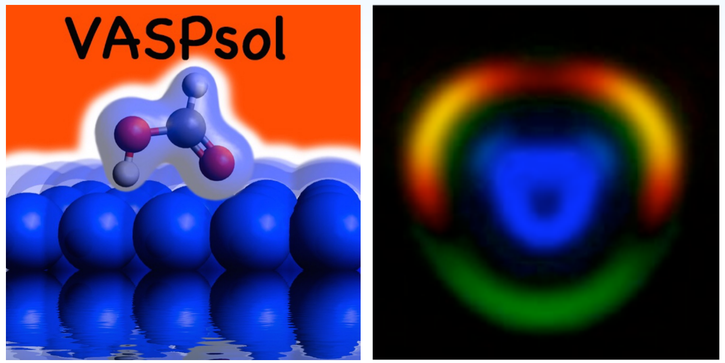
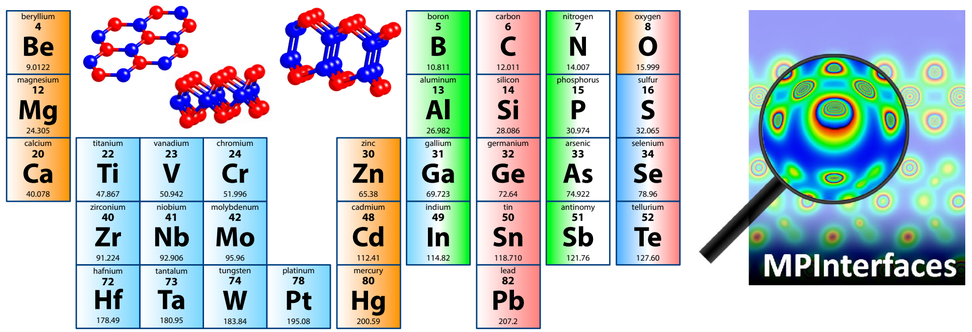
Machine-learning for materials and physics discovery through symbolic regression and kernel methods

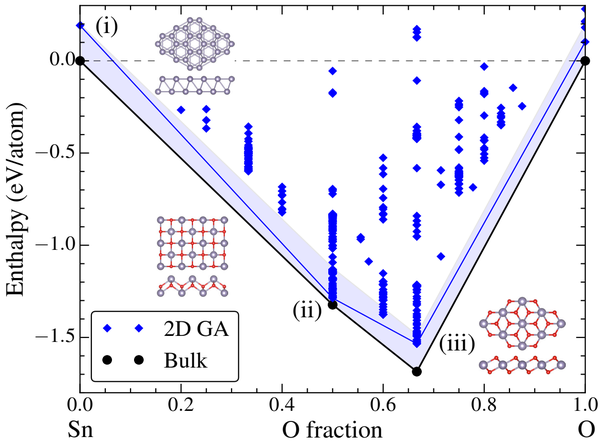
Stephen R. Xie, Shreyas Honrao, and Richard G. Hennig, University of Florida



VASPSol - Ab initio methods for solid/liquid interfaces

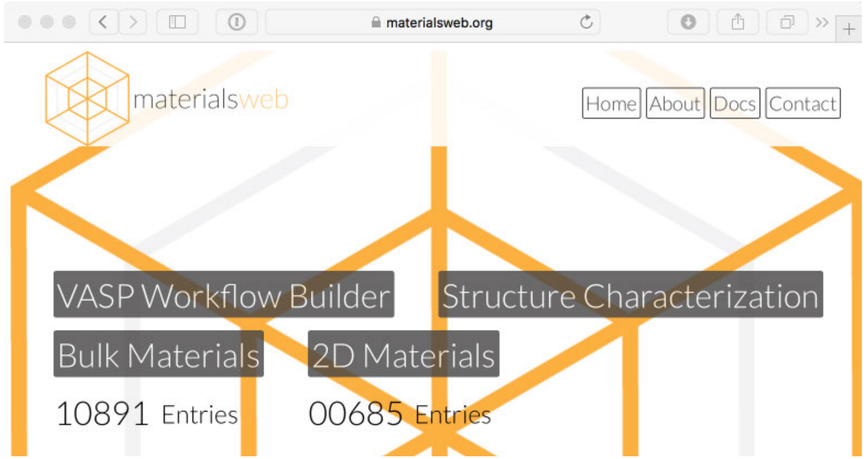
MPInterfaces - High throughput framework for 2D materials





GASP - Genetic algorithm and machine learning for structure predictions

Data available at http://materialsweb.org

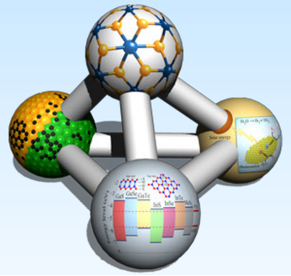


Powered by
MPInterfaces & materialsweb

Open source available at https://github.com/henniggroup

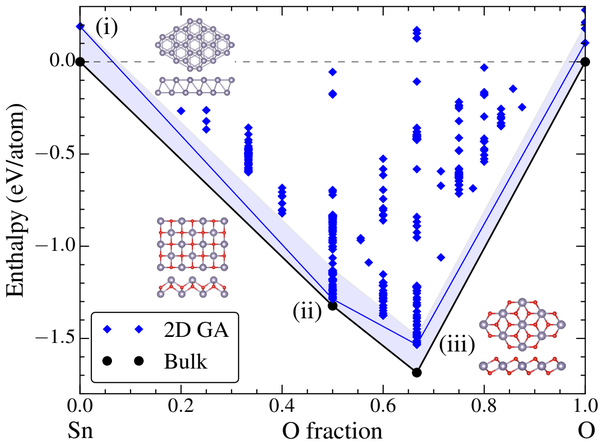
Machine-learning for materials and physics discovery through symbolic regression and kernel methods

Stephen R. Xie, Shreyas Honrao, and Richard G. Hennig, University of Florida



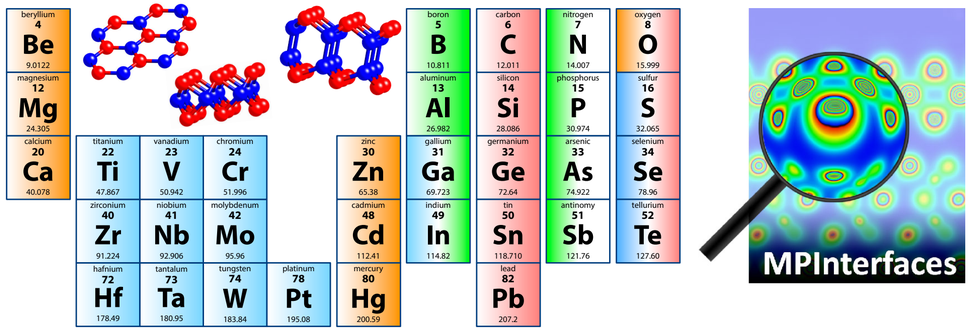
Machine Learning

- Machine learning of energy landscapes using distribution functions
- Reduction of error by learning atomic energies with local RDF descriptors
- Learning of analytic equations to predict superconductivity using small data sets

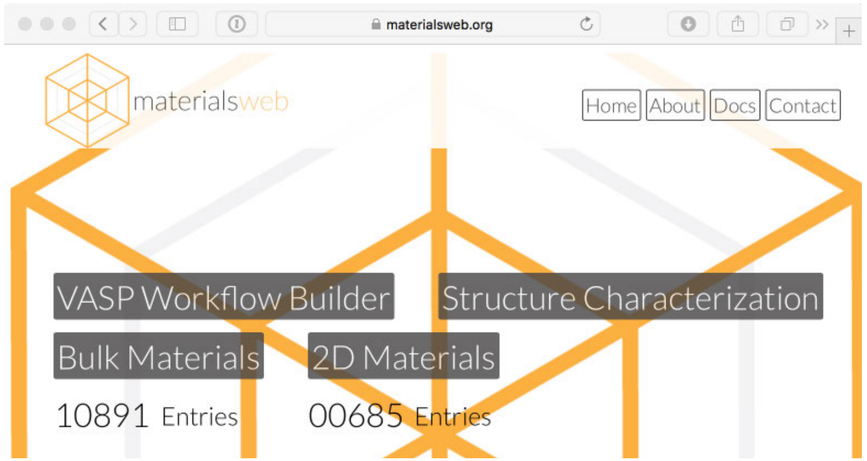


GASP - Genetic algorithm and machine learning for structure predictions

MPInterfaces - High throughput framework for 2D materials



Data available at http://materialsweb.org

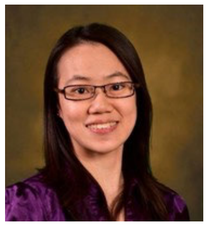


Powered by
MPInterfaces & materialsweb

Open source available at https://github.com/henniggroup

Acknowledgment

- GASP genetic algorithm: B. Revard, W. Tipton, A. Yesupenko, H. Lester,
- **Machine learning of physics and materials: S. Honrao**, **S. Xie**, B. Antonio
- Collaborators: Hao Li (UTAustin), Graeme Henkelman (UTAustin), Dallas R. Trinkle (UIUC)
- Machine learning of superconductivity: S. Xie, P. Hirschfeld, J. Hamlin, G. Stewart
- Financial support by NSF, DOE, NIST
- Computational resources: HiPerGator@UF, NSF XSEDE, and Google Cloud Platform



Powered by
MPInterfaces & materialsweb

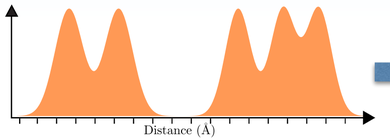
rhennig@ufl.edu
http://hennig.mse.ufl.edu

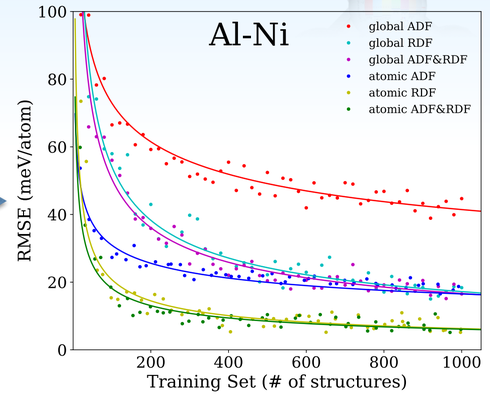
Machine Learning in Materials Science

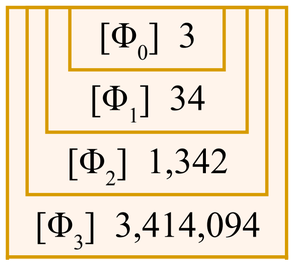
Supervised Learning

Big data
- High-throughput screening for properties
- Surrogate models for structure exploration

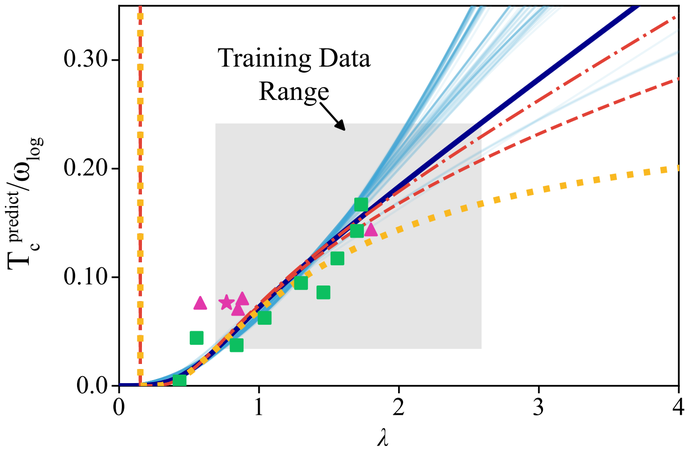
Small data
- Symbolic regression for analytic expressions





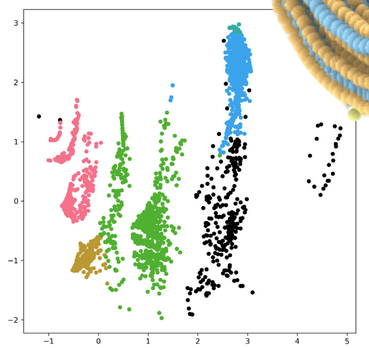


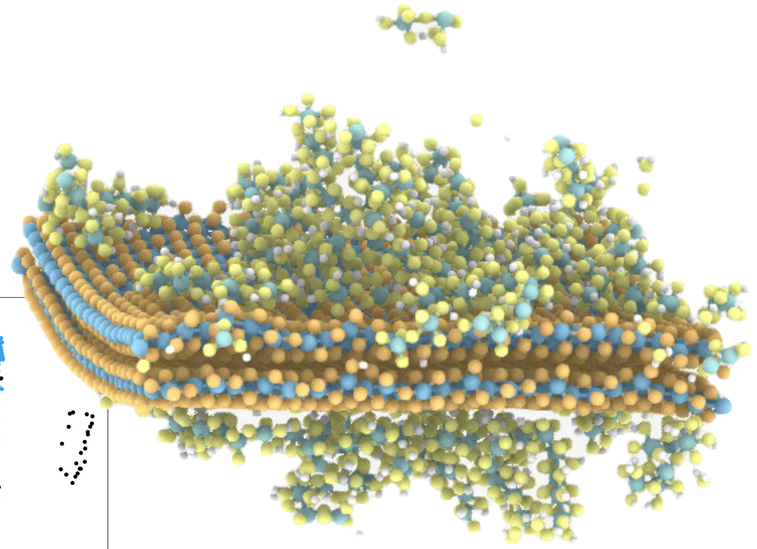
$$\omega_{\log}\ ,\ \mu^*\ ,\ \lambda\ ,\ \ldots$$
$$\omega_{\log}\times\lambda\ ,\ \sqrt{\mu^*}\ ,\ \lambda^3\ ,\ \ldots$$
$$\lambda^3\times(\omega_{\log}\times\lambda)\ ,\ \lambda^3+\sqrt{\mu^*}\ ,\ \ldots$$
$$\lambda^3\times(\omega_{\log}\times\lambda)\ /\ \left(\lambda^3+\sqrt{\mu^*}\right),\ \ldots$$



Unsupervised Learning

- Identification of new configurations in MD simulations





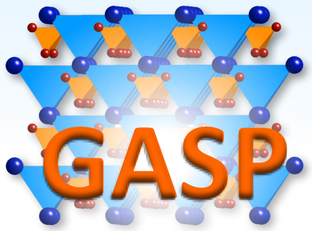
Powered by
MPInterfaces & materialsweb

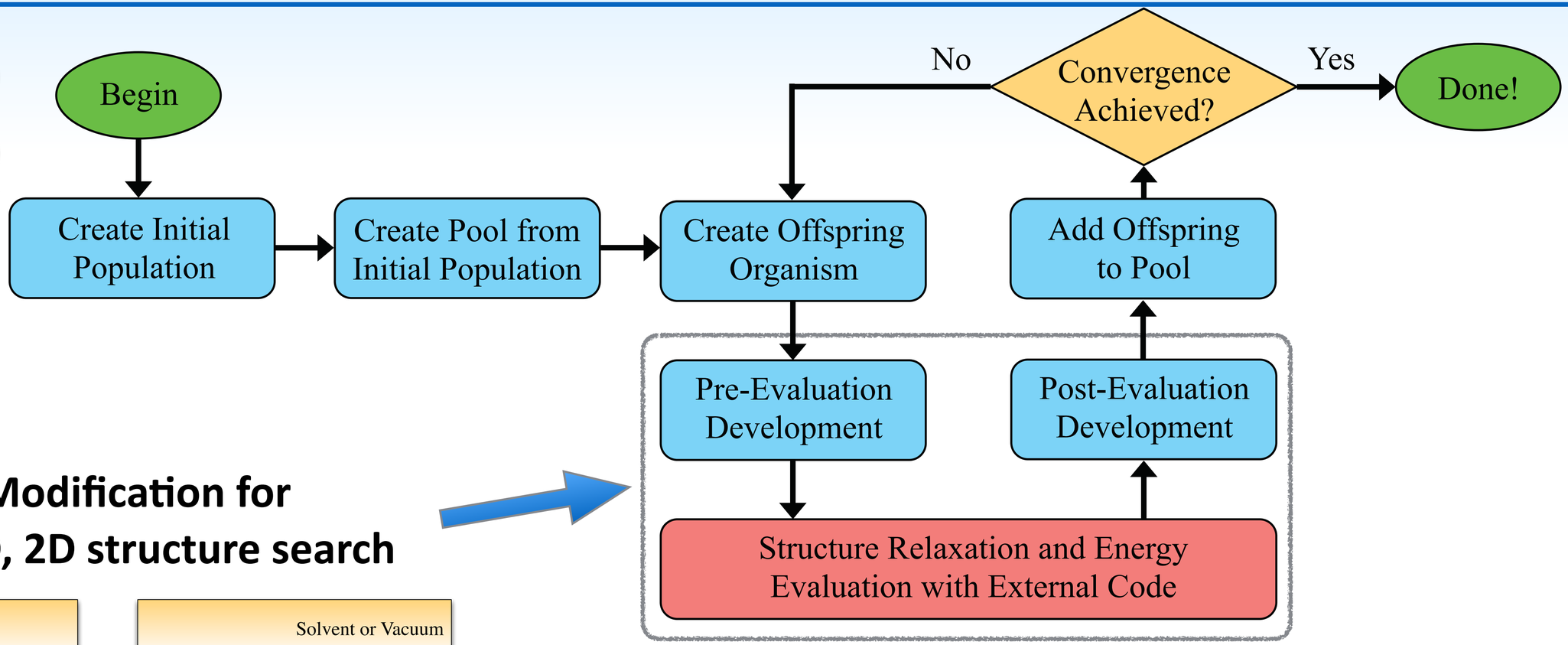
rhennig@ufl.edu
http://hennig.mse.ufl.edu

Part I: Exploration of Materials Energy Landscapes by Evolutionary/Genetic Algorithms

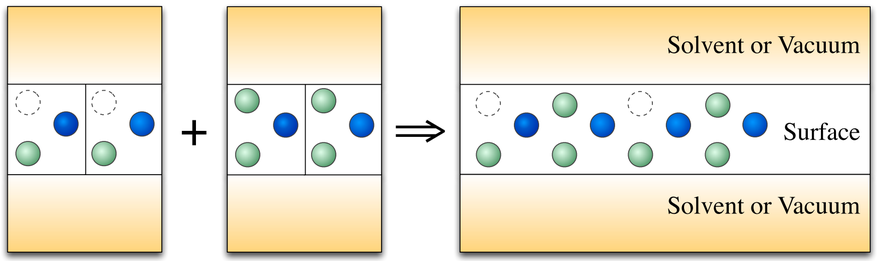
rhennig@ufl.edu
http://hennig.mse.ufl.edu

Genetic Algorithm Search for Crystalline Materials





Modification for
0D, 1D, 2D structure search



Grand canonical genetic algorithm
Variable number of atoms and composition

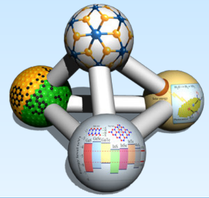
https://github.com/henniggroup/gasp-python

W. W. Tipton, RGH, J. Phys.: Cond. Matter 25, 495401 (2013)

B. C. Revard, W. W. Tipton, A. Yesypenko, R. G. Hennig, PRB 93, 054117 (2016)

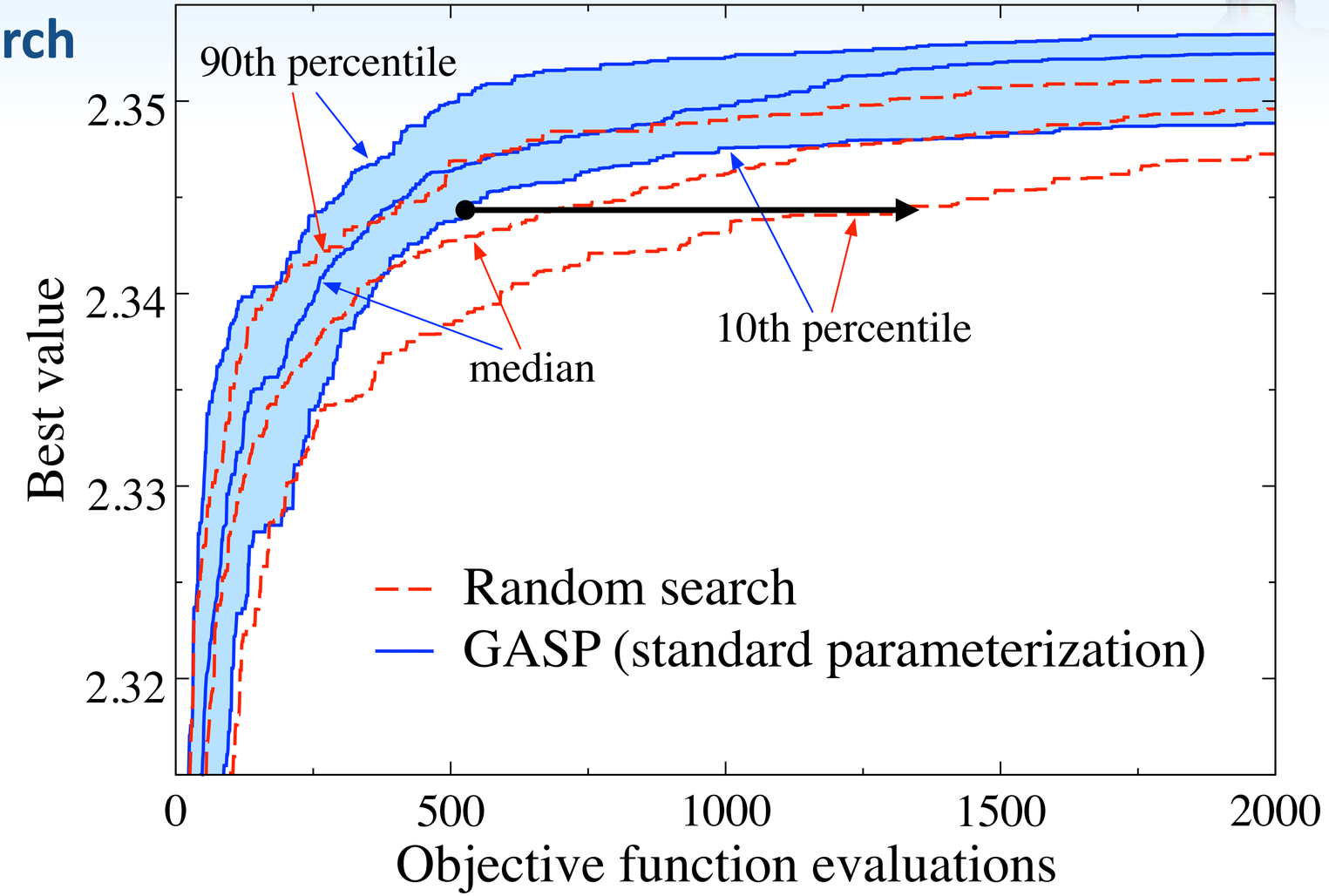
Powered by
MPInterfaces & materialsweb

Efficiency of Genetic Algorithm



Efficiency compared to random search

- Random search requires 2-3x more structure relaxations
- Genetic algorithm learns from previous structures

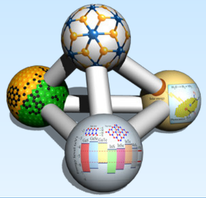


https://github.com/henniggroup/gasp-python

W. W. Tipton, RGH, J. Phys.: Cond. Matter 25, 495401 (2013)

B. C. Revard, W. W. Tipton, A. Yesypenko, R. G. Hennig, PRB 93, 054117 (2016)

Phase Diagram Searching

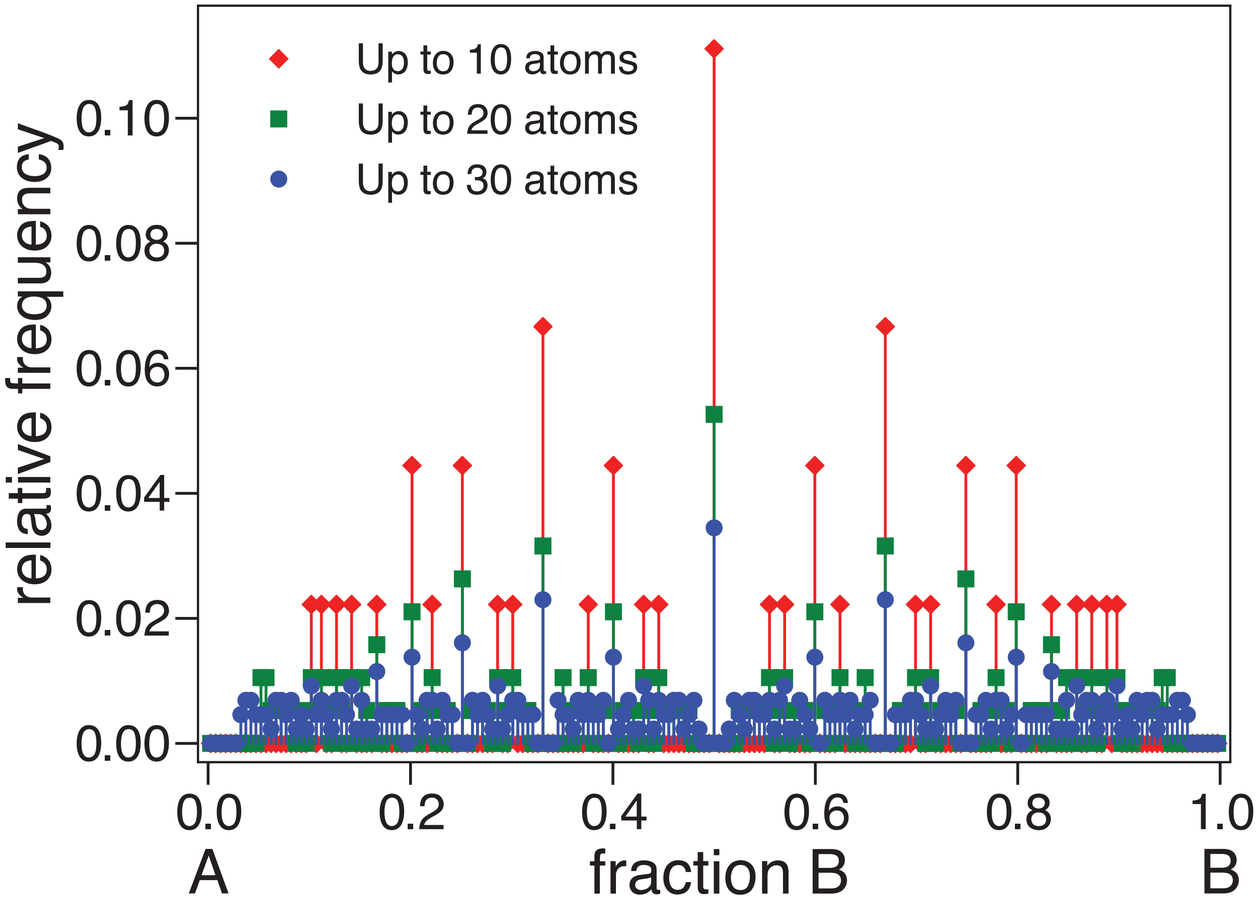


Problem:

- Naïve algorithm oversamples average compositions

Solutions

1. Use larger endpoint structures
   - Works but expensive
2. Preferentially select parents with similar compositions
   - Needs metric for distance of structures in composition space



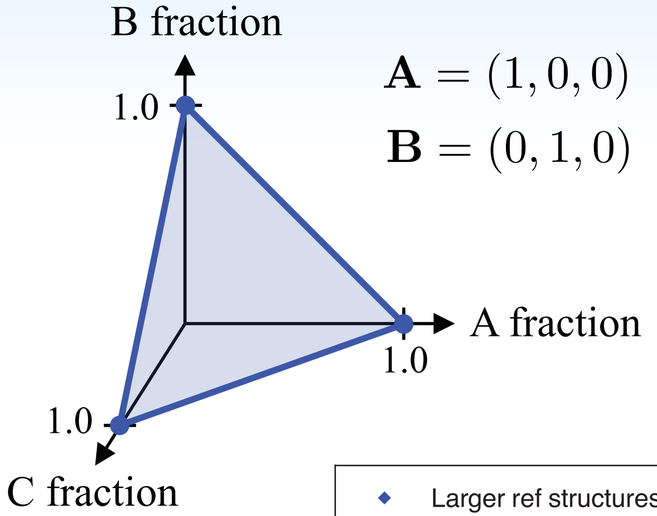
Phase Diagram Searching

Metric for distance in composition space

- Express composition as a vector
- Use L_1 Norm to define distance:

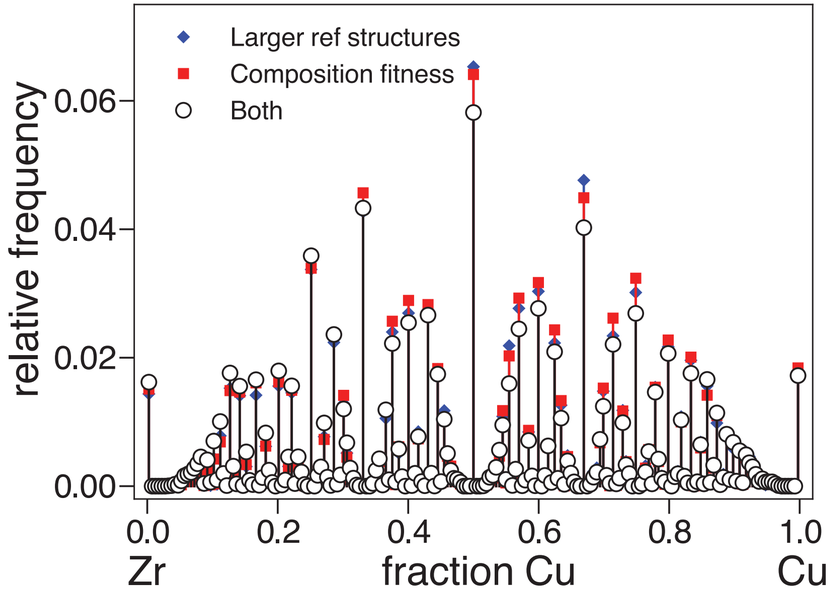
$$d_{XY} = \frac{1}{2}\left|\left|\mathbf{X} - \mathbf{Y}\right|\right|_1$$

$$\left|\left|\mathbf{A} - \mathbf{B}\right|\right|_1 = \left|\left|(1, -1, 0)\right|\right|_1 = |1| + |-1| + |0| = 2$$



Fitness for structures

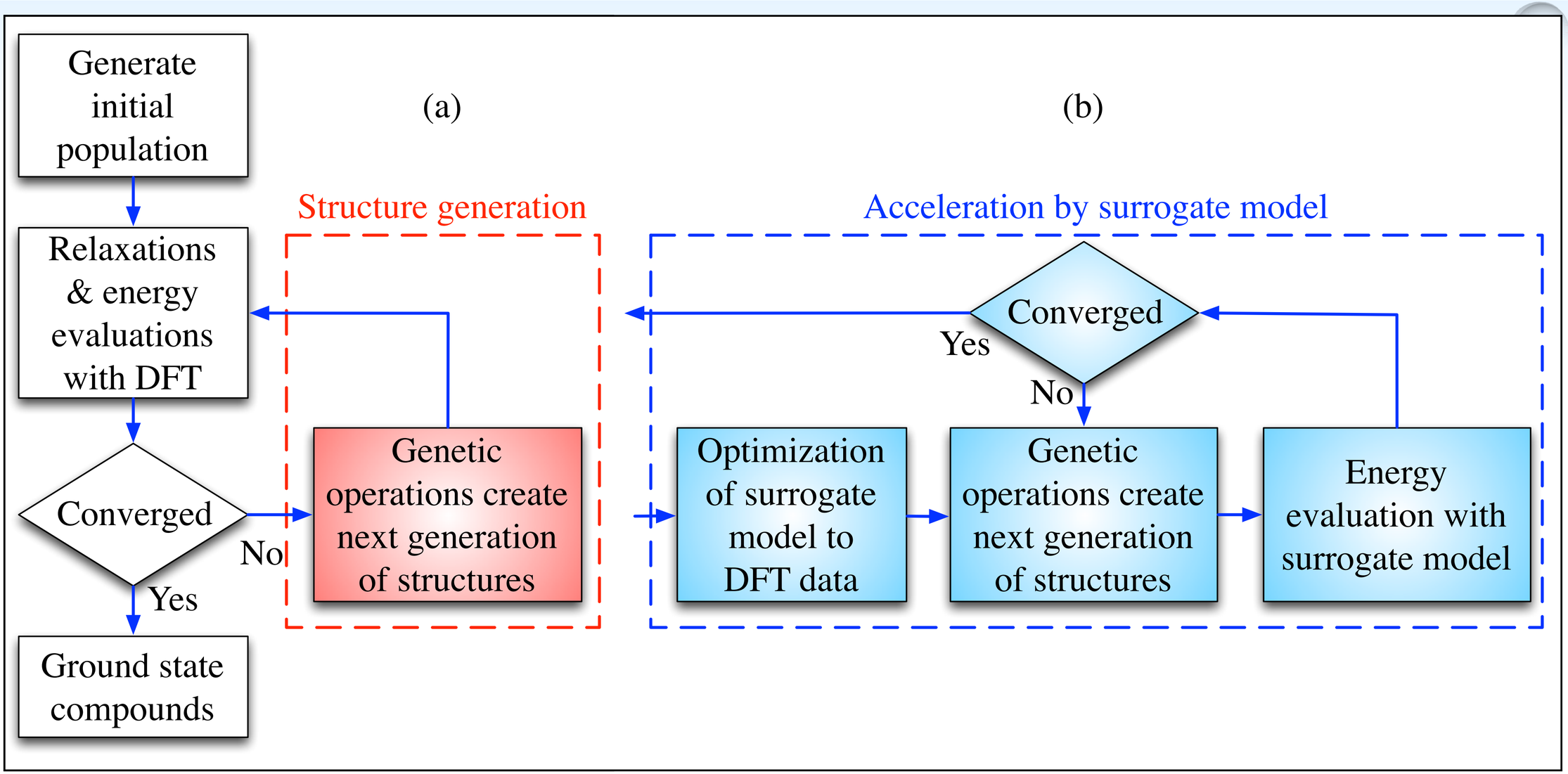
- Composition fitness $f_{\text{comp}} = 1 - d$
- Relative fitness $f_{\text{rel}} = w_{\text{comp}} f_{\text{comp}} + (1 - w_{\text{comp}}) f_{\text{reg}}$



Fitness for structures

- Sampling distribution improved but not uniform
- Use partial phase diagram searches if needed

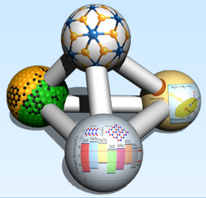
Acceleration of Genetic Algorithm



Part II: Machine Learning of Energy Landscapes

rhennig@ufl.edu
http://hennig.mse.ufl.edu

Machine Learning Surrogate Models



Machine Learning Regression

- Takes a vector $x \in R^n$ as input and return a scalar y
- Must first construct a vector-based data representation of the crystal structure that encodes relevant physical information, *i.e*. chemical identity and position of the atoms

Structure representation (features) should ideally fulfill three criteria

(i) **Invariance** with respect to choice of unit cell and crystal symmetry

(ii) **Uniqueness**, so no two different crystal structures have the same vector representation

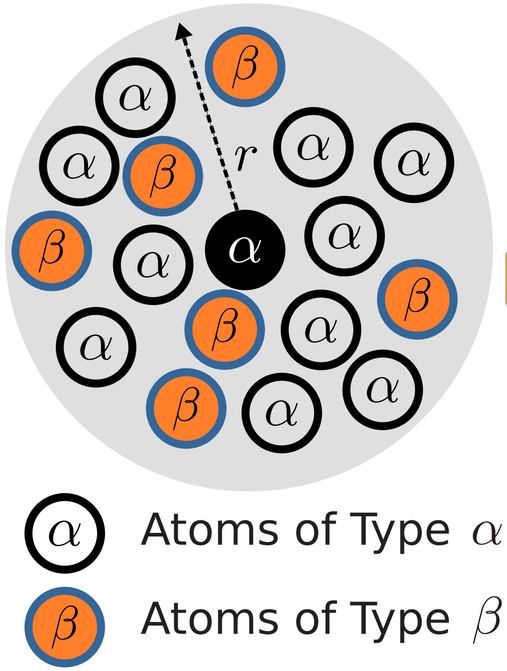
(iii) **Continuity**, such that the energy difference between two crystal structures with vector representations x_1 and x_2 goes to zero in the limit $\| x_1 - x_2 \| \rightarrow 0$

Partial Radial Distribution Functions

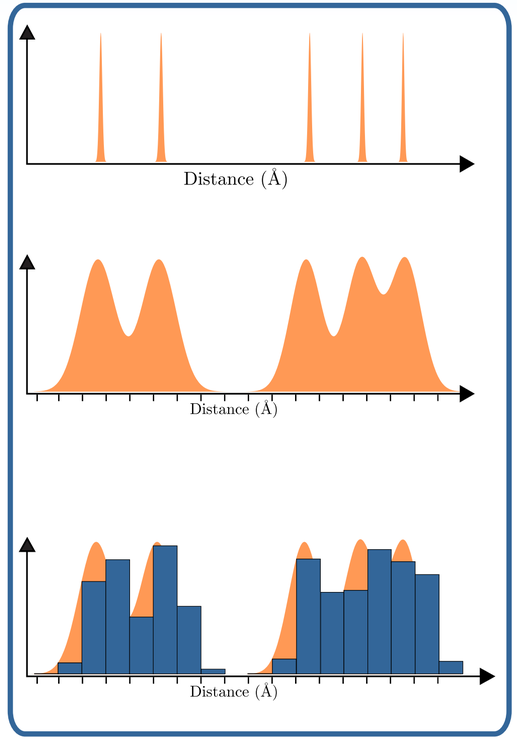
$$g_{AB}(r) = \frac{1}{N_A} \sum_{i=1}^{N_A} \sum_{j=1}^{\infty} \frac{1}{r^n} \exp\left[-\frac{\left(r - d_{ij}^{AB}\right)^2}{2\sigma_{\mathrm{g}}^2} \right] \Theta(d_{\mathrm{c}} - d_{ij}^{AB})$$

- Captures primary distance dependence of bonds
- Criteria:
  - + **Invariance**
  - + **Continuity**
  - - **Uniqueness**
- Cannot distinguish between *homometric structures*, *i.e.* structures of identical atoms that exhibit the same set of interatomic distances

Atomic Structure



Descriptors

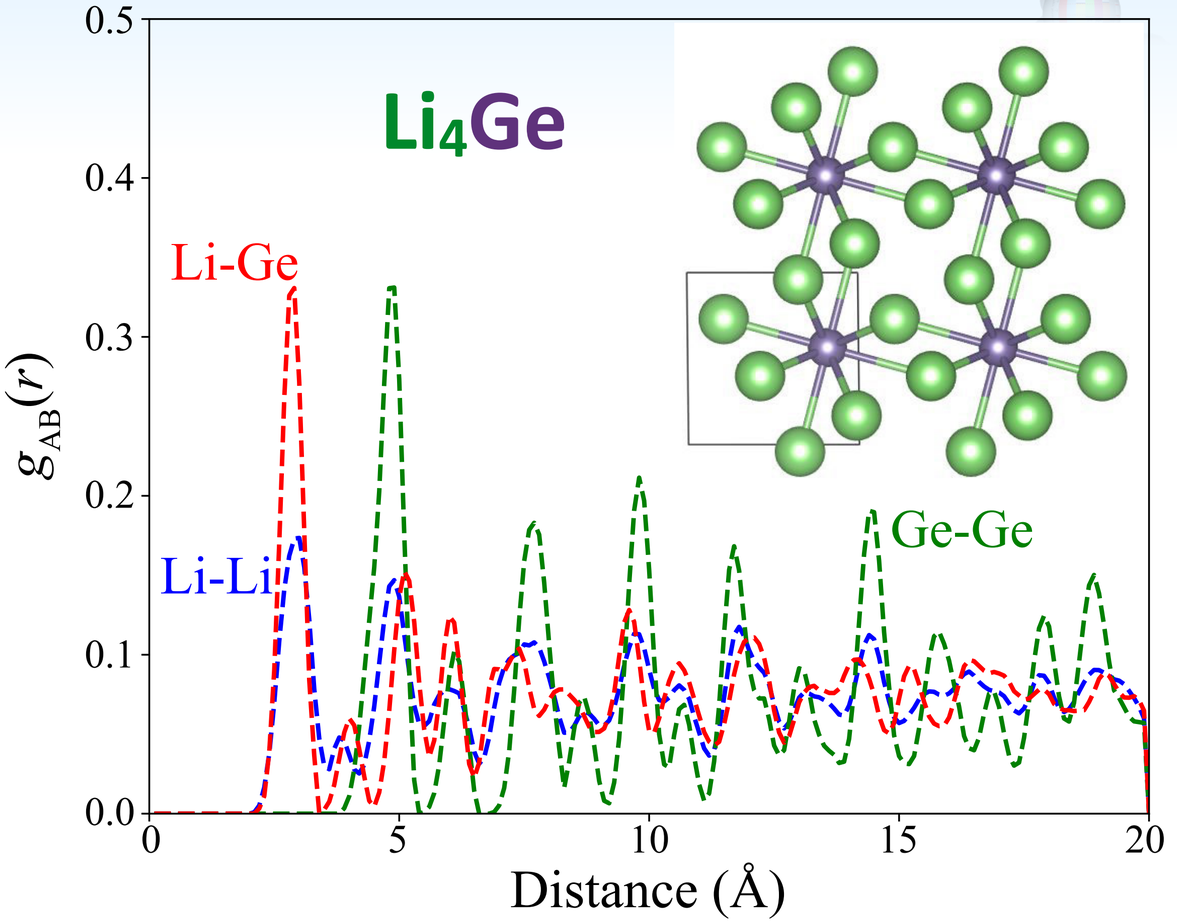


S. Honrao, RGH at al., submitted (2018)

Partial Radial Distribution Functions

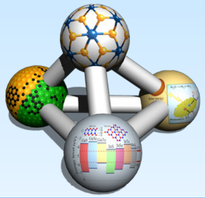
$$g_{AB}(r) = \frac{1}{N_A} \sum_{i=1}^{N_A} \sum_{j=1}^{\infty} \frac{1}{r^n} \exp\left[-\frac{\left(r - d_{ij}^{AB}\right)^2}{2\sigma_g^2} \right] \Theta(d_c - d_{ij}^{AB})$$

- Captures primary distance dependence of bonds
- Criteria:
  - + **Invariance**
  - + **Continuity**
  - - **Uniqueness**
- Cannot distinguish between *homometric structures*, *i.e.* structures of identical atoms that exhibit the same set of interatomic distances



S. Honrao, RGH at al., submitted (2018)

Data Set of Li-Ge Structures



Formation energies of Li-Ge structures

- 14,168 Li-Ge structures from genetic algorithm search for novel Li-Ge compounds
- Includes relaxed and unrelaxed structures from DFT relaxations of structure search
- Formation energy relative to crystal structure of pure components

$$E_{\mathrm{f}} = E_{\mathrm{tot}} - X_{\mathrm{Li}} E_{\mathrm{Li}} - X_{\mathrm{Ge}} E_{\mathrm{Ge}}$$

- **E_f is not simply counting bonds, sensitive to small changes in bonding character**

Structure groups

- Group the structures according to the basin of attraction ⇒ 679 basin groups
- Splitting of data significantly reduces the correlation between testing and training set
- Provides more stringent and realistic test of the ML methods

Machine Learning

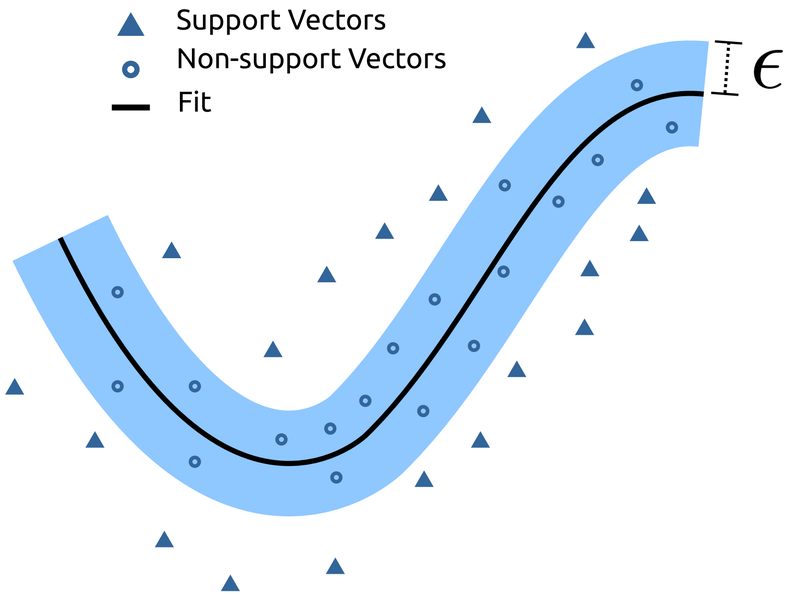
Machine-learning models

- Use kernel-ridge regression (KRR), ε-support vector regression (SVR), and neural networks

Input data preprocessing

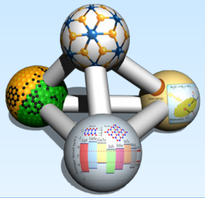
- Feature scaling of components of input vector $\mathbf{x}_i = g^i_{\mathrm{AB}}(r)$ in training set to obtain zero mean and unit standard deviation
- Standardizing each set of components avoids the norm from being biased towards vector components with higher variance
- 30% of data for learning, 70% for testing
- 10-fold cross validation for learning

$$x'_{ij} = \frac{x_{ij} - \mu_i}{\sigma_i}$$



S. Honrao, RGH at al., Comp. Mater. Sci. 158, 414 (2019)

Machine Learning



Hyper parameter selection

- Determined from 10-fold cross validation
- ε for SVR: negligible changes for ε < 10 meV/atom, use ε = 10 meV/atom
- Cutoff distance varied from 5 - 40 Å, larger errors for 5 Å, select 10 Å

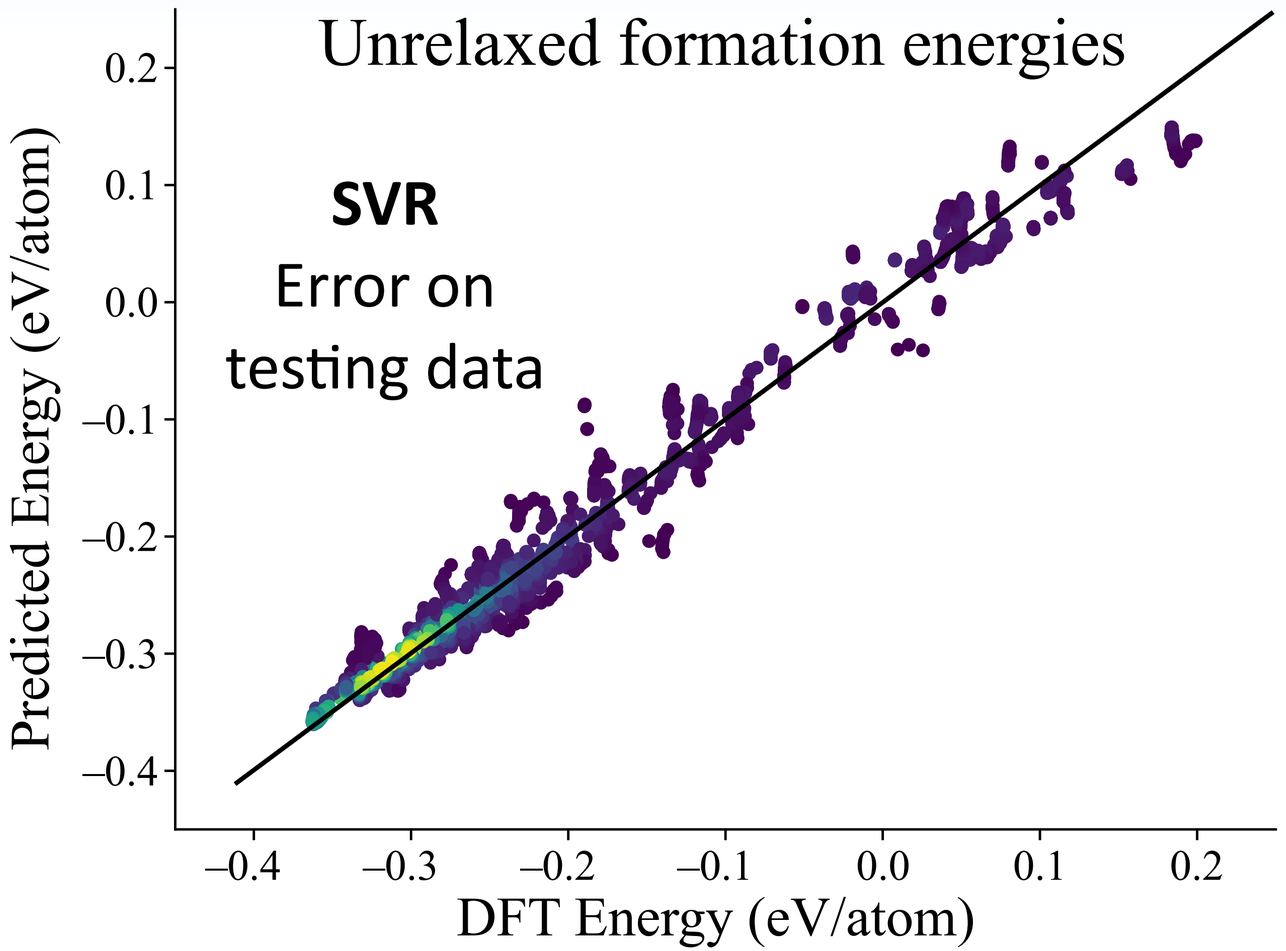
Algorithm	Kernel width	Average regularization parameter
KRR	21.2	0.015 (unit less)
SVR	54.3	10.5 meV/atom

- For neural network, 1 hidden layer and RELU function, 20 trials to estimate RMSE

Machine Learning Methods

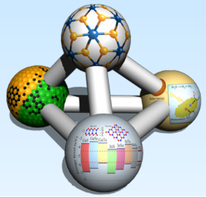
Algorithm	MAE	RMSE	R^2
KRR	12.7	20.4	0.98
SVR	13.6	20.8	0.98
NN	12.8	20.2	0.98

- Similar prediction errors (meV/atom) for different machine-learning techniques
- NN most demanding
- SVR is computationally most efficient and provides best tradeoff between complexity and prediction error



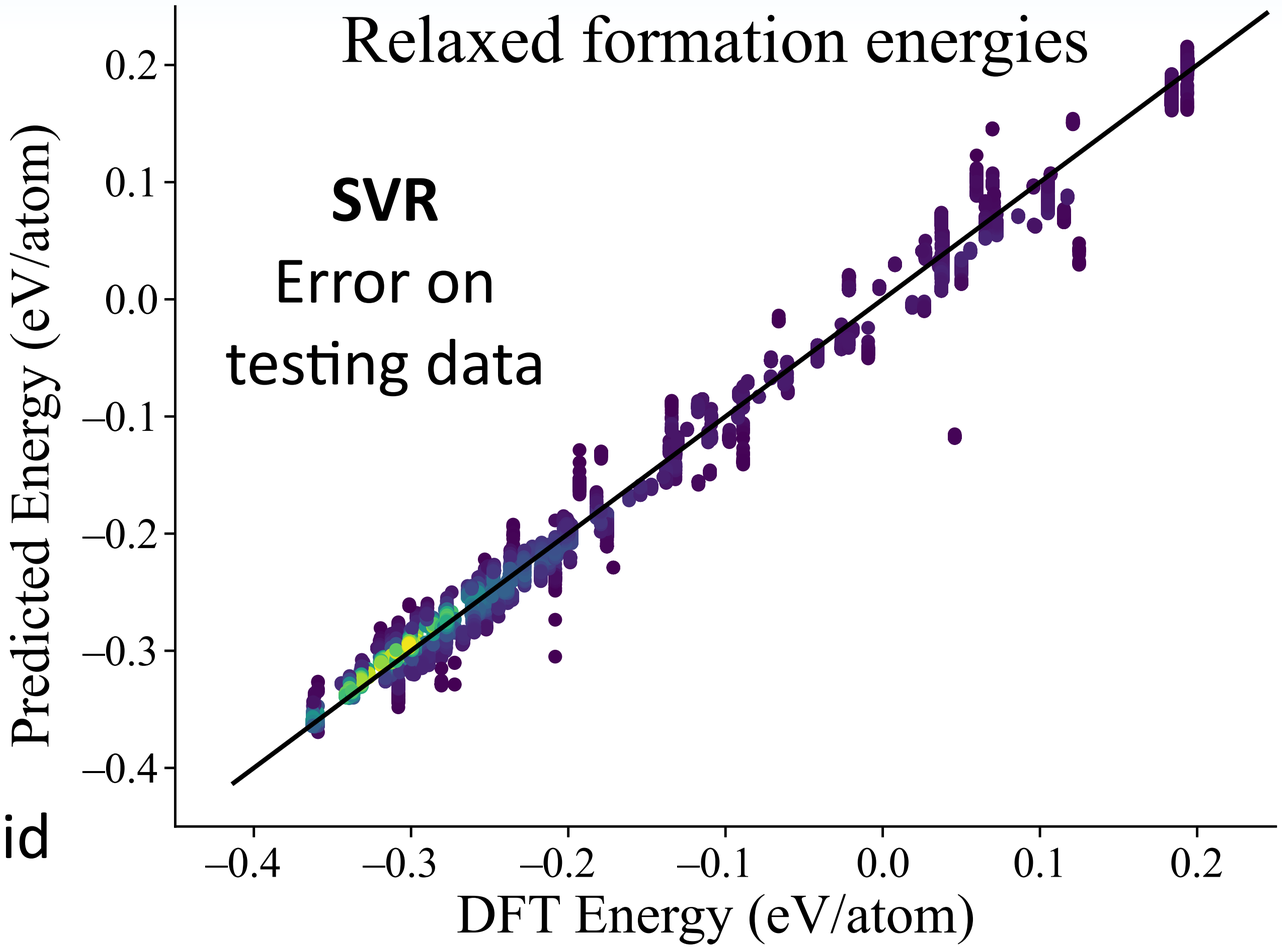
Chemical accuracy for learning of energy landscape

Learning of Basin of Attractions from Unrelaxed Structures



- Prediction of the relaxed energies (minima) from unrelaxed structures

Algorithm	MAE	RMS	R^2
KRR	12.7	20.4	0.98
SVR	13.6	20.8	0.98
NN	12.8	20.2	0.98
KRR-min	**11.8**	**20.3**	**0.98**
SVR-min	**13.4**	**20.9**	**0.98**

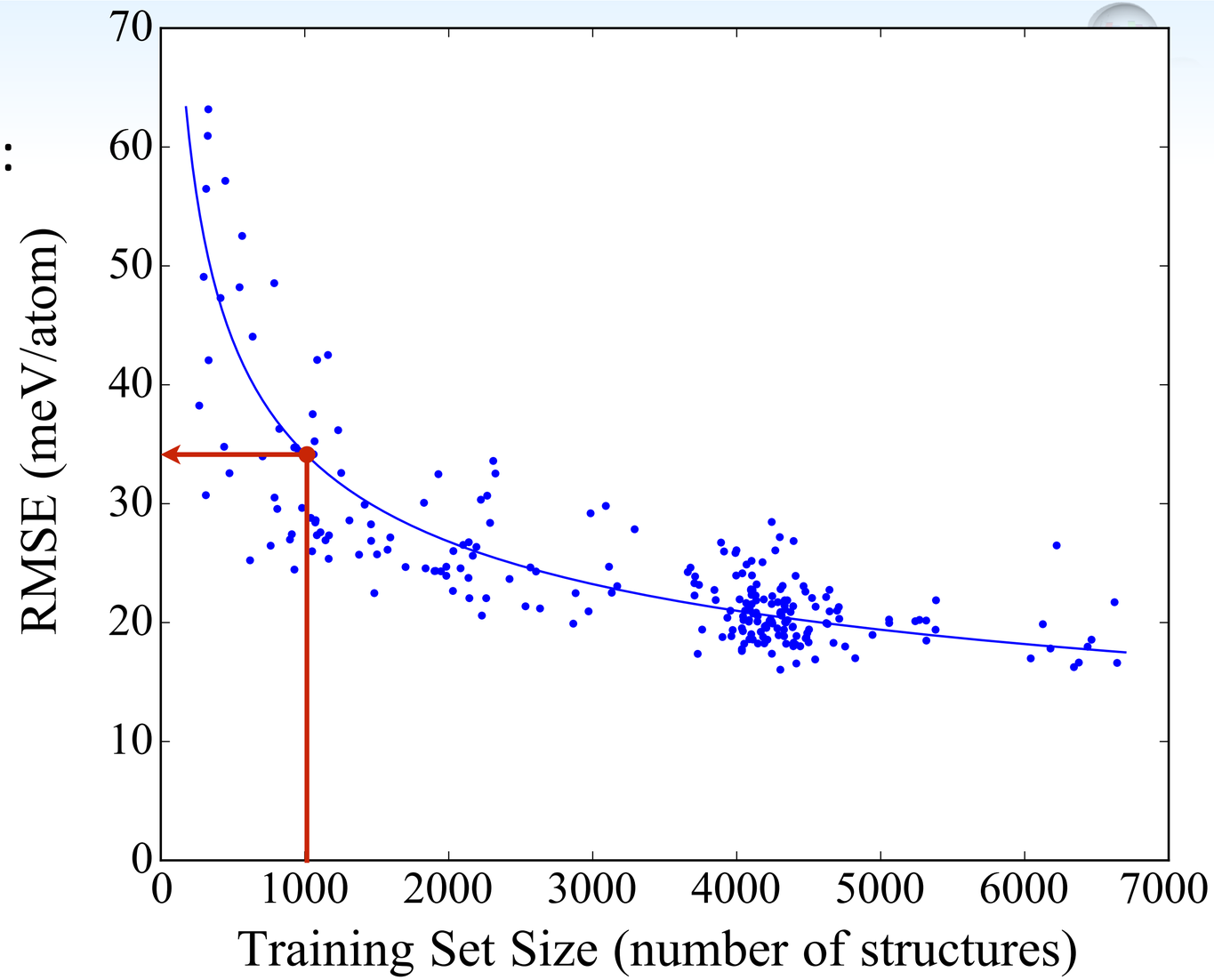


- Similar accuracy for minima prediction
- Useful for screening of GA structures to avoid costly DFT relaxations

ML learning of minima has similar accuracy as learning of energy landscape

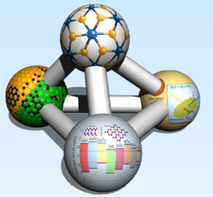
Big Data in Materials Predictions?

- Often insufficient amount of data for ML
- Early stage of genetic algorithm searches:
  - Few dozen relaxed structure
  - Maybe 1,000 configurations
  - RMSE ≈ 35 meV/atom



How can we improve the prediction error for small data sets?

Data Augmentation - Use Local Descriptors and Information

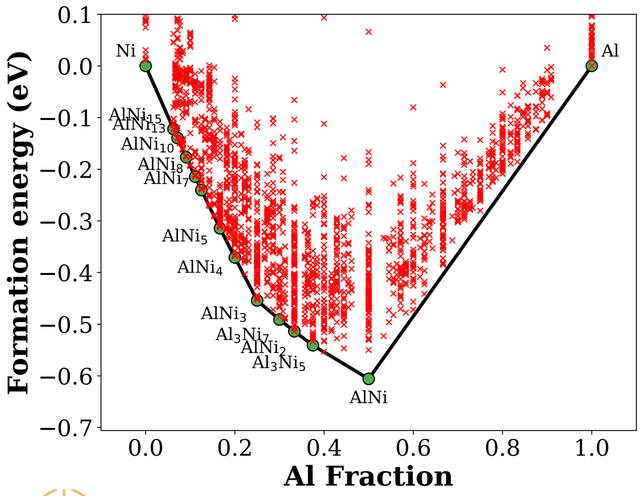


- Learn energy of individual atoms, separate ML models for each species
- Use local radial and angular distribution functions

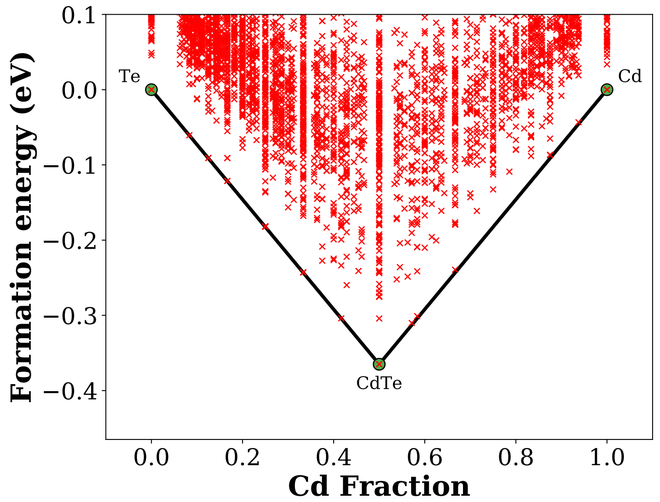
$$g_i^{AB}(r) = \sum_{j=1}^{\infty} \frac{1}{r^2} \exp\left[-\frac{\left(r - d_{ij}^{AB}\right)^2}{2\sigma_g^2}\right] f(d_{ij}^{AB}) \qquad q_i^{ABC}(x) = \sum_{j=1}^{\infty}\sum_{k=1}^{\infty} \exp\left[-\frac{\left(x - cos\theta_{jik}\right)^2}{2\sigma_g^2}\right] f(d_{ij}^{AB})\, f(d_{ik}^{AC})$$

- Dataset: GASP run with Ni-Al EAM and Cd-Te Stillinger-Weber potential

2,161 relaxed Al-Ni structures

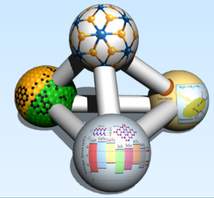


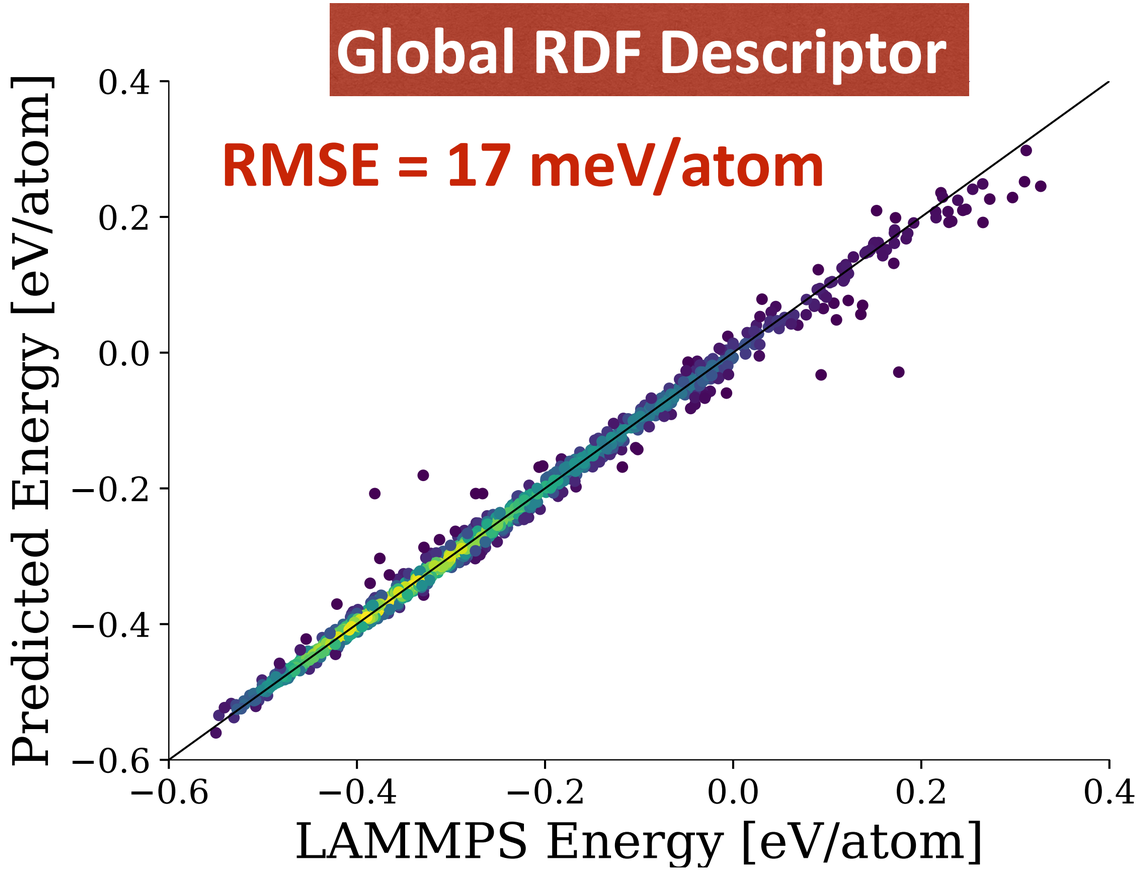
4,673 relaxed Cd-Te structures

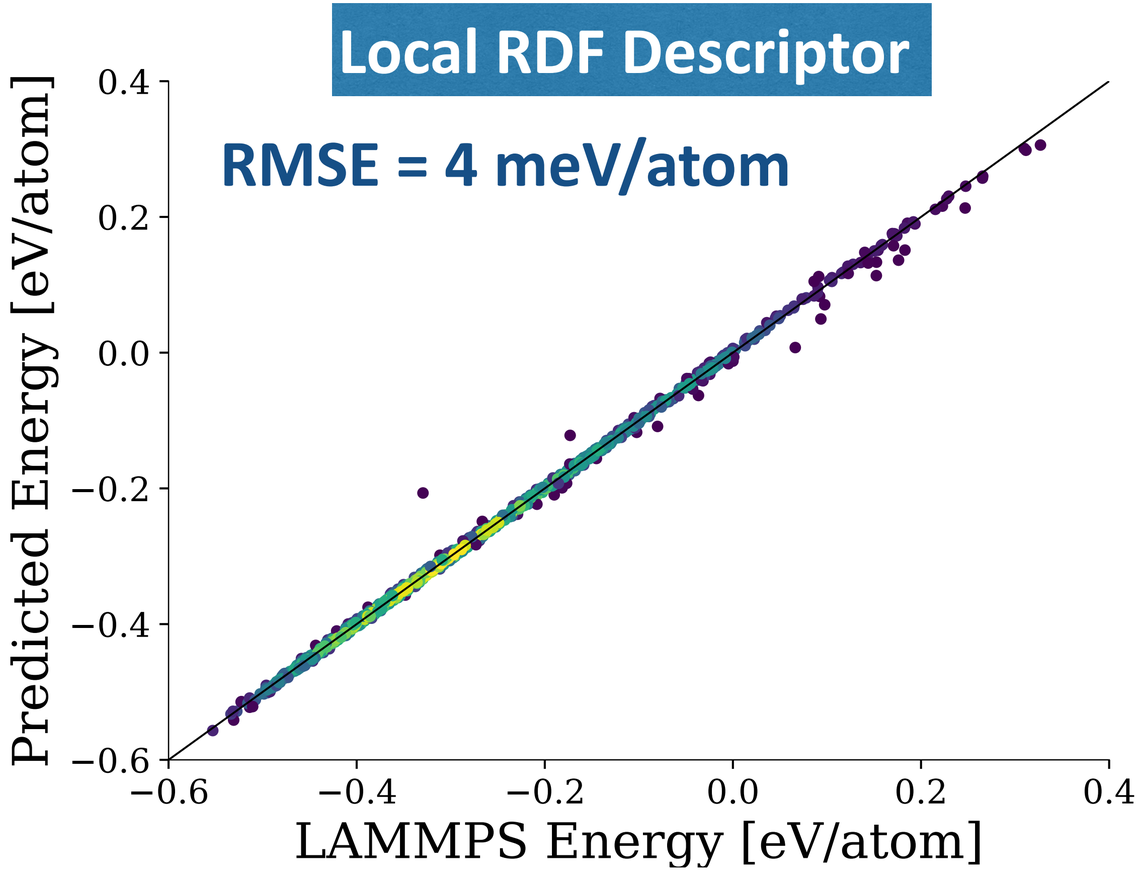


rhennig@ufl.edu
http://hennig.mse.ufl.edu

Prediction of Total Energy for Al-Ni



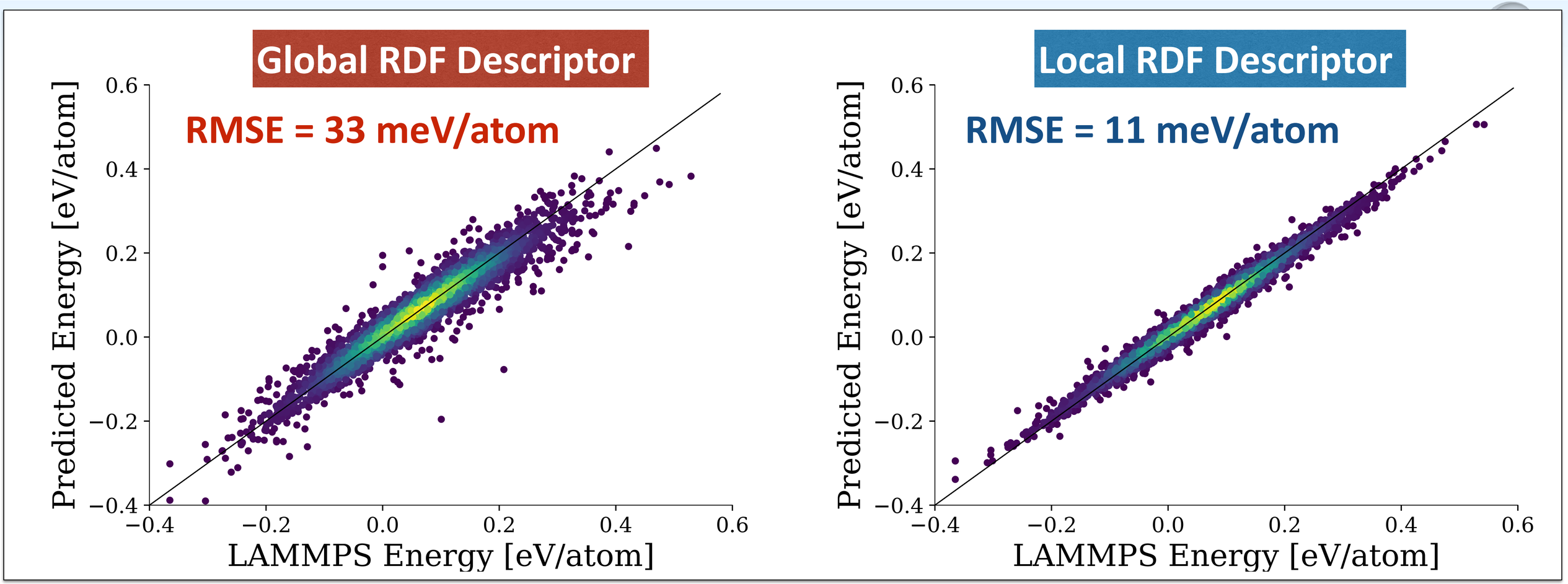




- Angular terms do not change prediction error for Al-Ni, expected for EAM pair functional form

Significant improvement of prediction error for same amount of data.

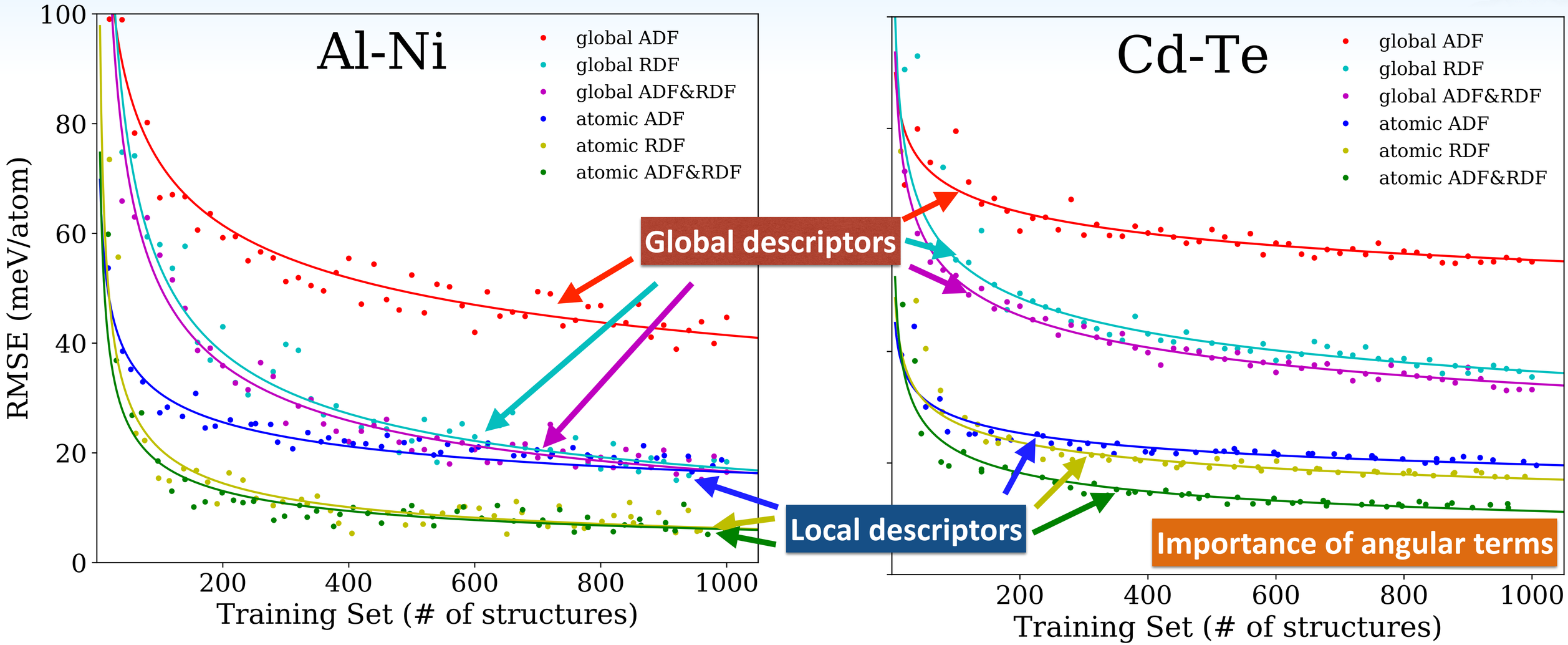
Prediction of Total Energy for Cd-Te



- Angular terms reduce prediction error for Cd-Te, expected for SW potential

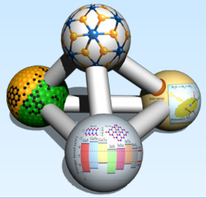
Significant improvement of prediction error for same amount of data.

Learning Curves



Local RDF descriptor reduces prediction error even for small datasets.

Comparison to Other Descriptors for Cd-Te

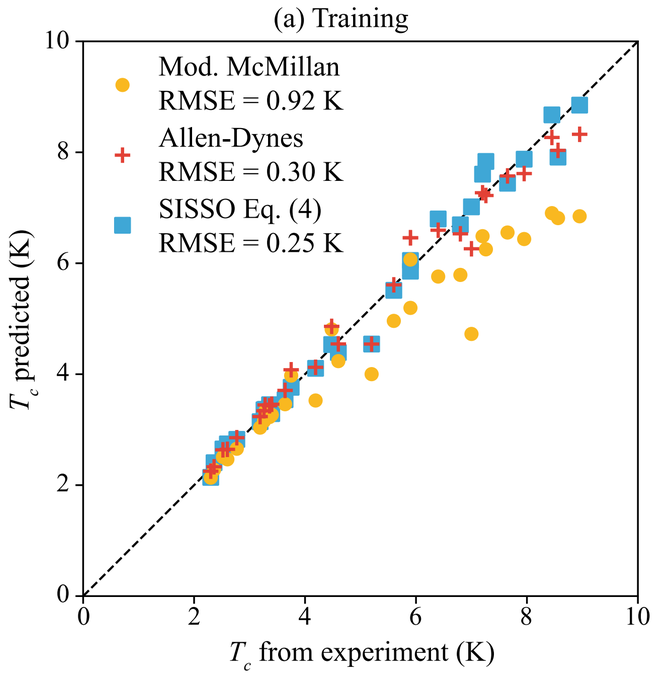


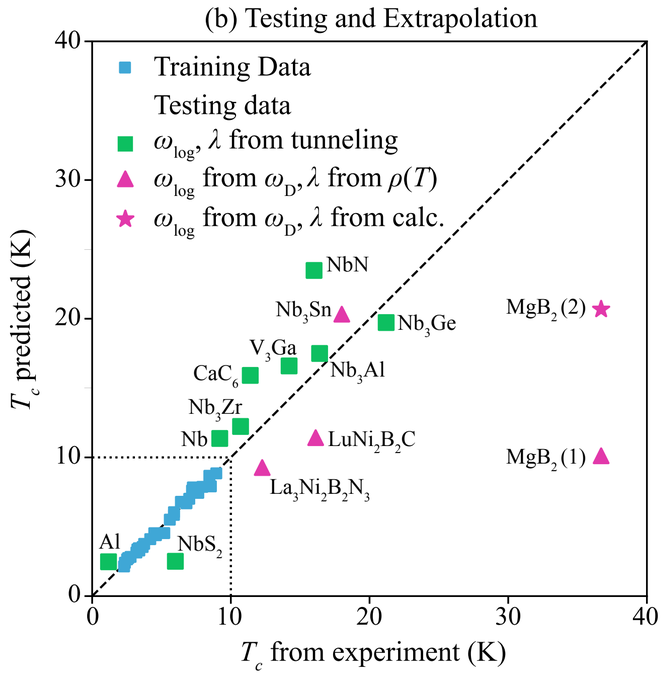
Data Representation	RMSE (meV/atom)	MAE (meV/atom)
Local RDF & ADF	11	8
Global RDF & ADF	33	24
Baseline model = mean E_f	109	85
Coulomb matrix	88	64
Orbital-field matrix	64	47
Bag of bonds	77	57
JARVIS-CFID	47	35
AGNI fingerprints	108	82

Importance of local RDF descriptors to capture change in chemical bonding

Part III: Functional Form of the Superconducting Critical Temperature from Machine Learning

Stephen R. Xie, James Hamlin, Gregory R. Stewart, Peter J. Hirschfeld, Richard G. Hennig



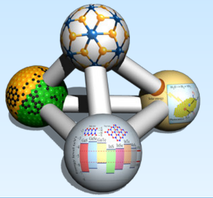


$$T_c^{\mathrm{SISSO}} = 0.095441 \frac{\lambda^4 \omega_{\log}}{\lambda^3 + \sqrt{\mu^*}}$$

Powered by
MPInterfaces & materialsweb

rhennig@ufl.edu
http://hennig.mse.ufl.edu

Motivation

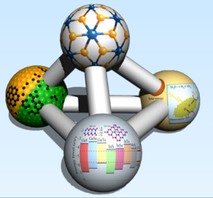


- Vast space of possible superconducting materials
- Significant efforts to apply computational methods with theory to screen materials
  - 2015: Prediction and discovery of T_c in H_3S at 200K and one megabar pressure
  - 2018: Prediction and discovery of T_c in Lanthanum Hydrides between 215-260K (approaching room temperature)

- What are the correct material "descriptors" that reflect the underlying mechanism of superconductivity?
- Machine-learning of materials often suffers from a small-data problem. What methods can overcome this issue?
- Can we use machine learning to identify analytical relationships between descriptors and T_c?

rhennig@ufl.edu
http://hennig.mse.ufl.edu

Training Data Set



- ***Training data*** from Table I by Allen & Dynes (1975)
  - 29 materials with parameters derived from tunneling
  - Parameters to use: ω_{log}, λ, μ^*
  - Allen-Dynes: Most widely used Eq. to predict T_c

PHYSICAL REVIEW B VOLUME 12, NUMBER 3 1 AUGUST 1975

Transition temperature of strong-coupled superconductors reanalyzed

P. B. Allen*
Department of Physics, State University of New York, Stony Brook, New York 11790

R. C. Dynes
Bell Laboratories, Murray Hill, New Jersey 07974
(Received 20 January 1975)

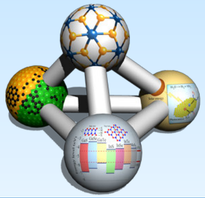
$$T_c = \frac{f_1 f_2 \omega_{log}}{1.20} \exp\left(-\frac{1.04(1+\lambda)}{\lambda - \mu^* - 0.62\lambda\mu^*}\right)$$

TABLE I. Parameters[a] of superconductors derived from tunneling measurements. The value of μ^* is renormalized from previously reported values as described in the text.

Material	ω_{log} (K)	$\overline{\omega}_1$ (K)	$\overline{\omega}_2$ (K)	λ	ω_{ph} (K)	$\mu^*(\omega_{ph})$	T_c (K)	η (eV/Å^2)
Pb	56	60	65	1.55	110	0.105	7.2	2.4
In	68	79	89	0.805	179	0.097	3.40	1.3
Sn	99	110	121	0.72	209	0.092	3.75	2.2
Hg	29	38	49	1.6	162	0.098	4.19	1.4
Tl	52	58	64	0.795	127	0.111	2.36	1.2
Ta	132	140	148	0.69	228	0.093	4.48	4.9
a-Ga	55	77	101	1.62	291	0.095	8.56	2.1
β-Ga	87	108	129	0.97	285	0.092	5.90	2.0
$Tl_{0.9}Bi_{0.1}$	48	55	62	0.78	120	0.099	2.30	1.1
$Pb_{0.4}Tl_{0.6}$	48	56	62	1.15	121	0.094	4.60	1.6
$Pb_{0.6}Tl_{0.4}$	50	57	62	1.38	119	0.103	5.90	2.0
$Pb_{0.8}Tl_{0.2}$	50	56	61	1.53	116	0.101	6.80	2.1
$Pb_{0.6}Tl_{0.2}Bi_{0.2}$	48	53	58	1.81	112	0.111	7.26	2.3
$Pb_{0.9}Bi_{0.1}$	50	56	60	1.66	108	0.081	7.65	2.2
$Pb_{0.8}Bi_{0.2}$	46	52	57	1.88	109	0.093	7.95	2.3
$Pb_{0.7}Bi_{0.3}$	47	52	57	2.01	110	0.092	8.45	2.4
$Pb_{0.65}Bi_{0.35}$	45	50	55	2.13	110	0.093	8.95	2.4
$In_{0.9}Tl_{0.1}$	63	75	86	0.85	176	0.103	3.28	1.4
$In_{0.73}Tl_{0.27}$	55	67	77	0.93	166	0.110	3.36	1.4
$In_{0.67}Tl_{0.33}$	57	68	79	0.90	167	0.110	3.26	1.4
$In_{0.57}Tl_{0.43}$	53	64	74	0.85	165	0.117	2.60	1.3
$In_{0.5}Tl_{0.5}$	53	64	73	0.83	163	0.110	2.52	1.3
$In_{0.27}Tl_{0.73}$	42	53	63	1.09	151	0.094	3.64	1.4
$In_{0.17}Tl_{0.83}$	45	55	63	0.98	144	0.101	3.19	1.3
$In_{0.07}Tl_{0.93}$	49	56	63	0.89	131	0.107	2.77	1.3
In_2Bi	46	57	67	1.40	174	0.096	5.6	1.6
Sb_2Tl_7	37	48	58	1.43	134	0.102	5.2	1.6
Bi_2Tl	47	53	59	1.63	120	0.101	6.4	2.1
a-$Pb_{0.45}Bi_{0.55}$	29	38	47	2.59	128	0.116	7.0	2.1

[a]Tabulation of the data used to derive these parameters is available in J. M. Rowell, W. L. McMillan, and R. C. Dynes, J. Phys. Chem. Ref. Data (to be published).

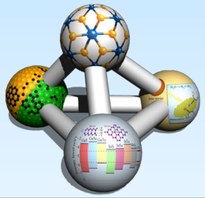
Symbolic Regression Machine-Learning



- Apply **Sure Independence Screening and Sparsifying Operator (SISSO)** method to generate predictive models from **training data** for 29 materials
  - Identifies ***analytical relations*** between a minimal set of descriptors and desired properties
  - Yields stable results with small training sets
- Compare equations by evaluating with ***testing data***
  - 13 superconductors from literature including A15 phases
  - Measure interpolative and extrapolative capacity, i.e., **transferability**
  - **Avoid overfitting**, i.e., artificially low error in training data with high error in testing data

rhennig@ufl.edu
http://hennig.mse.ufl.edu

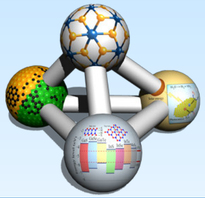
SISSO Definitions



- Feature: Quantity that is hypothesized to be relevant for describing target property, T_c
- Primary features: ω_{log}, λ, μ^*
- Sure Independence Screening – SIS
  1. Expand feature space Φ by recursively applying and combining algebraic/functional operation
     - +, -, ×, ÷, exp, log, √ , $^{-1}$, ^{2}, ^{3}, (sin, cos not used)
     - Respect units with dimensional reduction
  2. Rank features by their correlation magnitude (dot product of feature and T_c)
- Sparsifying Operator – SO
  - Use ℓ_0-norm regularized minimization to find sparse solution of linear equations
  - 1D solution is trivially the first-ranked feature
  - *n*-dimensional descriptors are used for classification models

rhennig@ufl.edu
http://hennig.mse.ufl.edu

Sure Independence Screening



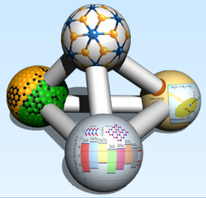
$$[\Phi_0] \quad \omega_{\log}, \quad \mu^*, \quad \lambda$$

$$3$$

$$\{ \quad +, -, \quad , \quad \times, \div, \quad , \quad \exp, \log, \quad , \quad {}^{-1}, {}^{2}, {}^{3}, \quad , \quad \sqrt{}, \sqrt[3]{} \quad \}$$

rhennig@ufl.edu
http://hennig.mse.ufl.edu

Sure Independence Screening



$[\Phi_0]$
3

$\omega_{\log}$, μ^* , λ

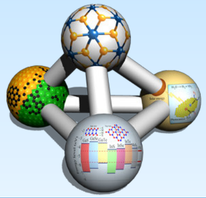
{ $+, -$, , $\times, \div$, , $\exp, \log$, , $^{-1}, ^2, ^3$, , $\sqrt{}, \sqrt[3]{}$ }

$[\Phi_1]$
34

$\mu^* + \lambda$

rhennig@ufl.edu
http://hennig.mse.ufl.edu

Sure Independence Screening



$$[\Phi_0] \quad \omega_{\log} \quad , \quad \mu^* \quad , \quad \lambda$$

$$3$$

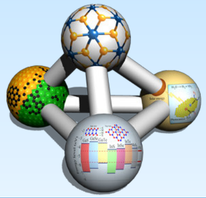
$$\{ \quad +, -, \quad , \quad \times, \div, \quad , \quad \exp, \log, \quad , \quad ^{-1}, ^{2}, ^{3}, \quad , \quad \sqrt{}, \sqrt[3]{} \quad \}$$

$$[\Phi_1] \quad \mu^* + \lambda \quad , \quad \omega_{\log} \times \lambda$$

$$34$$

rhennig@ufl.edu
http://hennig.mse.ufl.edu

Sure Independence Screening



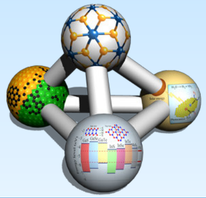
$[\Phi_0]$ 3 $\quad \omega_{\log}, \quad \mu^*, \quad \lambda$

$\{ \quad +, -, \quad , \quad \times, \div, \quad , \quad \exp, \log, \quad , \quad ^{-1}, ^2, ^3, \quad , \quad \sqrt{}, \sqrt[3]{} \quad \}$

$[\Phi_1]$ 34 $\quad \mu^* + \lambda, \quad \omega_{\log} \times \lambda, \quad \exp(\mu^*)$

rhennig@ufl.edu
http://hennig.mse.ufl.edu

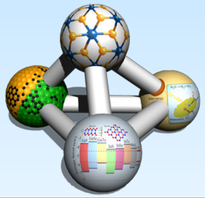
Sure Independence Screening



$$
\begin{array}{lllllllllll}
[\Phi_0] & & & & & & & & & & \\
3 & \omega_{\log} & , & \mu^* & , & \lambda & & & & & \\
\{ & +, -, & , & \times, \div, & , & \exp, \log, & , & {}^{-1}, {}^{2}, {}^{3}, & , & \sqrt{}, \sqrt[3]{} & \} \\
[\Phi_1] & & & & & & & & & & \\
34 & \mu^* + \lambda & , & \omega_{\log} \times \lambda & , & \exp(\mu^*) & , & \lambda^3 & & &
\end{array}
$$

rhennig@ufl.edu
http://hennig.mse.ufl.edu

Sure Independence Screening



$$[\Phi_0] \quad \omega_{\log} \quad , \quad \boxed{\mu^*} \quad , \quad \lambda$$

$$3$$

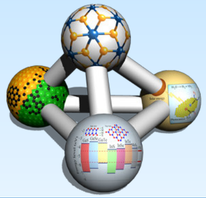
$$\{ \quad +, -, \quad , \quad \times, \div, \quad , \quad \exp, \log, \quad , \quad {}^{-1}, {}^{2}, {}^{3}, \quad , \quad \sqrt{\,}, \sqrt[3]{\,} \quad \}$$

$$[\Phi_1] \quad \mu^* + \lambda \quad , \quad \omega_{\log} \times \lambda \quad , \quad \exp(\mu^*) \quad , \quad \lambda^3 \quad , \quad \boxed{\sqrt{\mu^*}} \quad , \ldots$$

$$34$$

rhennig@ufl.edu
http://hennig.mse.ufl.edu

Sure Independence Screening



$[\Phi_0]$ 3 $\quad \omega_{\log} \quad , \quad \mu^* \quad , \quad \lambda$

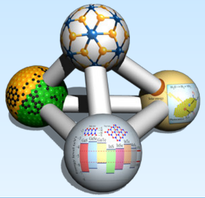
$\{ \quad +, -, \quad , \quad \times, \div, \quad , \quad \exp, \log, \quad , \quad ^{-1}, ^2, ^3, \quad , \quad \sqrt{}, \sqrt[3]{} \quad \}$

$[\Phi_1]$ 34 $\quad \mu^* + \lambda \quad , \quad \omega_{\log} \times \lambda \quad , \quad \exp(\mu^*) \quad , \quad \lambda^3 \quad , \quad \sqrt{\mu^*} \quad , \ldots$

$[\Phi_2]$ 1,342 $\quad \omega_{\log} \div \exp(\mu^*)$

rhennig@ufl.edu
http://hennig.mse.ufl.edu

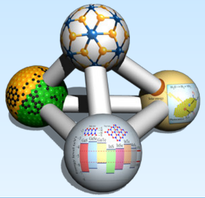
Sure Independence Screening



$$
\begin{array}{lccccccccc}
[\Phi_0] & & & & & & & & & \\
3 & \omega_{\log} & , & \mu^* & , & \lambda & & & & \\
\{ & +, -, & , & \times, \div, & , & \exp, \log, & , & {}^{-1}, {}^{2}, {}^{3}, & , & \sqrt{}, \sqrt[3]{} \quad \} \\
[\Phi_1] & & & & & & & & & \\
34 & \mu^* + \lambda & , & \omega_{\log} \times \lambda & , & \exp(\mu^*) & , & \lambda^3 & , & \sqrt{\mu^*} \quad , \ldots \\
[\Phi_2] & & & & & & & & & \\
1{,}342 & \omega_{\log} \div \exp(\mu^*) & , & \lambda^3 + \sqrt{\mu^*} & , & \sqrt[3]{\mu^* + \lambda} & , & \lambda^3 \times (\omega_{\log} \times \lambda) & , & \exp(\lambda^3) \quad , \ldots
\end{array}
$$

rhennig@ufl.edu
http://hennig.mse.ufl.edu

Sure Independence Screening



$[\Phi_0]$ 3: $\omega_{\log}$, μ^* , λ

{ $+, -,$, $\times, \div,$, exp, log, , $^{-1}, ^2, ^3,$, $\sqrt{}, \sqrt[3]{}$ }

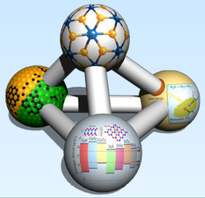
$[\Phi_1]$ 34: $\mu^* + \lambda$, $\omega_{\log} \times \lambda$, $\exp(\mu^*)$, λ^3 , $\sqrt{\mu^*}$, …

$[\Phi_2]$ 1,342: $\omega_{\log} \div \exp(\mu^*)$, $\lambda^3 + \sqrt{\mu^*}$, $\sqrt[3]{\mu^* + \lambda}$, $\lambda^3 \times (\omega_{\log} \times \lambda)$, $\exp(\lambda^3)$, …

$[\Phi_3]$ 3,414,094: $\exp(\lambda^3 + \sqrt{\mu^*})$,

rhennig@ufl.edu
http://hennig.mse.ufl.edu

Sure Independence Screening



$[\Phi_0]$ 3: $\omega_{\log}$, μ^* , λ

{ $+, -,$, $\times, \div,$, $\exp, \log,$, $^{-1}, ^2, ^3,$, $\sqrt{}, \sqrt[3]{}$ }

$[\Phi_1]$ 34: $\mu^* + \lambda$, $\omega_{\log} \times \lambda$, $\exp(\mu^*)$, λ^3 , $\sqrt{\mu^*}$, …

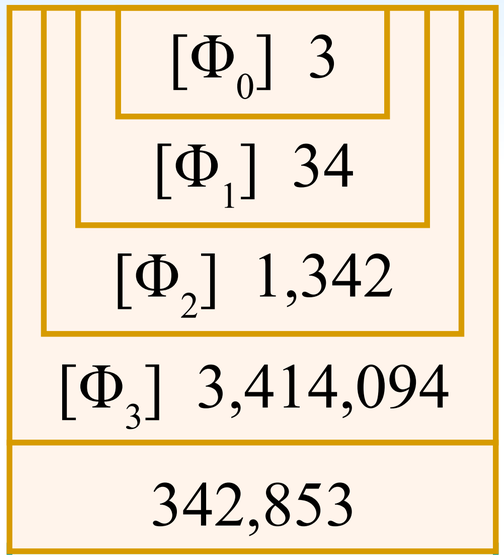
$[\Phi_2]$ 1,342: $\omega_{\log} \div \exp(\mu^*)$, $\lambda^3 + \sqrt{\mu^*}$, $\sqrt[3]{\mu^* + \lambda}$, $\lambda^3 \times (\omega_{\log} \times \lambda)$, $\exp(\lambda^3)$, …

$[\Phi_3]$ 3,414,094: $\exp(\lambda^3 + \sqrt{\mu^*})$, $\mu^* + \exp(\lambda^3)$, $\dfrac{\lambda^3 \times (\omega_{\log} \times \lambda)}{\lambda^3 + \sqrt{\mu^*}}$, …

Results in 3,414,094 analytic equations

rhennig@ufl.edu
http://hennig.mse.ufl.edu

Sure Independence Screening



$$\omega_{\log} \ , \ \mu^* \ , \ \lambda \ , \ \ldots$$

$$\omega_{\log} \times \lambda \ , \ \sqrt{\mu^*} \ , \ \lambda^3 \ , \ \ldots$$

$$\lambda^3 \times (\omega_{\log} \times \lambda) \ , \ \lambda^3 + \sqrt{\mu^*} \ , \ \ldots$$

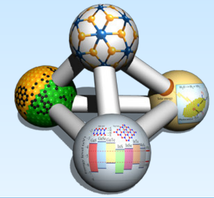
$$\lambda^3 \times (\omega_{\log} \times \lambda) \ / \ (\lambda^3 + \sqrt{\mu^*}) \ , \ \ldots$$

Sure Independence Screening

Select the equations with the highest linear correlation to T_c, inner product > 0.5

rhennig@ufl.edu
http://hennig.mse.ufl.edu

Use Physical Constraints to Reduce Feature Space

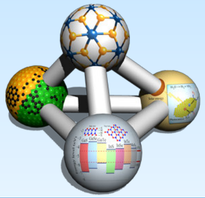


$[\Phi_0]$ 3	$\omega_{\log}\ ,\ \mu^*\ ,\ \lambda\ ,\ \ldots$
$[\Phi_1]$ 34	$\omega_{\log} \times \lambda\ ,\ \sqrt{\mu^*}\ ,\ \lambda^3\ ,\ \ldots$
$[\Phi_2]$ 1,342	$\lambda^3 \times (\omega_{\log} \times \lambda)\ ,\ \lambda^3 + \sqrt{\mu^*}\ ,\ \ldots$
$[\Phi_3]$ 3,414,094	$\lambda^3 \times (\omega_{\log} \times \lambda) \,/\, \left(\lambda^3 + \sqrt{\mu^*}\right)\ ,\ \ldots$
342,853	Sure Independence Screening
22,552	Dimensions
15,886	$\lambda \rightarrow 0$ Limit
10,839	Strictly Positive
6,021	Finite, Continuous, Real, Monotonic

Tools available at https://github.com/henniggroup/symbolic-regression-utilities

rhennig@ufl.edu
http://hennig.mse.ufl.edu

Evaluating Model Performance

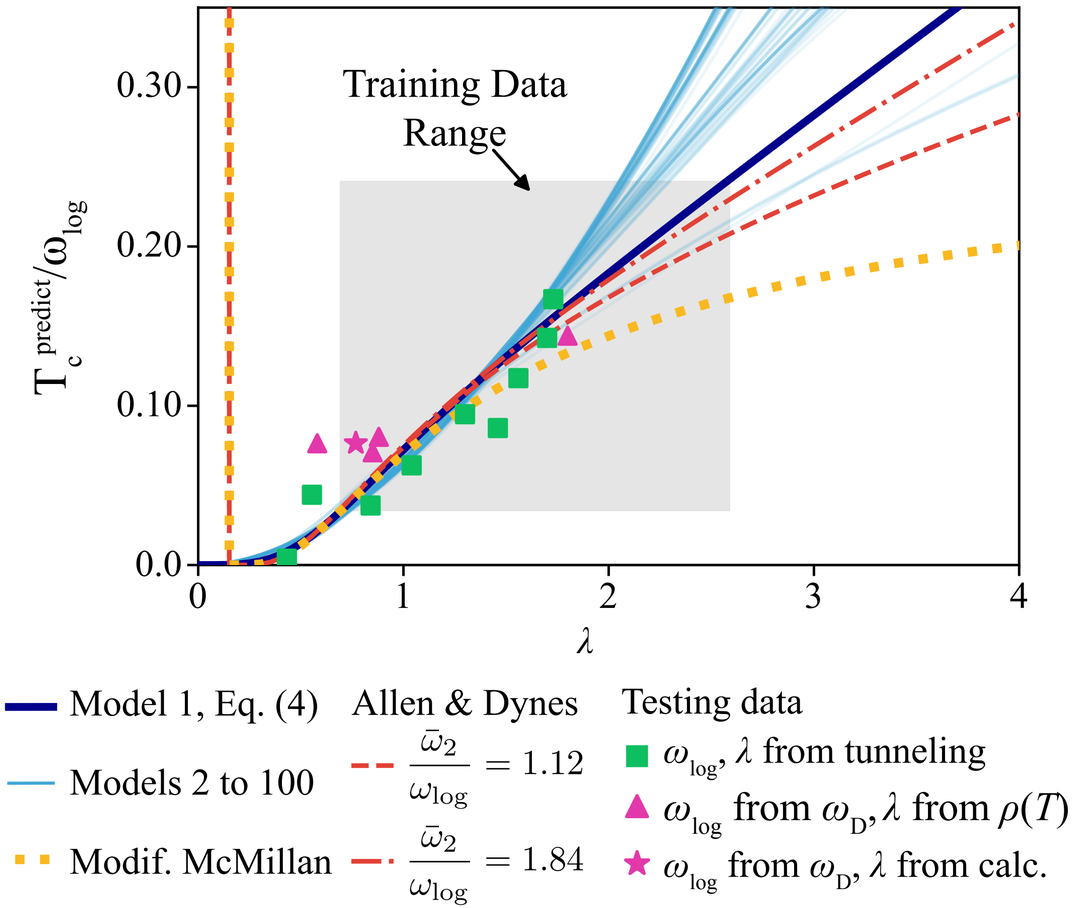


$$6{,}021 \text{ features: } \exp(\lambda^3 + \sqrt{\mu^*}), \quad \mu^* + \exp(\lambda^3) \quad , \frac{\lambda^3 \times (\omega_{\log} \times \lambda)}{\lambda^3 + \sqrt{\mu^*}} \quad , \quad \ldots$$

- Fit multiplicative factor to features, e.g.,

$$y' = C\frac{\lambda^4 \omega_{\log}}{\lambda^2 + \sqrt{\mu^*}}$$

- Evaluate root-mean square (relative) error with leave-one-out cross-validation
- Figure shows distribution of resulting functions for $\mu^* = 0.1$
- Narrow distribution of training and testing data
- Data for large and small λ would be helpful

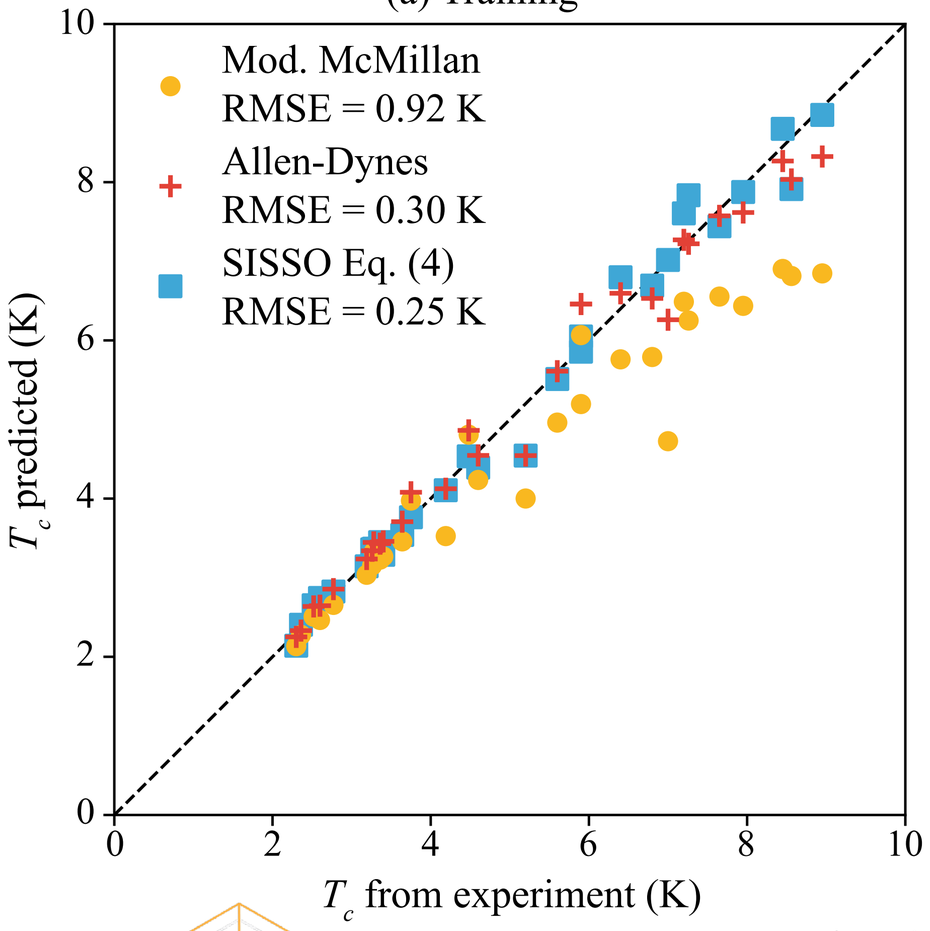


rhennig@ufl.edu
http://hennig.mse.ufl.edu

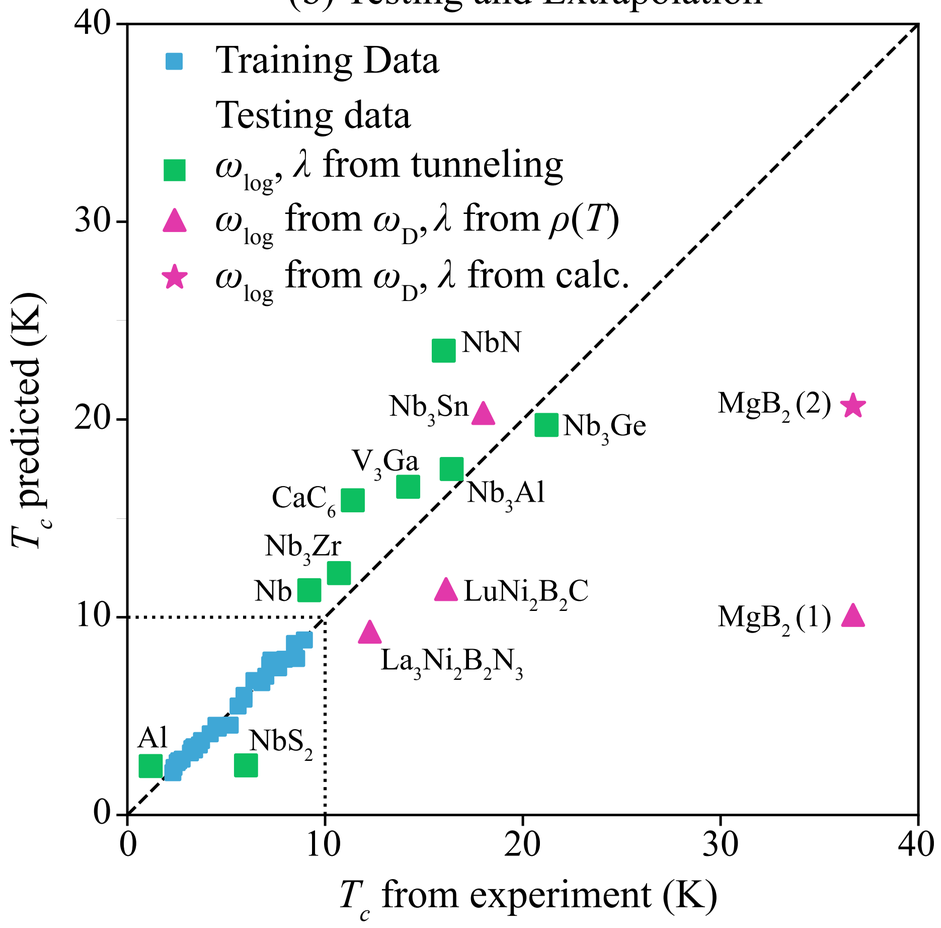
Best Model Performance

$$T_c^{\text{SISSO}} = 0.09525 \frac{\lambda^4 \omega_{\log}}{\lambda^3 + \sqrt{\mu^*}}$$

(a) Training



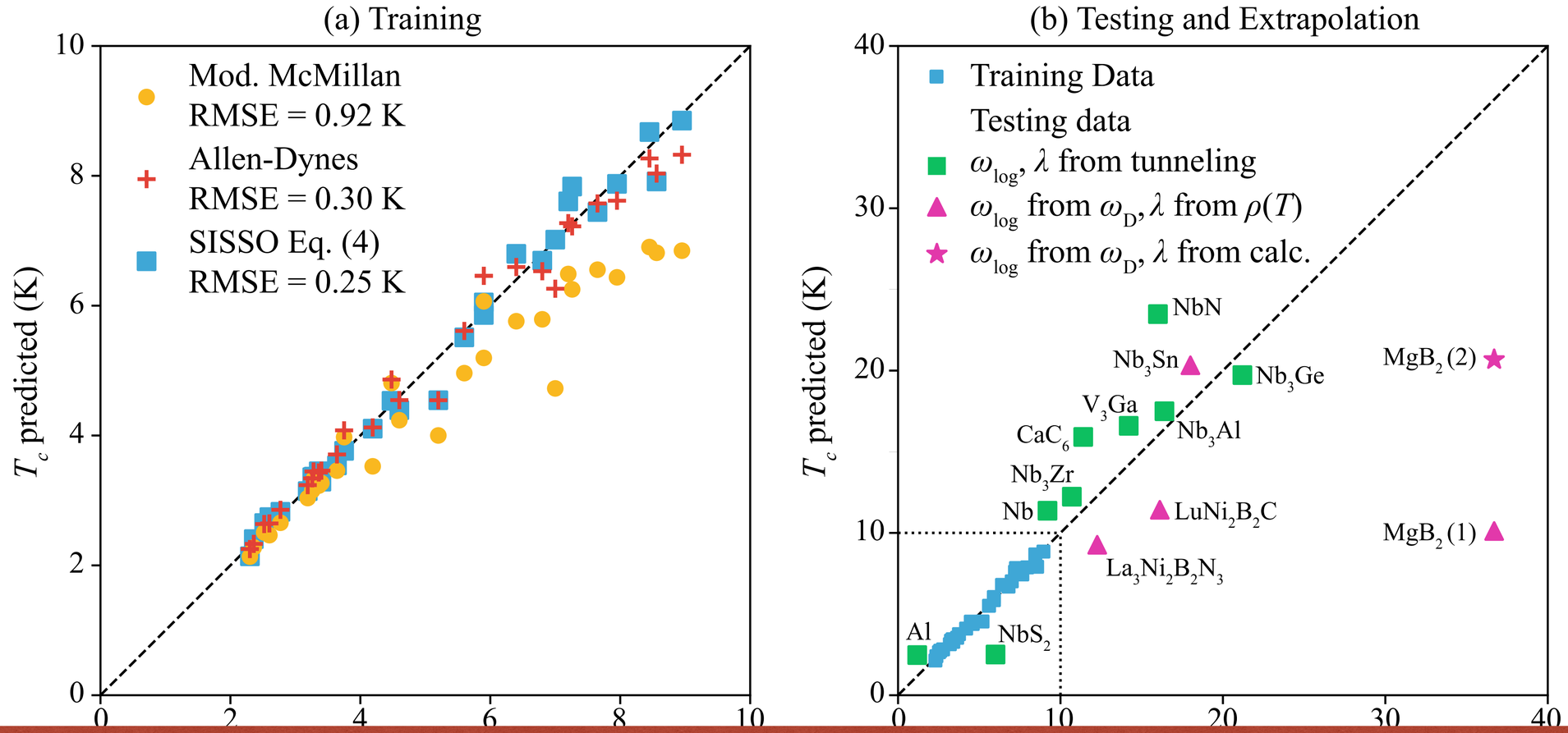
(b) Testing and Extrapolation



rhennig@ufl.edu
http://hennig.mse.ufl.edu

Best Model Performance

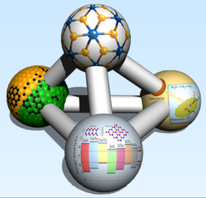
$$T_c^{\text{SISSO}} = 0.09525 \frac{\lambda^4 \omega_{\log}}{\lambda^3 + \sqrt{\mu^*}}$$



Machine-learning predicts isotropic electron-phonon superconductors

Outliers (MgB_2, NbS_2) indicate importance of anisotropic electron-phonon coupling.

Dimensionality and Complexity



Model	CV-RMSE (K) Training	RMSE (K) Testing	Materials Parameters	Numerical Coefficients
SISSO	0.25	3.2	3	1
mod. McMillan	0.92	–	3	4
Allen-Dynes	0.30	–	4	7

Machine-learned model has small RMSE and low computational complexity.

rhennig@ufl.edu
http://hennig.mse.ufl.edu

Dimensionality and Complexity

Can the machine-learning identify relevant materials parameters?

- Fit models with up to 7 materials parameters for 29 materials
- Best model:

$$T_c^{\mathrm{SISSO}} = 0.09525 \frac{\lambda^4 \omega_{\log}}{\lambda^3 + \sqrt{\mu^*}}$$

CV-RMSE = 0.25 K

- Second best model:

$$T_c = -0.059 \left(\omega_2 - \omega_1 - \frac{\omega_2}{\lambda} \right) \frac{\lambda^3}{\sqrt[3]{\lambda}}$$

CV-RMSE = 0.27K

- Adding parameters beyond $\omega_{\log}$, λ, μ^* does not improve description

TABLE I. Parameters[a] of superconductors derived from tunneling measurements. The value of μ^* is renormalized from previously reported values as described in the text.

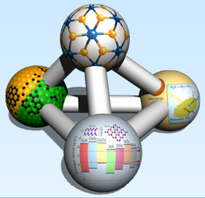
Material	$\omega_{\log}$ (K)	$\overline{\omega}_1$ (K)	$\overline{\omega}_2$ (K)	λ	ω_{ph} (K)	$\mu^*(\omega_{ph})$	T_c (K)	η (eV/Å^2)
Pb	56	60	65	1.55	110	0.105	7.2	2.4
In	68	79	89	0.805	179	0.097	3.40	1.3
Sn	99	110	121	0.72	209	0.092	3.75	2.2
Hg	29	38	49	1.6	162	0.098	4.19	1.4
Tl	52	58	64	0.795	127	0.111	2.36	1.2
Ta	132	140	148	0.69	228	0.093	4.48	4.9
α-Ga	55	77	101	1.62	291	0.095	8.56	2.1
β-Ga	87	108	129	0.97	285	0.092	5.90	2.0
$Tl_{0.9}Bi_{0.1}$	48	55	62	0.78	120	0.099	2.30	1.1
$Pb_{0.4}Tl_{0.6}$	48	56	62	1.15	121	0.094	4.60	1.6
$Pb_{0.6}Tl_{0.4}$	50	57	62	1.38	119	0.103	5.90	2.0
$Pb_{0.8}Tl_{0.2}$	50	56	61	1.53	116	0.101	6.80	2.1
$Pb_{0.6}Tl_{0.2}Bi_{0.2}$	48	53	58	1.81	112	0.111	7.26	2.3
$Pb_{0.9}Bi_{0.1}$	50	56	60	1.66	108	0.081	7.65	2.2
$Pb_{0.8}Bi_{0.2}$	46	52	57	1.88	109	0.093	7.95	2.3
$Pb_{0.7}Bi_{0.3}$	47	52	57	2.01	110	0.092	8.45	2.4
$Pb_{0.65}Bi_{0.35}$	45	50	55	2.13	110	0.093	8.95	2.4
$In_{0.9}Tl_{0.1}$	63	75	86	0.85	176	0.103	3.28	1.4
$In_{0.73}Tl_{0.27}$	55	67	77	0.93	166	0.110	3.36	1.4
$In_{0.67}Tl_{0.33}$	57	68	79	0.90	167	0.110	3.26	1.4
$In_{0.57}Tl_{0.43}$	53	64	74	0.85	165	0.117	2.60	1.3
$In_{0.5}Tl_{0.5}$	53	64	73	0.83	163	0.110	2.52	1.3
$In_{0.27}Tl_{0.73}$	42	53	63	1.09	151	0.094	3.64	1.4
$In_{0.17}Tl_{0.83}$	45	55	63	0.98	144	0.101	3.19	1.3
$In_{0.07}Tl_{0.93}$	49	56	63	0.89	131	0.107	2.77	1.3
In_2Bi	46	57	67	1.40	174	0.096	5.6	1.6
Sb_2Tl_7	37	48	58	1.43	134	0.102	5.2	1.6
Bi_2Tl	47	53	59	1.63	120	0.101	6.4	2.1
α-$Pb_{0.45}Bi_{0.55}$	29	38	47	2.59	128	0.116	7.0	2.1

[a]Tabulation of the data used to derive these parameters is available in J. M. Rowell, W. L. McMillan, and R. C. Dynes, J. Phys. Chem. Ref. Data (to be published).

Machine-learning identifies most relevant materials parameters in agreement with McMillan and Allen & Dynes

rhennig@ufl.edu
http://hennig.mse.ufl.edu

Dimensionality and Complexity



Can increasing the number of numerical coefficients improve description

- Include multiplicative and additive numerical coefficient to every materials parameter
- Best model:

$$T_c = \omega_{\log}\left(\frac{0.233 - 0.0170\lambda}{1.28\mu^* + 0.00784} + (0.791\lambda - 1.408)^3\right)\left(0.0655\lambda + 0.00530 - \frac{0.000780}{1.206\mu^* - 0.0725}\right)$$

CV-RMSE = 0.19 K

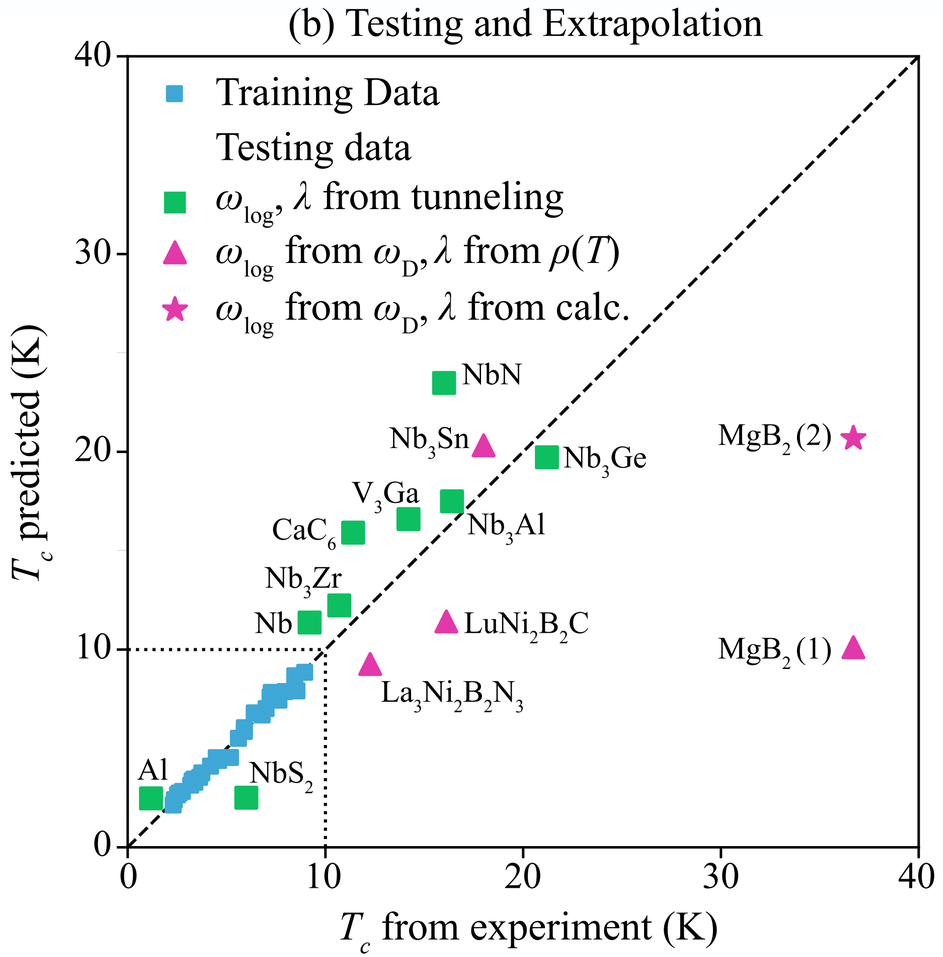
- Significant improvement from 0.25 to 0.19 K, however,
  at cost of significant more complexity (increase from 1 to 11 numerical coefficients)
  and reduced physical interpretability (high power of λ)

rhennig@ufl.edu
http://hennig.mse.ufl.edu

Conclusions

- Machine-learning of analytic equation with fewer parameters
- Identification of relevant physical parameters
- Use of analytic expressions and physical constraints can help overcome small-data problem
- Predict known superconductors of same type as the original Allen-Dynes dataset
- Anomalous outliers suggests need for new descriptors anisotropy of the electron-phonon interaction

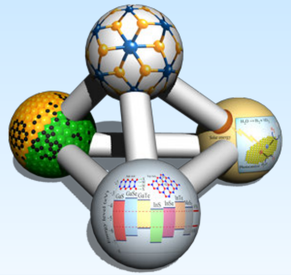
$$T_c^{\mathrm{SISSO}} = 0.09525 \frac{\lambda^4 \omega_{\log}}{\lambda^3 + \sqrt{\mu^*}}$$



rhennig@ufl.edu
http://hennig.mse.ufl.edu

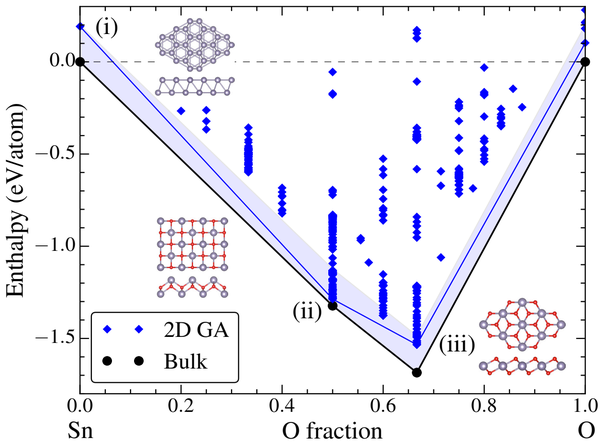
Machine-learning for materials and physics discovery through symbolic regression and kernel methods

Stephen R. Xie, Shreyas Honrao, and Richard G. Hennig, University of Florida



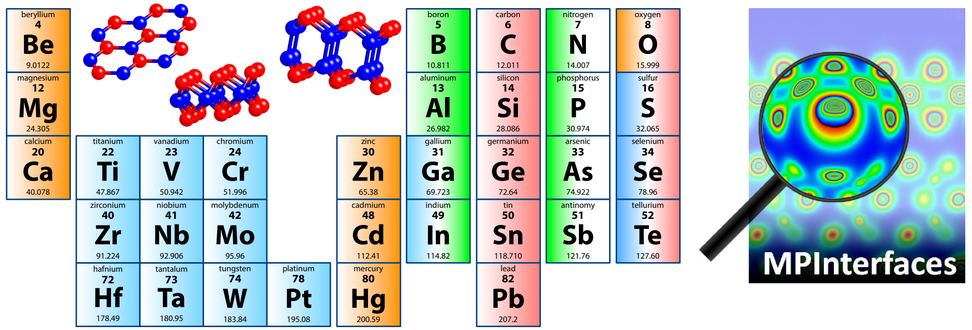
Search for materials

- Structure prediction by genetic algorithms
- Machine learning of energy landscapes using distribution functions
- Learning of Allen-Dynes equation for T_c
- Use of analytic equations and physical constraints to overcome small data problem

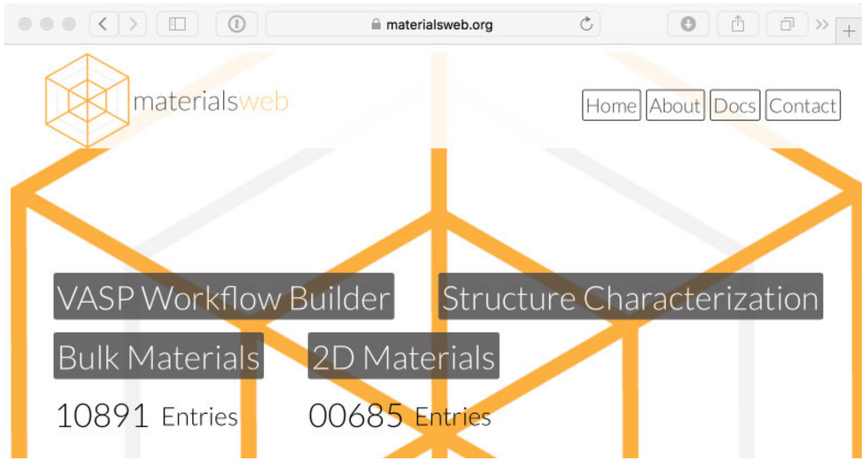


GASP - Genetic algorithm and machine learning for structure predictions

MPInterfaces - High throughput framework for 2D materials



Data available at http://materialsweb.org



Powered by
MPInterfaces & materialsweb

Open source available at https://github.com/henniggroup