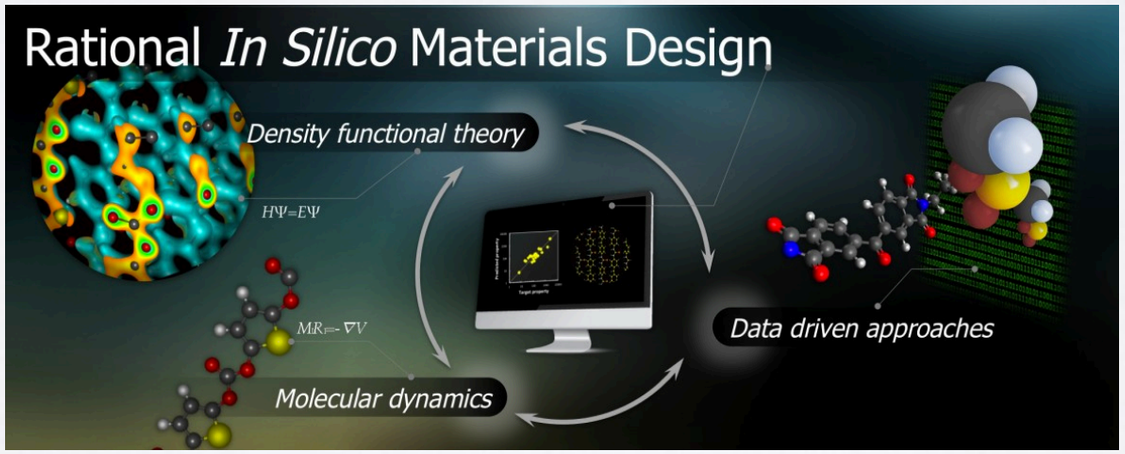
MACHINE LEARNING IN MATERIALS SCIENCE

Rampi Ramprasad
University of Connecticut



UCONN
UNIVERSITY OF CONNECTICUT

http://rampi.ims.uconn.edu

RAMPRASAD GROUP

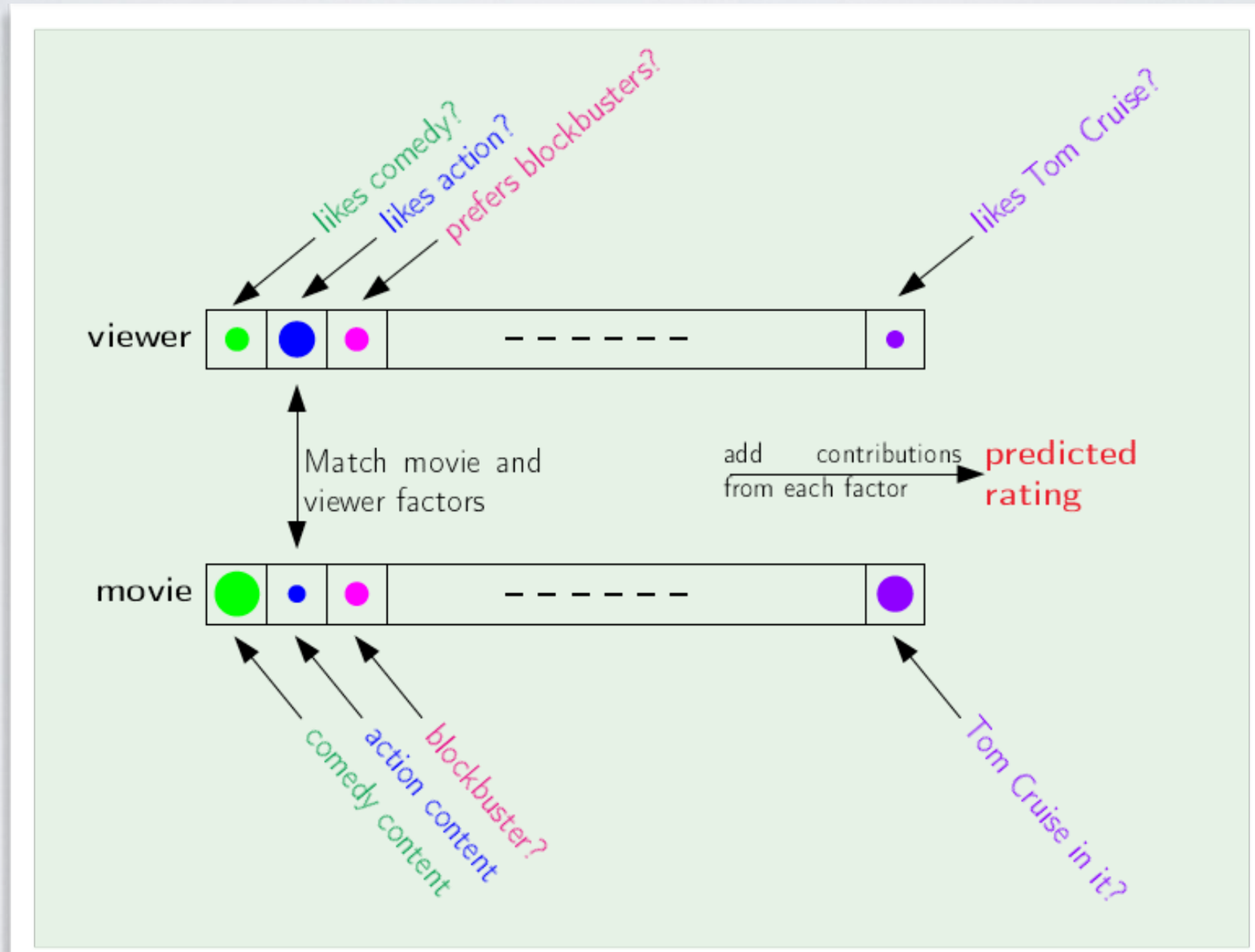
WHAT IS MACHINE LEARNING?

- **The context**: A pattern exists … but we cannot pin it down mathematically … we have data on it though …

- **The concept**: Build a heuristic, predictive model … based purely on past experience/data … rather than explicitly solving equations underlying known laws

- The learning process should be "adaptive", i.e., as more data accumulates, the predictive power should increase

EXAMPLE: MOVIE RATING



Courtesy: Learning from data, Abu-Mustafa, Magnon-Ismail & Lin

EXAMPLE: CREDIT APPROVAL

Applicant information

age	23 years
gender	male
annual salary	$30,000
years in residence	1 year
years in job	1 year
current debt	$15,000
...	...

Approve credit?

Courtesy: Learning from data, Abu-Mustafa, Magnon-Ismail & Lin

LEARNING COMPONENTS

Input: $\mathbf{x}$	(customer application)
Output: y	(good/bad customer?)
Data: $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)$	(historical records)
Hypothesis: $g(\mathbf{x}) = y$	(To be determined, refined, and used in the future)

"Training set" versus "test set"

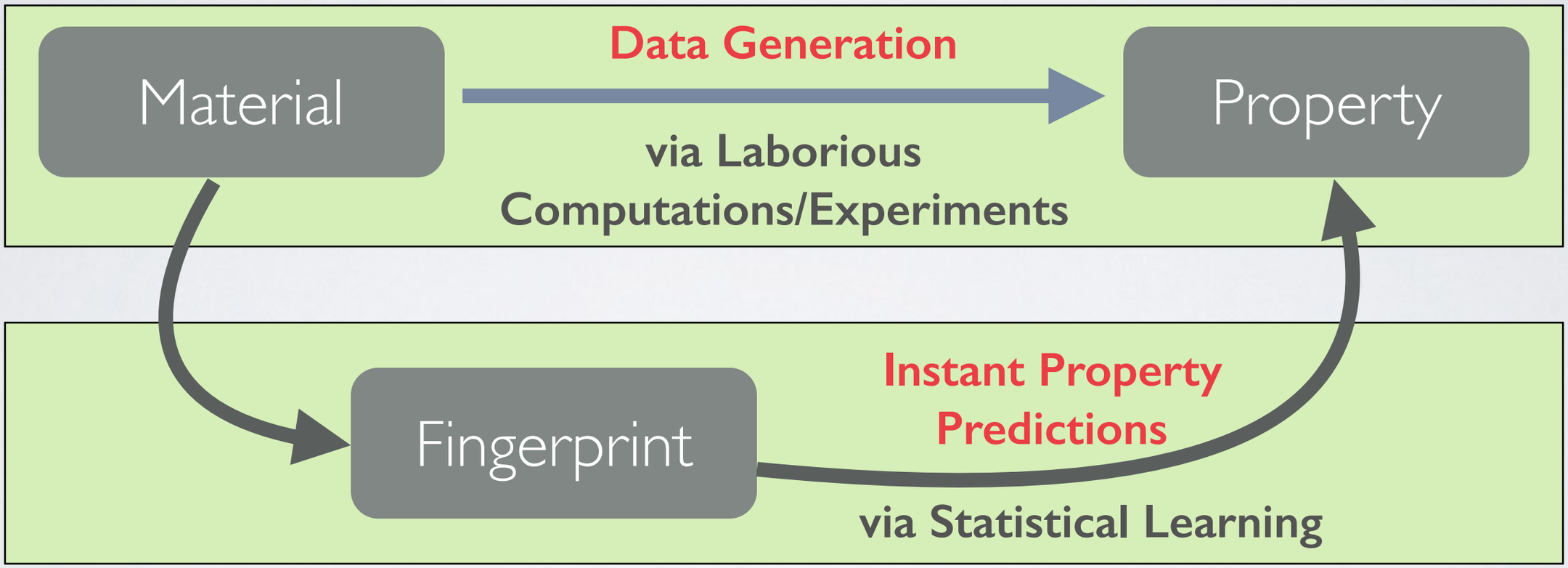
WHEN TO USE MACHINE LEARNING?

- When fundamental laws underlying a process don't exist [e.g., social science problems]

- When such fundamental laws may exist, but are enormously complex [e.g., weather prediction]

- When we have a lot of data and we are looking for simple rules and correlations [e.g., Hall-Petch equation]

MACHINE LEARNING IN MATERIALS SCIENCE



"Machine learning in materials science: Recent progress and emerging applications,"
Mueller, Kusne & Ramprasad, *Reviews of Computational Chemistry* (2016)

"Atomistic calculations and materials informatics: A review"
Ward & Wolverton, *Current Opinion in Solid State and Materials Science* (2016)

FINGERPRINT IS KEY!

- The "fingerprint" is a numerical representation of the material
- It should be defined based on the application & domain knowledge
- It should be intuitive, and inexpensive to compute
- It should be invariant to transformations of the material, such as translation, rotation, and permutations of like elements
- Fingerprints can be macroscopic or microscopic

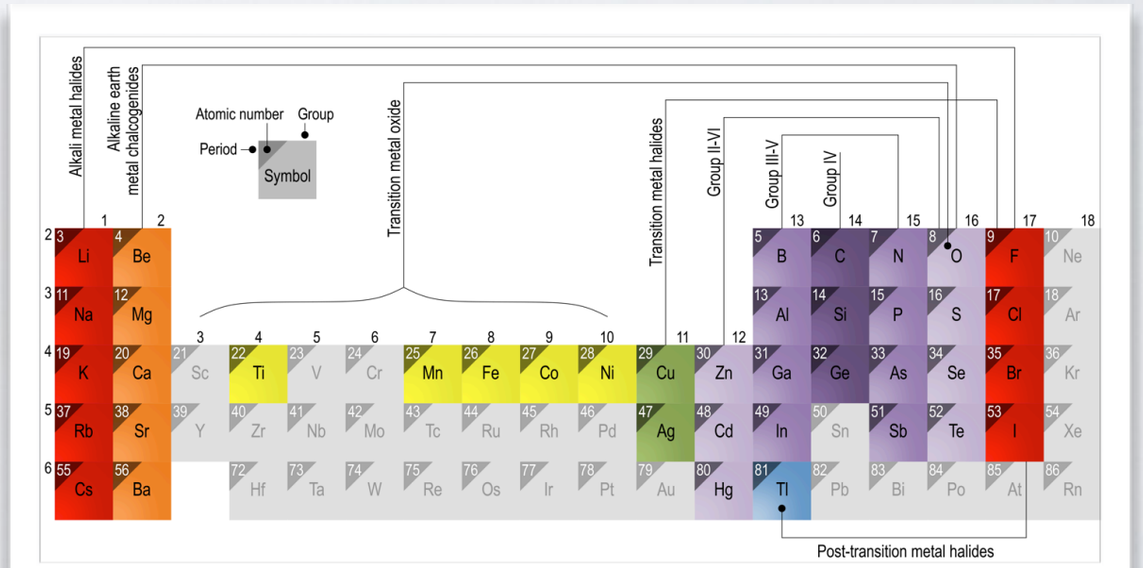
EXAMPLE 1: COARSE FINGERPRINTS
PREDICTING ELECTRICAL BREAKDOWN

Predicting the *intrinsic* electrical breakdown field of an insulator from first principles is difficult …

… it is determined by the balance between energy gained by an electron from the field and energy lost to phonons …

… but can the breakdown field be estimated rapidly using a simple heuristic model ?

Dependence on Chemistry?

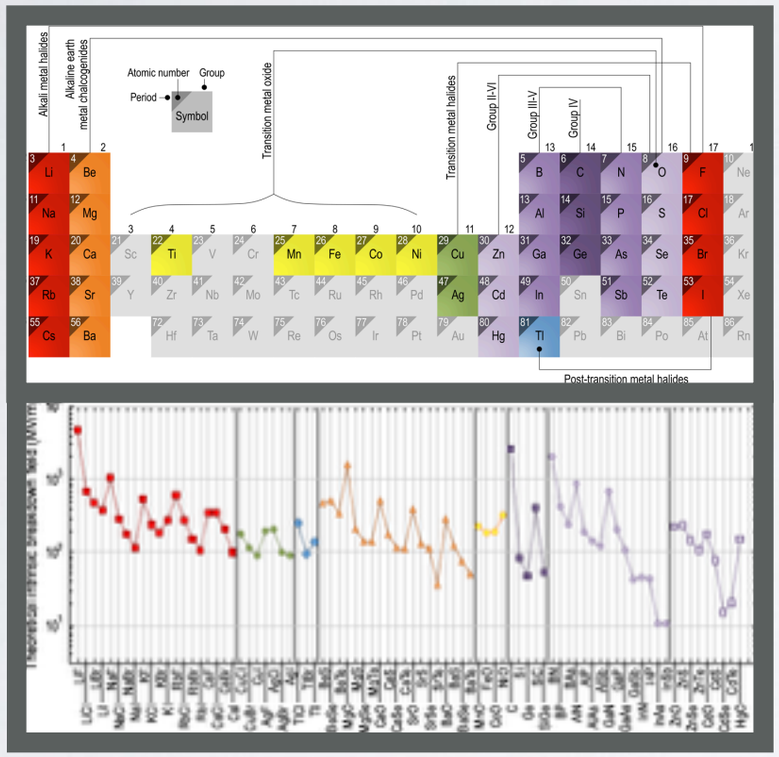


… consider 82 binary octets

Frohlich, *Nature* 151, 339 (1943)
Sun, Boggs & Ramprasad, *Appl. Phys. Lett.* 101, 132906 (2012)
Kim, Pilania & Ramprasad, *Chemistry of Materials* 28, 1304 (2016)

LEARNING FROM DATA

Intrinsic breakdown field of 82 binary octets



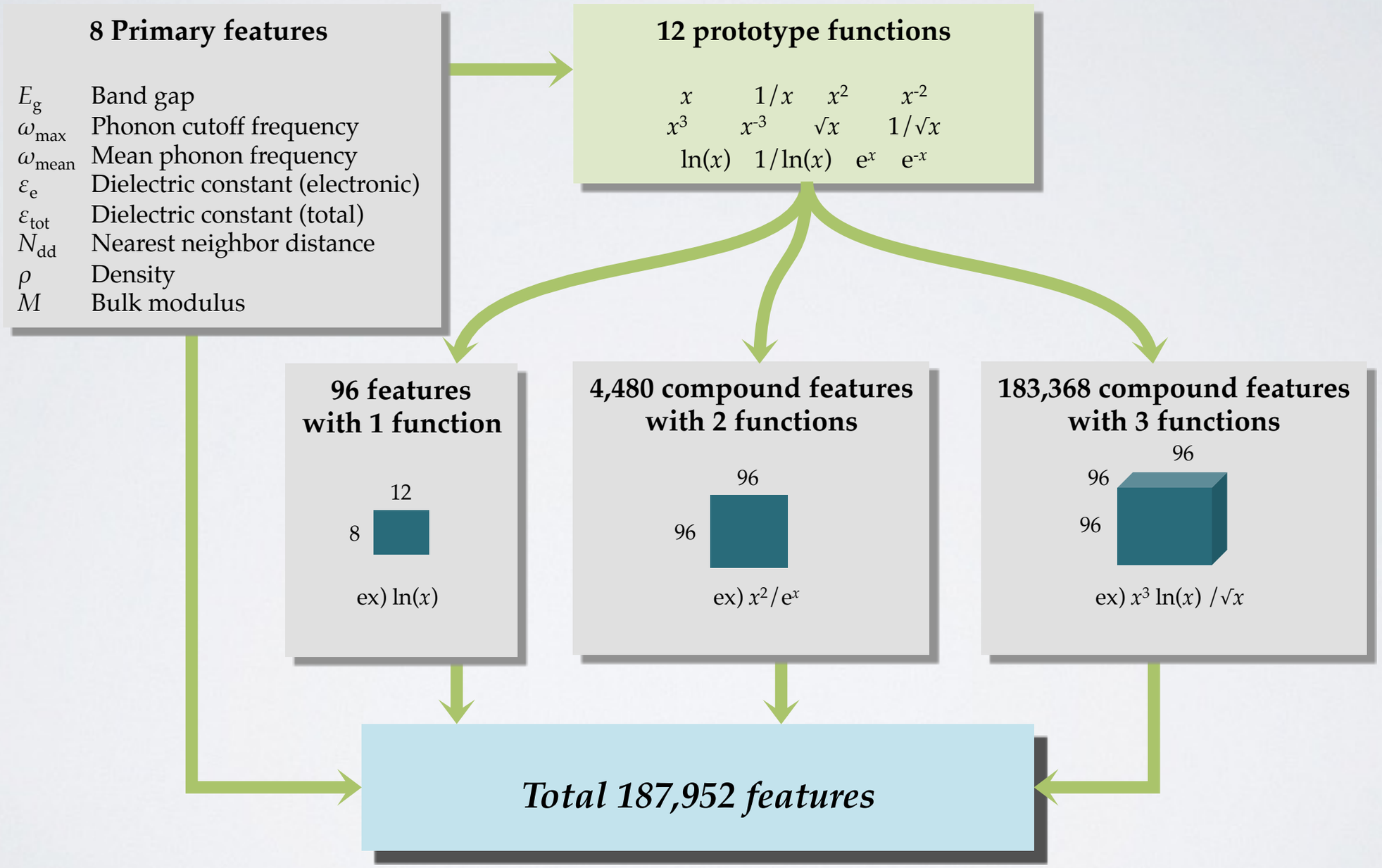
Easily accessible material properties

Band gap
Phonon cutoff frequency
Mean phonon frequency
Dielectric constant (electronic)
Dielectric constant (total)
Nearest neighbor distance
Density
Bulk modulus

$F_b = f(A,B,\ldots)?$

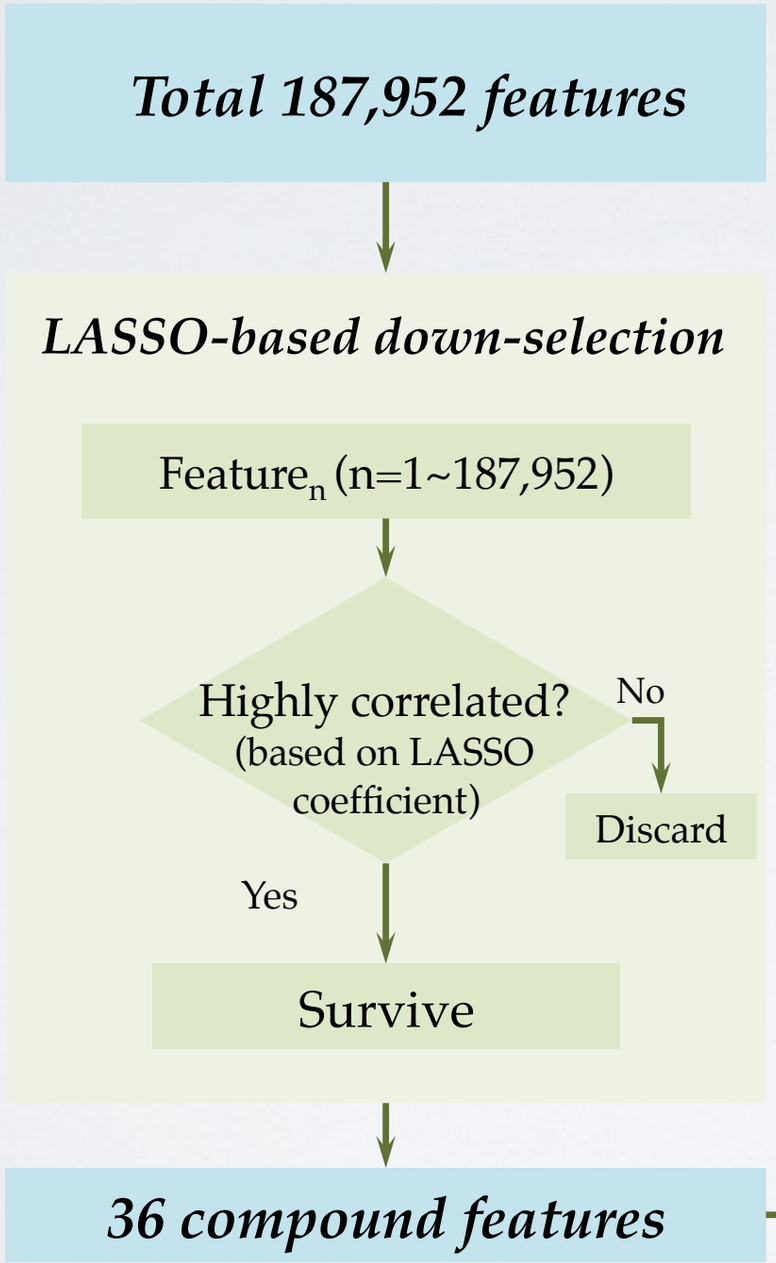
Correlation analysis & Machine learning

FEATURE CREATION



See also: Ghiringhelli, et al, *Phys. Rev. Lett.* (2015)

FEATURE SELECTION WITH "LASSO"



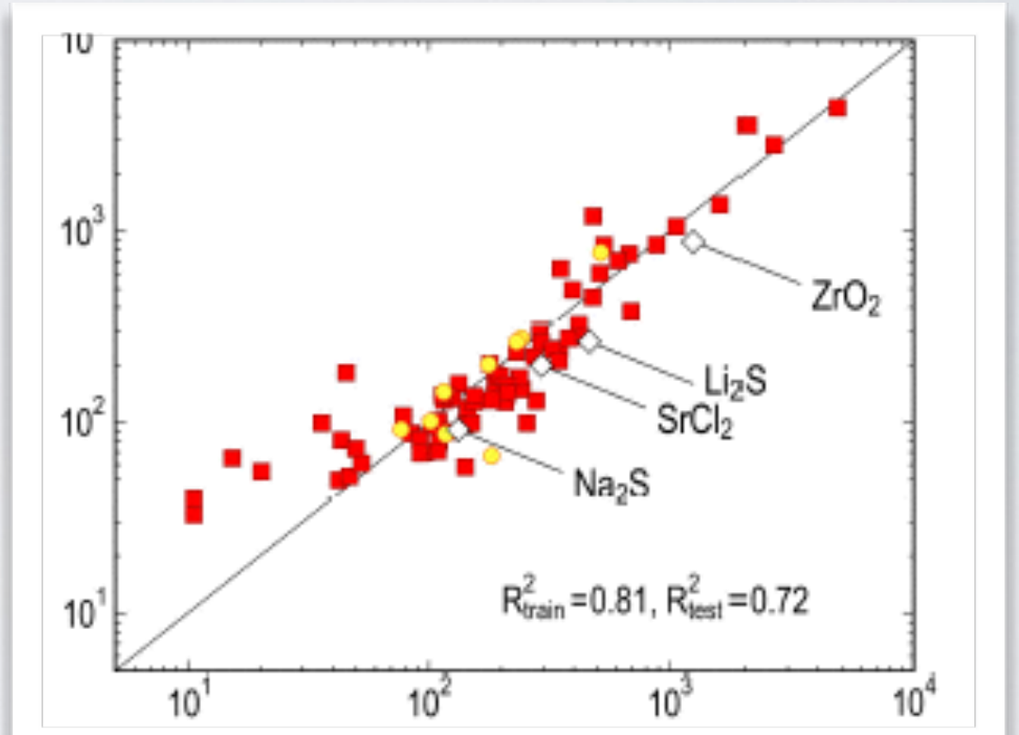
Ranking	Compound Feature	Absolute Pearson correlation /w $\ln F_b$
1	$\ln E_g \ln\omega_{max} / \sqrt{d_{NN}}$	0.899
2	$\sqrt{\omega_{max}} \sqrt{E_g}$	0.890
3	$\sqrt{\omega_{max}} \ln E_g$	0.890
4	$\sqrt{E_g} \ln\omega_{max}$	0.889
5	$\sqrt{E_g} / d_{NN}$	0.885
6	$\ln E_g / d_{NN}^2$	0.883
7	$\ln E_g / \exp(d_{NN})$	0.880
8	$\sqrt{E_g} / \ln d_{NN}$	0.879
9	$\omega_{max} \sqrt{E_g} / \ln\omega_{mean}$	0.871
10	$\sqrt{\omega_{max}} / \sqrt{\varepsilon_e}$	0.869
…	…	…
36	$\sqrt{\varepsilon_{tot}} \sqrt{E_g}$	0.480

PREDICTION (& DESIGN?)

The finding

$$F_b = 24.442\exp(0.315\sqrt{E_g\omega_{max}})$$

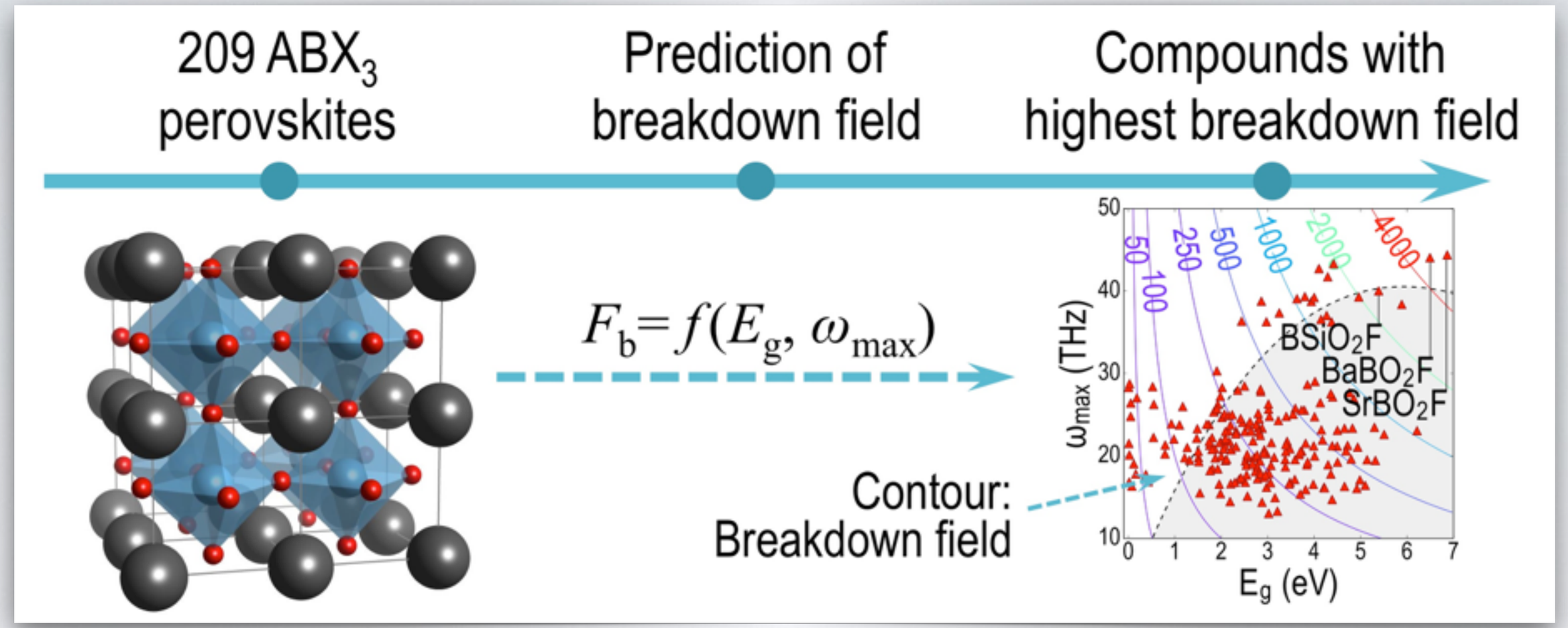
Machine Learning



DFT

Note: This is a heuristic, not a law; we cannot take it too seriously!

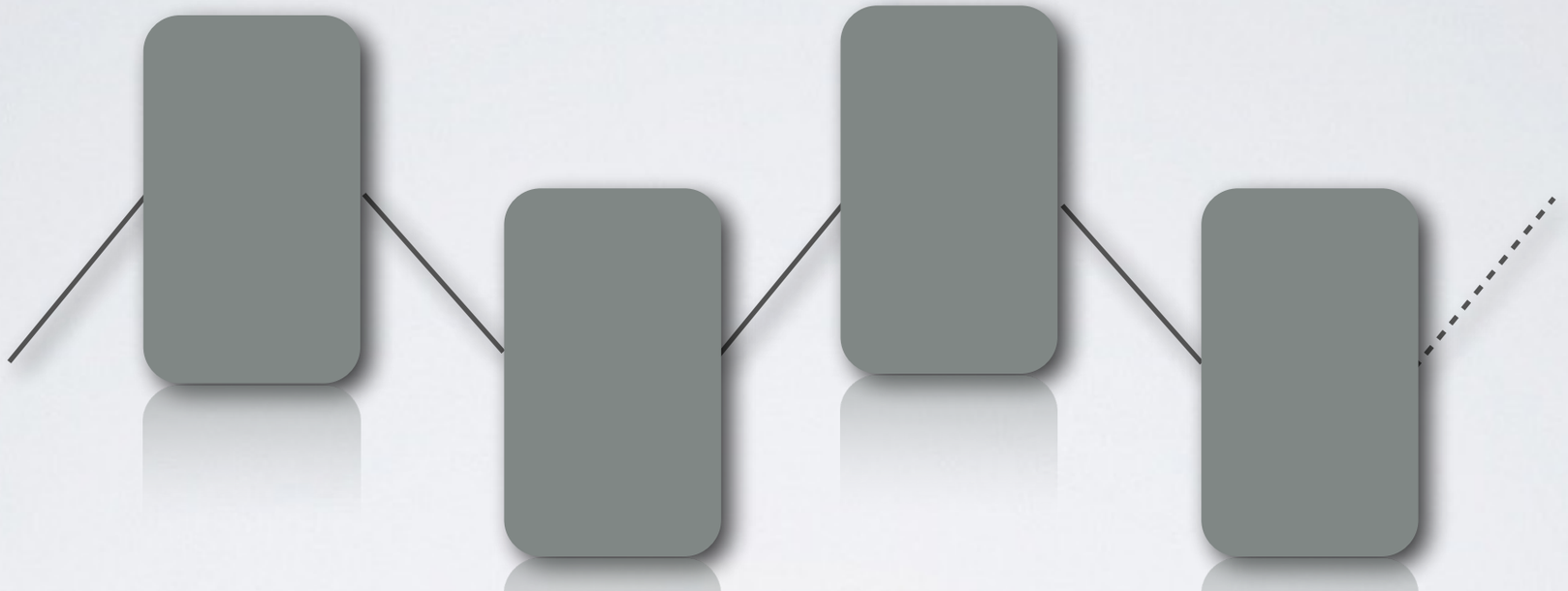
AN APPLICATION



Kim, Pilania & Ramprasad, *J. Phys. Chem. C* 120, 14575 (2016)
Kim, Pilania & Ramprasad, *Chemistry of Materials* 28, 1304 (2016)

EXAMPLE 2: MEDIUM FINGERPRINTS

PREDICTING POLYMER PROPERTIES



Organic Blocks: CH_2, CO, CS, O, NH, C_6H_4, C_4H_2S

Fingerprint Choice 1: "Singles" - number of CH_2, NH, CO, C_6H_4, etc.

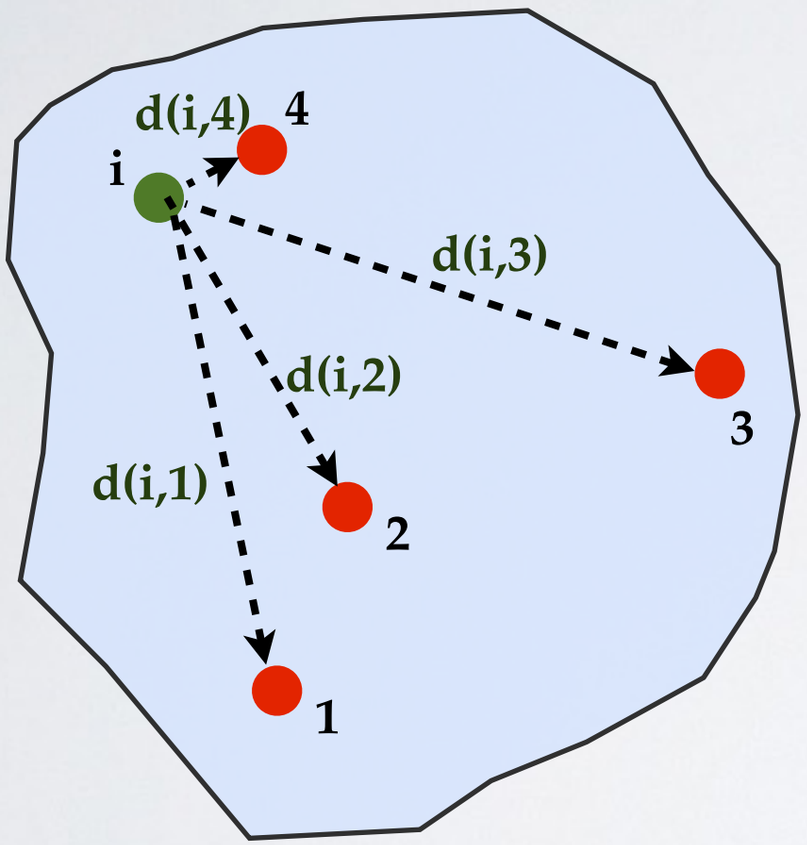
Fingerprint Choice 2: "Doubles" - number of CH_2-CH_2, NH-CH_2, CO-NH, etc.

Fingerprint Choice 3: "Triples" - number of CH_2-CH_2-CH_2, NH-CH_2-CO, etc.

Mannodi-Kanakkithodi et al., *Sci. Rep.* 6, 20952 (2016)
Tran et al., *Phys. Rev. B* 92, 014106 (2015)
Pilania et al., *Sci. Rep.* 3, 2810 (2013)

THE LEARNING FRAMEWORK

Kernel ridge regression



Measure of similarity: Euclidean distance

$$d(i,j) = \sqrt{\left(\left|x_{i_1} - x_{j_1}\right|^2 + \left|x_{i_2} - x_{j_2}\right|^2 + \ldots + \left|x_{i_p} - x_{j_p}\right|^2\right)}$$

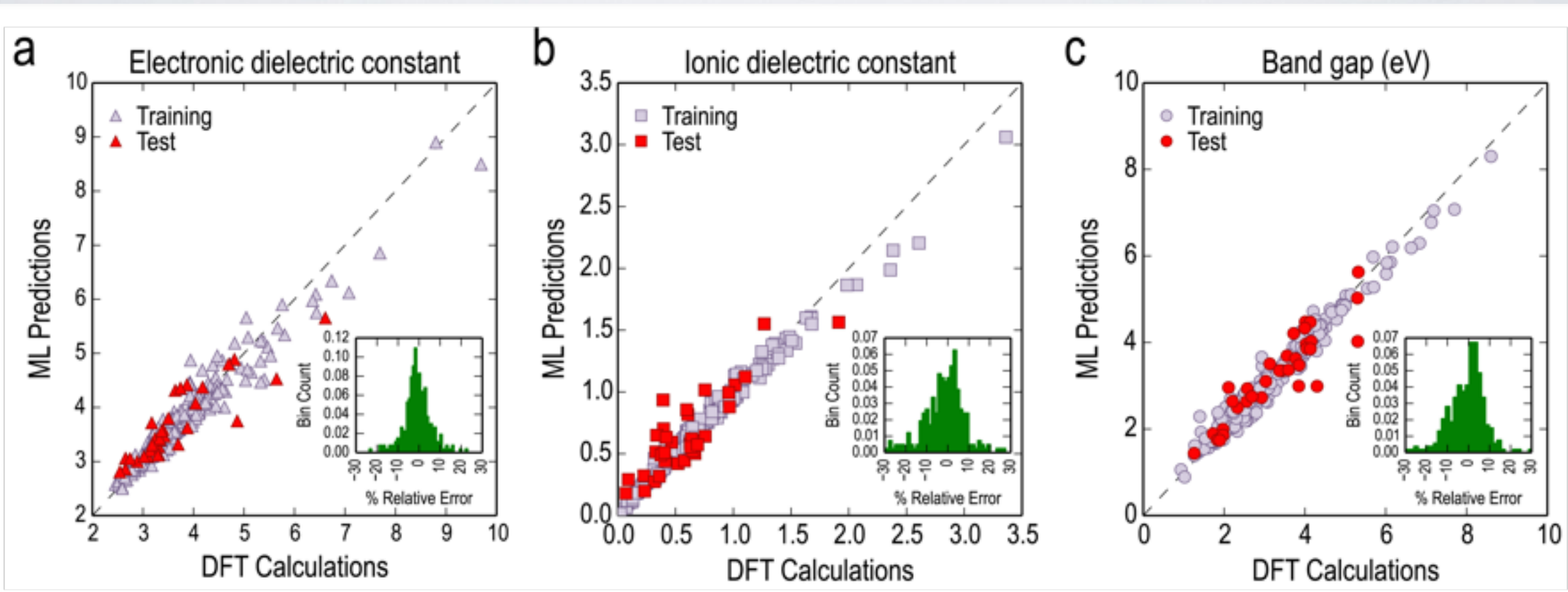
Property Estimation:
Sum of weighted Gaussians

$$Y^{est}(P) = \sum_{i=1}^{N} \alpha_i \exp\left[-\frac{1}{2\sigma^2} d(P,P_i)^2\right]$$

Rupp, Tkatchenko, Muller, Von Lilienfeld, "Fast and accurate modeling of molecular atomization energies with machine learning", *Phys. Rev. Lett.* 108, 058301 (2012)

INSTANT PROPERTY PREDICTIONS

… using a DFT dataset for about 300 organic polymers



Mannodi-Kanakkithodi et al., *Scientific Reports* 6, 20952 (2016)
Tran et al., *Phys. Rev. B* 92, 014106 (2015)

Polymer Dielectrics Property Prediction

Obtain properties predicted by DFT and/or machine learning [1,2]

Choose the building blocks (co-polymers) of polymer repeat unit.

NH - CO - NH - C6H4 - Block 5 - Block 6 - Block 7 - Block 8 - Block 9 - Block 10 -

Predict Properties

References

1. A. Mannodi-Kanakkithodi, G. Pilania, T. D. Huan, T. Lookman, R. Ramprasad Machine learning strategy for accelerated design of polymer dielectrics Scientific Reports, 6, 20952 (2016). Article

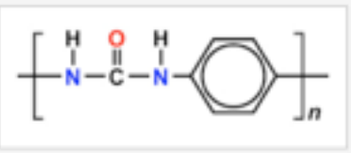
2. T. D. Huan, A. Mannodi-Kanakkithodi, R. Ramprasad Accelerated materials property predictions and design using motif-based fingerprints Physical Review B, 92, 014106 (2015). Article

Target polymer summary

Repeat unit: -(-NH-CO-NH-C_6H_4-)$_n$-

Number of building blocks (co-polymers): 4

Chemical formula: $C_7H_6ON_2$



Predicted properties using machine learning model

Method	Polymer	Dielectric Constant	Refractive Index	Band Gap, HSE06 (eV)
Machine Learning [1]	$C_7H_6ON_2$ NH-CO-NH-C6H4	4.95	1.98	3.30
Machine Learning [2]	$C_7H_6ON_2$ NH-CO-NH-C6H4	4.90	1.93	3.28
DFT (ID: 0299)	$O_2C_{14}H_{12}N_4$ NH-CO-NH-C6H4	4.91	1.97	3.45

Polymer Dielectrics Property Prediction

Obtain properties predicted by DFT and/or machine learning [1,2]

Choose the building blocks (co-polymers) of polymer repeat unit.

NH - CO - NH - C6H4 - CH2 - C4H2S - CS - O - CH2 - Block 10 -

Predict Properties

References

1. A. Mannodi-Kanakkithodi, G. Pilania, T. D. Huan, T. Lookman, R. Ramprasad Machine learning strategy for accelerated design of polymer dielectrics Scientific Reports, 6, 20952 (2016). Article

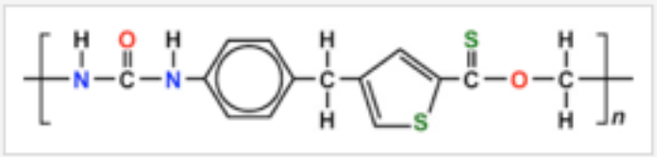
2. T. D. Huan, A. Mannodi-Kanakkithodi, R. Ramprasad Accelerated materials property predictions and design using motif-based fingerprints Physical Review B, 92, 014106 (2015). Article

Target polymer summary

Repeat unit: -(-NH-CO-NH-C_6H_4-CH_2-C_4H_2S-CS-O-CH_2-)$_n$-

Number of building blocks (co-polymers): 9

Chemical formula: $C_{14}H_{12}O_2N_2S_3$



Predicted properties using machine learning model

Method	Polymer	Dielectric Constant	Refractive Index	Band Gap, HSE06 (eV)
Machine Learning [1]	$C_{14}H_{12}O_2N_2S_3$ NH-CO-NH-C6H4-CH2-C4H2S-CS-O-CH2	4.51	1.92	3.92
Machine Learning [2]	$C_{14}H_{12}O_2N_2S_3$ NH-CO-NH-C6H4-CH2-C4H2S-CS-O-CH2	4.46	1.92	3.35

EXAMPLE 3: FINE FINGERPRINTS
PREDICTING ATOMIC FORCES

- Given only the atomic configuration, can we directly predict the atomic forces rapidly & accurately?

- Atoms respond to forces, and force is a local quantity (unlike energy)

AUGUST 15, 1939 PHYSICAL REVIEW VOLUME 56

Forces in Molecules

R. P. FEYNMAN
Massachusetts Institute of Technology, Cambridge, Massachusetts
(Received June 22, 1939)

Formulas have been developed to calculate the forces in a molecular system directly, rather than indirectly through the agency of energy. This permits an independent calculation of the slope of the curves of energy *vs.* position of the nuclei, and may thus increase the accuracy, or decrease the labor involved in the calculation of these curves. The force on a nucleus in an atomic system is shown to be just the classical electrostatic force that would be exerted on this nucleus by other nuclei and by the electrons' charge distribution. Qualitative implications of this are discussed.

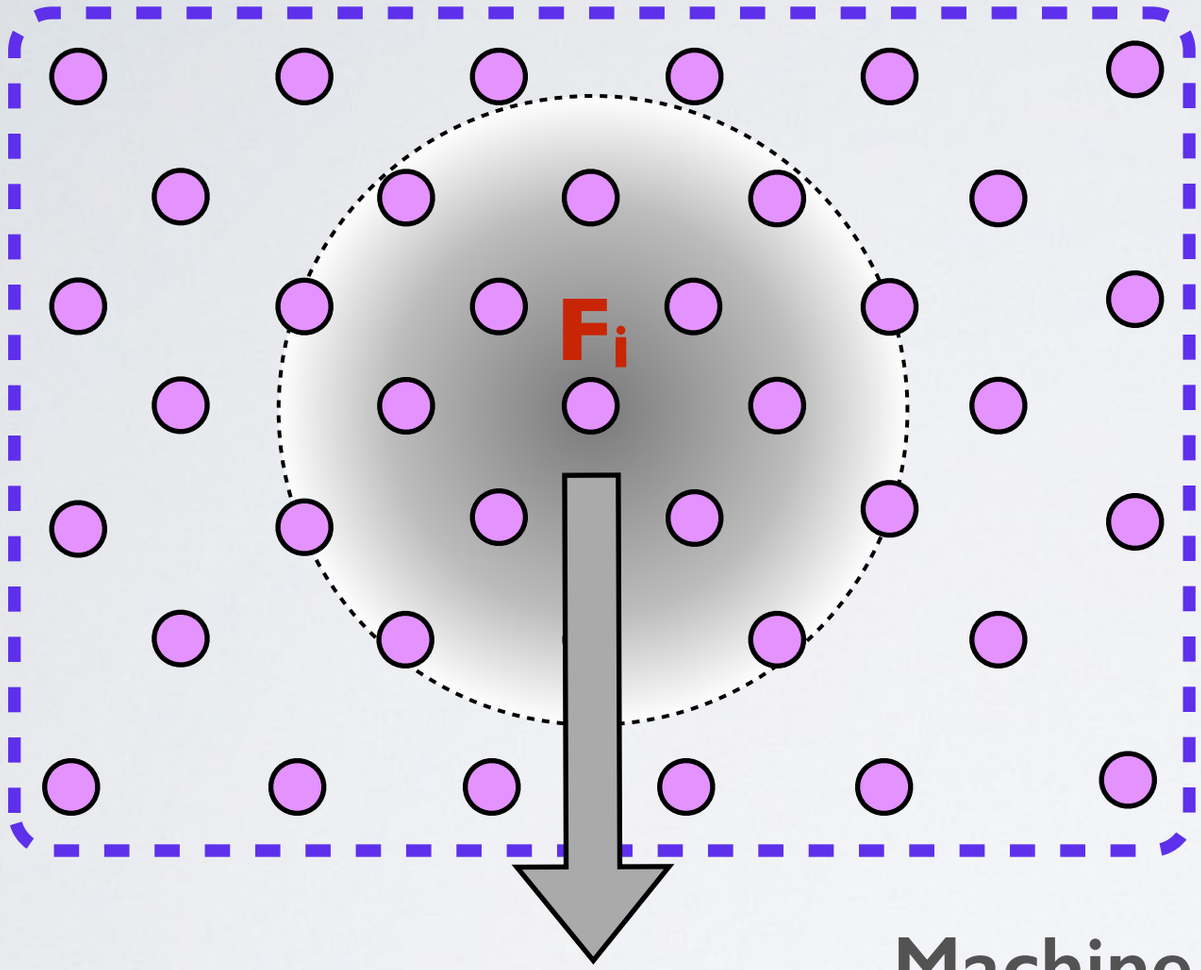
MANY of the problems of molecular structure are concerned essentially with forces. The stiffness of valence bonds, the distortions in geometry due to the various repulsions and attractions between atoms, the tendency of valence bonds to occur at certain definite angles with each other, are some examples of the kind of problem in which the idea of force is paramount.

Usually these problems have been considered through the agency of energy, and its changes with changing configuration of the molecule. The reason for this indirect attack through energy, rather than the more qualitatively illuminating one, by considerations of force, is perhaps twofold. First it is probably thought that force is a quantity that is not easily described or calculated by wave mechanics, while energy is, and second, the first molecular problem to be solved is the analysis of band spectra, strictly a problem of energy as such. It is the purpose of this paper

entire process is repeated for a new nuclear position, and the new value of energy calculated. Proceeding in this way, a plot of energy *vs.* position is obtained. The force on a nucleus is of course the slope of this curve.

The following method is one designed to obtain the forces at a given configuration, when only the configuration is known. It does not require the calculations at neighboring configurations. That is, it permits a calculation of the slope of the energy curve as well as its value, for any particular configuration. It is to be emphasized that this allows a considerable saving of labor of calculations. To obtain force under the usual scheme the energy needs to be calculated for two or more different and neighboring configurations. Each point requires the calculation of the wave functions for the entire system. In this new method, only one configuration, the one in question, need have its wave functions computed in detail. Thus the labor is consider-

THE CONCEPT



Potential energy

global property of entire system

Atomic force

local environment-dependent property

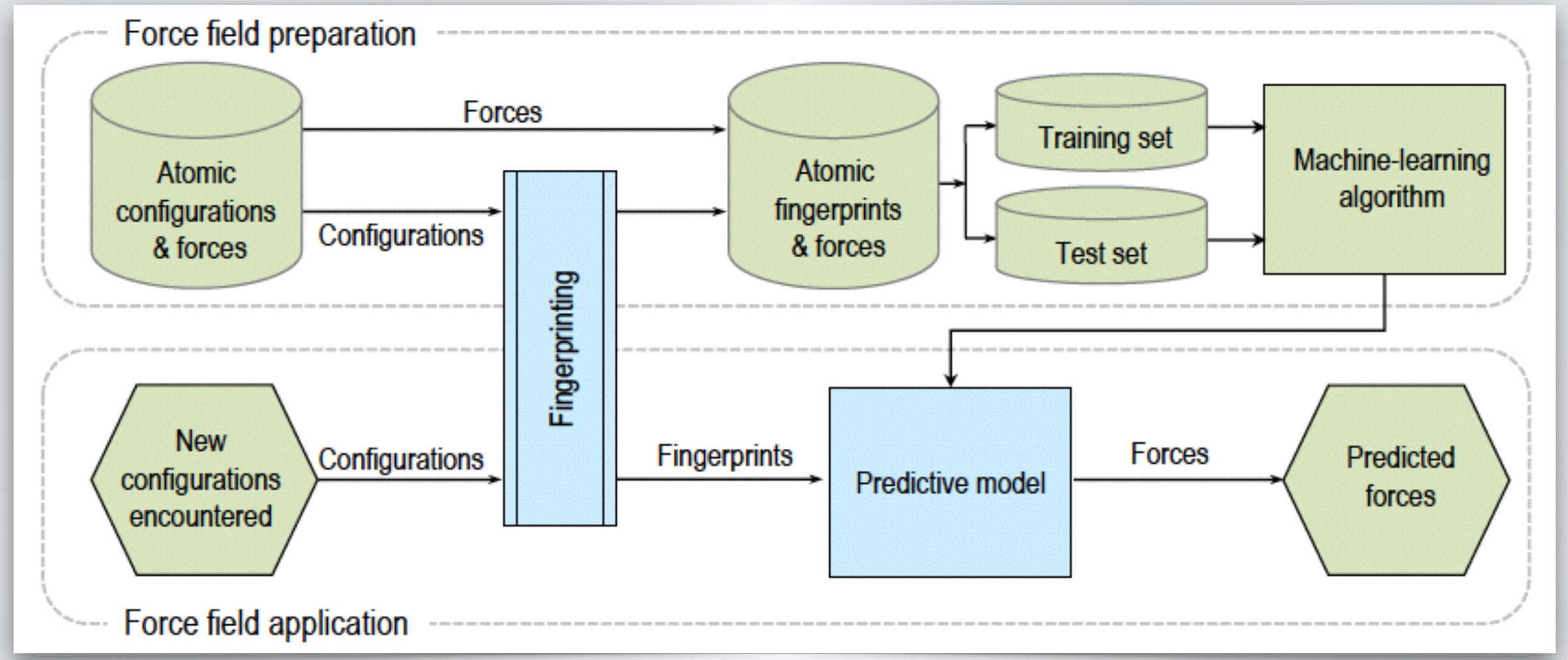
"Fingerprint"

Machine Learning

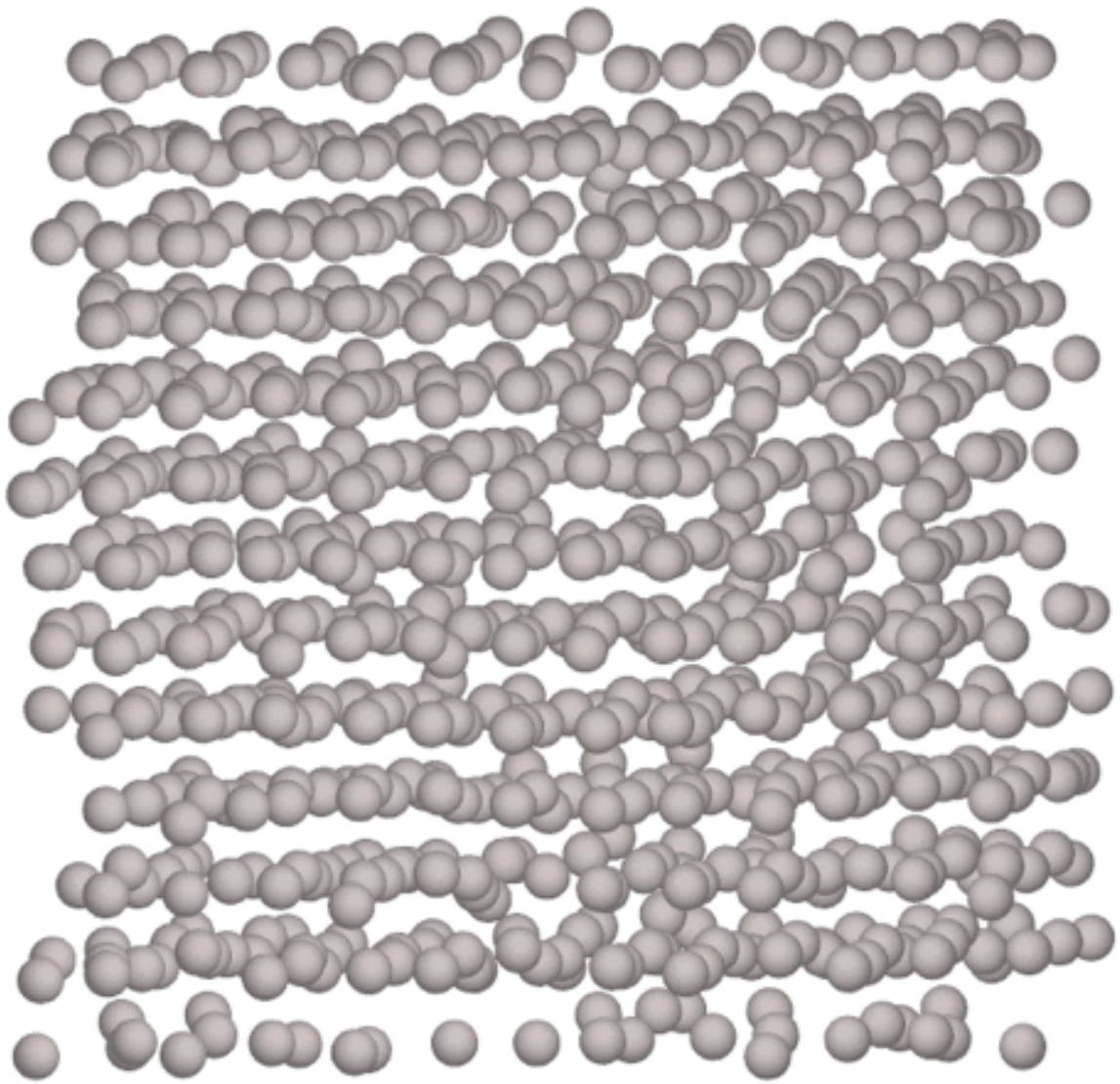
Force Field

Botu & Ramprasad, IJQC (2014) & *Phys. Rev. B* (2015)
Also see: Li, et al, *Phys. Rev. Lett.* (2015)

THE ML FORCE FIELD **(AGNI)**



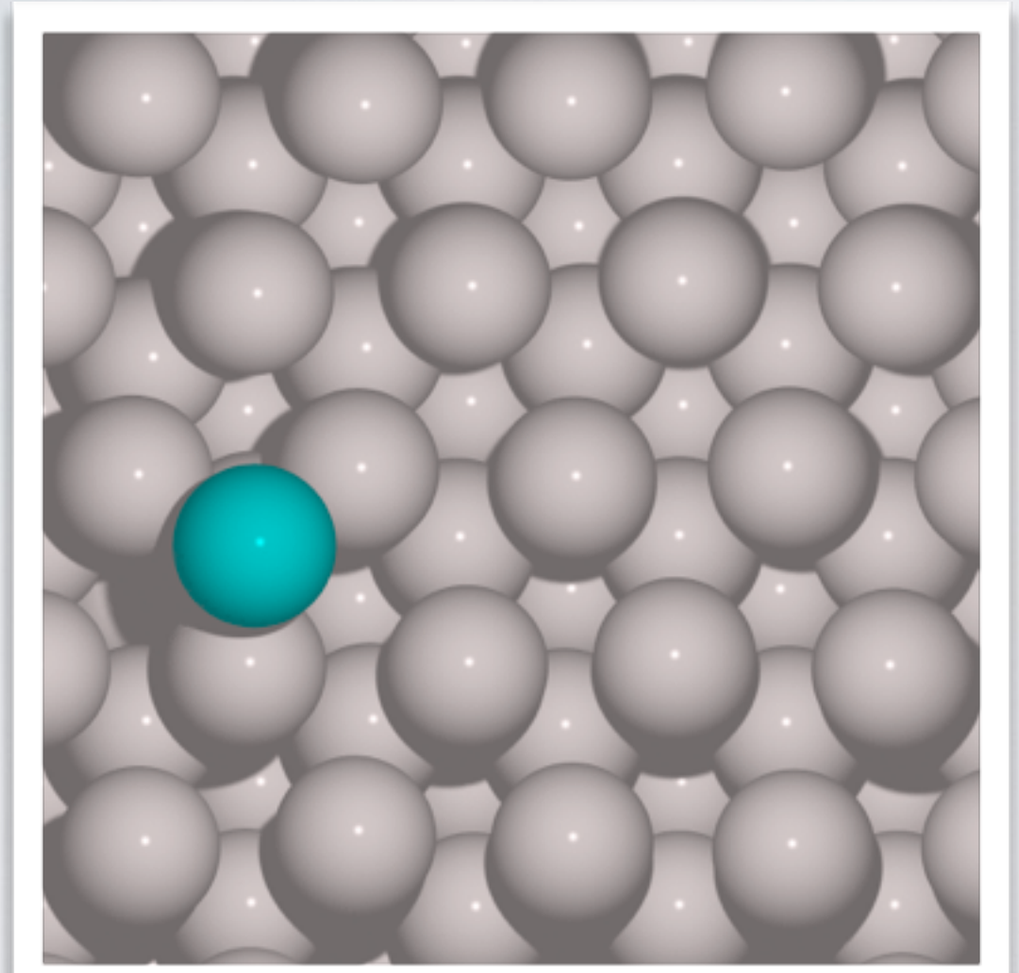
GEOMETRY OPTIMIZATION

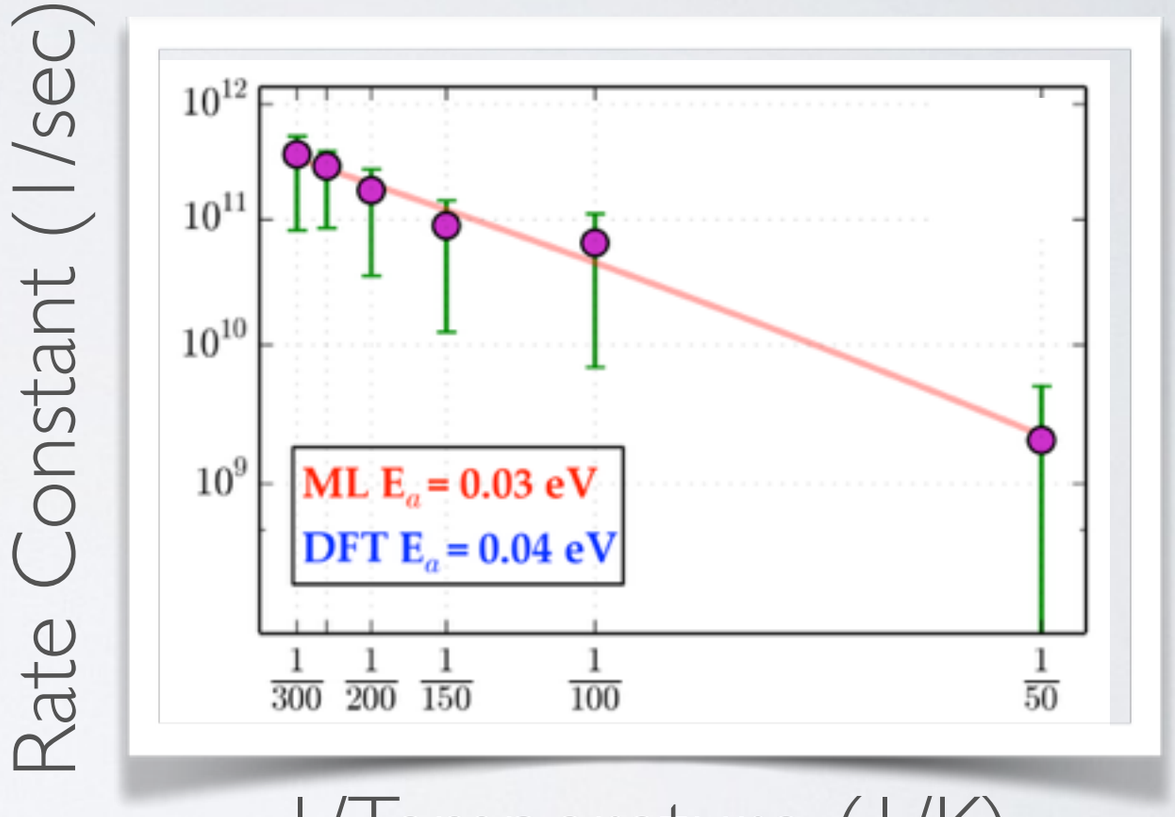


~ 850 Al atoms
in FCC phase
with vacancies

LARGE-SCALE MOTIONS

Adatom migration on Al (111)

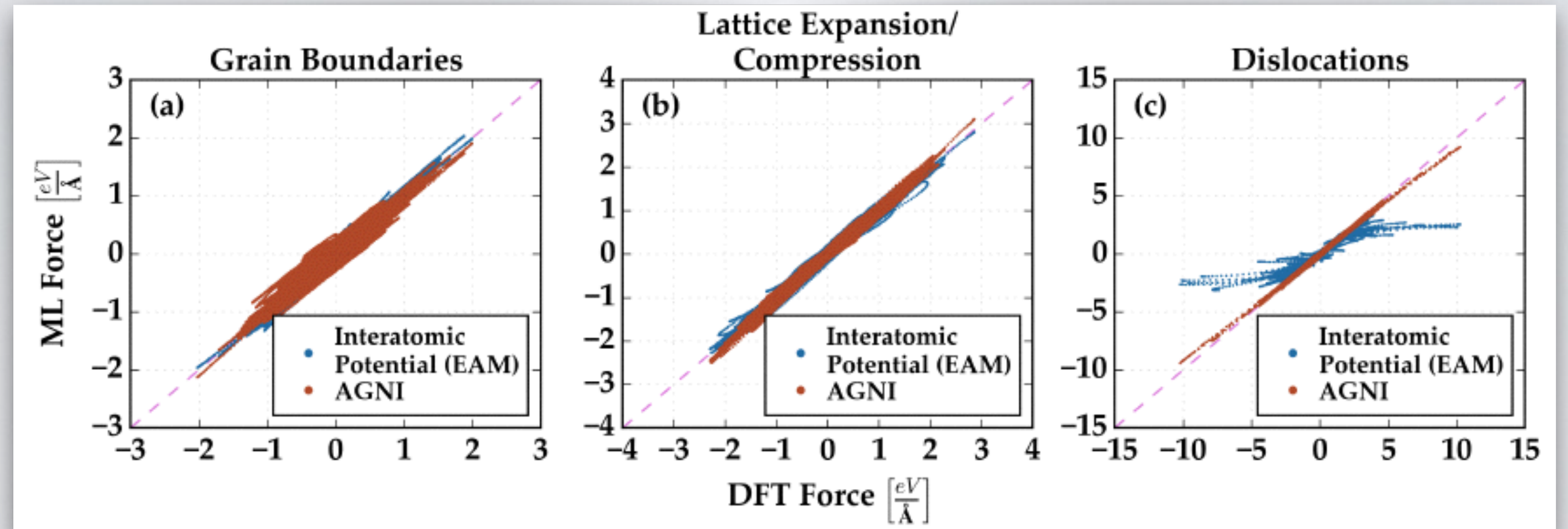




Is the physics underlying dynamics preserved?

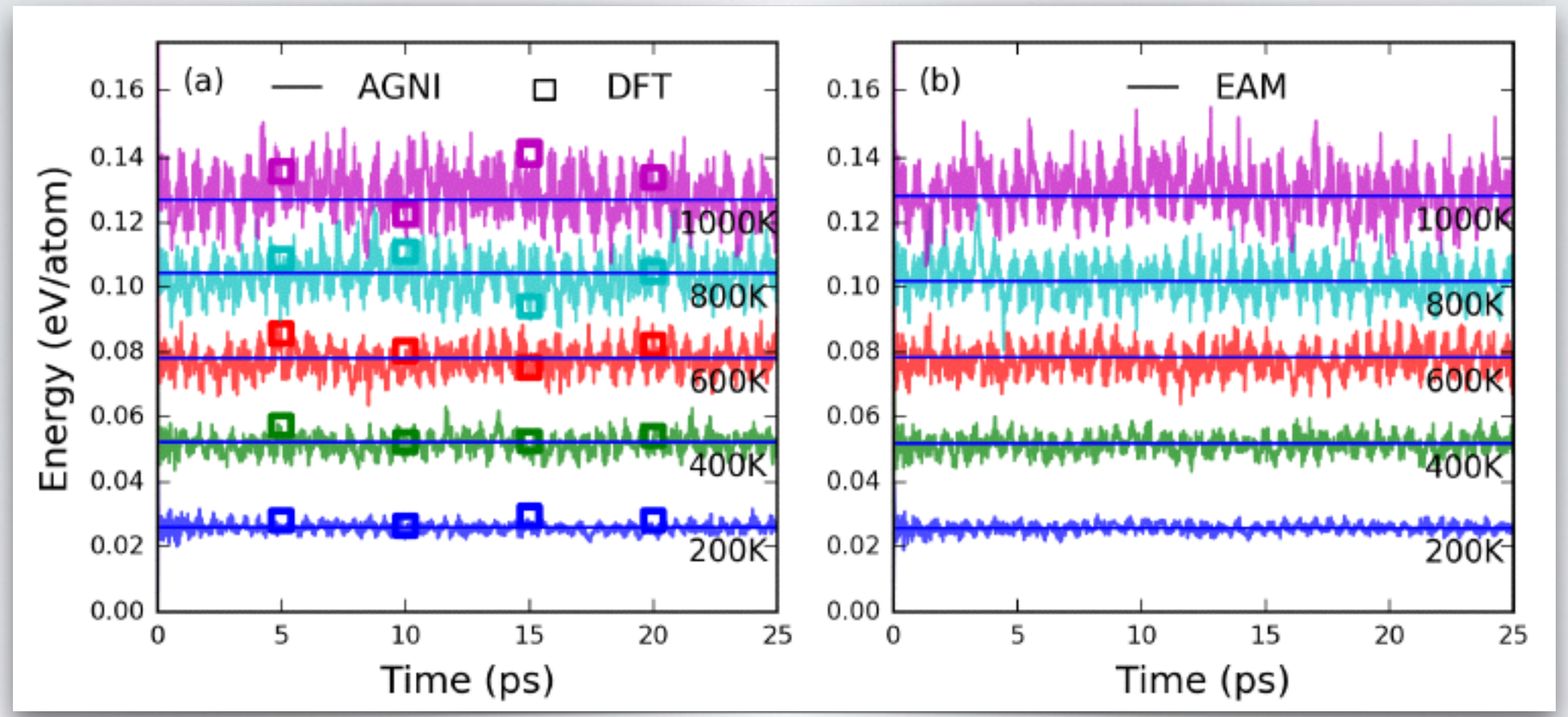
FURTHER VALIDATION

… on situations not used during "training"



ENERGY FROM FORCES

MD simulations of bulk Al



HIERARCHY OF FINGERPRINTS

Scale	Illustrative Examples	Accuracy
Coarse	**Material constituent features ⟶ breakdown strength, phase stability, catalytic activity, friction coefficient, etc.** Kim, et al, Chem. Mat. (2016); Pilania, et al, Sci. Rep. (2016); Ghiringhelli, et al, PRL (2015); Bucholtz, et al, Tribol Lett (2012)	Moderate
Medium	**Building units, Substructures, Motifs ⟶ Properties** Mannodi-Kanakkithodi, et al, Sci. Rep. (2016); Tran Huan, et al, PRB (2015); Yang, Ceder, et al, PRB (2014)	Moderate
Fine	**Atomic/electronic arrangements (Coulomb matrix, SOAP, Symmetry functions) ⟶ Properties, Energies, Forces** Botu, Ramprasad, PRB (2015) & IJQC (2014); Hansen, et al, JPC Lett (2015); Bartok, et al, IJQC (2015); Behler, IJQC (2015); Li, et al, PRL (2015)	High

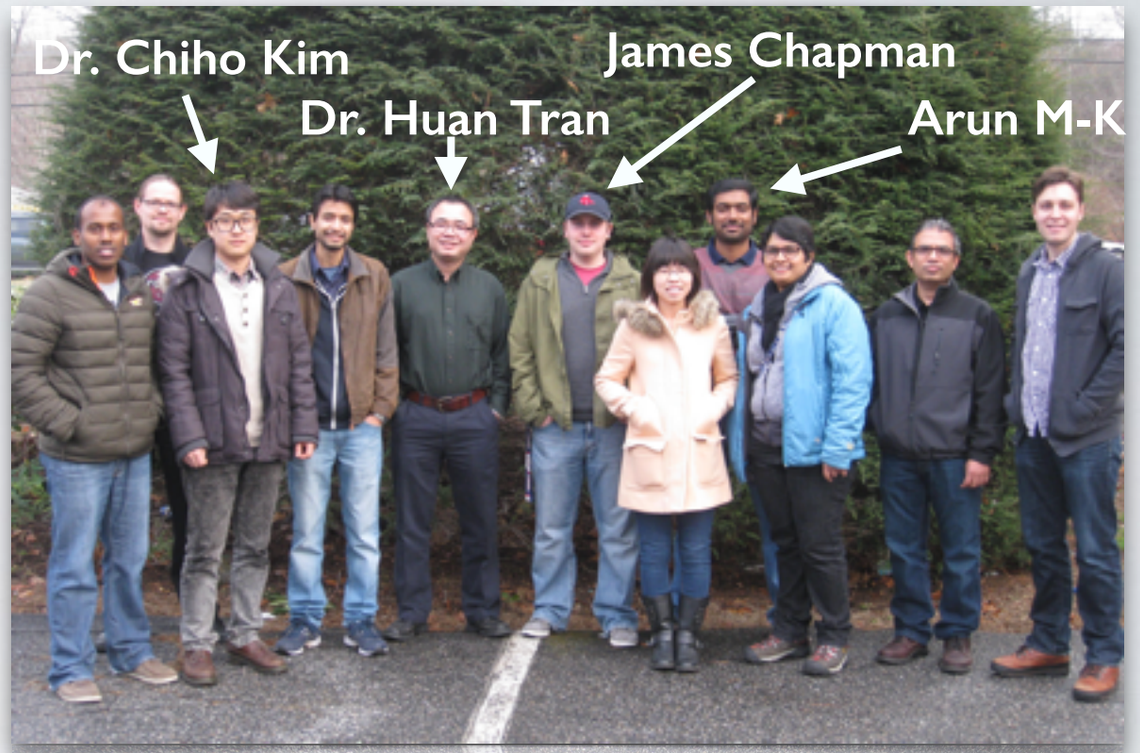
CRITICAL NEXT STEPS

- Decision on whether to use machine learning
- Choice of fingerprints
- Uncertainty (out-of-domain) quantification
- Automatic & adaptive improvement of model
- Show the value!

CREDITS

2013

2015



TURING'S "LEARNING MACHINES"

Instead of trying to produce a program to simulate the adult mind, why not rather try to produce one which simulates the child's ? If this were then subjected to an appropriate course of education one would obtain the adult brain …

"Computing Machinery and Intelligence"
A. M. Turing
Mind, Vol. LIX, No. 236, p. 433 (1950)