Karsten W. Jacobsen
Department of Physics
Technical University of Denmark

Machine learning and computational screening

Outline

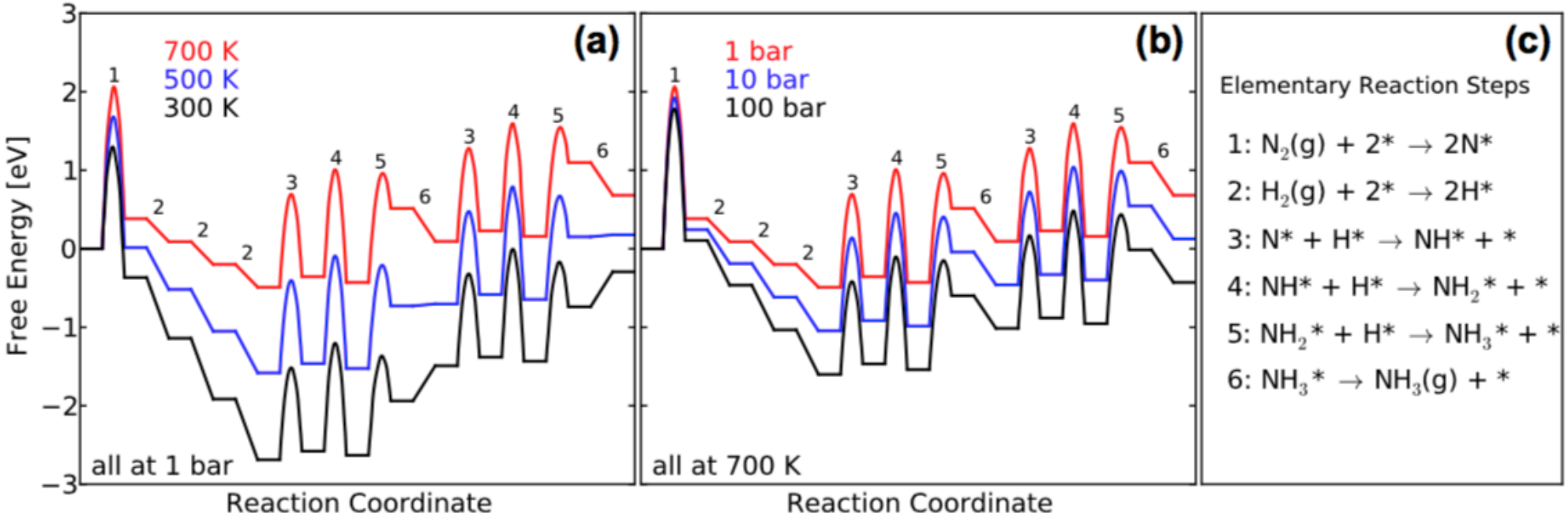
- Motivation
- Machine-learning background
  - Bayes theorem
  - Kernel regression
  - Gaussian processes
- Structures and energetics
  - Binding sites
  - Transition states
  - Global structure search
- Computational screening
  - Water splitting
  - Organic solar cells

Ammonia synthesis

Can we evaluate these quantities faster than standard approaches today?

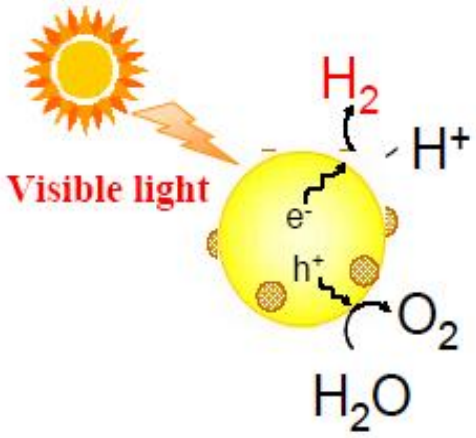
$$N_2 + 3H_2 \rightarrow 2NH_3$$

Descriptors: Adsorption energies and reaction barriers



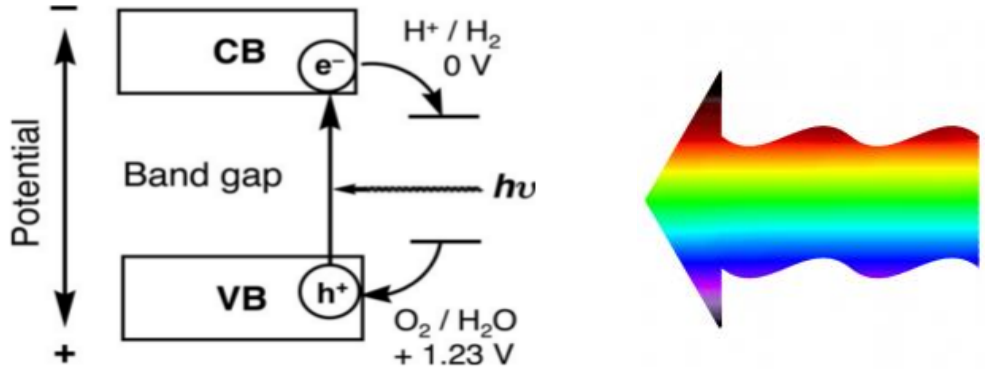
Vojvodic, Medford, Studt, Abild-Pedersen, Khan, Bligaard, and Nørskov, *Chemical Physics Letters*, **598**, 108 (2014)

Computational screening: light-induced water splitting



Descriptors:

- **Stability of material**
  - Heat of formation
- **Good light absorption**
  - Bandgap in the visible range
- **Photogenerated charges at right potentials**
  - Band edges straddle the water redox potentials

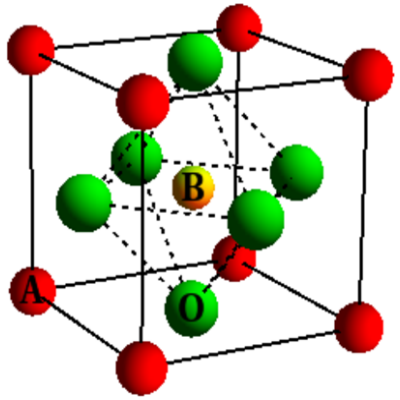


Principle of water splitting using semiconductor photocatalysts.

(Fujishima and Honda, Nature 1972)

Cubic perovskites ABX_3

(X_3= O_3, O_2N, ON_2, N_3, O_2S, O_2F, OFN)

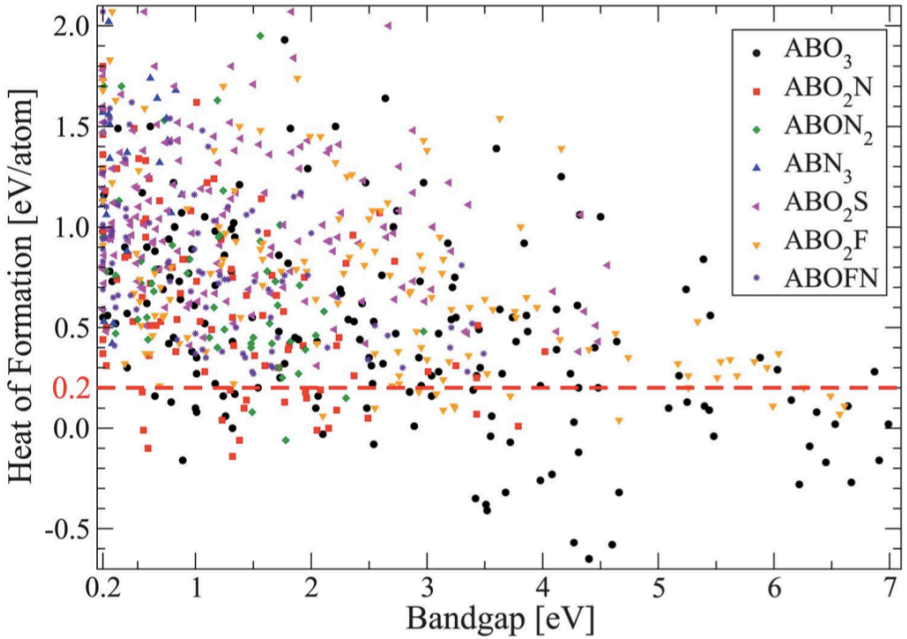


Can we avoid all the useless DFT calculations?

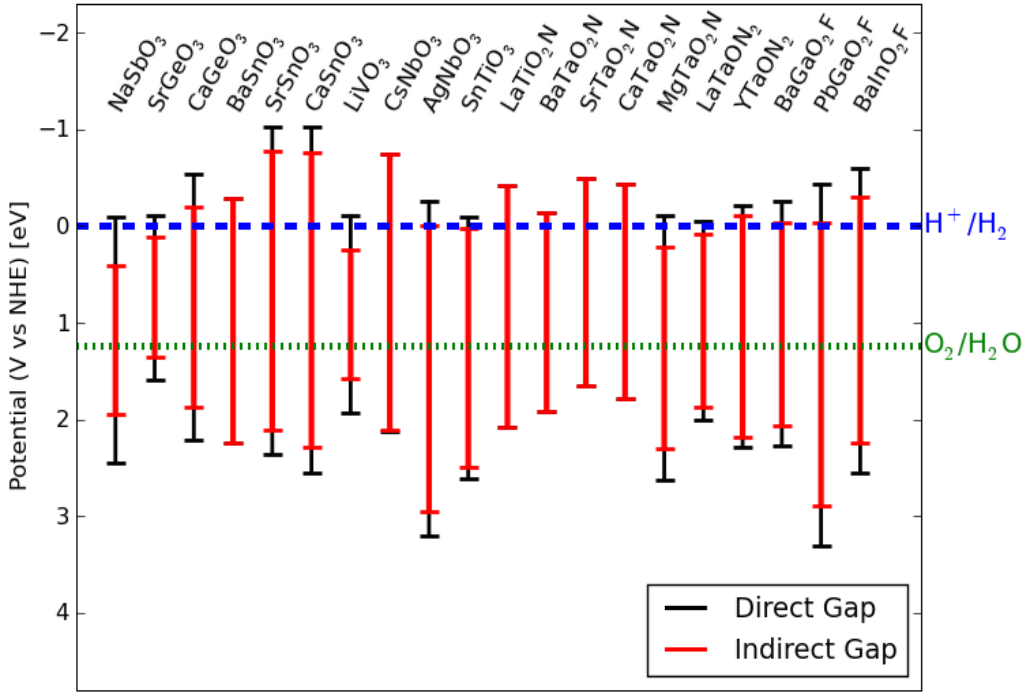
Screening criteria:

- Stability (heat of formation)
- Band gap
- Band alignment

~19000 materials



20 candidate materials
About half are known



Castelli, Olsen, Datta, Landis, Dahl, Thygesen, Jacobsen, *Energy & Environmental Science*, *5*(2), 5814 (2012).
Castelli, Landis, Thygesen, Dahl, Chorkendorff, Jaramillo, Jacobsen, Energy Environ Sci **5**, 9034 (2012)

Bayes' theorem

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

Prior probability distribution

$$P(\text{Model}|\text{Data}) = \frac{1}{P(\text{Data})} P(\text{Data}|\text{Model}) P_0(\text{Model})$$

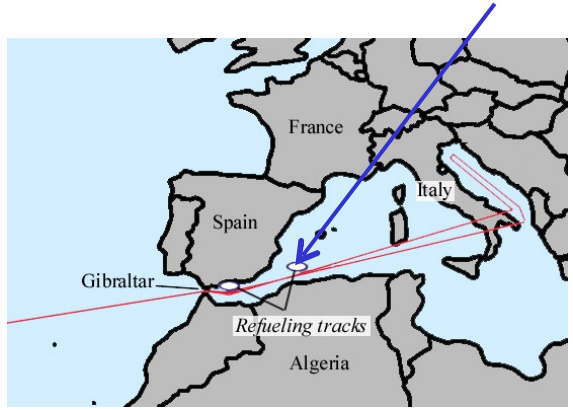
Update of model as new data are introduced.

Why do we believe Newton's second law?

Bayesian Search Theory in Practice: The 1966 Palomares B-52 crash

B-52G collided with KC-135 tanker when fueling

Refueling



4 H-bombs dropped
3 on land
1 in the Mediterranean Sea

Bayesian search theory applied:
Assign probabilities to different areas of the sea based on available information
(a local fisherman "Bomb Frankie" saw the bomb dropping)
Update your probability depending on your search.

Bomb recovered!

Bayes theorem example: Screening for diseases

You get a positive test. What is the risk that you are ill?

$$\frac{P(ill|positive)}{P(healthy|positive)} = \frac{P(positive|ill)}{P(positive|healthy)} \frac{P(ill)}{P(healthy)}$$

$$\frac{1}{10} = \frac{\approx 1}{1/100} \frac{1}{1000}$$

Probability of being ill before test

Machine learning: Kernel regression

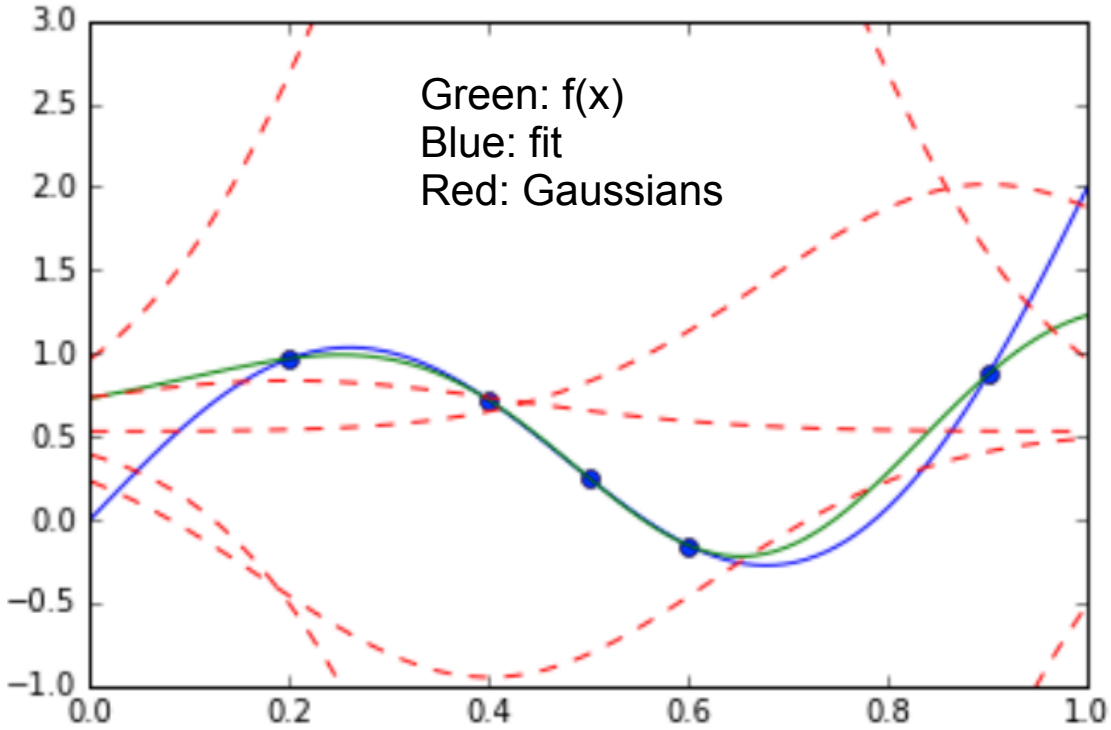
Fitting a function f(x) based on data points $y_i = f(x_i)$

Drop a Gaussian on each data point:

$$k(x, x_i) = \exp(-|x - x_i|^2/2\rho^2)$$

Interpolation: $y(x) = \sum_i k(x, x_i)\alpha_i$

Coefficients determined by data points:



$$y_j = \sum_i k(x_j, x_i)\alpha_i = \sum_i K_{ji}\alpha_i \rightarrow \mathbf{y} = \mathbf{K}\boldsymbol{\alpha} \rightarrow \boldsymbol{\alpha} = \mathbf{K}^{-1}\mathbf{y}$$

Interpolation: $y(x) = \mathbf{k}^T\mathbf{K}^{-1}\mathbf{y}$

with $k_i = k(x, x_i)$

The Bayesian approach: Gaussian process

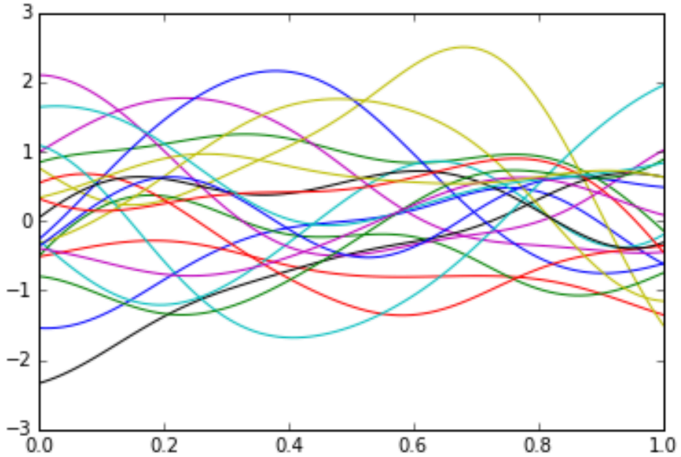
Prior probability

Probabilistic approach based on Bayes theorem:

$$P(\text{Model}|\text{Data}) = \frac{1}{P(\text{Data})} P(\text{Data}|\text{Model}) P_0(\text{Model})$$

"Reinterpretation" of kernel function as correlation:

$$\rho^2 = 0.1$$

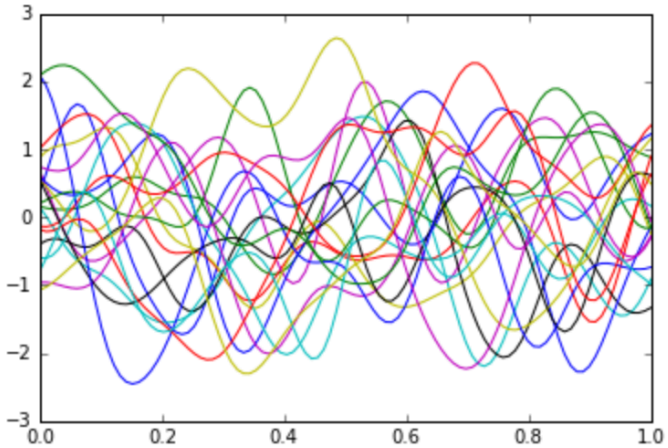


$$K_{ij} = \langle y(x_i) y(x_j) \rangle = k(x_i, x_j) = \exp(-|x_i - x_j|^2 / 2\rho^2)$$

Prior probability distribution (i.e. *without* data points):

$$P_0(\mathbf{y}) = \frac{1}{\sqrt{2\pi det(\mathbf{K})}} \exp\left(-\frac{1}{2}\mathbf{y^T K^{-1} y}\right)$$

$$\rho^2 = 0.01$$

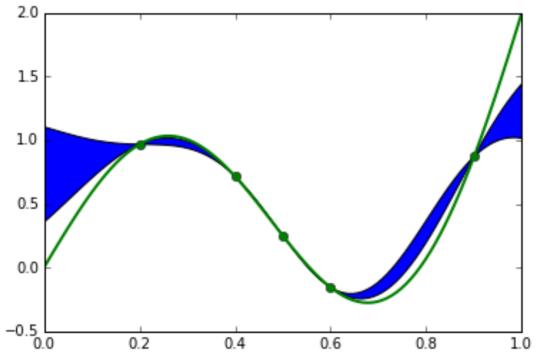


$$\mathbf{y}^T = (y(x_1), y(x_2), \ldots, y(x_N))$$

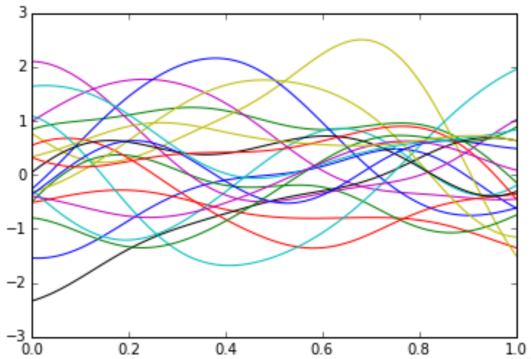
Gaussian process

Fitting a function f(x) based on data points $y_i = f(x_i)$

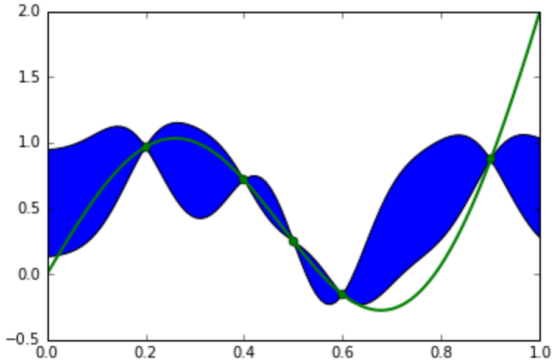
$$P(\text{Model}|\text{Data}) = \frac{1}{P(\text{Data})} P(\text{Data}|\text{Model}) P_0(\text{Model})$$

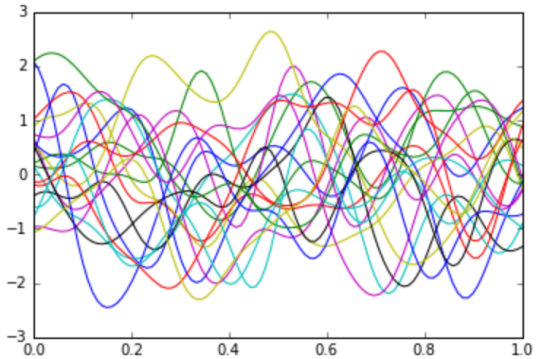


Update of model because of data



$\rho^2 = 0.1$





$\rho^2 = 0.01$

The value of ρ can be addressed by cross validation

Gaussian process with gradient information

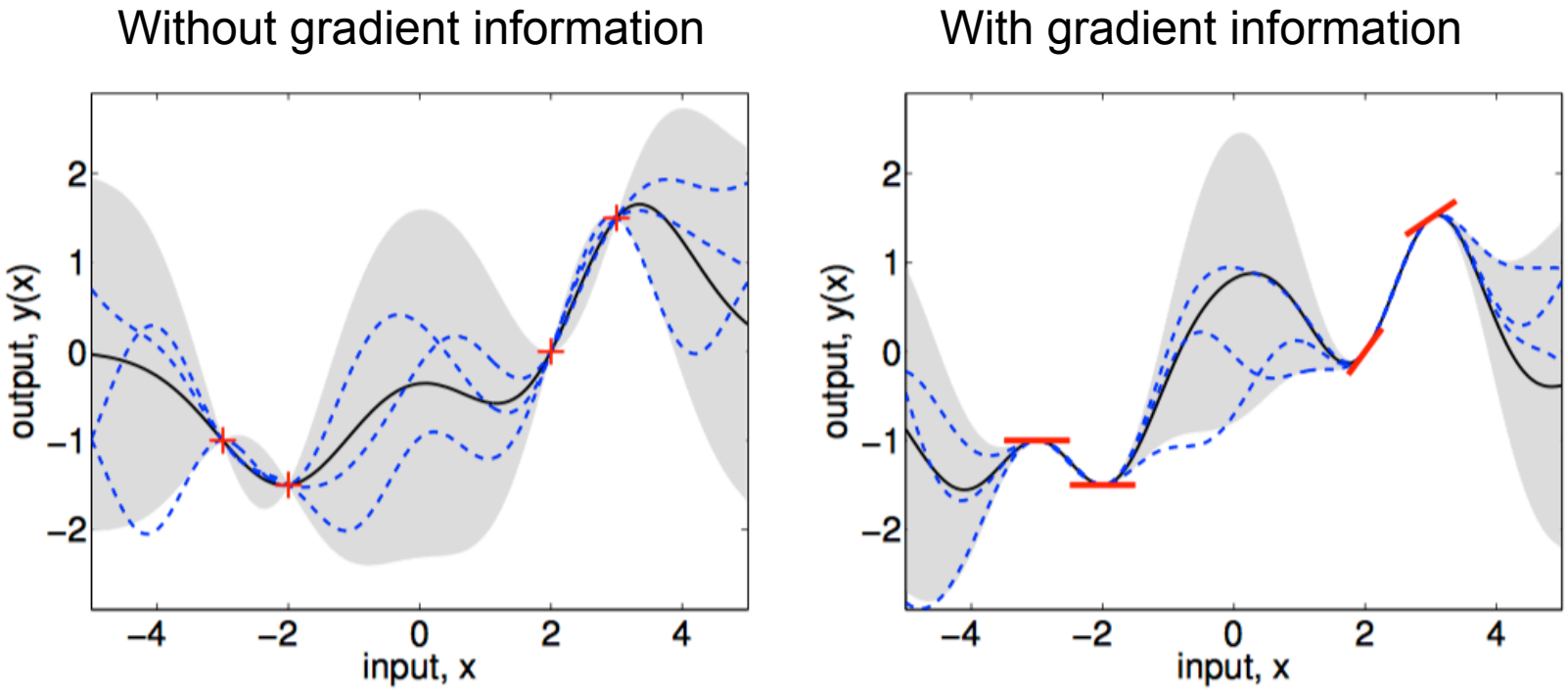
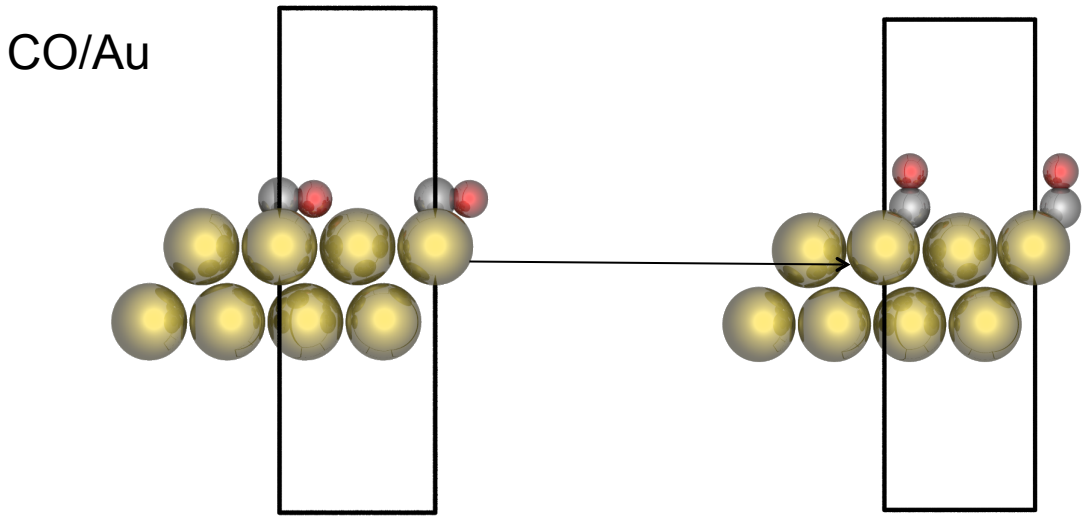


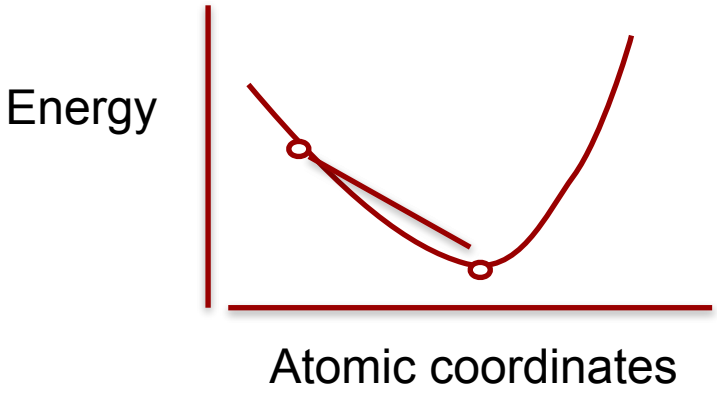
Figure from
C. E. Rasmussen and
C. K. I. Williams,
Gaussian Processes
for Machine Learning.
The MIT Press, 2006.

$$y(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i \, e^{-\|\mathbf{x}^{(i)} - \mathbf{x}\|^2/2l^2} + \sum_{i=1}^{n} \sum_{j=1}^{D} \beta_{ij} \, \frac{x_j^{(i)} - x_j}{l^2} e^{-\|\mathbf{x}^{(i)} - \mathbf{x}\|^2/2l^2}$$

Example: Local structure optimization of atomic systems

- Multidimensional local optimization
- A number of well-developed techniques are available: Conjugate Gradients, BFGS, …
- Takes up a large fraction of CPU hours on supercomputers performing electronic structure calculations

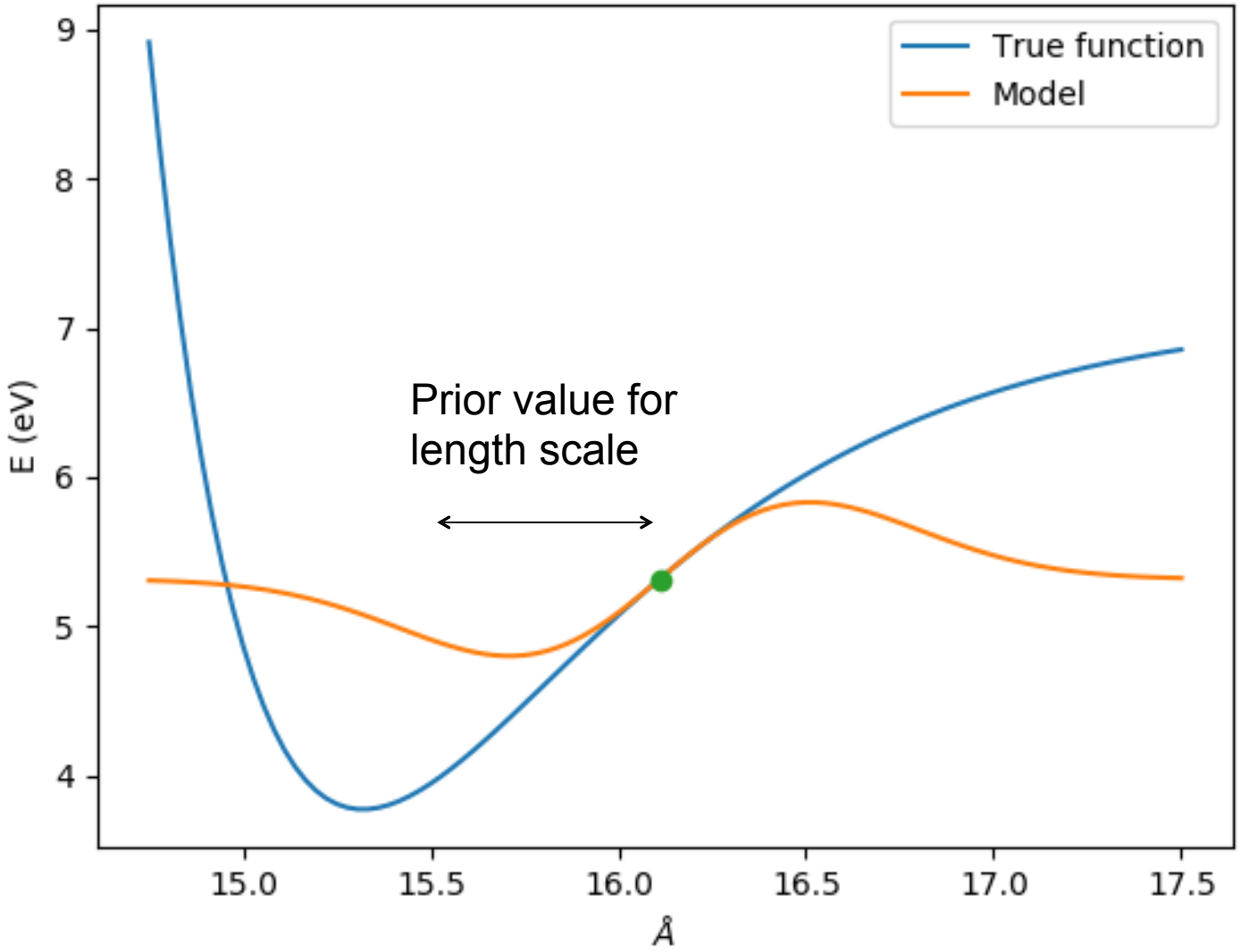




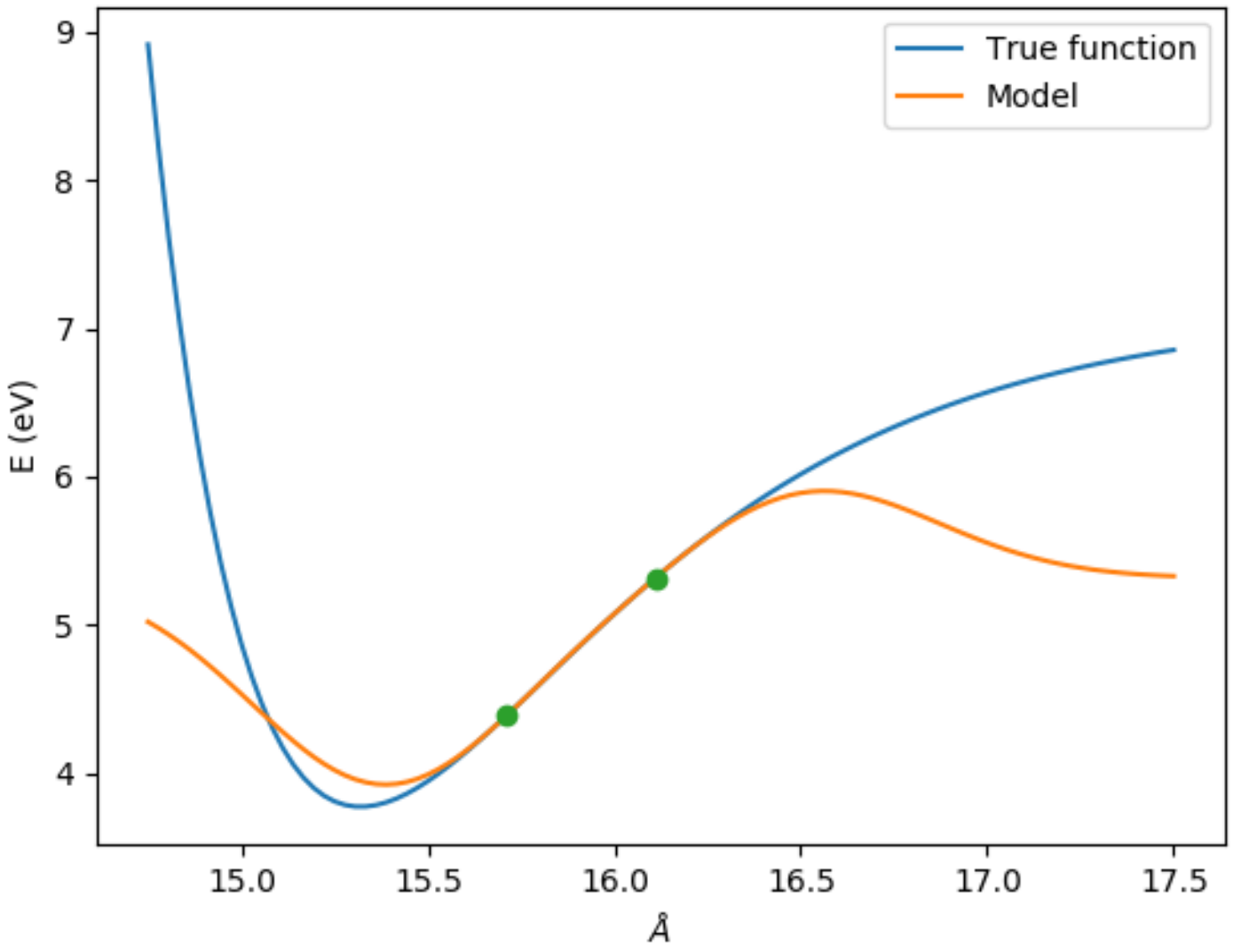
Use kernel regression on energy surface!

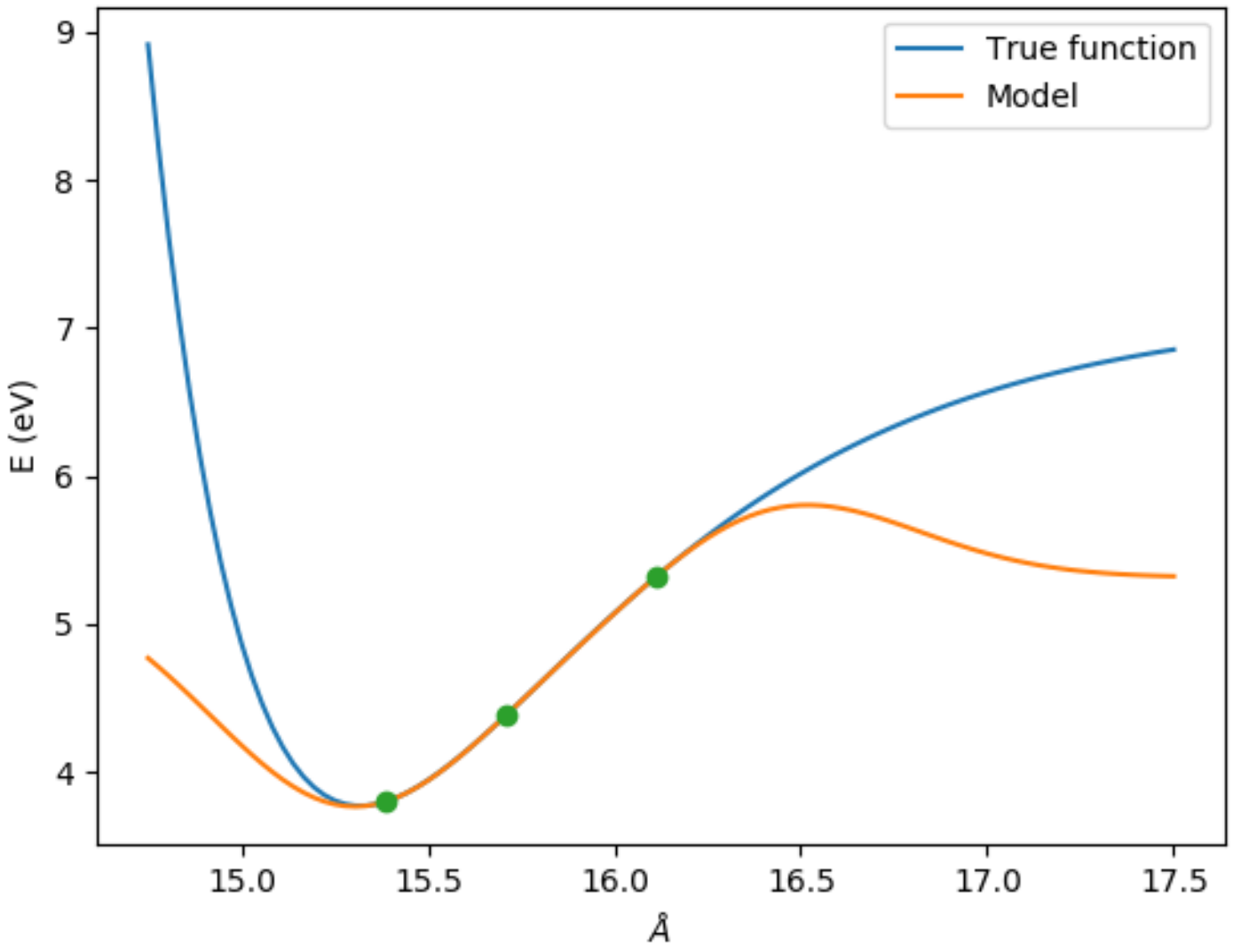
(E. Garijo del Río, J. J. Mortensen, and K. W. Jacobsen, Phys. Rev. B, **100**, 104103 (2019))

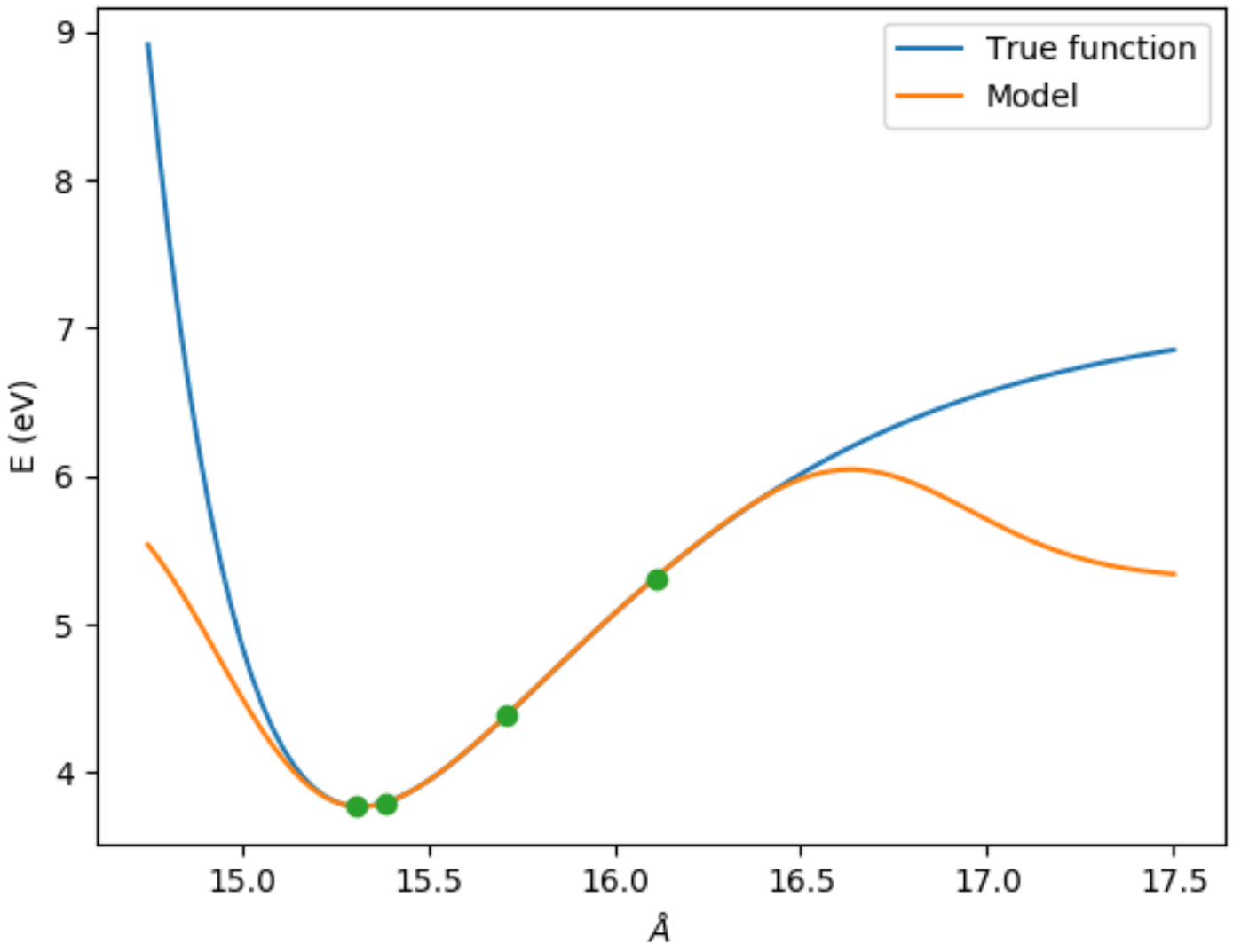
Construction of ”surrogate” or ”model” potential energy surface

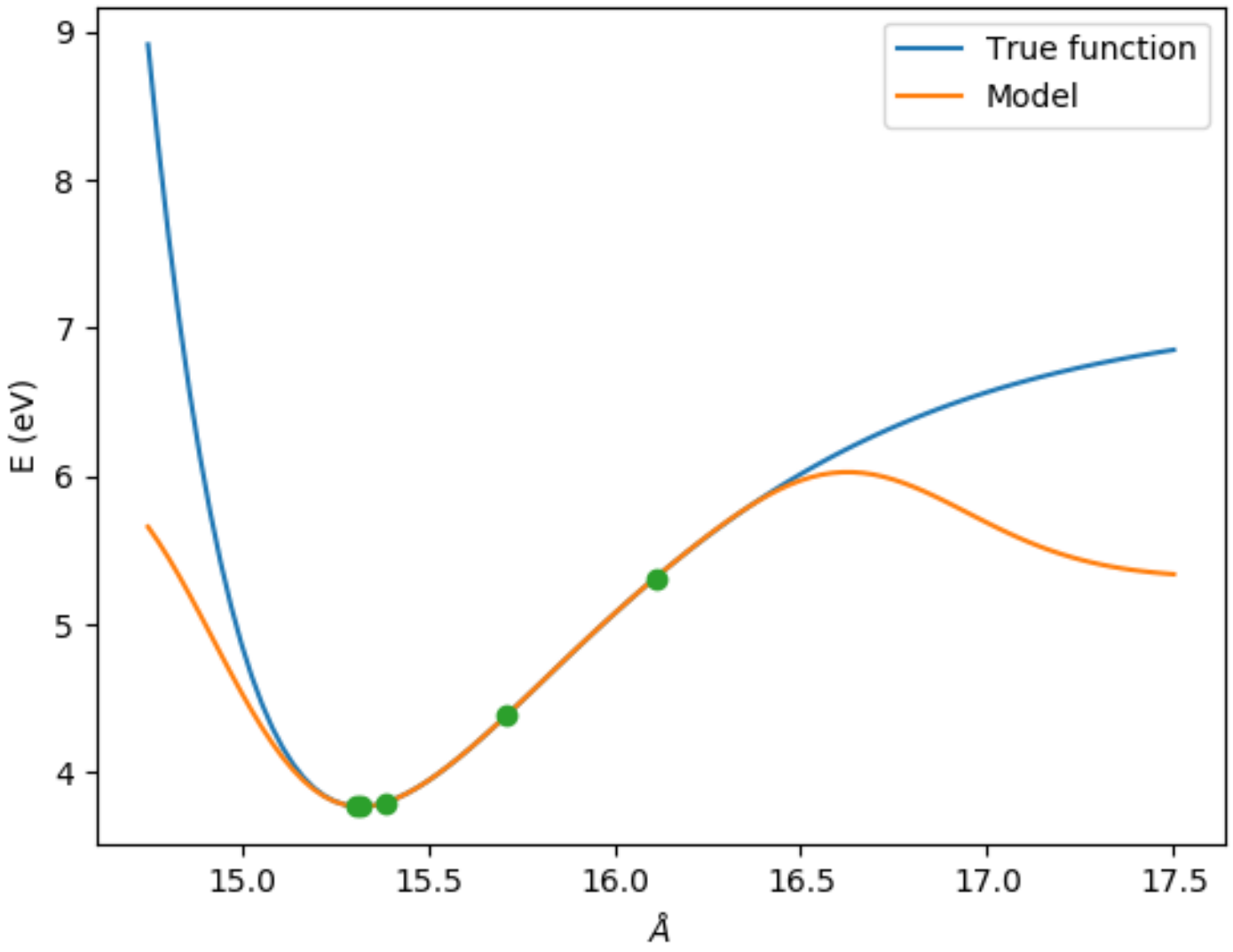


Prior energy value


True function
Model
E (eV)
Å
9
8
7
6
5
4
15.0
15.5
16.0
16.5
17.0
17.5

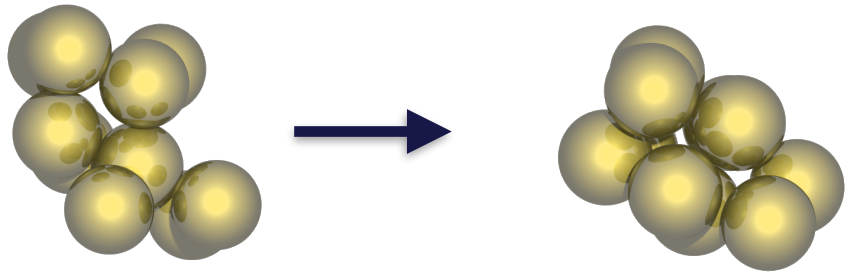

True function
Model
E (eV)
Å
9
8
7
6
5
4
15.0
15.5
16.0
16.5
17.0
17.5


True function
Model
E (eV)
Å
9
8
7
6
5
4
15.0
15.5
16.0
16.5
17.0
17.5

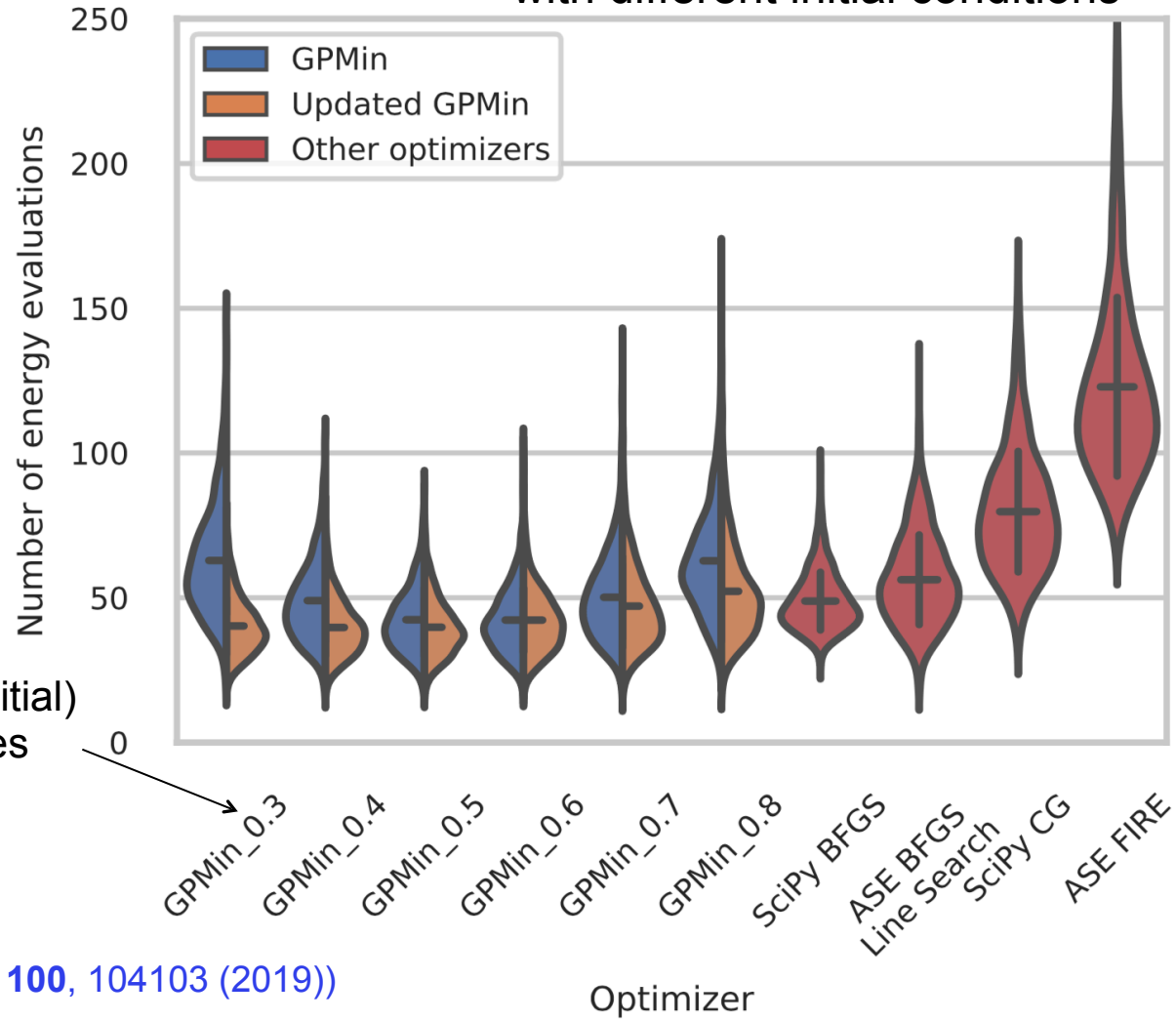

True function
Model
E (eV)
Å
9
8
7
6
5
4
15.0
15.5
16.0
16.5
17.0
17.5

Test case: Optimize the structure of a 10 atom Au cluster

10 atom Au cluster with Effective Medium Theory interatomic potential.



1000 energy minimizations with different initial conditions



Different (initial) length scales

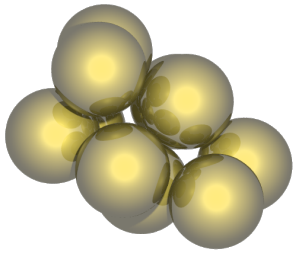
Update of length scale:

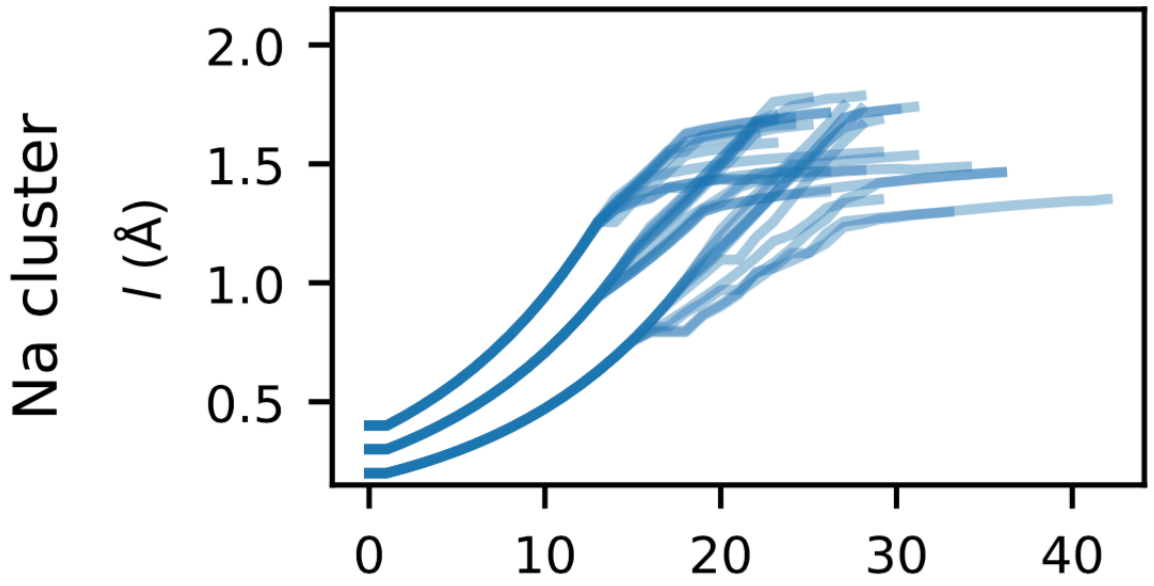
Maximize:

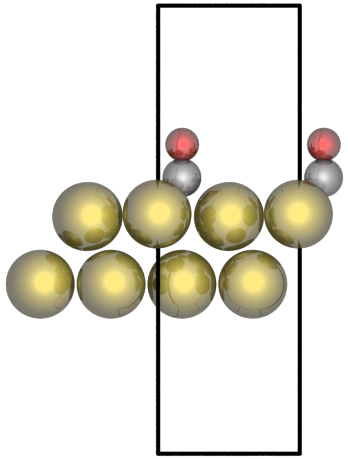
$$P(length|data) \propto P(data|length)P_0(length)$$

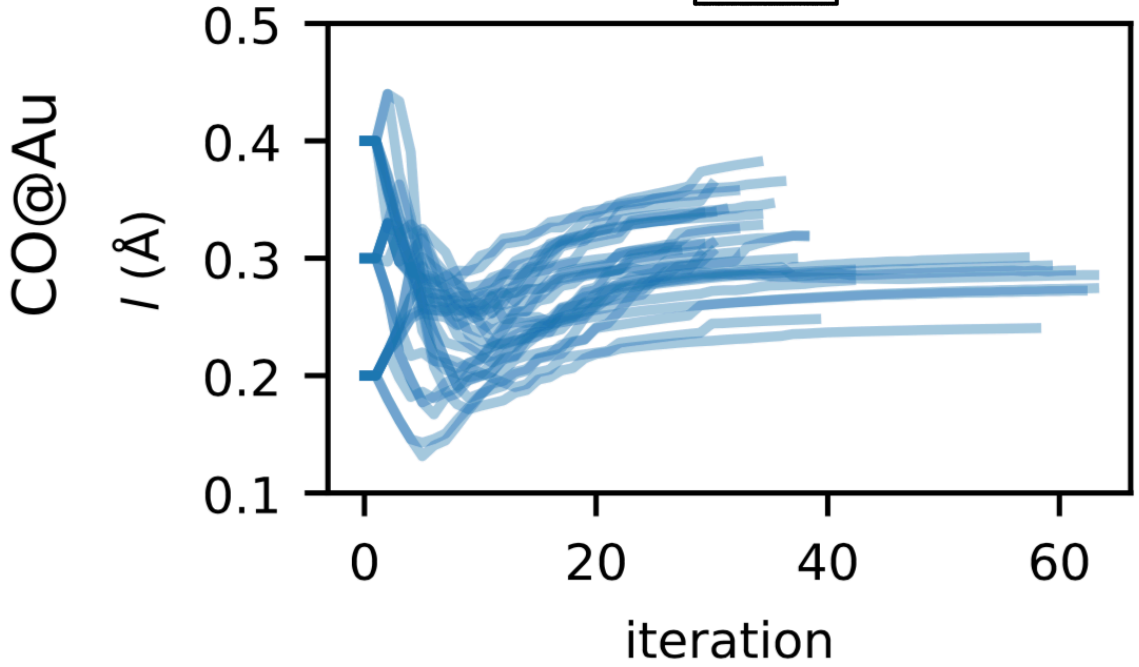
(E. Garijo del Río, J. J. Mortensen, and K. W. Jacobsen, Phys. Rev. B, **100**, 104103 (2019))

Updating the length scale

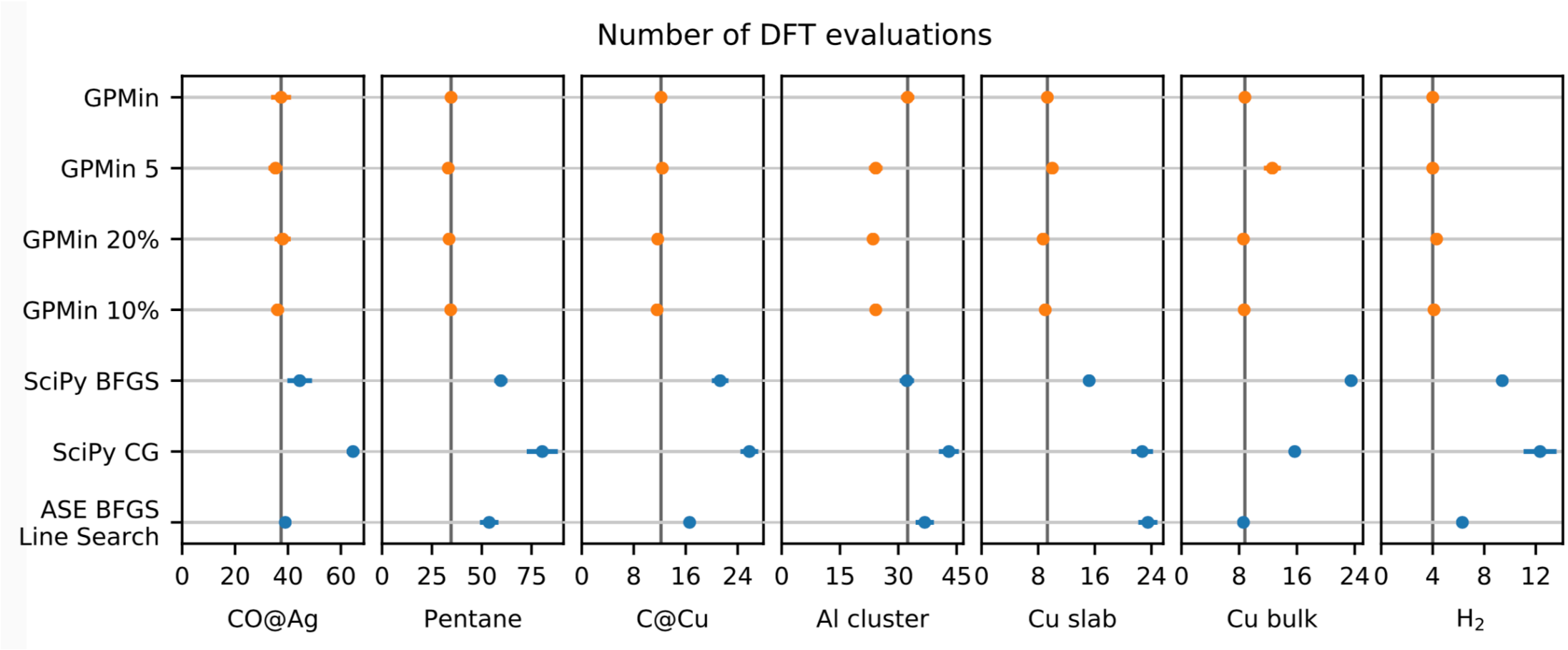








Testing the GP optimizer



(E. Garijo del Río, J. J. Mortensen, and K. W. Jacobsen, Phys. Rev. B, **100**, 104103 (2019))

Finding transition states: Nudged Elastic Band (NEB)

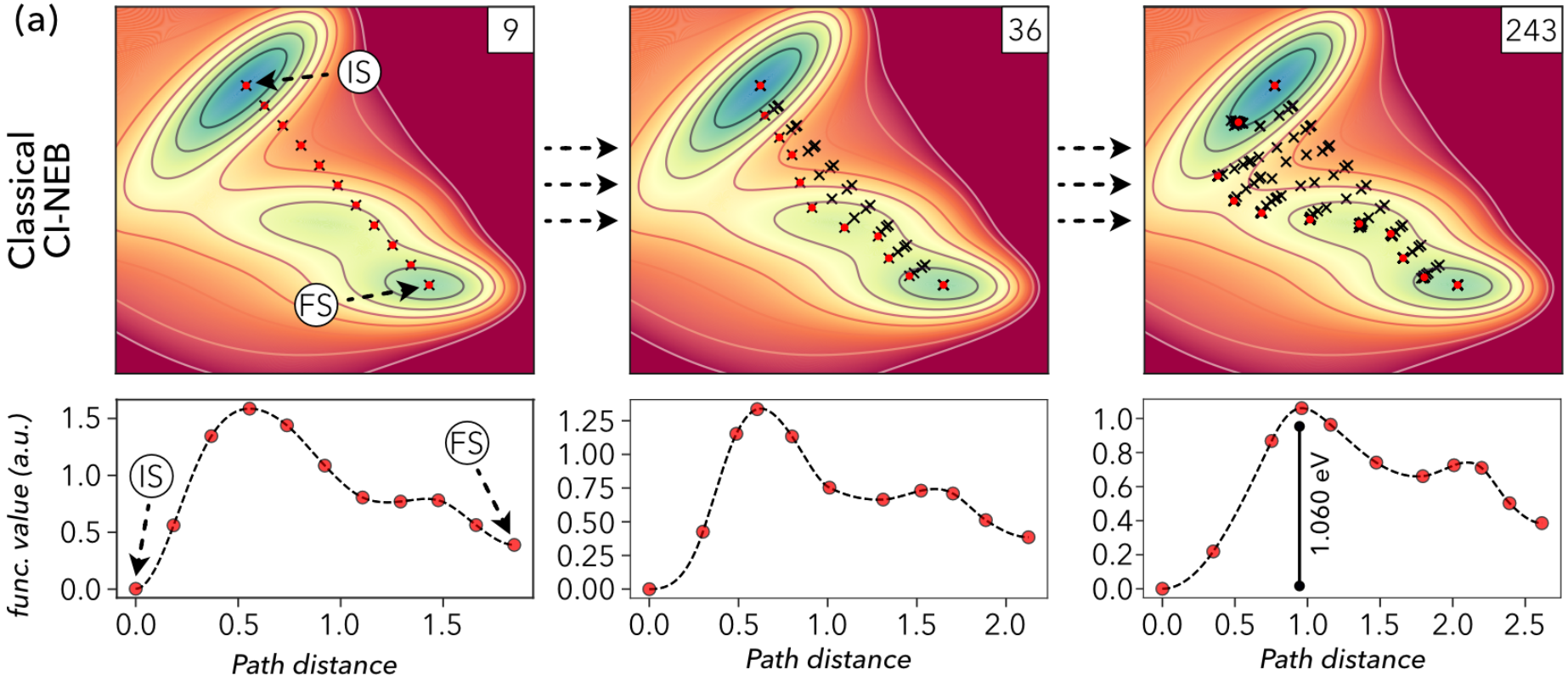
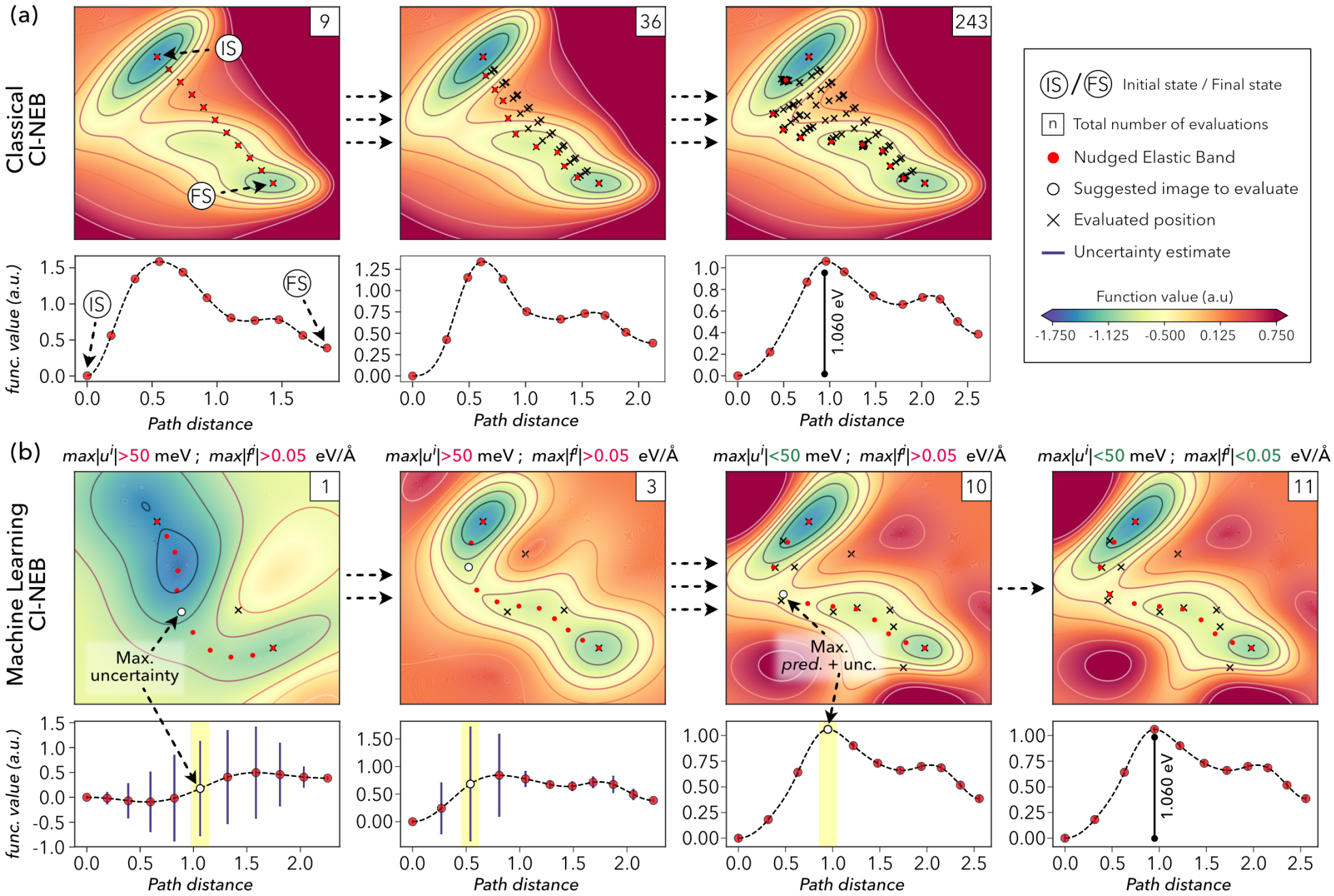


Figure from
Torres, Jennings, Hansen, Boes, Bligaard, Phys. Rev. Lett., **122**, 156001 (2019)

Mills, Jonsson PRL **72**, 1124 (1994)
Mills, Jonsson, Schenter Surf Sci **324**, 305f (1995)

Jónsson, Mills, Jacobsen, in Nudged Elastic Band Method for Finding Minimum Energy Paths of Transitions (World Scientific, Singapore, 1998), pp. 385– 404.



Torres, Jennings, Hansen, Boes, Bligaard, Phys. Rev. Lett., **122**, 156001 (2019)

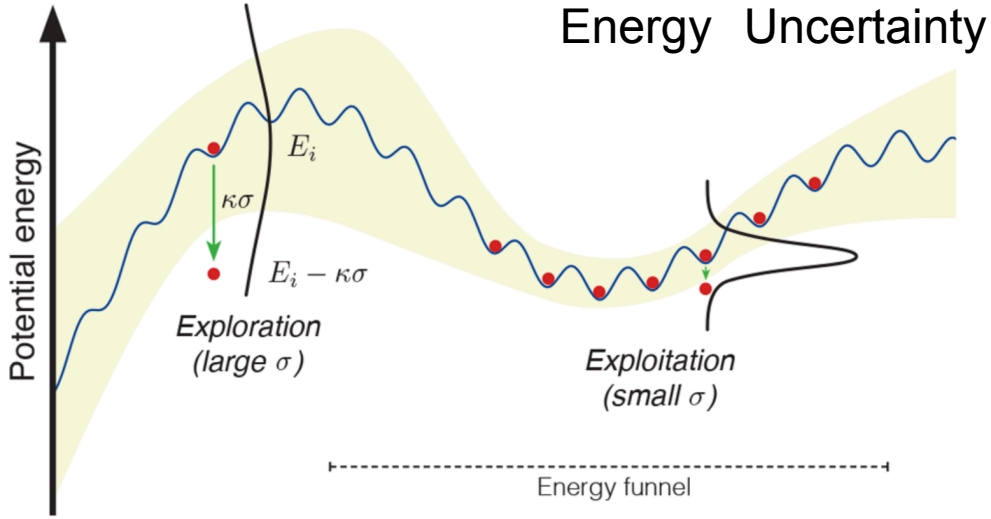
See also: Koistinen, Dagbjartsdóttir, Ásgeirsson, Vehtari, Jónsson, *J Chem Phys*, **147,** 152720 (2017)

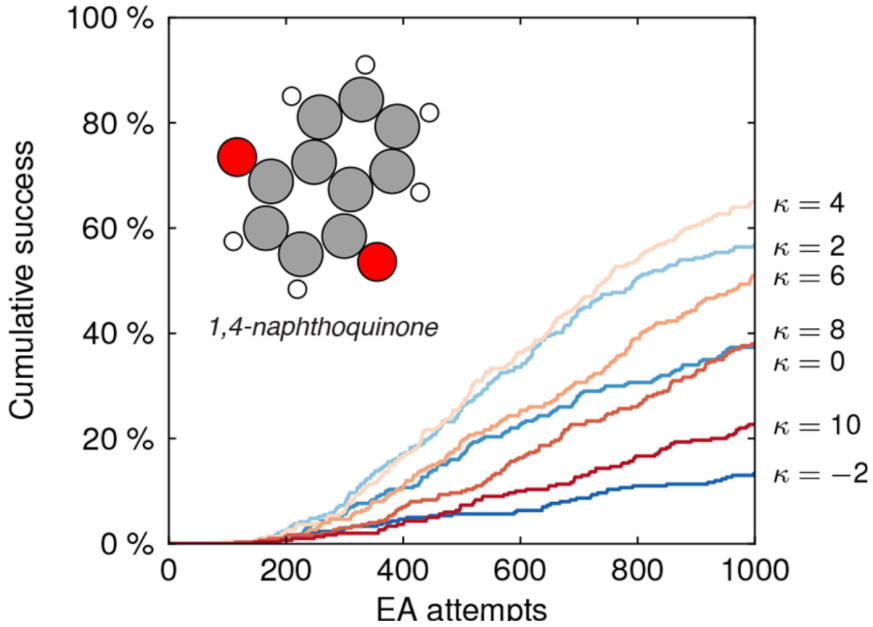
Global structure search: exploitation vs. exploration

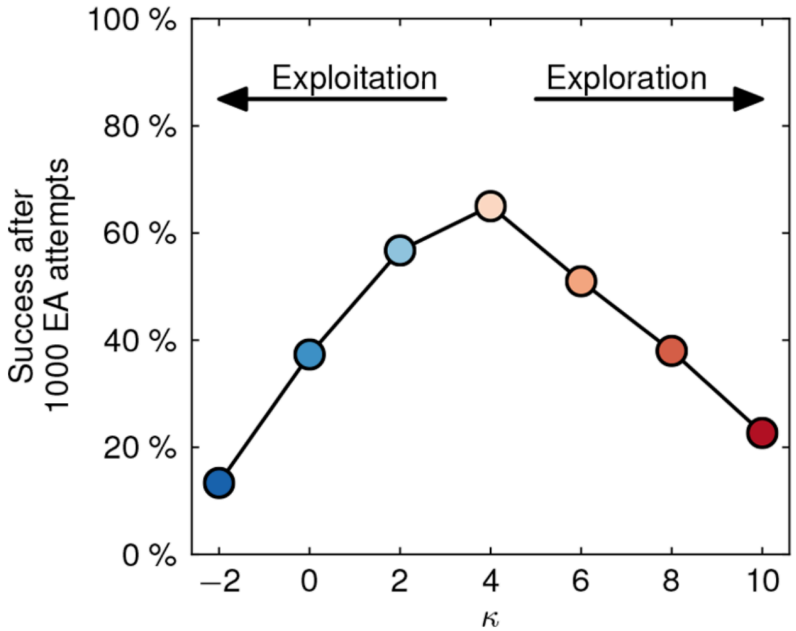
Fitness function in evolutionary algorithm

$$F_i = E_i - \kappa\sigma_i$$

Energy Uncertainty





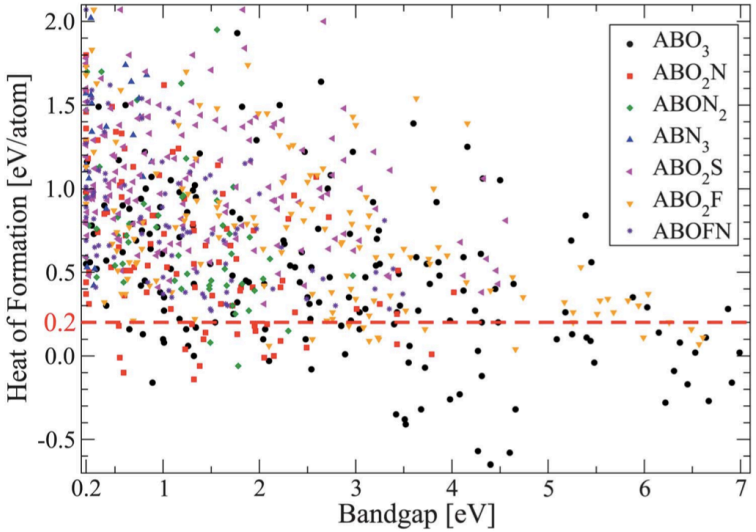


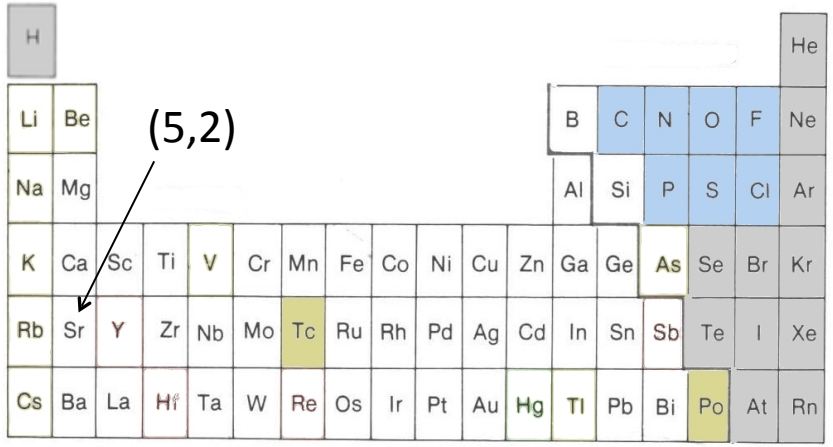
(Jørgensen, Larsen, Jacobsen, and Hammer, *J. Phys. Chem A*, **122**, 1504 (2018))

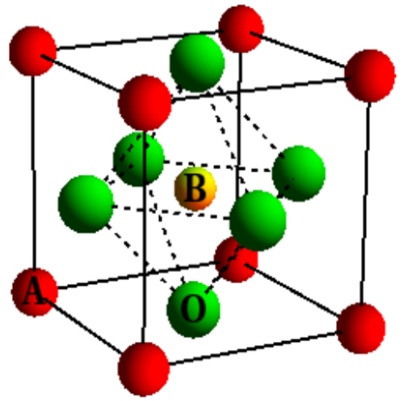
(See also "BOSS" Todorović, Gutmann, Corander, Rinke, *npj Comput Mater* **5**, 35 (2019))

Back to computational screening with machine learning: water splitting

About 19000 cubic perovskites oxides, oxynitrides, oxysulfides, oxyfluorides, oxyfluornitrides







ABO_3, $ABON_2$, ..

Fingerprint (x-vector):

$$x(\mathrm{SrTaO_2N}) = (5, 2, 6, 5, 2, 1, 0, 0)$$

Sr "coordinates"

O, N, S, F

Kernel function:

$$k(x_i, x_j) = \exp(-|x_i - x_j|^2/2\rho^2)$$

Water splitting with Gaussian process

Training on 500 perovskites (~2.6 % of the total dataset).

Prediction: $y(x) = \mathbf{k}^T \mathbf{K}^{-1} \mathbf{y}$ with

Only determined by "metric" (not by data)

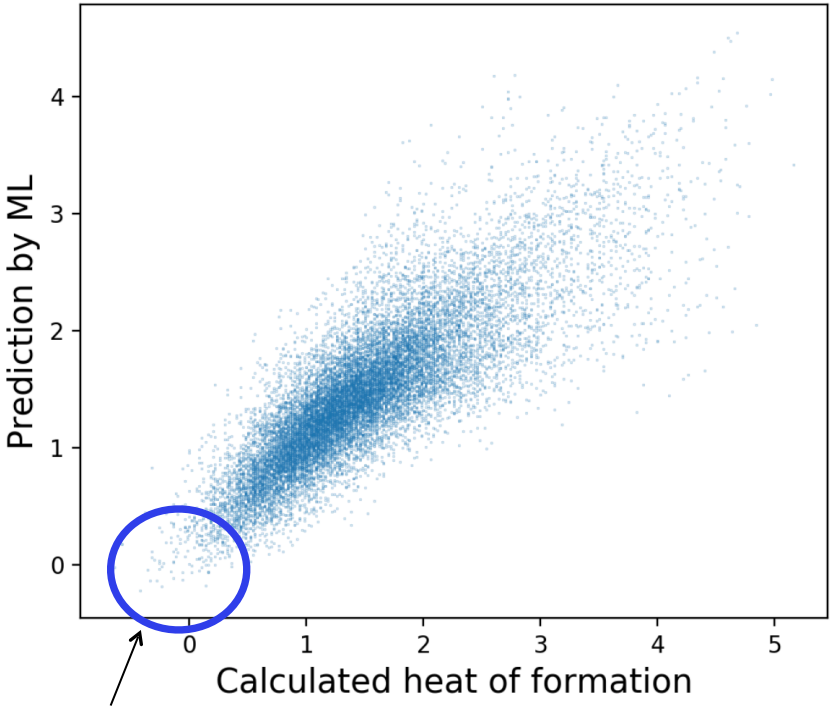
Data

$$k_i = k(x, x_i)$$

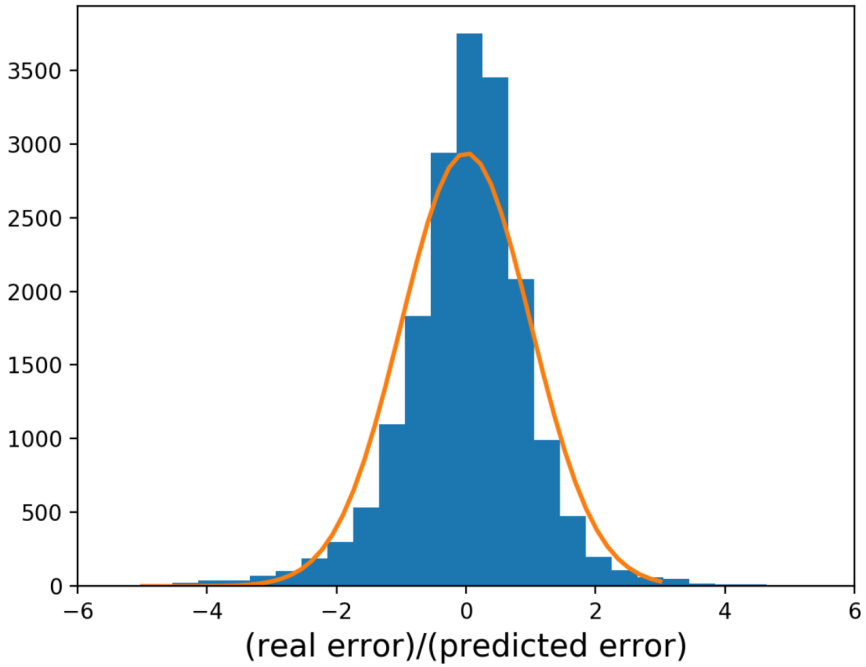
+ error prediction

Example: Heat of formation

Mean Absolute Error: 0.28 eV
Mean Absolute Predicted Error: 0.38 eV



Stable compounds



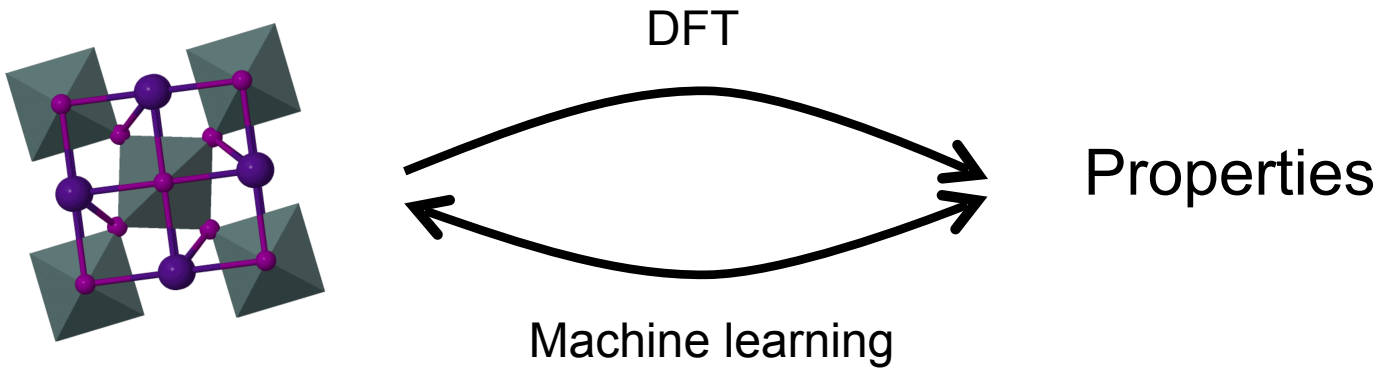
Large reduction in the number of necessary DFT calculations for stable compounds!

Machine learning accelerated computational screening of *new* materials

Two challenges:

1) Can we predict material properties for materials in many different structures where the detailed atomic positions are not known?

2) Can we invert the process so we go directly from properties to material? (to avoid evaluation of properties of maybe billions of (irrelevant) materials)

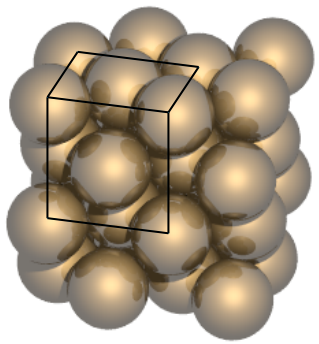


How do we classify materials without using atomic positions?

Composition, symmetry and prototypes:

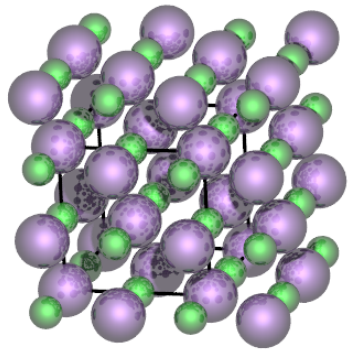
FCC Cu

Space group 225
Only variable is lattice parameter



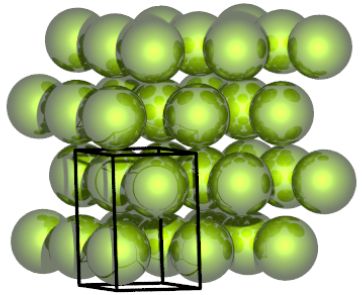
Rocksalt NaCl

Space group 225
Only variable is lattice parameter



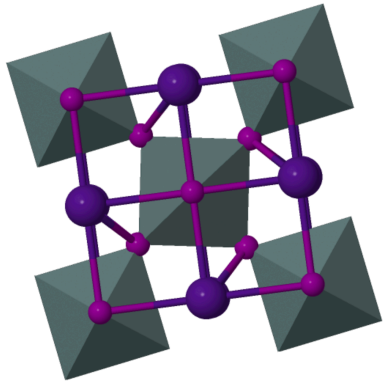
HCP Mg

Space group 194
Two lattice parameters
c = 1.624*a



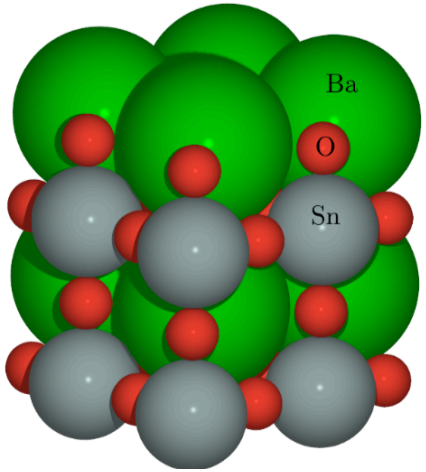
$CsSnI_3$

Space group 127
Two lattice parameters
Rotation angle

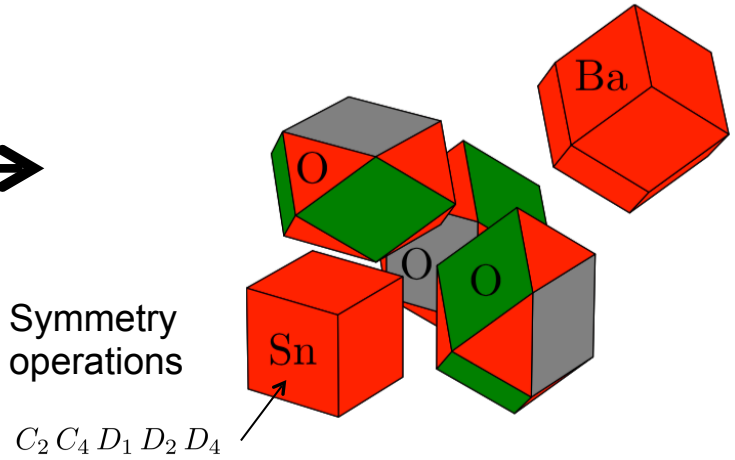


Voronoi cells and graphs

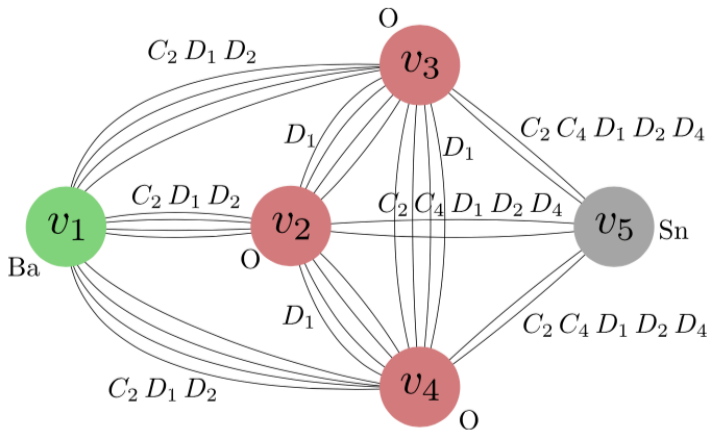
$BaSnO_3$ cubic perovskite



Voronoi (Wigner-Seitz) cells



Symmetry-labeled graph



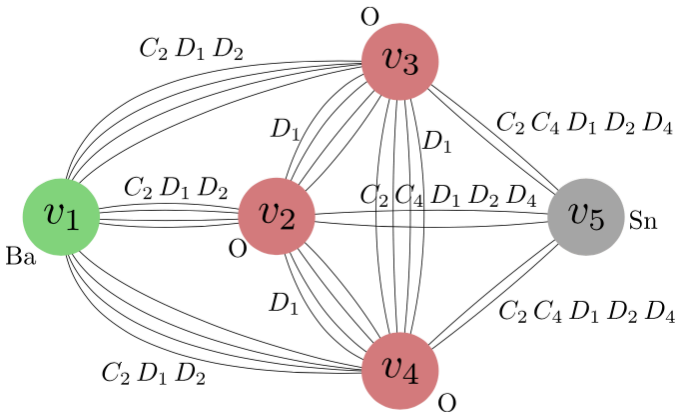
Schütt, Arbabzadah, Chmiela, Müller, Tkatchenko, Nat Commun, **8**, 13890 (2017)
Xie and Grossman, PRL (2018)
Peter B. Jørgensen, Estefanía Garijo del Río, Mikkel N. Schmidt, Karsten W. Jacobsen, arXiv:1905.06048 (2019)

For an alternative approach using Wyckoff sites see Jain, Bligaard, *Phys. Rev. B*, **98**, 1–7 (2018)

Quotient graph

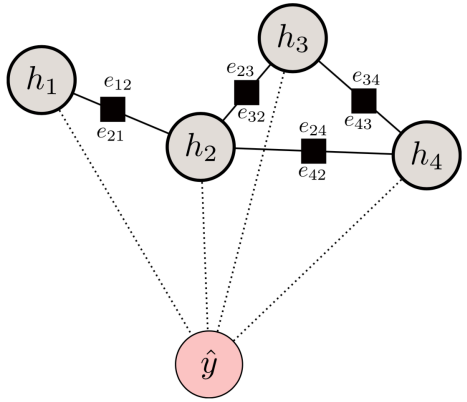
Message passing neural network

Only input:
Atomic numbers Z
Symmetry-labeled quotient graph



atom -> node
bond -> edge

Message passing neural network



$$m_v^{t+1} = \sum_{w \in N(v)} M_t(h_v^t, h_w^t, e_{vw}^t)$$

$$h_v^{t+1} = S_t\left(h_v^t, m_v^{t+1}\right)$$

$$e_{vw}^{t+1} = E_t\left(h_v^{t+1}, h_w^{t+1}, e_{vw}^t\right)$$

$$\hat{y} = R\left(\{h_v^T \in G\}\right)$$

Peter B. Jørgensen, Estefanía Garijo del Río, Mikkel N. Schmidt, Karsten W. Jacobsen, arXiv:1905.06048 (2019)

Neural networks for materials

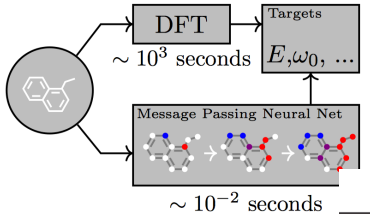
arXiv:1704.01212

Neural Message Passing for Quantum Chemistry

Justin Gilmer [1] Samuel S. Schoenholz [1] Patrick F. Riley [2] Oriol Vinyals [3] George E. Dahl [1]

Abstract

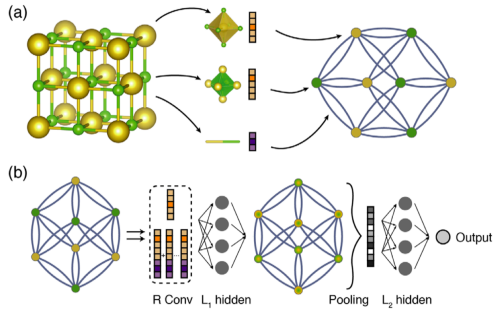
Supervised learning on molecules has incredible potential to be useful in chemistry, drug discovery, and materials science. Luckily, several promising and closely related neural network models invariant to molecular symmetries have already been described in the literature. These models learn a message passing algorithm and aggregation procedure to compute a function of their entire input graph. At this point, the next step is to find a particularly effective variant of



PHYSICAL REVIEW LETTERS **120,** 145301 (2018)

Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties

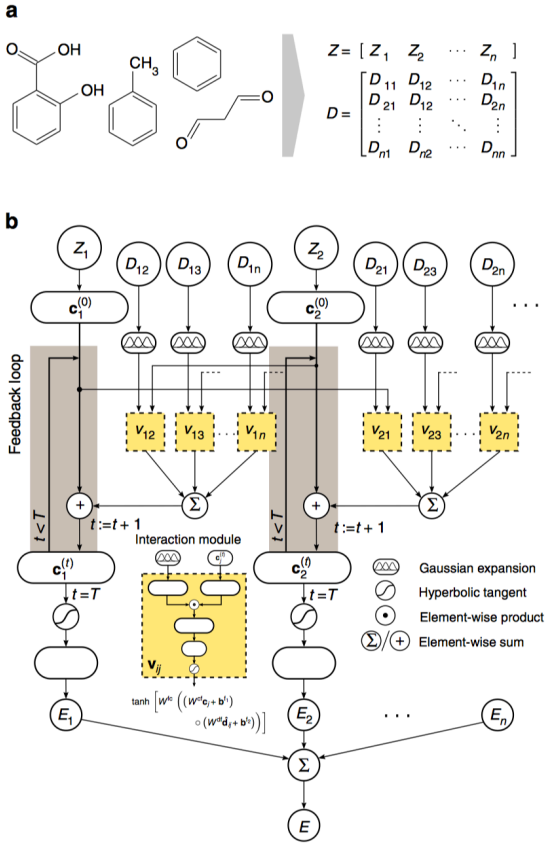
Tian Xie and Jeffrey C. Grossman
Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA



NATURE COMMUNICATIONS | DOI: 10.1038/ncomms13890

Quantum-chemical insights from deep tensor neural networks

Kristof T. Schütt[1], Farhad Arbabzadah[1], Stefan Chmiela[1], Klaus R. Müller[1,2] & Alexandre Tkatchenko[3,4]



Predictions on the Open Quantum Materials Database (OQMD)

~500000 DFT calculations for inorganic materials

Heat of formation:
Mean absolute error

Dataset	Dist.	Sym	No sym	V-RF
OQMD all	14	22	26	85

meV!

Ward, Liu, Krishna, Hegde, Agrawal, Choudhary, Wolverton, Phys. Rev. B Condens. Matter 96, 024104 (2017).
(Random forest algorithm based on Voronoi construction, but with some distance information.)

Only Voronoi graph (+symmetry)

5-fold cross validation

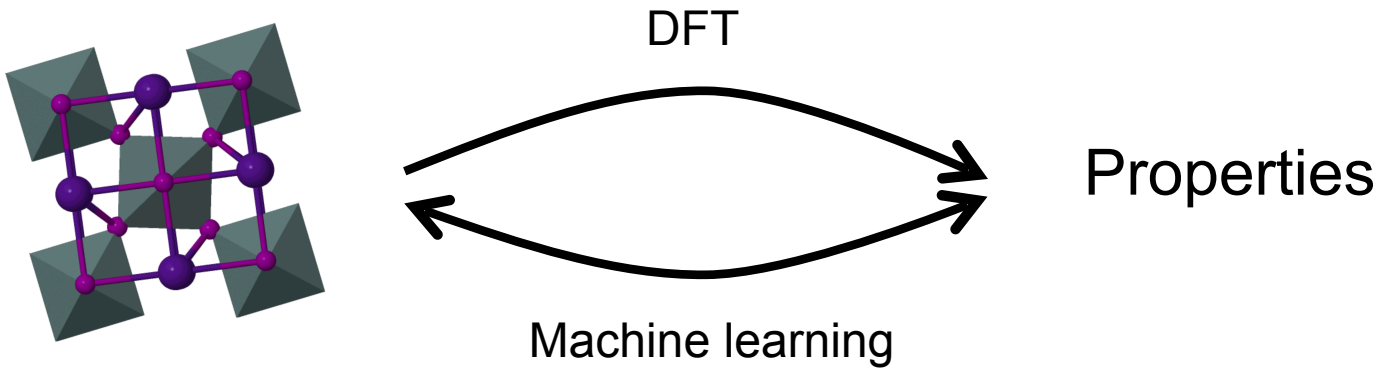
Mean absolute error on formation energies 20 meV!
Accuracy of DFT ~100-200 meV.

Machine learning accelerated computational screening of *new* materials

Two challenges:

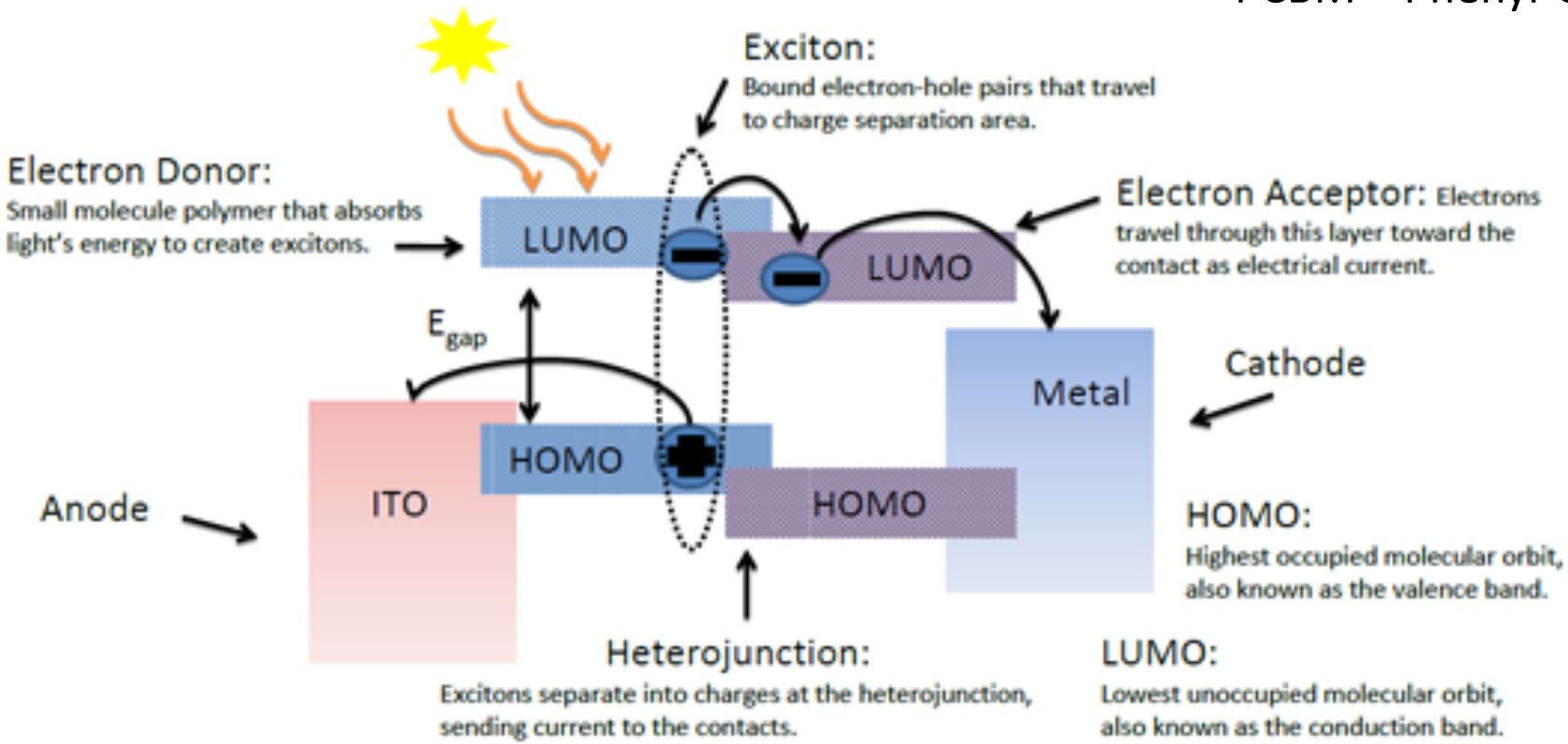
1) Can we predict material properties for materials in many different structures where the detailed atomic positions are not known?

2) Can we invert the process so we go directly from properties to material? (to avoid evaluation of properties of maybe billions of (irrelevant) materials)



Organic solar cell
(PCBM-based blended polymer solar cell)

PCBM = Phenyl-C’61-Butyric-Acid-Methyl-Ester



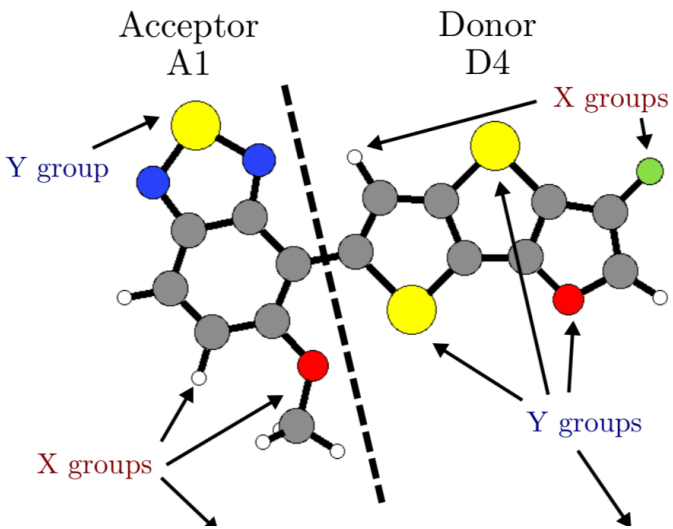
Peter Bjørn Jørgensen, Murat Mesta, Suranjan Shil, Juan Maria García Lastra, Karsten Wedel Jacobsen, Kristian Sommer Thygesen, and Mikkel N. Schmidt
The Journal of Chemical Physics **148**, special issue (2018)

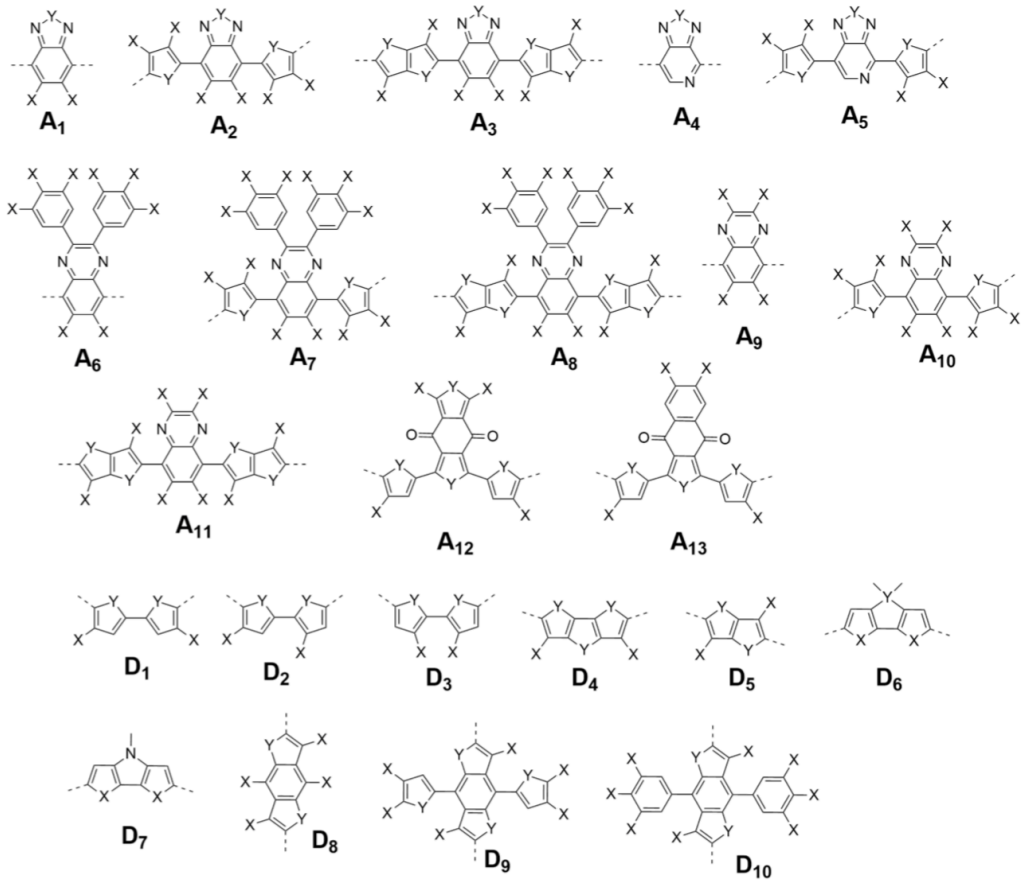
Donor-acceptor molecules (polymer units)

What is the position of the LUMO and the optical gap for these molecules?

Training set with 3989 molecules (Gaussian, B3LYP)

In principle 10^{14} molecules!
One prediction 1ms
-> Total 10^{11} sec
~ 3000 years





A(1-13) = Acceptors
D(1-10) = Donors

X = H, F, CH_3, OCH_3, SCH_3
Y (divalent) = O, S, Se, NCH_3
Y (tetravalent) = C, Si, Ge

Data representation

String representation of molecules.

Grammatical production rules.

No specification of atomic coordinates.

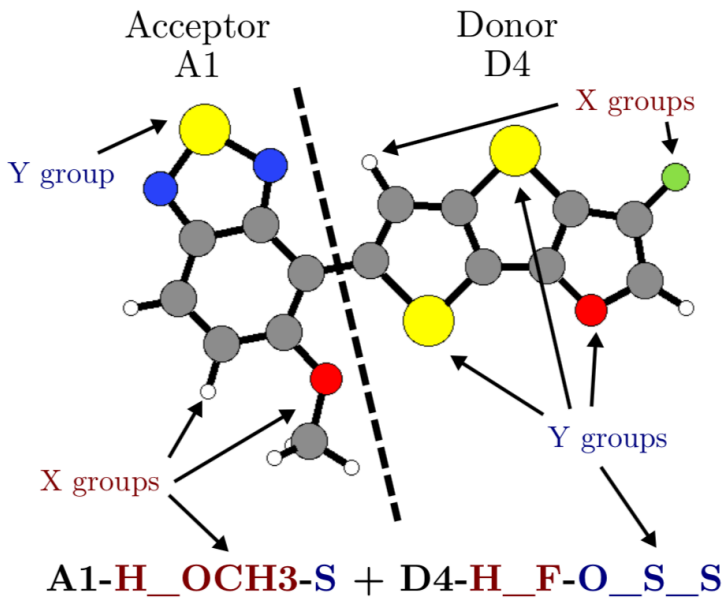


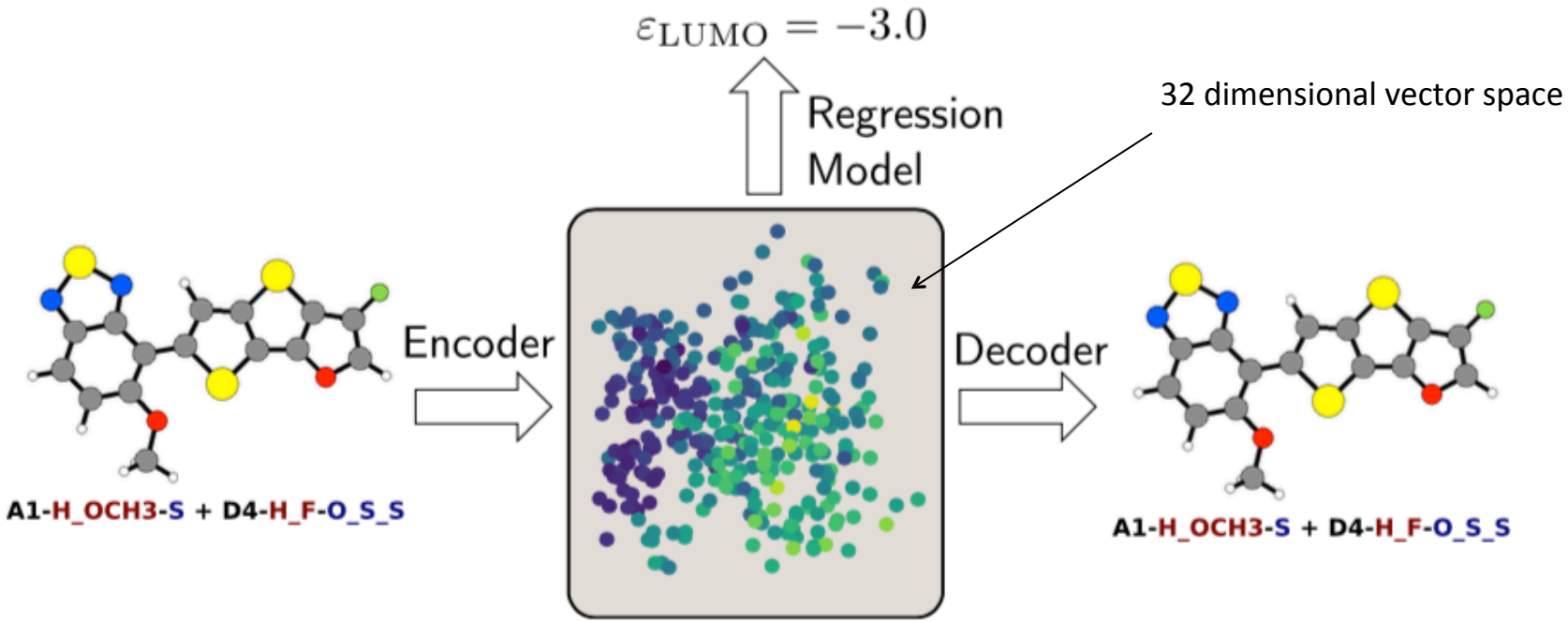
FIG. 2: String representation of one of the molecules of the solar cell dataset: "Acceptor backbone"-"X groups"-"Y groups"+"Donor backbone"-"X groups"-"Y groups". Whenever no side groups are present "*" character is used instead.

Earlier work uses SMILES to represent molecules:

Gómez-Bombarelli et al. (2016), arXiv:1610.02415 [cs.LG].
Kusner et al. (2017), arXiv:1703.01925 [stat.ML].

Variational autoencoder

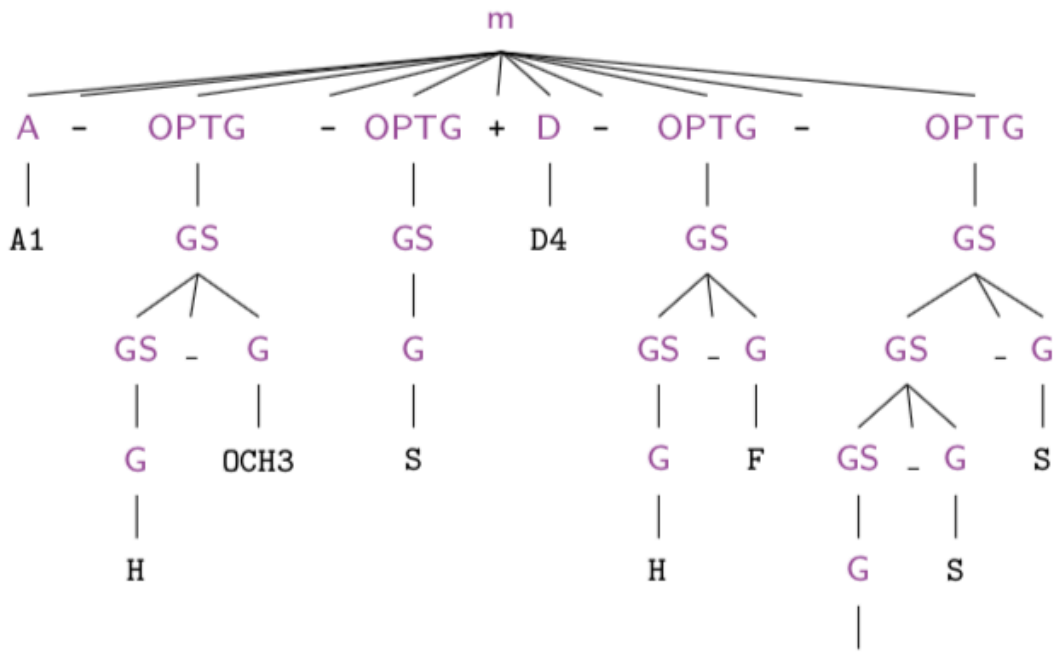
Kingma and Welling [2013], Rezende et al. [2014]



Grammar variational autoencoder

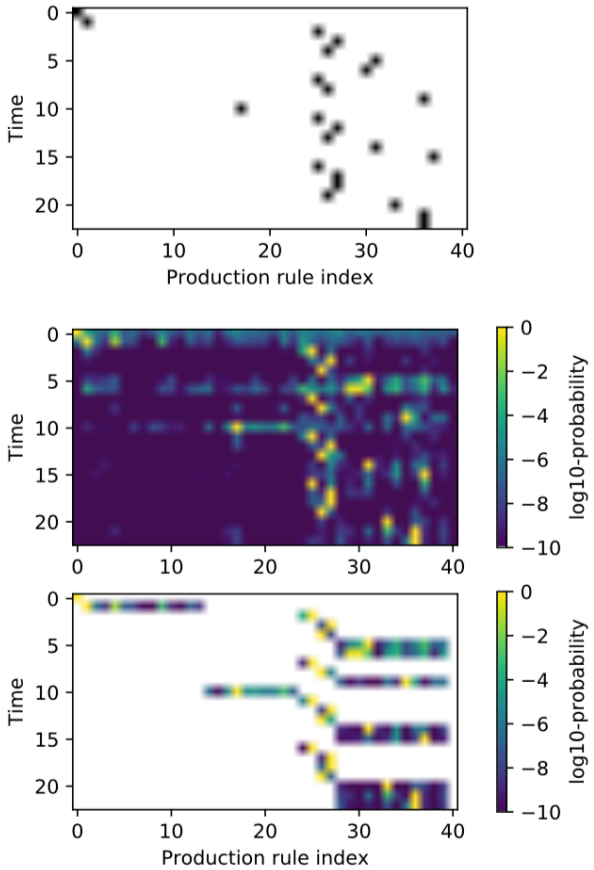
Production rules:

```
m    →  A - OPTG - OPTG + D - OPTG - OPTG
A    →  A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 | A9 | A10 | A11 | A12 | A13
D    →  D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 | D9 | D10
OPTG →  * | GS
GS   →  G | GS _ G
G    →  Ge | CH3 | OCH3 | H | C | O | SCH3 | NCH3 | S | F | Si | Se
```



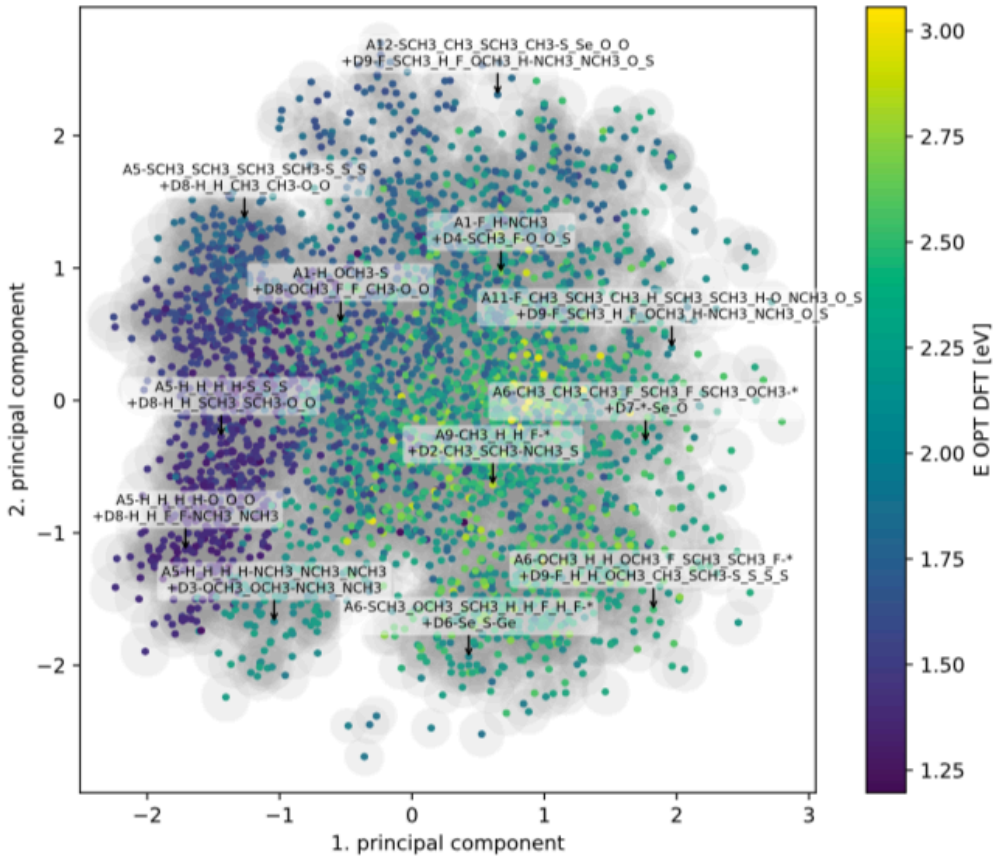
Production rule matrix encoding

m → A - OPTG - OPTG + D - OPTG - OPTG
A → A1
OPTG → GS
GS → GS _ G
GS → G
G → H
G → OCH3
OPTG → GS
GS → G
G → S
D → D4
OPTG → GS
GS → GS _ G
GS → G
G → H
G → F
OPTG → GS
GS → GS _ G
GS → GS _ G
GS → G
G → O
G → S
G → S

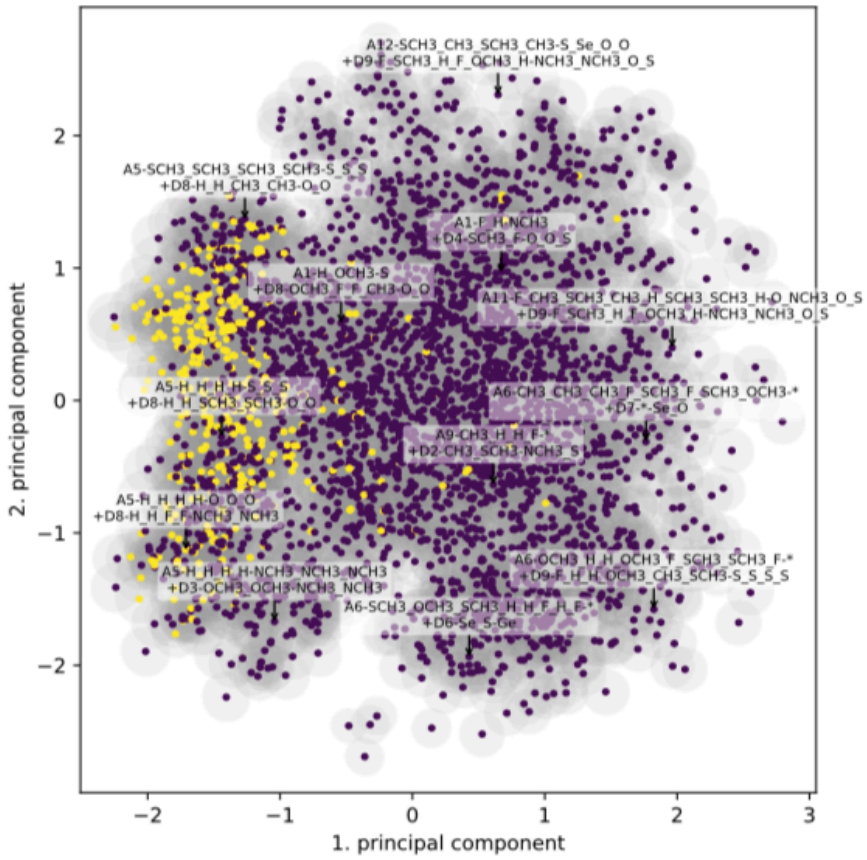


Latent space

First 2 principal components of 32-dimensional space



Colored according to optical gap



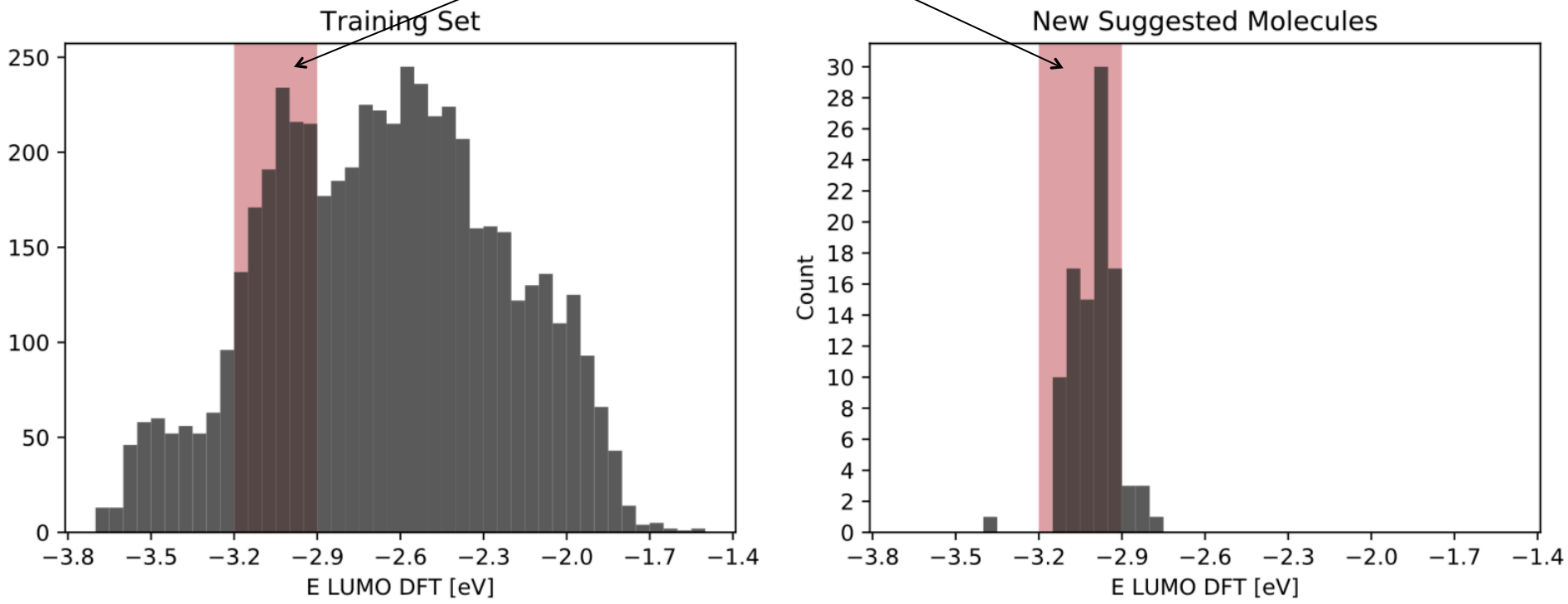
Bright points are within target range

Prediction of new molecules

LUMO energy

Target region

100 new molecules predicted



Acknowledgments

- **CAMD/DTU**
  Mohnish Pandey
  Korina Kuhar
  Estefanía Garijo del Río
  Ivano E. Castelli
  Thomas Olsen
  Kristian S. Thygesen
  Jens K. Nørskov
- **DTU COMPUTE**
  Peter Bjørn Jørgensen
  Mikkel N. Schmidt
- **Aarhus University**
  Bjørk Hammer
  Mathias S. Jørgensen
  Uffe F. Larsen
- **SURFCAT/DTU:**
  Andrea Crovetto
  Brian Seger
  Søren Dahl
  Peter Vesborg
  Ole Hansen
  Ib Chorkendorff
- **SUNCAT/SLAC/STANFORD**
  Thomas Bligaard
  Jose Garrido Torres

CASE
Catalysis for Sustainable Energy

neec
Center on Nanostructuring for Efficient Energy Conversion

The VILLUM Center for the Science of Sustainable Fuels and Chemicals