Predicting Properties of Molecules and Materials

Matthias Rupp

Fritz Haber Institute of the Max Planck Society, Berlin, Germany

2017 NYU Shanghai Summer School on
Machine Learning in the Molecular Sciences

June 12–16, Shanghai, China

Outline

1. Validation
   statistical validation, free parameters
2. Predicting experiments
   cheminformatics, examples
3. Predicting calculations
   QM/ML models, examples

Validation

Why?

- assess model performance
- optimize free parameters (hyperparameters)

Which statistics?

- root mean squared error (RMSE)
- mean absolute error (MAE)
- maximum error
- squared correlation coefficient (R^2)

What else can we learn from validation?

- distribution of errors, not only summary statistics
- convergence of error with number of samples

Validation

Golden rule:
Never use training data for validation

Violation of this rule leads to overfitting
by measuring flexibility in fitting instead of generalization ability
rote learner example

If there is sufficient data:

- divide data into two subsets, training and validation
- build model on training subset
- estimate error of trained model on validation subset

Sometimes an external validation set is used in addition.

Statistical validation

If too few data, statistical re-sampling methods can be used, such as cross-validation, bagging, bootstrapping, jackknifing

***k*-fold cross-validation:**

- divide data into k evenly sized subsets
- for $i = 1, \ldots, k$,
  build model on union of subsets $\{1, \ldots, k\} \setminus \{i\}$
  and validate on subset i

All model building steps must be repeated for data splits:

- all pre-processing such as feature selection and centering
- optimization of hyperparameters

Hansen et al, *J. Chem. Theor. Comput.*, 3404, 2103

Hyperparameters: physically motivated choices

Length scale σ:

$\sigma \approx \|\boldsymbol{x} - \boldsymbol{z}\|_1$

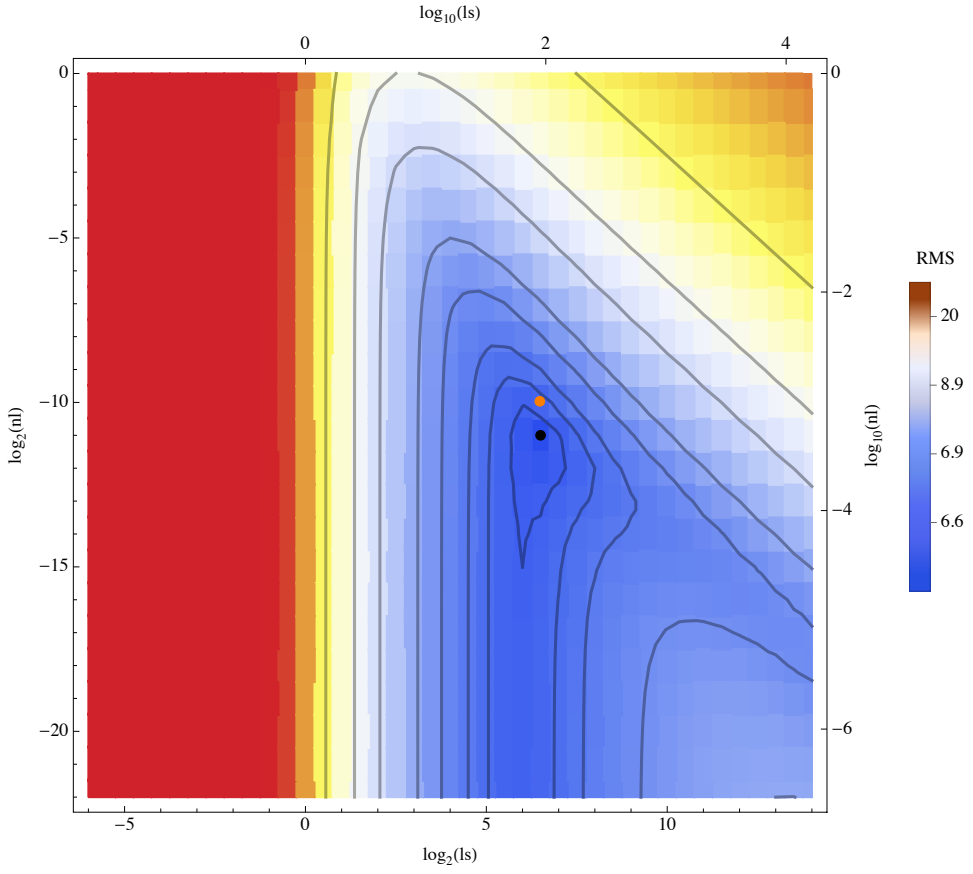
median nearest neighbor distance

Regularization strength λ:

$\hat{=}$ noise variance (Bayesian)

$\hat{=}$ leeway around y_i for fitting

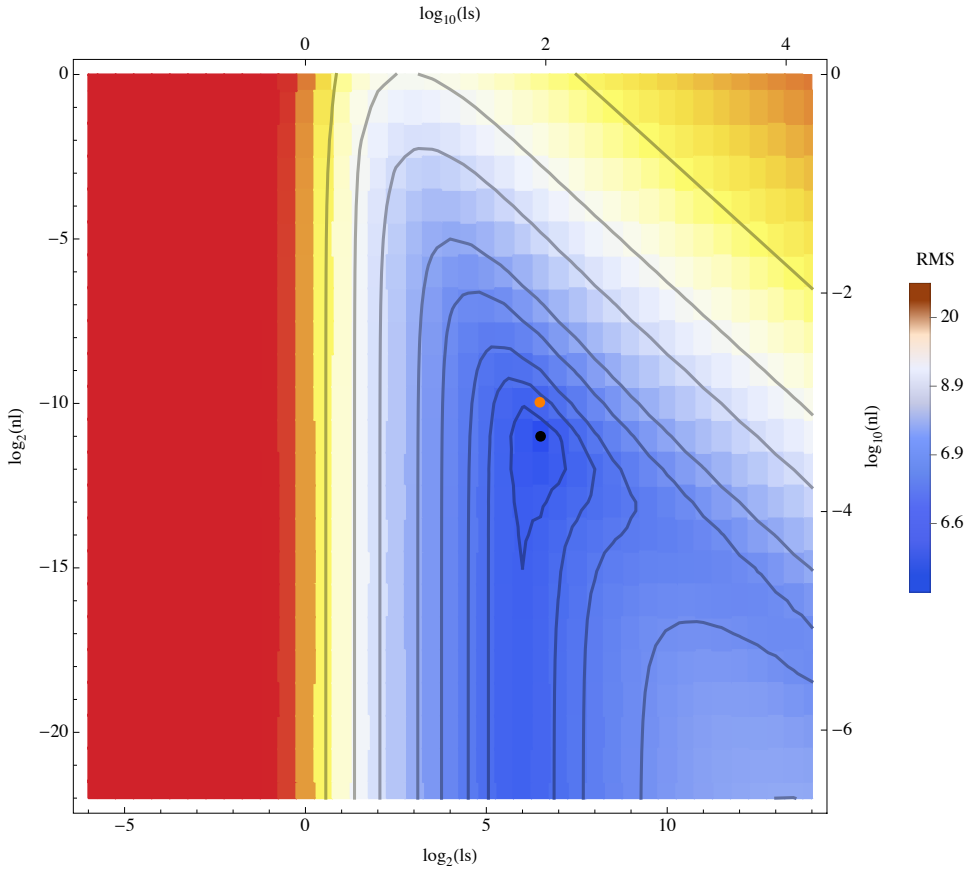
$\Rightarrow$ target accuracy



Rupp: *Int. J. Quant. Chem.*, 1058, 2015

Hyperparameters: statistically motivated choices

- data-driven method for choosing hyperparameters
- optimize using grid search or gradient descent
- use statistical validation to estimate error
- for validation and hyperparameter optimization, use nested data splits



Rupp: *Int. J. Quant. Chem.*, 1058, 2015

Nested data splits

- **never use data from training in validation**
- for performance assessment **and** hyperparameter optimization, use nested cross-validation or nested hold-out sets
- beware of overfitting

Example 1: plain overfitting

× train on all data, predict all data

✓ split data, train, predict

Example 2: centering

× center data, split data, train & predict

✓ split data, center training set, train, center test set, predict

Example 3: cross-validation with feature selection

× feature selection, cross-validation

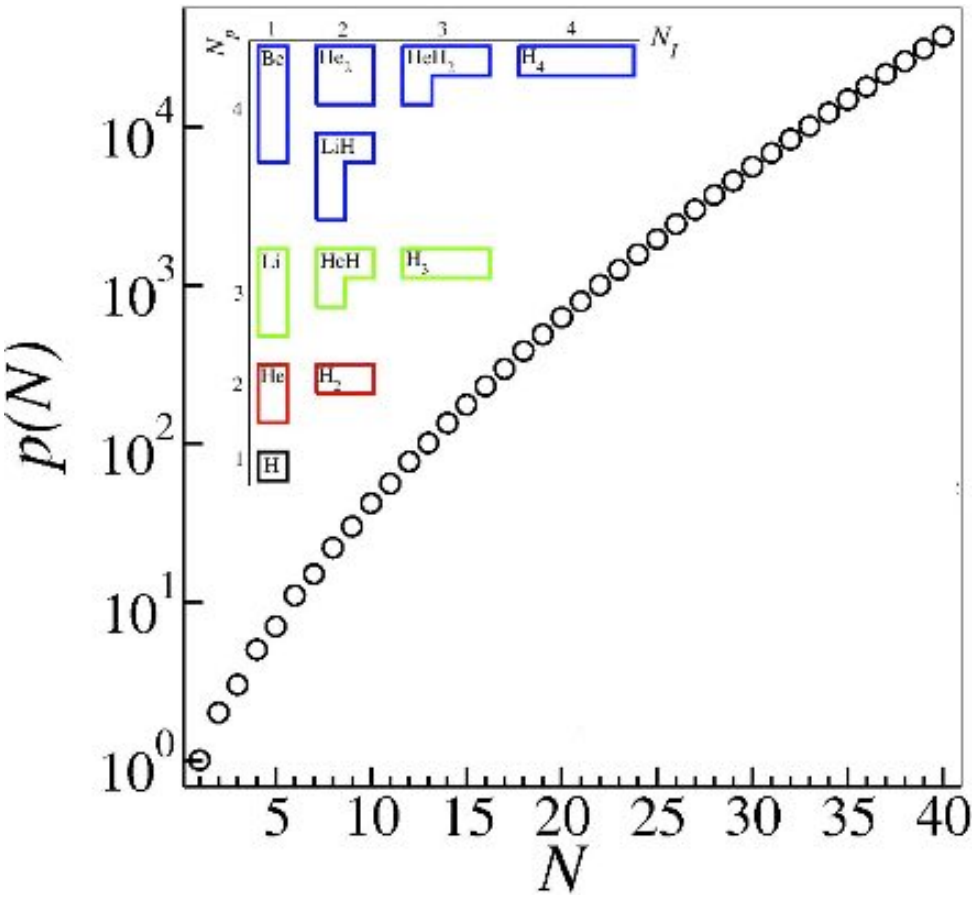
✓ feature selection for each split of cross-validation

Predicting experiments

The combinatorial nature of chemical/materials space

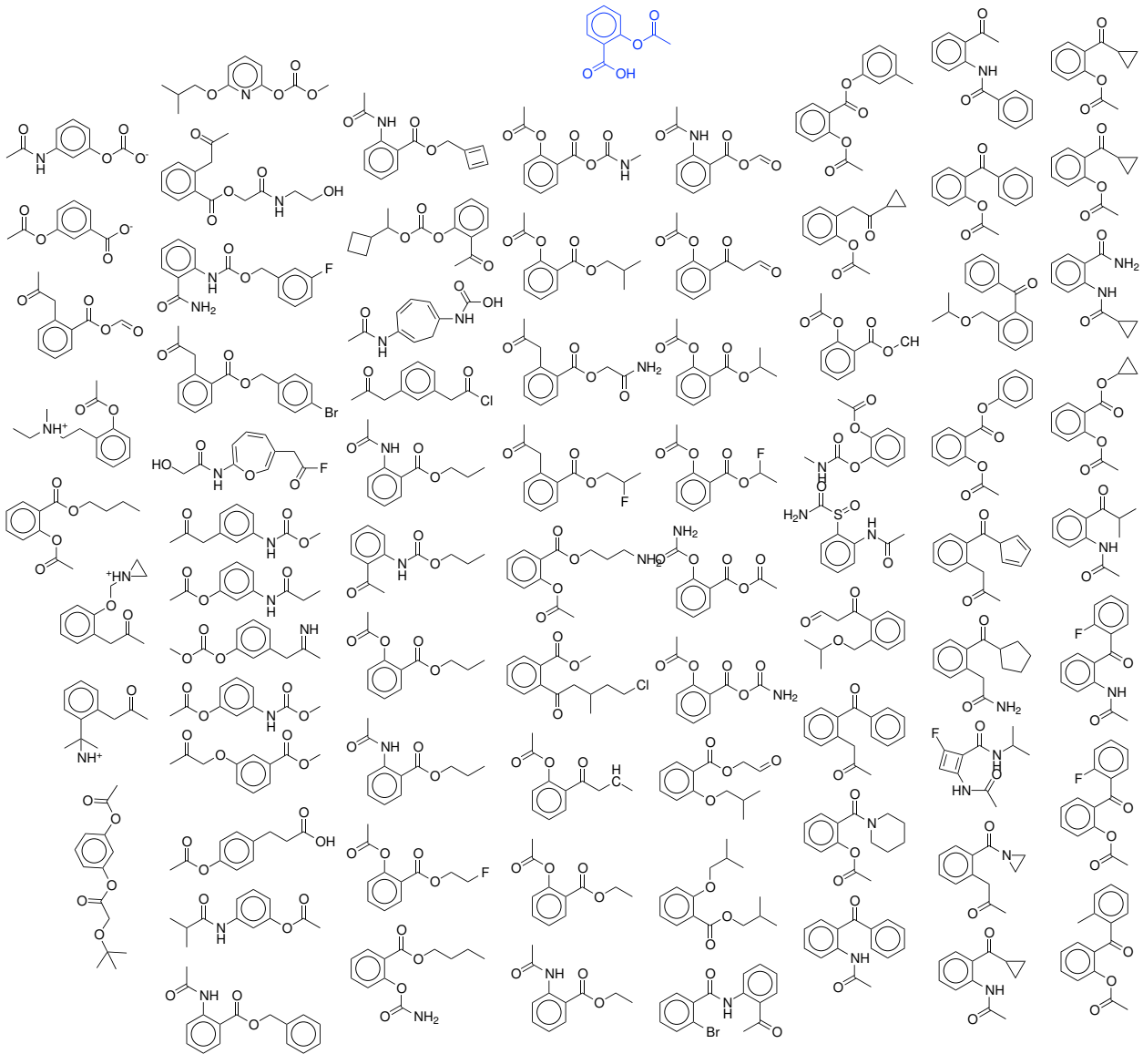
How large is chemical space?

- how many stoichiometries are possible with $N = 40$ protons?
- number of ways to write N as sum of positive integers
- Young-Ferrers diagrams
- $> 3.7 \cdot 10^4$ for $N = 40$



von Lilienfeld, *Int. J. Quant. Chem.* 113: 1676, 2013

The combinatorial nature of chemical/materials space

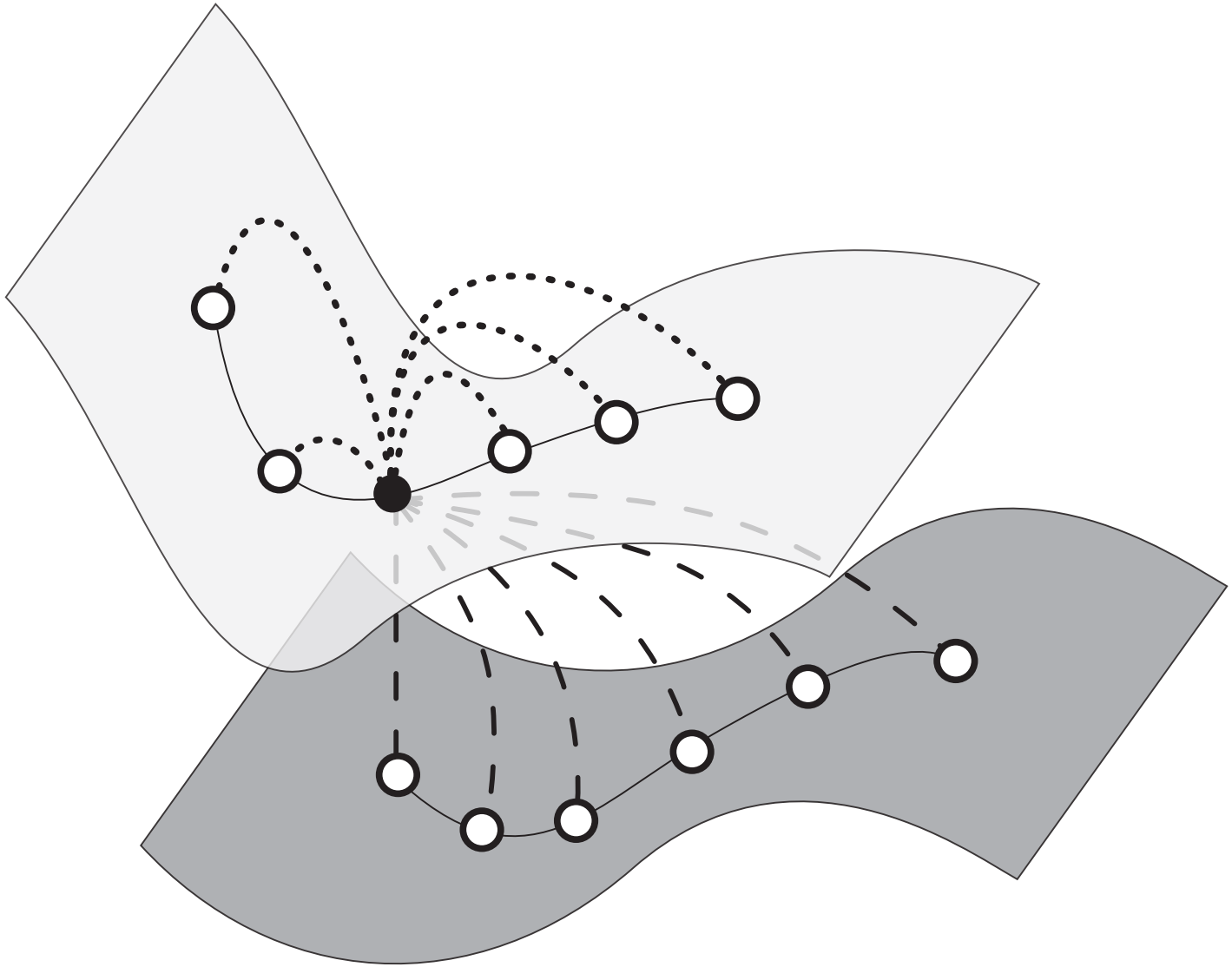


- molecule space: graph theory
- materials space: group theory
- combinatorial explosion

aspirin derivatives

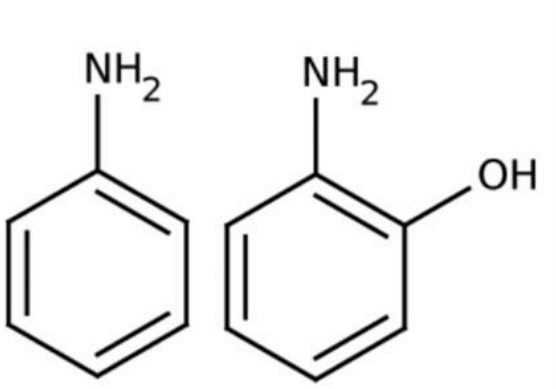
Gómez-Bombarelli et al, arXiv, 2016

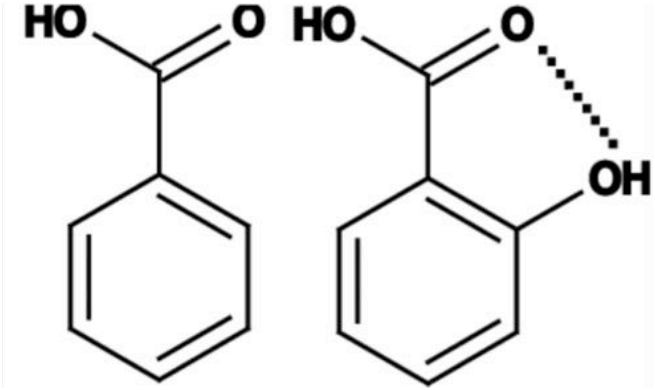
Learning across chemical space



Chang, von Lilienfeld: *CHIMIA* 68, 602, 2014
von Lilienfeld, *Int. J. Quant. Chem.* 113, 1676, 2013.

“Activity cliffs”





- $\Delta pK_a = 0.24$
- phenylamine
  2-aminophenol

- $\Delta pK_a = 1.22$
- benzoic acid
  2-hydroxybenzoic acid

Schönherr, Cernak: *Angew. Chem. Int. Ed.* 52, 12256, 2013

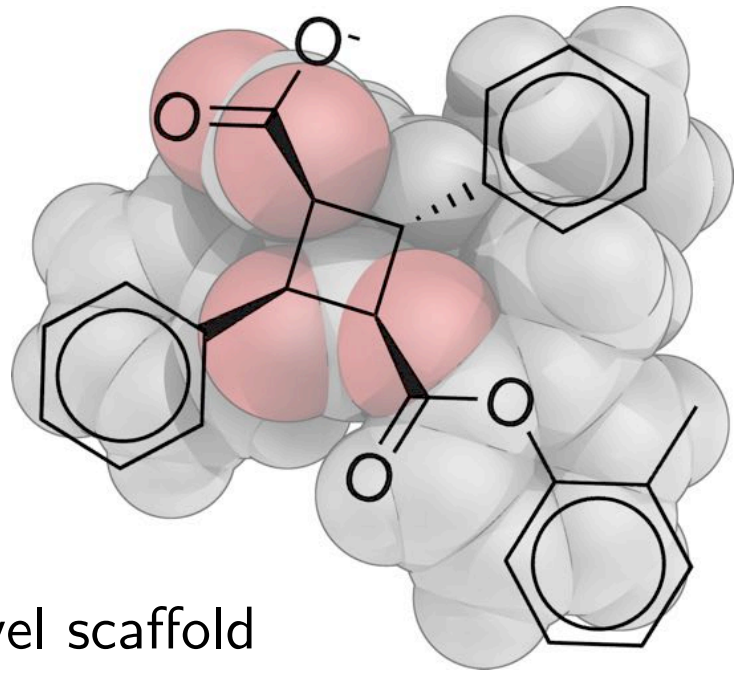
Activation of PPARγ

Target

- peroxisome proliferator-activated receptor γ (PPARγ)
- related to type 2 diabetes and dyslipidemia

Methods

- Gaussian process regression
- descriptors + graph kernel
- cellular reporter gene assay



Results

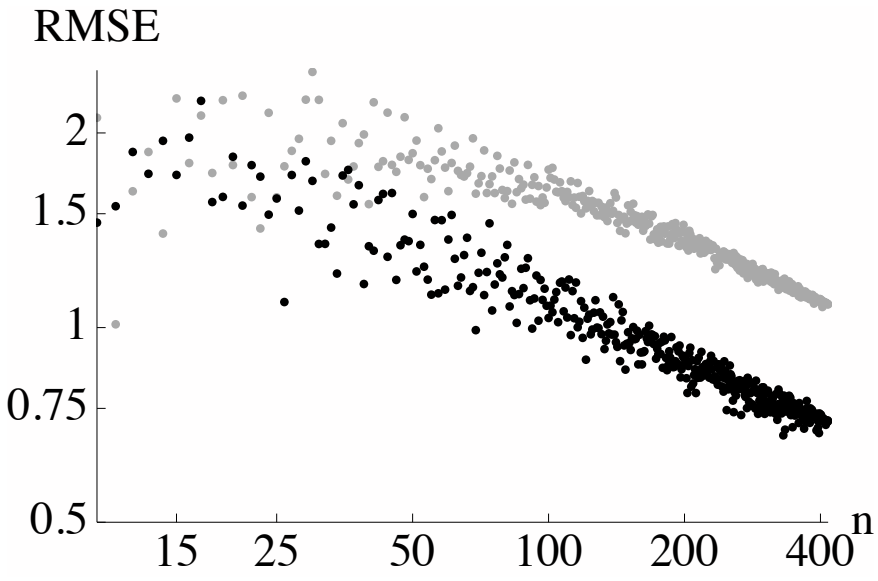
- 8 out of 15 compounds active
- one selective PPARγ agonist with novel scaffold (derivative of natural product truxillic acid), $EC_{50} = 10.03 \pm 0.2\,\mu M$

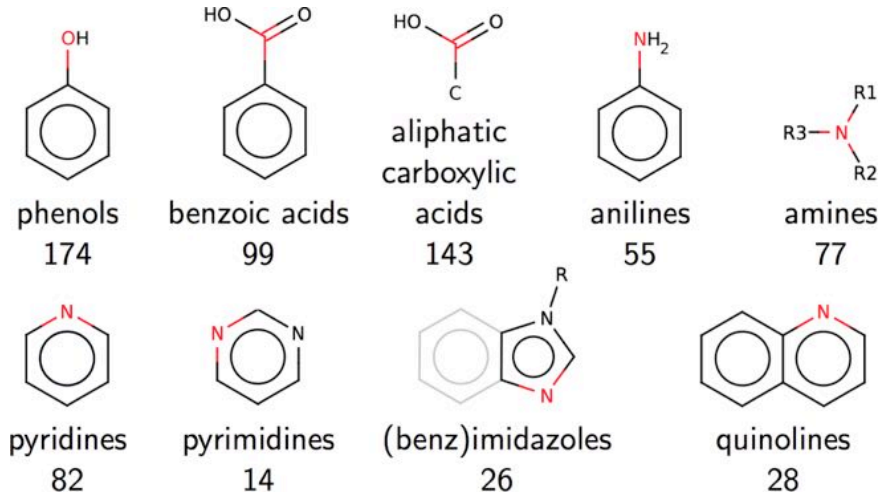
Rupp et al., ChemMedChem 5(2): 191, 2010.

Acid dissociation constants

- "An acid (base) is a species having a tendency to lose (add on) a proton." (Brönsted, 1923)
- pK_a expresses strength of acids and bases (pH where 50 % ionized)

- gray dots: descriptors from electron frontier theory
  black dots: graph kernel





Rupp, Körner, Tetko: *Mol. Inf.* 29, 731, 2010

Predicting calculations

Rationale

"The underlying physical laws necessary for [. . .] chemistry are thus completely known, and the difficulty is only that the exact application of these laws leads to equations much too complicated to be soluble. It therefore becomes desirable that approximate practical methods of applying quantum mechanics should be developed."

Paul A.M. Dirac

Dirac, *Proc. Roy. Soc. A* 123: 714, 1929.

The problem of computational cost

- systematic computational study and design of molecules and materials requires accurate atomic-scale treatment
- accurate numerical solutions have high **computational cost**

Numerical approximations to Schrödinger's equation:

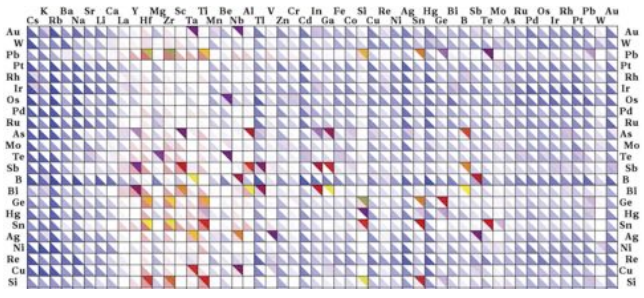
Abrv.	Method	Runtime
FCI	Full Configuration Interaction (CISDTQ)	$O(N^{10})$
CC	Coupled Cluster (CCSD(T))	$O(N^{7})$
FCI	Full Configuration Interaction (CISD)	$O(N^{6})$
MP2	Møller-Plesset second order perturbation theory	$O(N^{5})$
HF	Hartree-Fock	$O(N^{4})$
DFT	Density Functional Theory (Kohn-Sham)	$O(N^{3-4})$
TB	Tight Binding	$O(N^{3})$
MM	Molecular Mechanics	$O(N^{2})$

N = system size

The problem of computational cost

This **limits**

- number of screened systems



Castelli et al, *Energy Environ Sci* 12, 2013

- size of systems

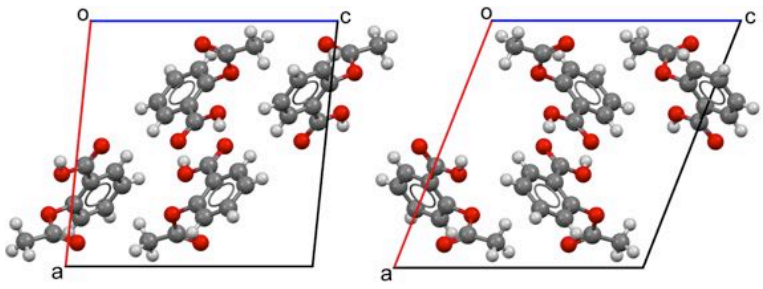


Image: Reilly & Tkatchenko, *Phys Rev Lett* 2014

- length of simulations

Liwo et al, *Proc Natl Acad Sci USA* 102: 2362, 2005

- phenomena studied

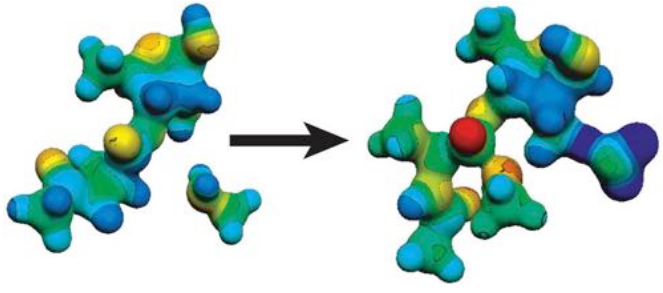
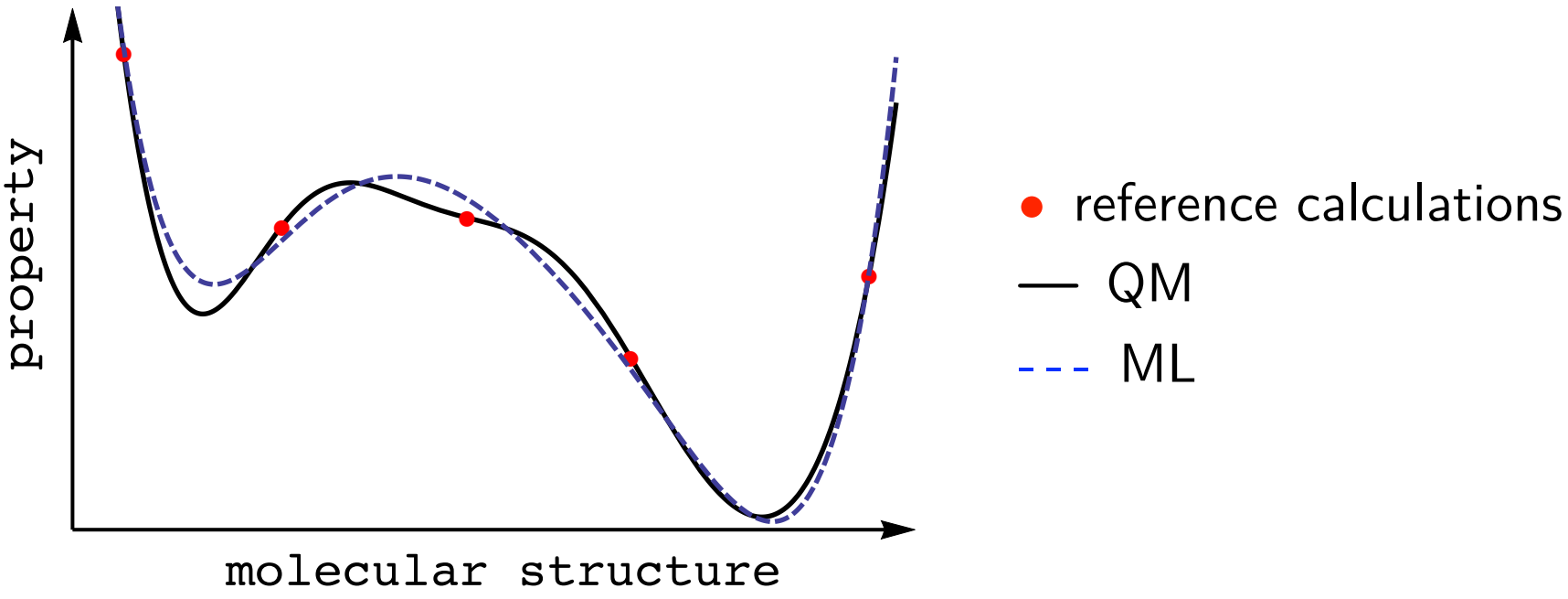


Image: Hiller et al, *Nature* 476: 236, 2011

Machine learning for quantum mechanics

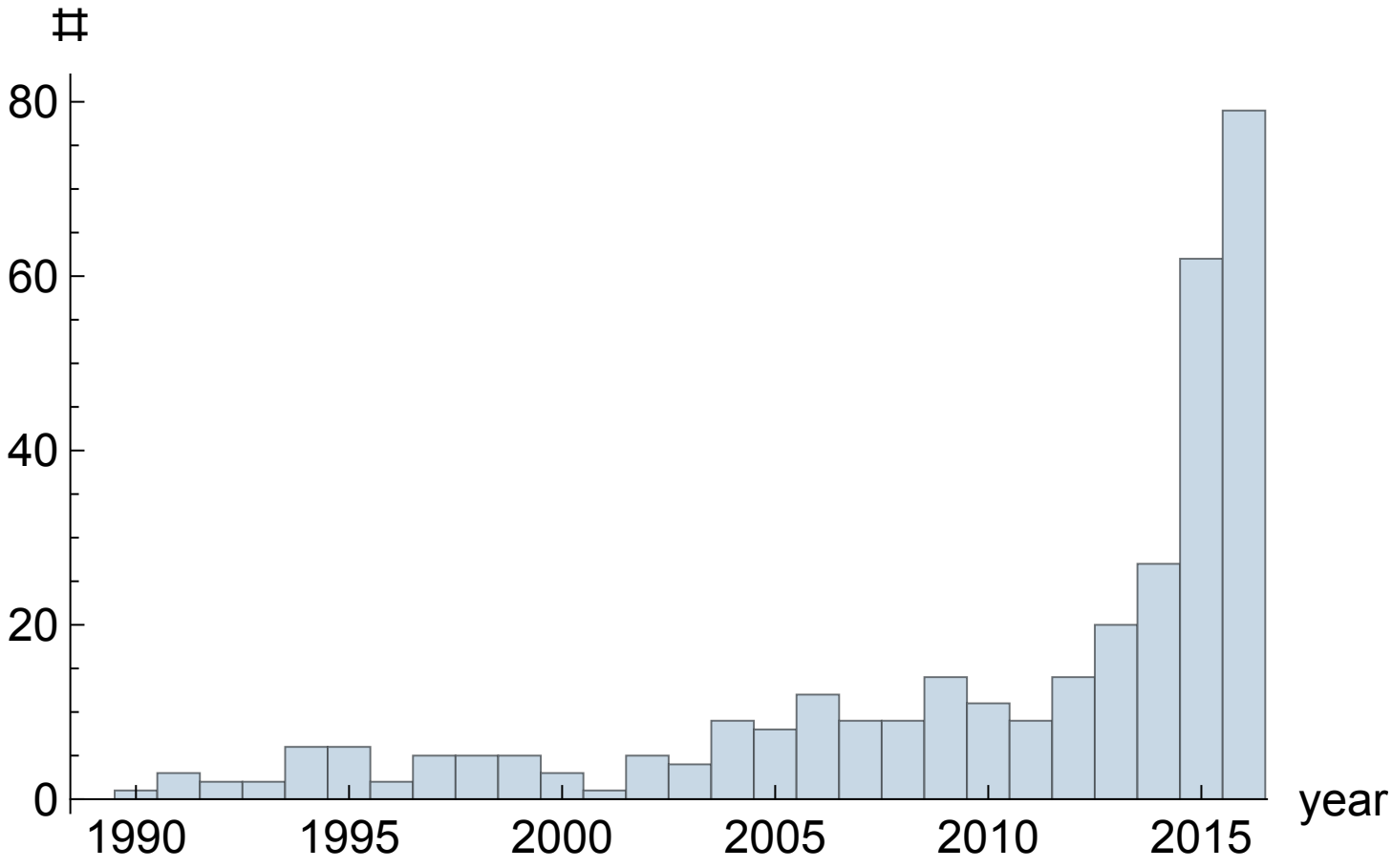
- correlated inputs yield correlated outputs
- exploit redundancy in related QM calculations
- **Interpolate** between QM calculations using ML
- smoothness assumption (regularization)



Relationship to other models

quantum chemistry	force fields	machine learning
generally applicable	limited domain	generally applicable
no or little fitting	fitting to one class	refitted to any dataset
form from physics	form from physics	form from statistics
deductive	mostly deductive	inductive
few or no parameters	some parameters	many parameters
slow	fast	in between
small systems	large systems	large systems

Publications



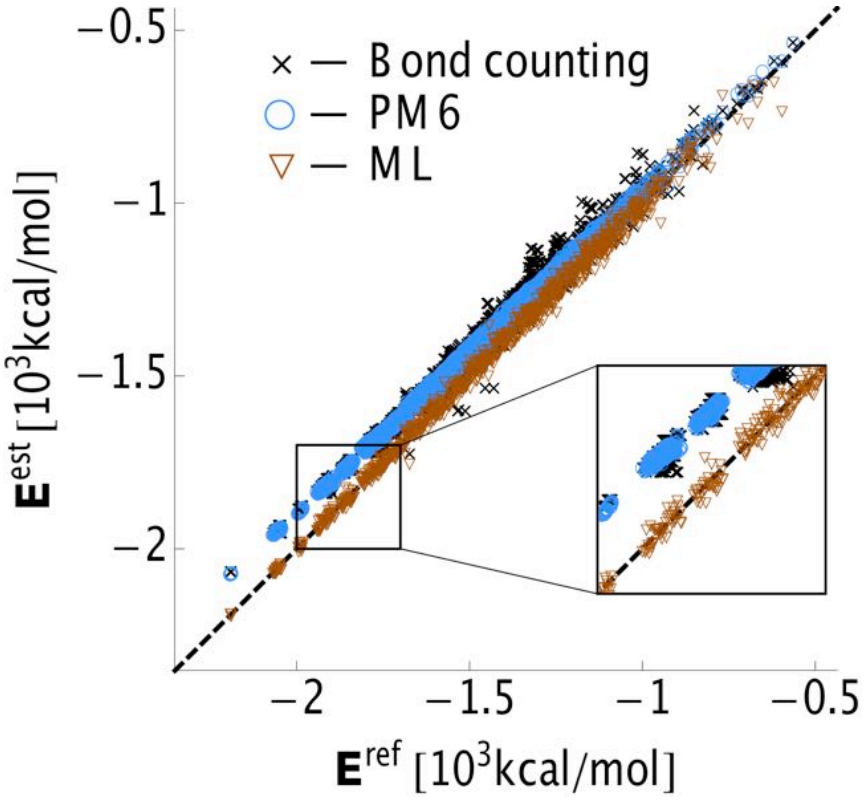
Interdisciplinary field with rapid growth in last years

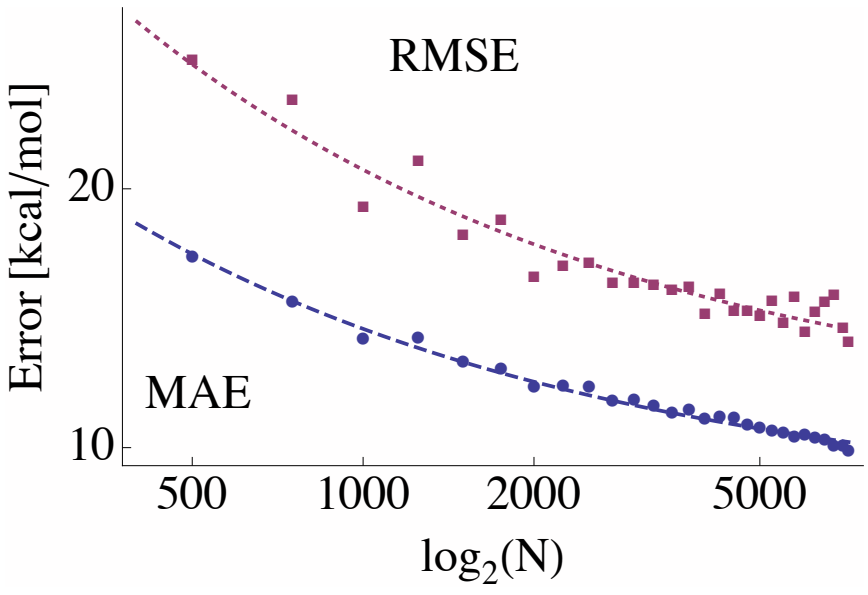
Examples

- screening: **chemical interpolation**
  Rupp *et al.*, Phys. Rev. Lett. 108(5): 058301, 2012
- molecular dynamics: potential energy surfaces
  Behler, Phys. Chem. Chem. Phys. 13(40): 17930, 2011
- dynamics simulations: crack propagation in silicon
  Li et al, *Phys Rev Lett* 114: 096405, 2015.
- crystal structure prediction: (meta)stable states
  Ghiringhelli *et al.*, Phys. Rev. Lett. 114(10): 105503, 2015
- density functional theory: kinetic energies
  Snyder *et al.*, Phys. Rev. Lett. 108(25): 253002, 2012
- transition state theory: dividing surfaces
  Pozun *et al.*, J. Chem. Phys. 136(17): 174101, 2012
- amorphous systems: relaxation in glassy liquids
  Schoenholz, Cubuk et al, *Nat. Phys.* 12(5): 469, 2016
- design: stable interface search
  Kiyohara, Oda, Tsuda, Mizoguchi, *Jpn. J. Appl. Phys.* 55(4): 045502, 2016

Predicting atomization energies

- 7 165 small organic molecules (H,C,N,O,S; 1–7 non-H atoms)
- DFT PBE0 atomization energies
- kernel ridge regression, Gaussian kernel $k(\mathbf{M}, \mathbf{M}') = \exp\left(-\frac{d^2(\mathbf{M},\mathbf{M}')}{2\sigma^2}\right)$



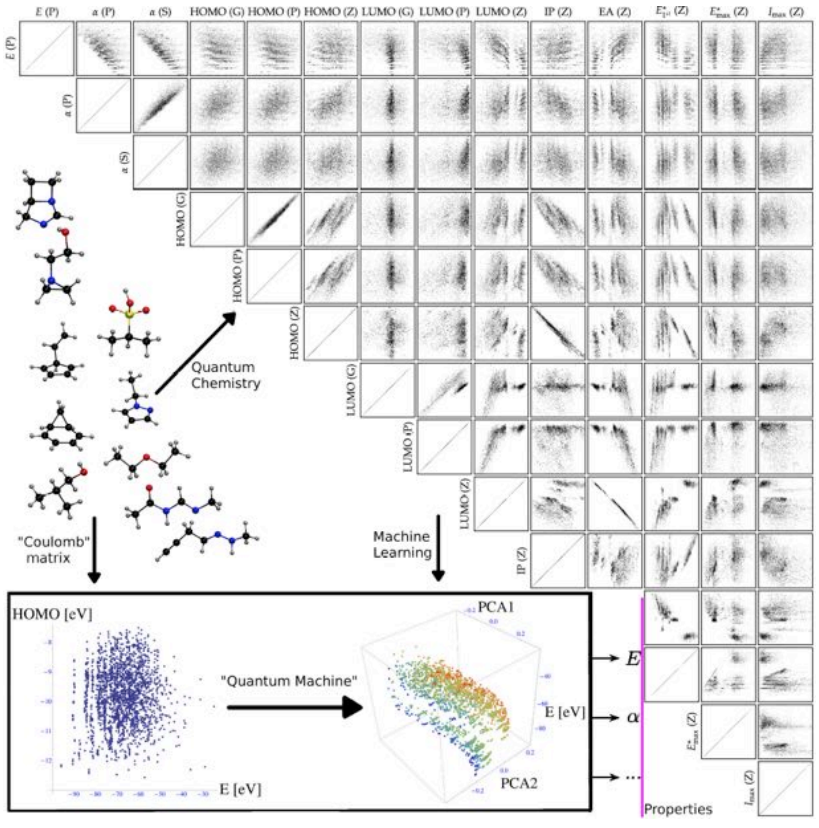


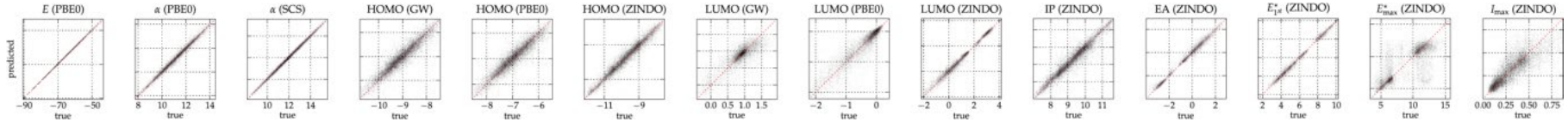
Rupp, Tkatchenko, Müller, von Lilienfeld: *Phys. Rev. Lett.*, 2012

Extension to other properties

Learning the map from molecular structure to molecular properties

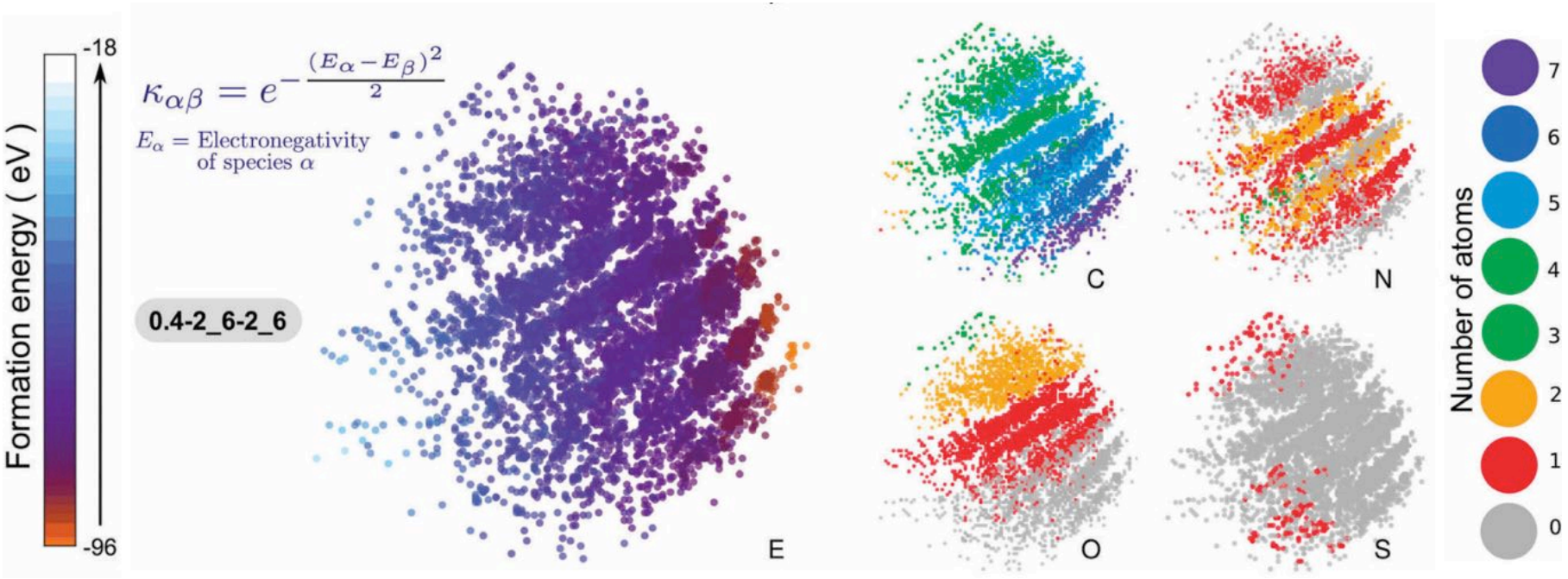
- various properties
- various levels of theory
- small organic molecules
- Coulomb matrix representations
- kernel learning,
  deep neural networks
- for 5k training molecules, errors
  are comparable to the reference



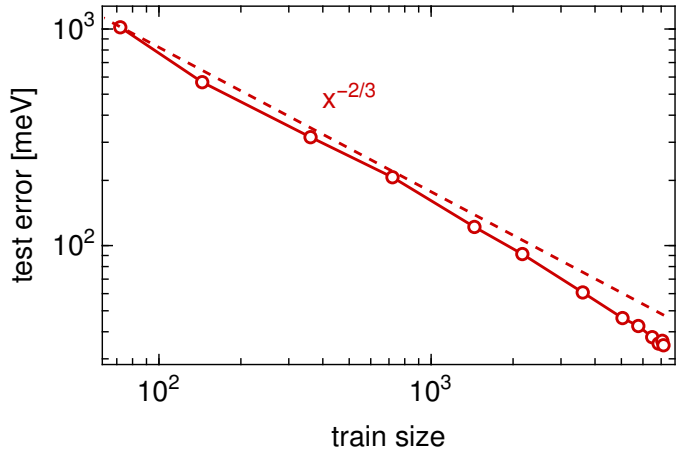


Montavon et al, *New. J. Phys.*, 2013; Hansen et al, *J. Chem. Theor. Comput.*, 2013.

Molecular energies

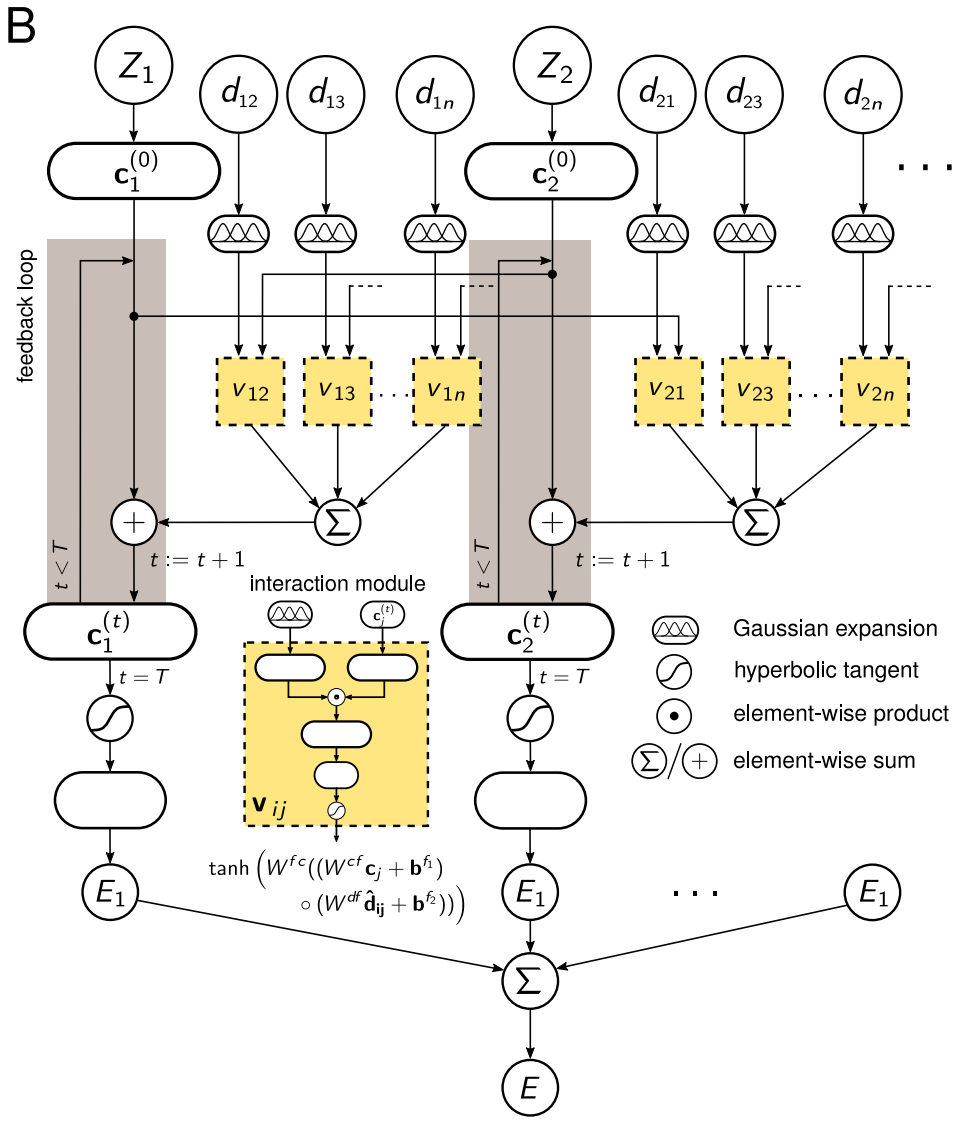


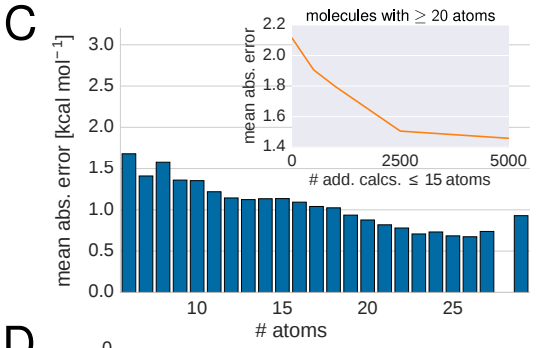
- Gaussian process regression
- regularized entropy match kernel (Sinkhorn distance) with smooth overlap of atomic positions representation
- MAE = 0.6 kcal/mol, RMSE = 0.9 kcal/mol

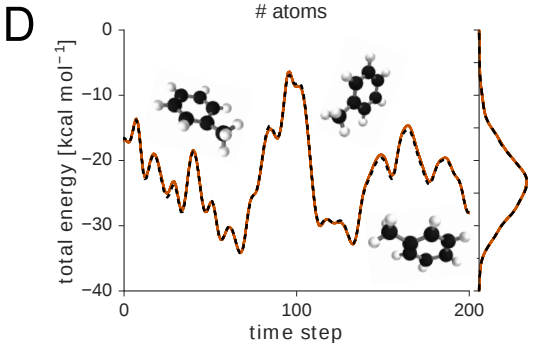


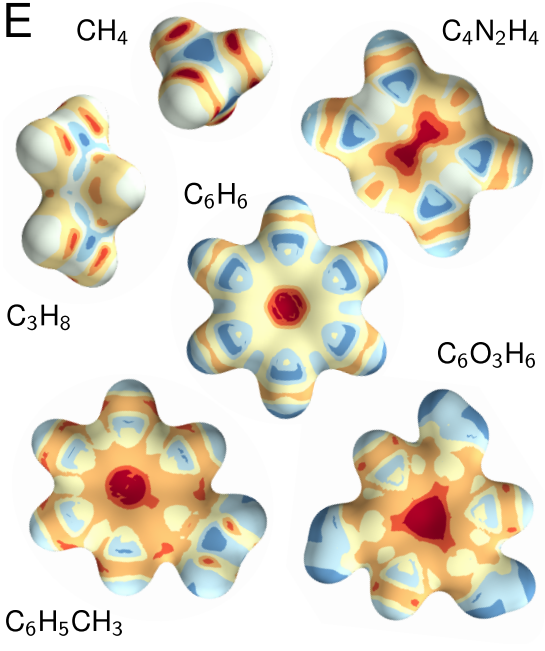
De, Bártok, Csányi, Ceriotti: *Phys. Chem. Chem. Phys.* 18, 13754, 2016

Deep tensor neural networks



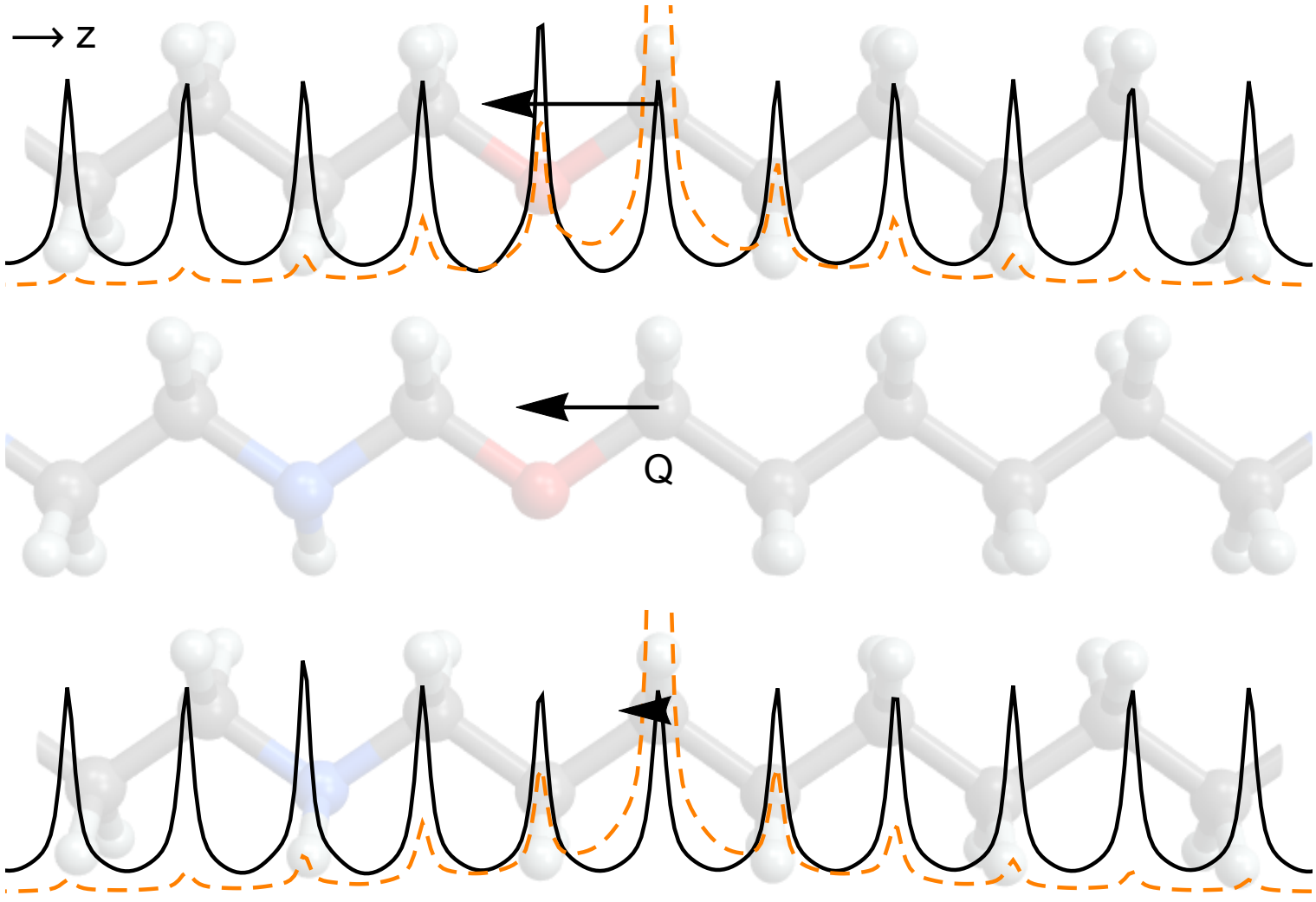






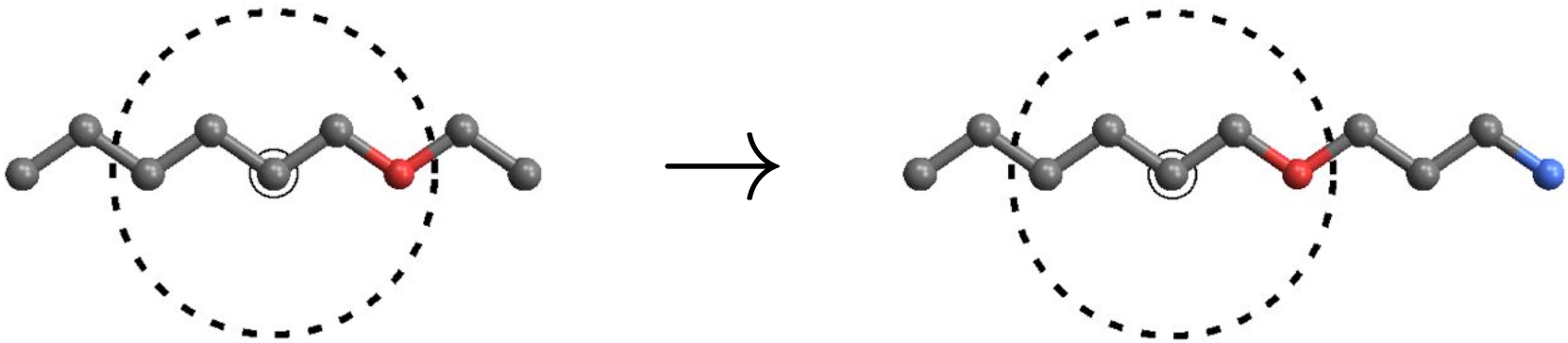
Schütt, Arbabzadah, Chmiela, Müller, Tkatchenko: *Nat. Comm.*, 13890, 2017

Local properties



Local properties

Local interpolation is global extrapolation.

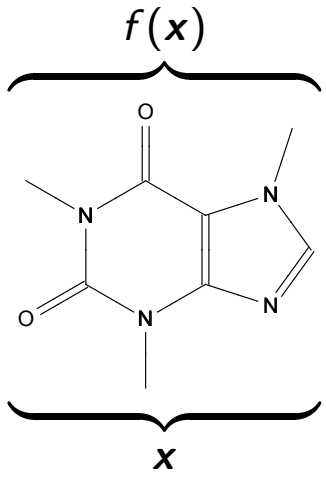


- **linear scaling** of computational effort with system size
- size consistent in the limit
- requires **partitioning** for global properties

Bartók et al, *Phys Rev Lett* 104, 2010
Rupp et al, *J Phys Chem Lett* 6, 2015
Behler, *J Phys Condens Matter* 26, 2014

Local properties

Molecular model

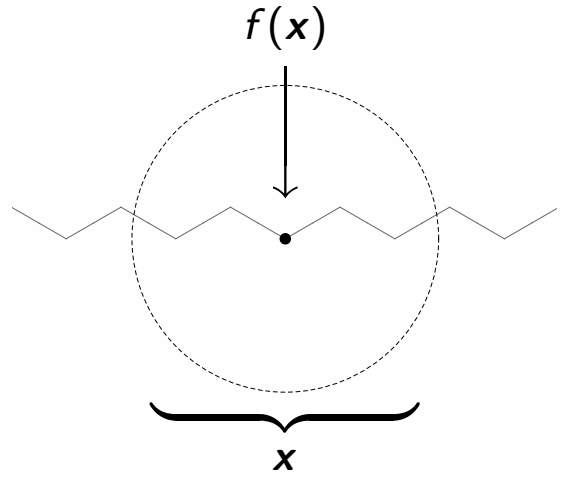


$$f(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i k(\boldsymbol{x}_i, \boldsymbol{x})$$

n = number of molecules

$\boldsymbol{x}$ = representation of molecule

Atomic model

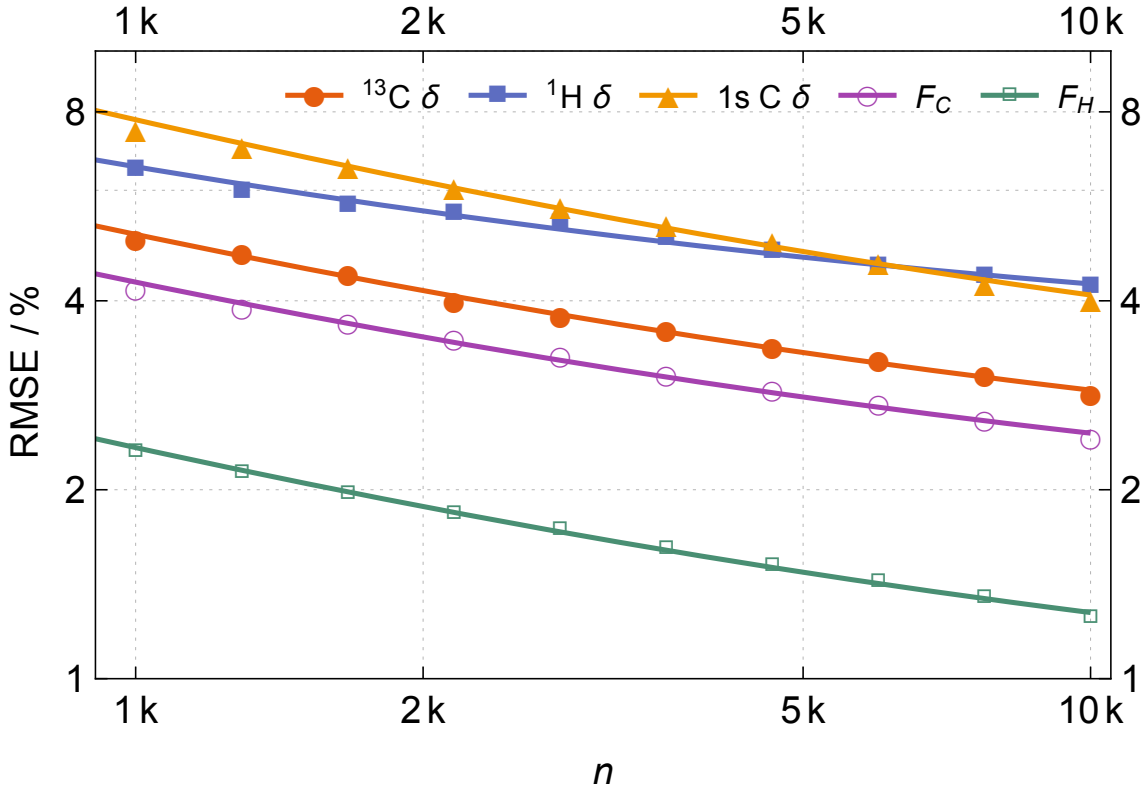


$$f(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i k(\boldsymbol{x}_i, \boldsymbol{x})$$

n = number of atoms

$\boldsymbol{x}$ = representation of atom

Local properties



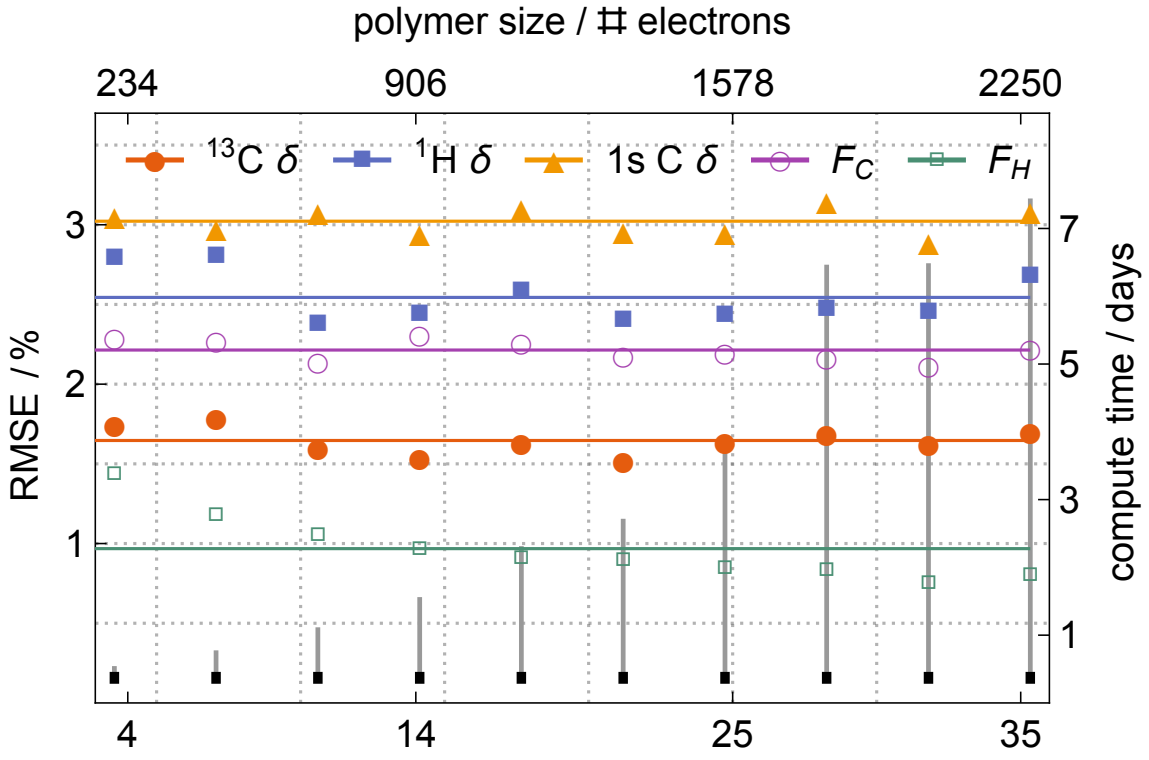
Property	Ref.	RMSE	maxAE	R^2
^{13}C δ / ppm	2.4	5.8 ± 0.3	36 ± 8	0.988 ± 0.001
^{1}H δ / ppm	0.11	0.42 ± 0.02	3.2 ± 1.1	0.954 ± 0.005
1s C δ / mE_h	7.5	6.5 ± 0.3	34 ± 17	0.971 ± 0.002
F_C / mE_h / a_0	1	4.7 ± 0.15	29 ± 5.5	0.983 ± 0.002
F_H / mE_h / a_0	1	1.1 ± 0.03	7.4 ± 2.6	0.996 ± 0.003

Local properties

Dataset

- linear polyethylene $(CH_2CH_2)_n$, doped with N and O
- varying length in multiples of basic unit (29 non-H atoms)
- DFT / PBE0 / def2TZVP using Gaussian 09

Scaling



- training on shortest polymers only
- prediction of polymers of increasing size (up to x10)

Summary

1. validation must follow the golden rule
2. examples of predicting experimental outcomes
3. examples of predicting computational outcomes

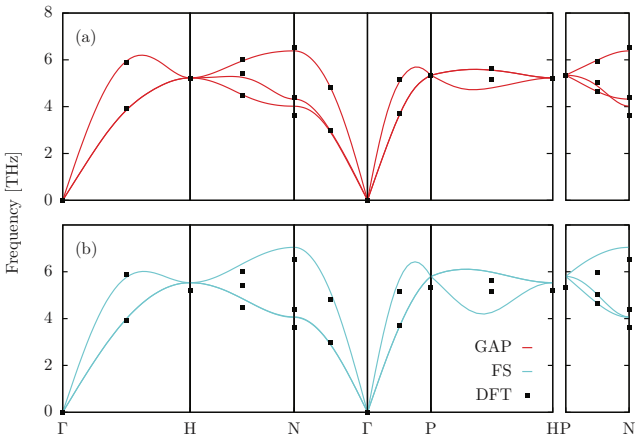
Molecular dynamics — adsorption on surfaces

- since early 1990s $>$ 35 studies on molecules
- many studies using artificial neural networks for potential energy surface interpolation

Year	Ref.	System	Reference method
1995	Blank *et al* [77]	CO @ Ni(1 1 1)	empirical PES
1995	Blank *et al* [77]	H_2 @ Si(1 0 0)-(2 $\times$ 1)	DFT (LDA)
2004	Lorenz *et al* [223]	H_2 @ K(2 $\times$ 2)/Pd(1 0 0)	DFT (PW91)
2005	Behler *et al* [123, 224, 225]	O_2 @ Al(1 1 1)	DFT (RPBE)
2006	Lorenz *et al* [226]	H_2 @ Pd(1 0 0)	empirical PES
2006	Lorenz *et al* [226]	H_2 @ (2 $\times$ 2)S/Pd(1 0 0)	empirical PES
2007	Ludwig and Vlachos [227]	H_2 @ Pt(1 1 1)	empirical PES
2007	Ludwig and Vlachos [227]	H_2 @ Pt(1 1 1)	DFT (PW91)
2008	Behler *et al* [225]	O_2 @ Al(1 1 1)	DFT (PBE)
2008	Latino *et al* [228]	ethanol @ Au(1 1 1)	DFT (B3LYP)
2008	Carbogno *et al* [229]	O_2 @ Al(1 1 1)	DFT (RPBE)
2009	Manzhos *et al* [97]	N_2O @ Cu(1 0 0)	DFT
2009	Carbogno *et al* [230]	O_2 @ Al(1 1 1)	DFT (RPBE)
2010	Latino *et al* [231]	ethanol @ Au(1 1 1)	DFT (B3LYP)
2010	Manzhos and Yamashita [232]	N_2O @ Cu(1 0 0)	DFT
2012	Goikoetxea *et al* [233]	O_2 @ Ag(1 1 1)	DFT (PBE)
2013	Liu *et al* [234]	HCl @ Au(1 1 1)	DFT (PW91)

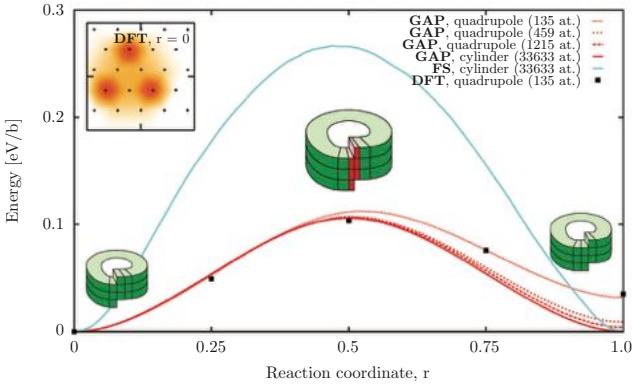
Behler, *J Phys Condens Matter* 26: 183001, 2014.

Molecular dynamics — tungsten

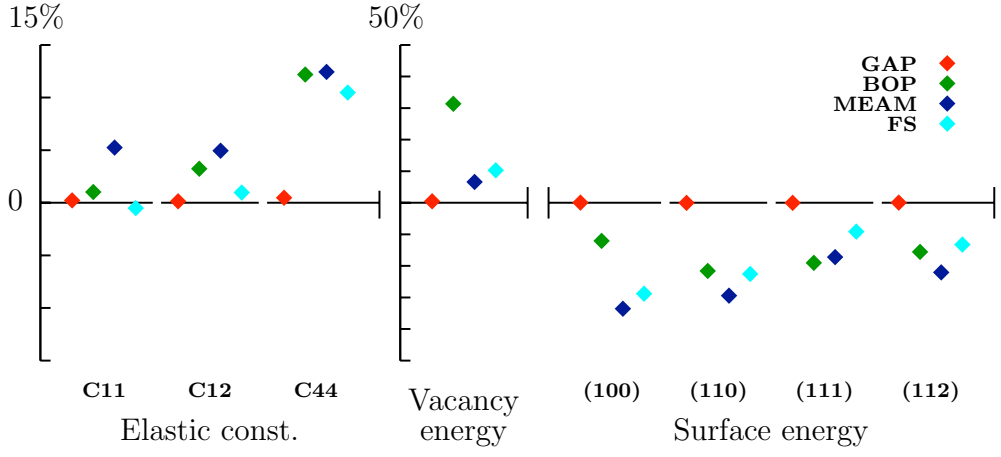


Phonon spectrum

- tungsten in bcc crystal phase
- Gaussian approximation potential
- DFT (PBE, plane waves, pseudopotentials) reference
- screw dislocation



Peierls barrier

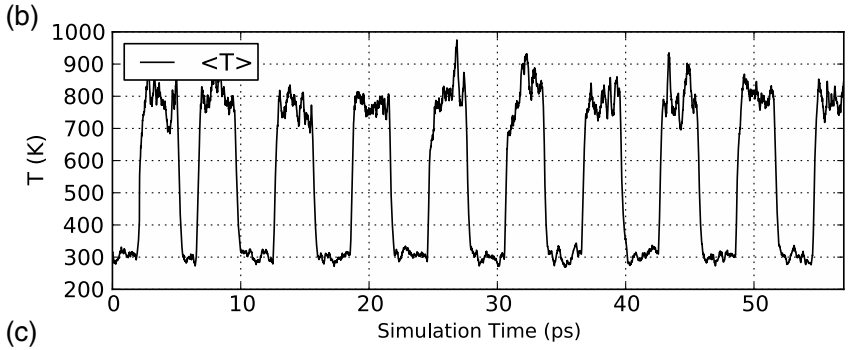


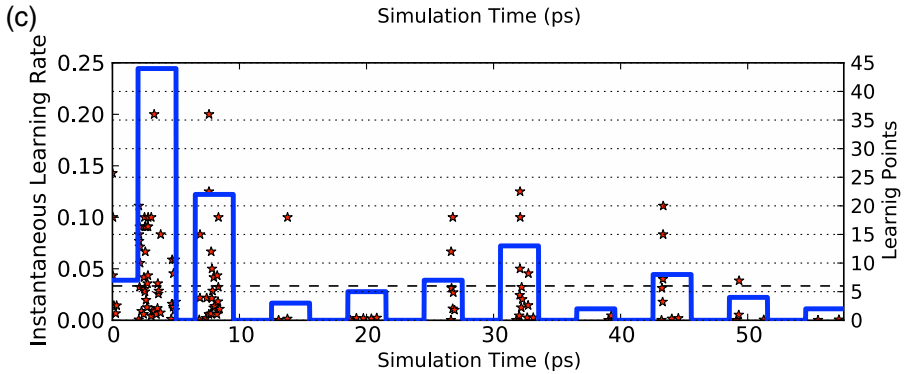
Errors on properties

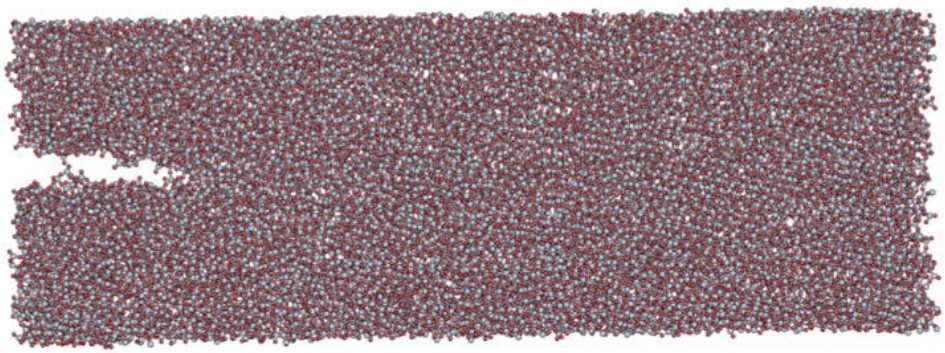
Bartók, et al, *Phys Rev Lett* 104: 136403, 2010. Szlachta et al, *Phys Rev B* 90: 104108, 2014.

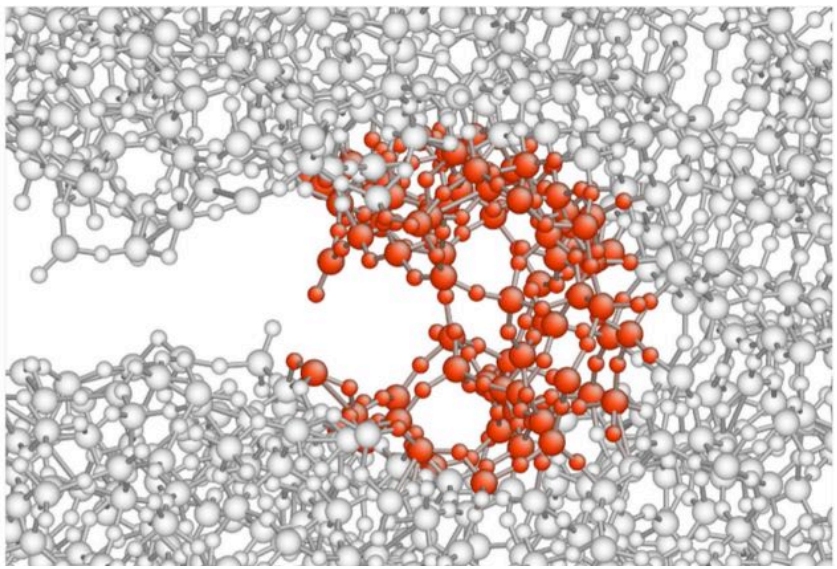
Molecular dynamics — crack propagation

- crack propagation in silicon
- learning on the fly (model is updated when leaving domain)
- form of active learning
- k-step predictor/corrector







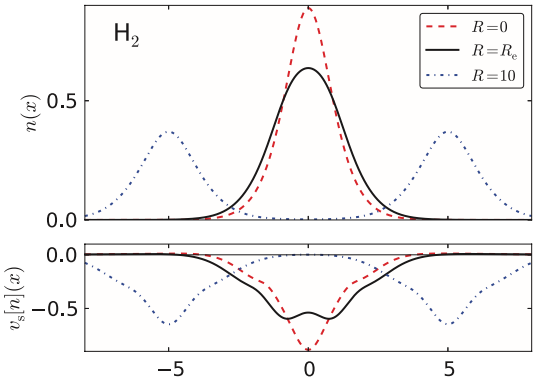


Caccin et al, *Int J Quant Chem* 115: 1129, 2015.
Li et al, *Phys Rev Lett* 114: 096405, 2015.

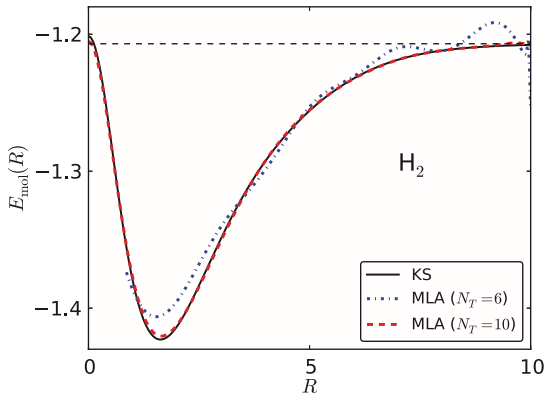
Density functional theory

Learning the map from electron density to kinetic energy

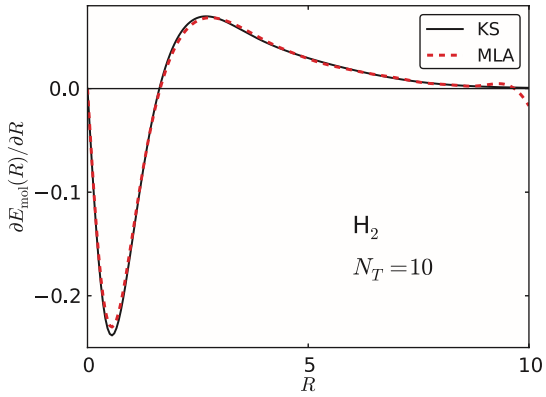
- orbital-free DFT
- 1D toy system
- DFT/LDA as reference
- error decays to zero
- self-consistent densities
- bond breaking and formation



H_2 potential



H_2 binding curve

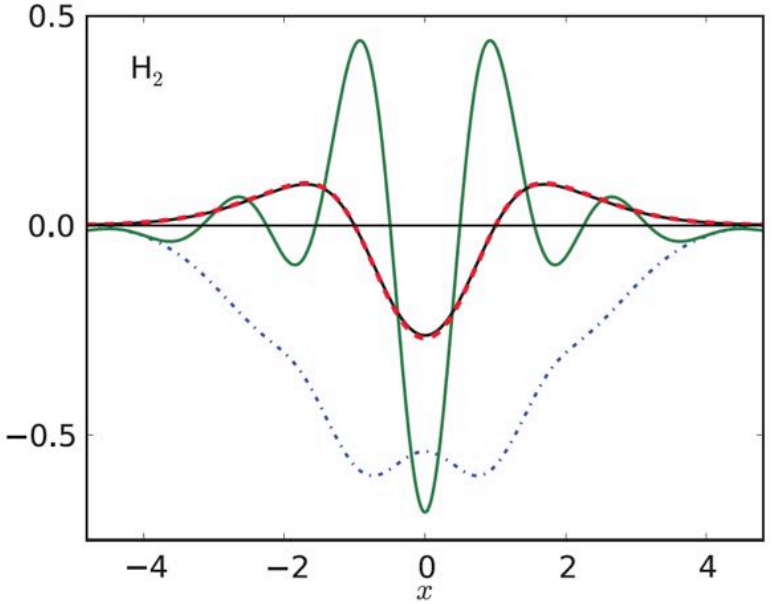


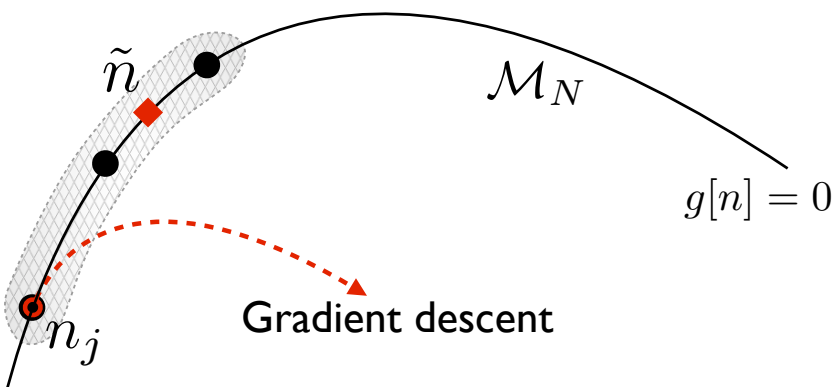
H_2 forces

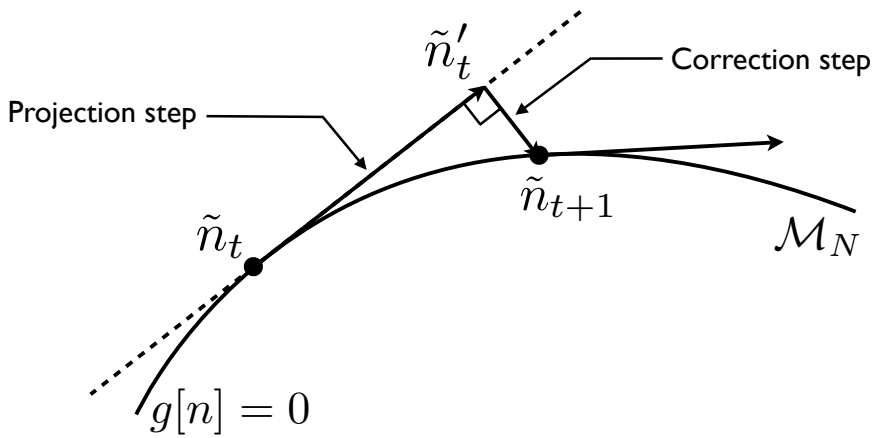
Snyder et al, *Phys. Rev. Lett.*, 2012. Snyder et al, *J. Chem. Phys.*, 2013.

Electron densities — projected gradients

- kinetic energy of electron densities
- Gaussian process regression
- orbital-free DFT, 1D toy system
- error decays to zero
- projected gradients for self-consistent densities ("non-linear gradient denoising")



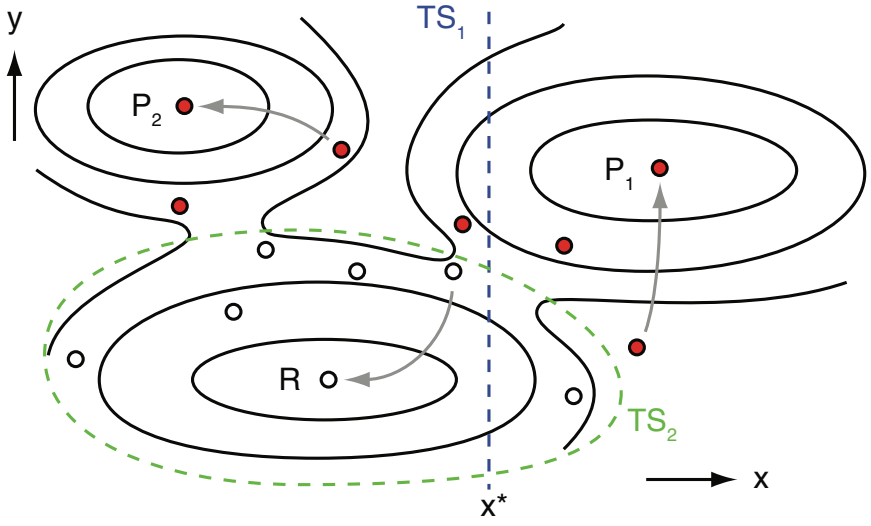


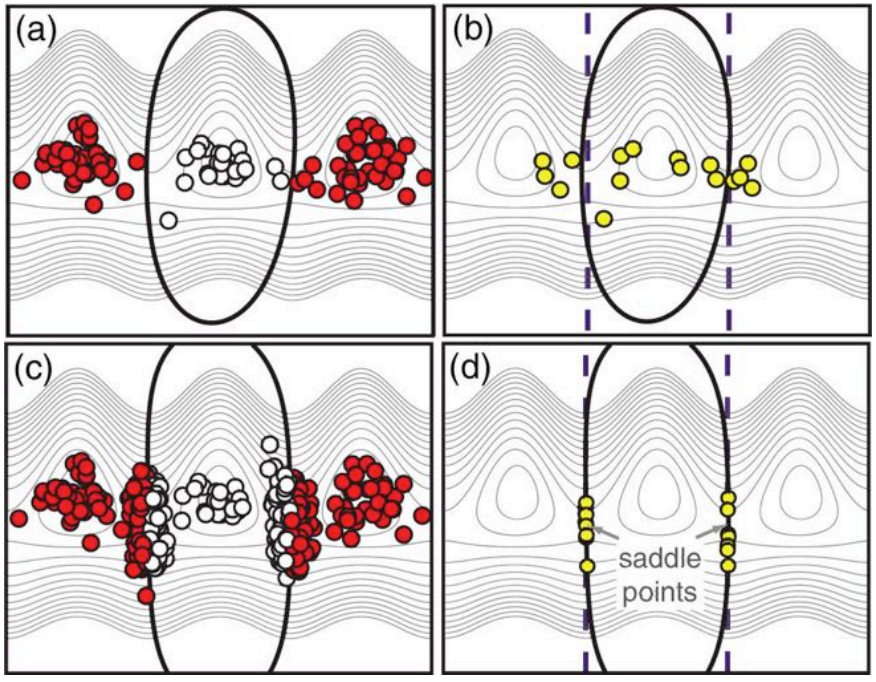


Snyder et al, *Phys Rev Lett* 108: 253002, 2012. Snyder et al, *J Chem Phys* 139: 224104, 2013. Li et al, *Int J Quant Chem*, in press, 2015.

Transition state theory

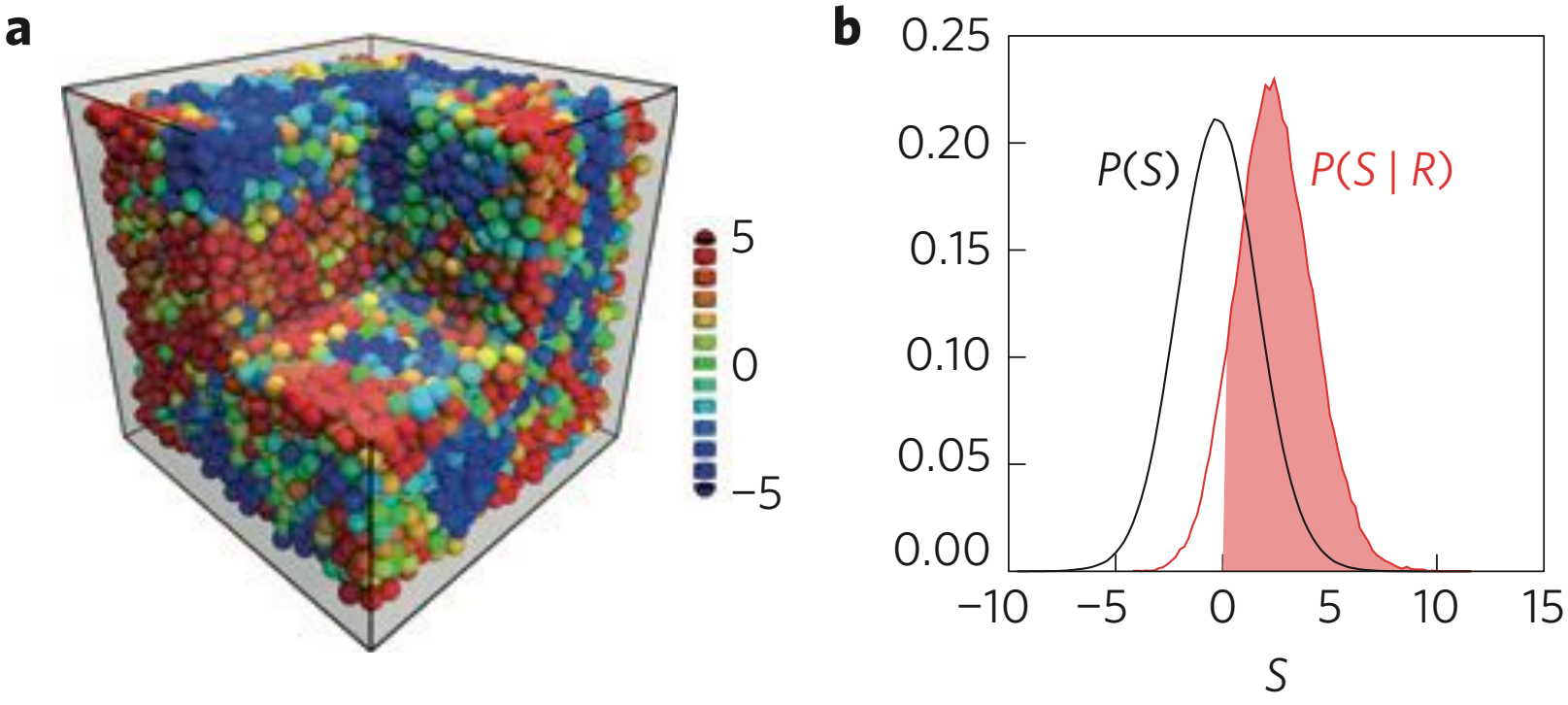
- characterization of dividing surfaces
- support vector machine for classification
- alternate between learning and sampling
- no prior information required
- iteratively refined by biased sampling along dividing surface





Pozun et al, *J. Chem. Phys.*, 174101, 2012.

Relaxation in glassy liquids

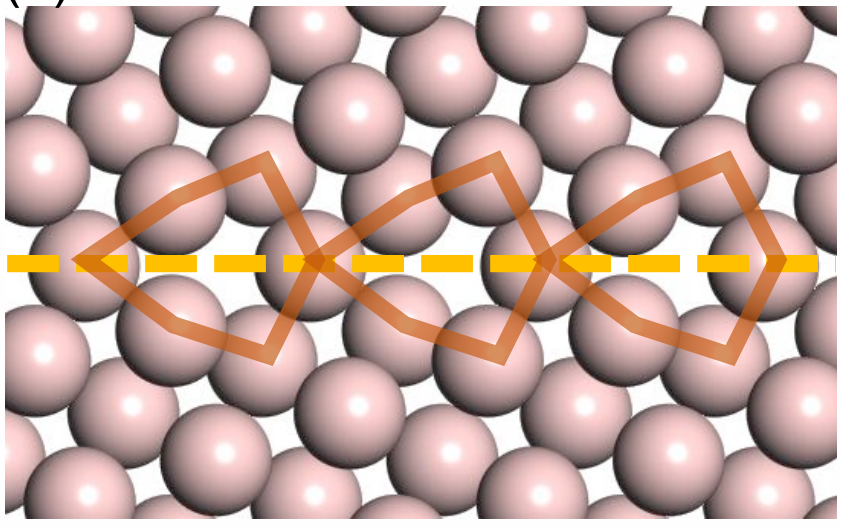


- identify subtle structural changes ("softness") in glassy dynamics
- softness correlates to probability of rearrangement in near future

Schoenholz, Cubuk et al, *Nat. Phys.*, 2016

Stable interface search

(a)

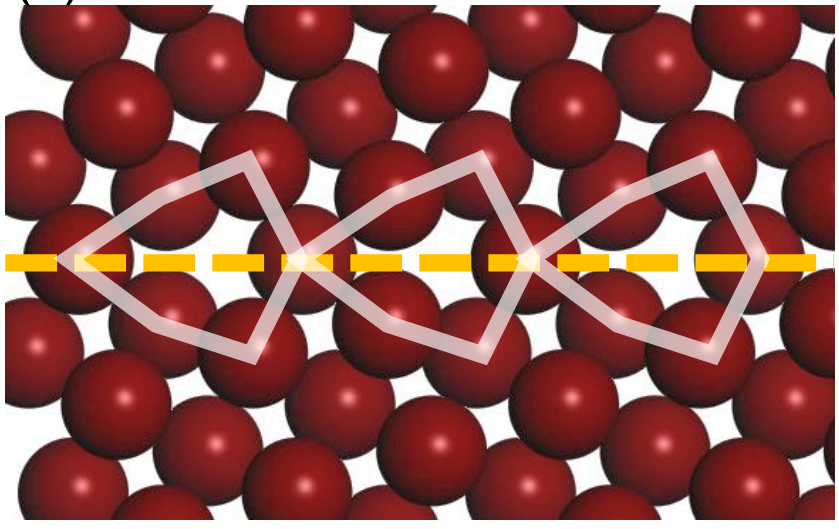


Exhaustive calculations

GB energy=0.96J/m^2

Number of energy calculations
＝16,983

(b)



Bayesian optimization

GB energy=0.96J/m^2

Number of energy calculations
＝69

Kiyohara, Oda, Tsuda, Mizoguchi, *Jpn. J. Appl. Phys.*, 2016