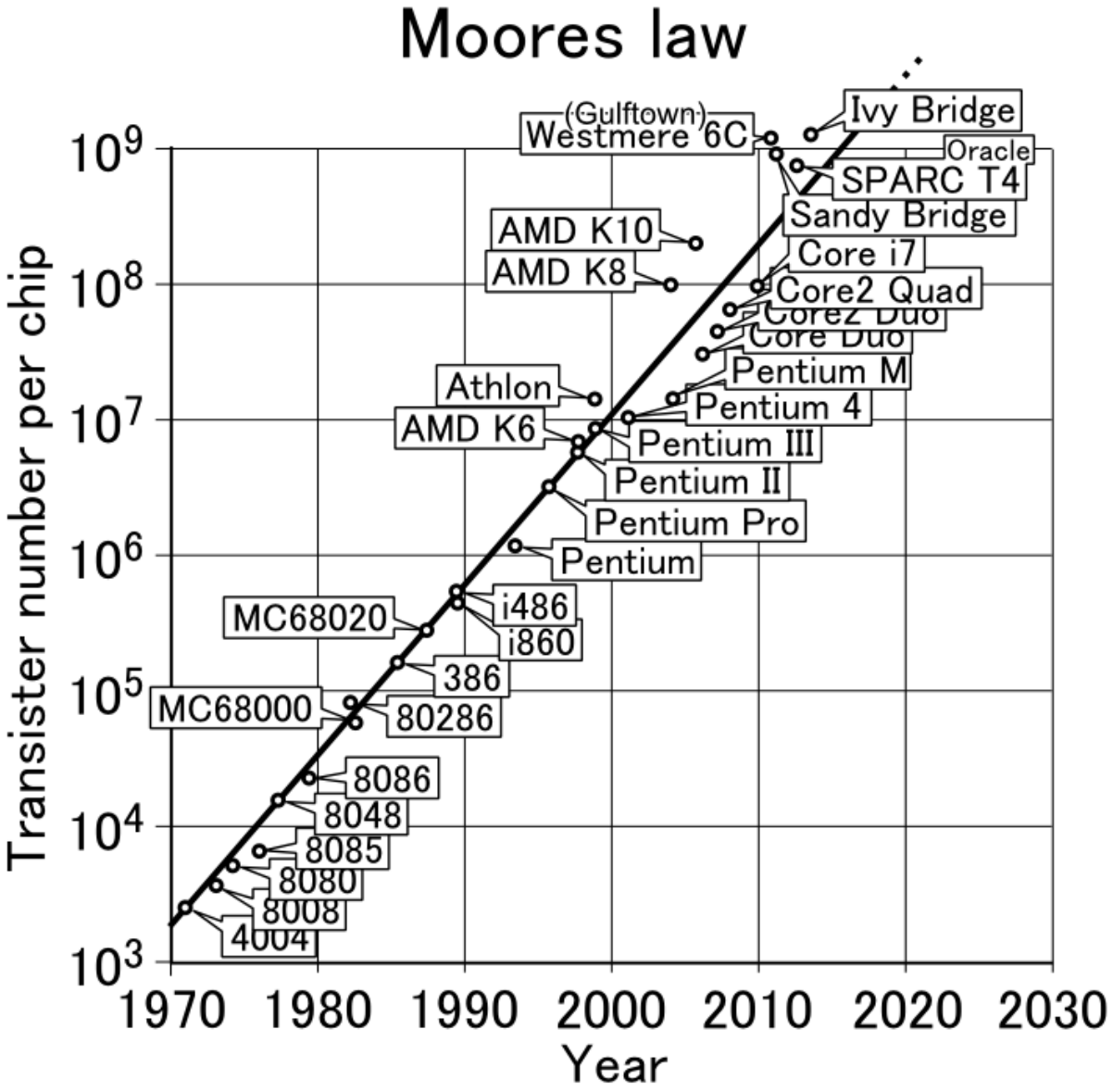
Machine Learning in Chemical Space

Anatole von Lilienfeld[1]
Institute of Physical Chemistry and National Center for Computational Design and Discovery of Novel Materials (MARVEL), Department of Chemistry, University Basel

Many of the most relevant chemical properties of matter depend explicitly on atomistic details, rendering a first principles approach mandatory. Alas, even when using high-performance computers, brute force high-throughput screening of compounds with electronic structure theory is beyond any capacity for all but the simplest systems and properties due to the combinatorial nature of chemical space, i.e. all the compositional, constitutional, and conformational isomers. Consequently, efficient exploration algorithms should exploit all implicit redundancies present in high-throughput approaches. In this talk, I will describe recently developed statistical approaches for interpolating (Kriging) quantum mechanical observables in composition space. Examples will be presented for predicting properties of out-of-sample molecules or solids with high accuracy and small computational cost.

There’s an on-going revolution …



VISUALIZING PROGRESS

If transistors were people

If the transistors in a microprocessor were represented by people, the following timeline gives an idea of the pace of Moore's Law.

2,300
Average music hall capacity

134,000
Large stadium capacity

32 Million
Population of Tokyo

1.3 Billion
Population of China

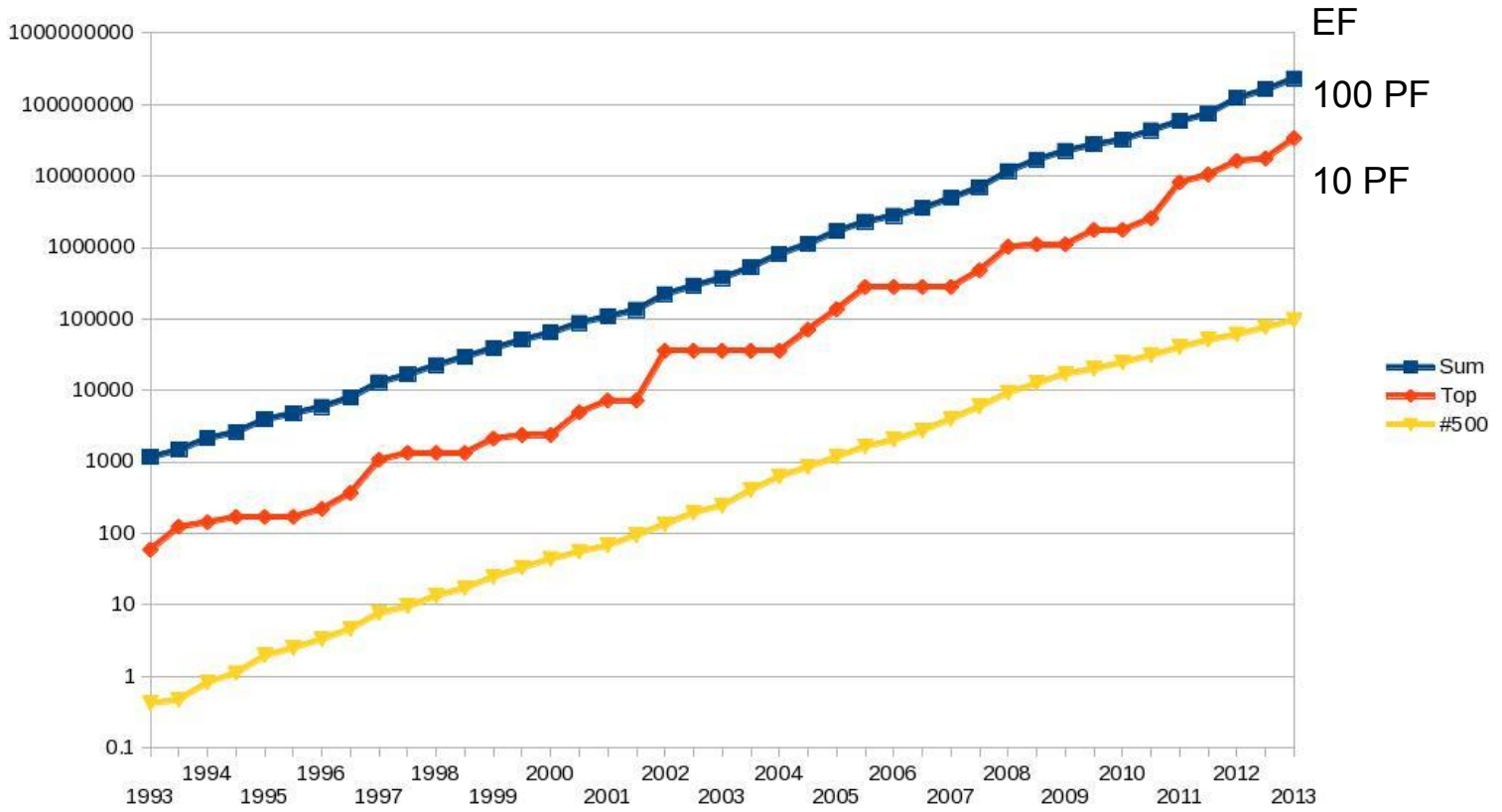
Now imagine that those 1.3 billion people could fit onstage in the original music hall. That's the scale of Moore's Law.

www.top500.org

RANK	SITE	SYSTEM	CORES	RMAX (TFLOP/S)	RPEAK (TFLOP/S)	POWER (KW)
1	National Super Computer Center in Guangzhou China	**Tianhe-2 (MilkyWay-2)** - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P NUDT	3,120,000	33,862.7	54,902.4	17,808
2	DOE/SC/Oak Ridge National Laboratory United States	**Titan** - Cray XK7 , Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x Cray Inc.	560,640	17,590.0	27,112.5	8,209
3	DOE/NNSA/LLNL United States	**Sequoia** - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom IBM	1,572,864	17,173.2	20,132.7	7,890
4	RIKEN Advanced Institute for Computational Science (AICS) Japan	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect Fujitsu	705,024	10,510.0	11,280.4	12,660
5	DOE/SC/Argonne National Laboratory United States	**Mira** - BlueGene/Q, Power BQC 16C 1.60GHz, Custom IBM	786,432	8,586.6	10,066.3	3,945
6	Swiss National Supercomputing Centre (CSCS) Switzerland	**Piz Daint** - Cray XC30, Xeon E5-2670 8C 2.600GHz, Aries interconnect , NVIDIA K20x Cray Inc.	115,984	6,271.0	7,788.9	2,325
7	King Abdullah University of Science and Technology Saudi Arabia	**Shaheen II** - Cray XC40, Xeon E5-2698v3 16C 2.3GHz, Aries interconnect Cray Inc.	196,608	5,537.		
8	Texas Advanced Computing Center/Univ. of Texas United States	**Stampede** - PowerEdge C8220, Xeon E5-2680 8C 2.700GHz, Infiniband FDR, Intel Xeon Phi SE10P Dell	462,462	5,168.		
9	Forschungszentrum Juelich (FZJ) Germany	**JUQUEEN** - BlueGene/Q, Power BQC 16C 1.600GHz, Custom Interconnect IBM	458,752	5,008.		
10	DOE/NNSA/LLNL United States	**Vulcan** - BlueGene/Q, Power BQC 16C 1.600GHz, Custom Interconnect IBM	393,216	4,293.		

Computing power 1993-2013

If brick-and-mortar labs were to follow ... a 1 yr experiment in 1986 → 1 s in 2015 ($3x10^7$ speed-up)

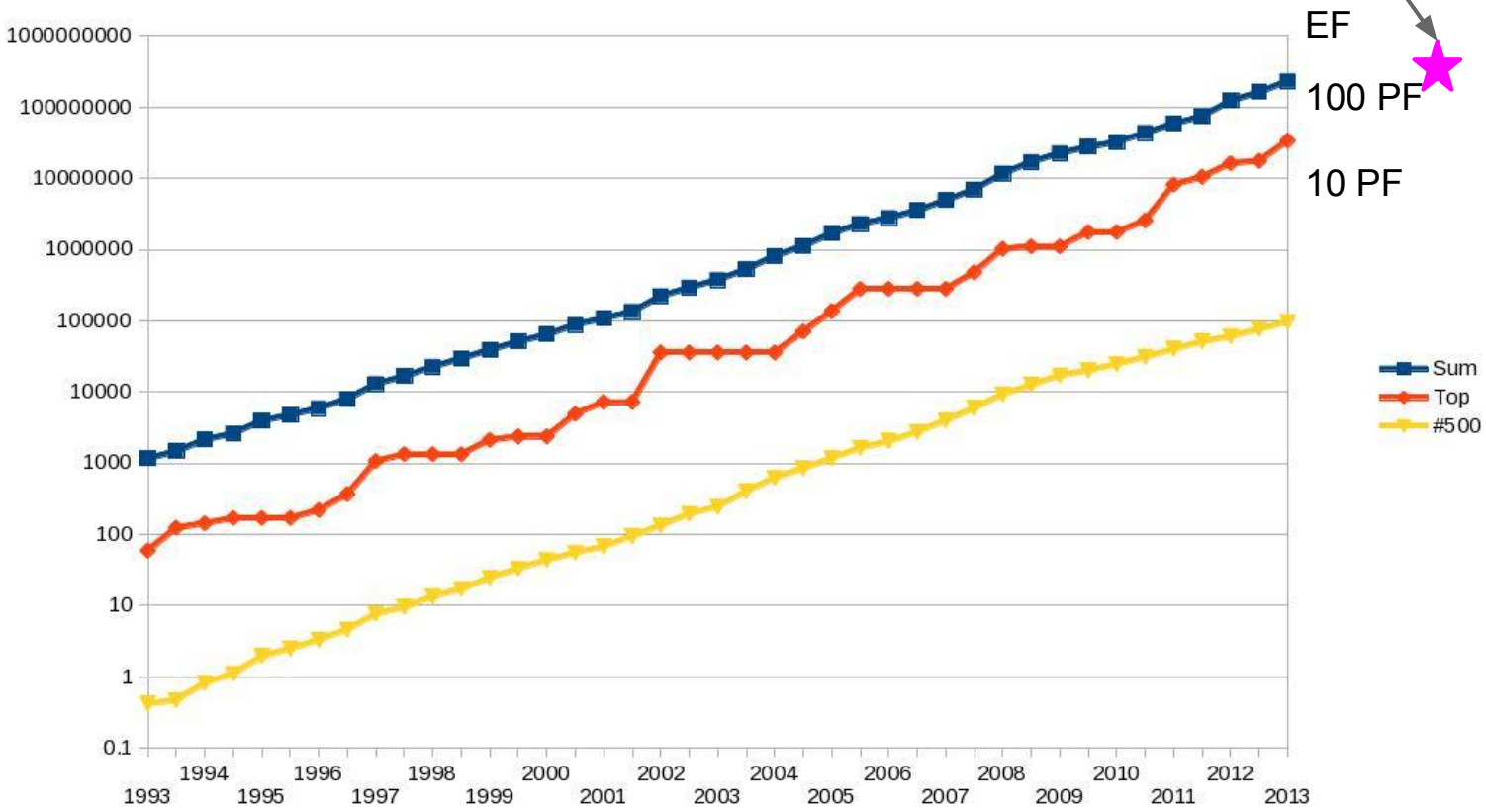


Apr 9 2015

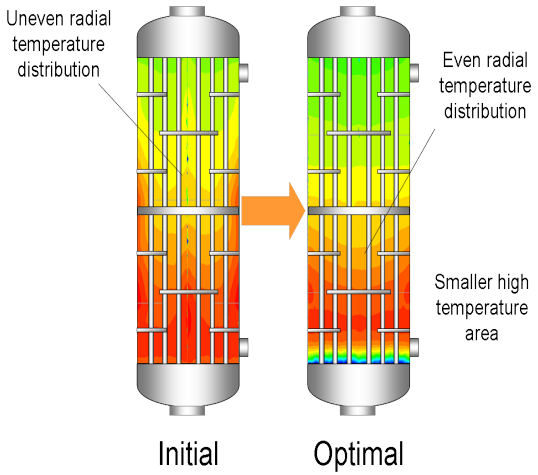
The U.S. Department of Energy announced a $200 million investment to deliver a next-generation supercomputer, known as Aurora, to the Argonne Leadership Computing Facility.

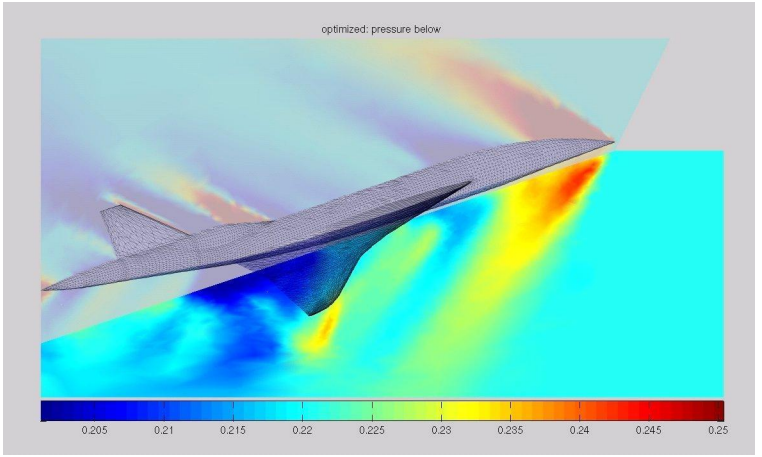
will use Intel's HPC scalable system framework to provide a peak performance of 180 PetaFLOP/s.

Early Science: 2016-2018
http://aurora.alcf.anl.gov/



Computational design ...





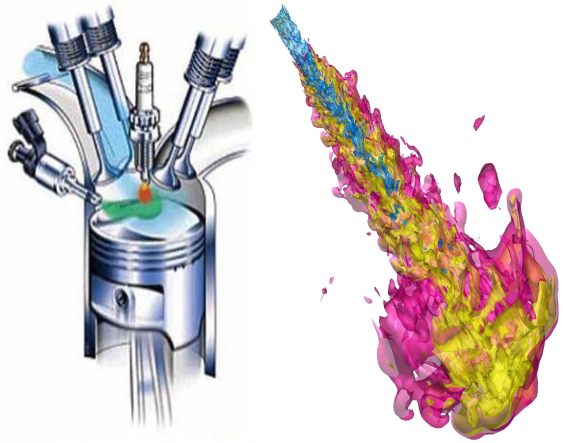


Figure 2: EcoBoost—with direct injection, fuel is injected into each cylinder of an engine in small, precise amounts.

Where is the design of new chemicals?

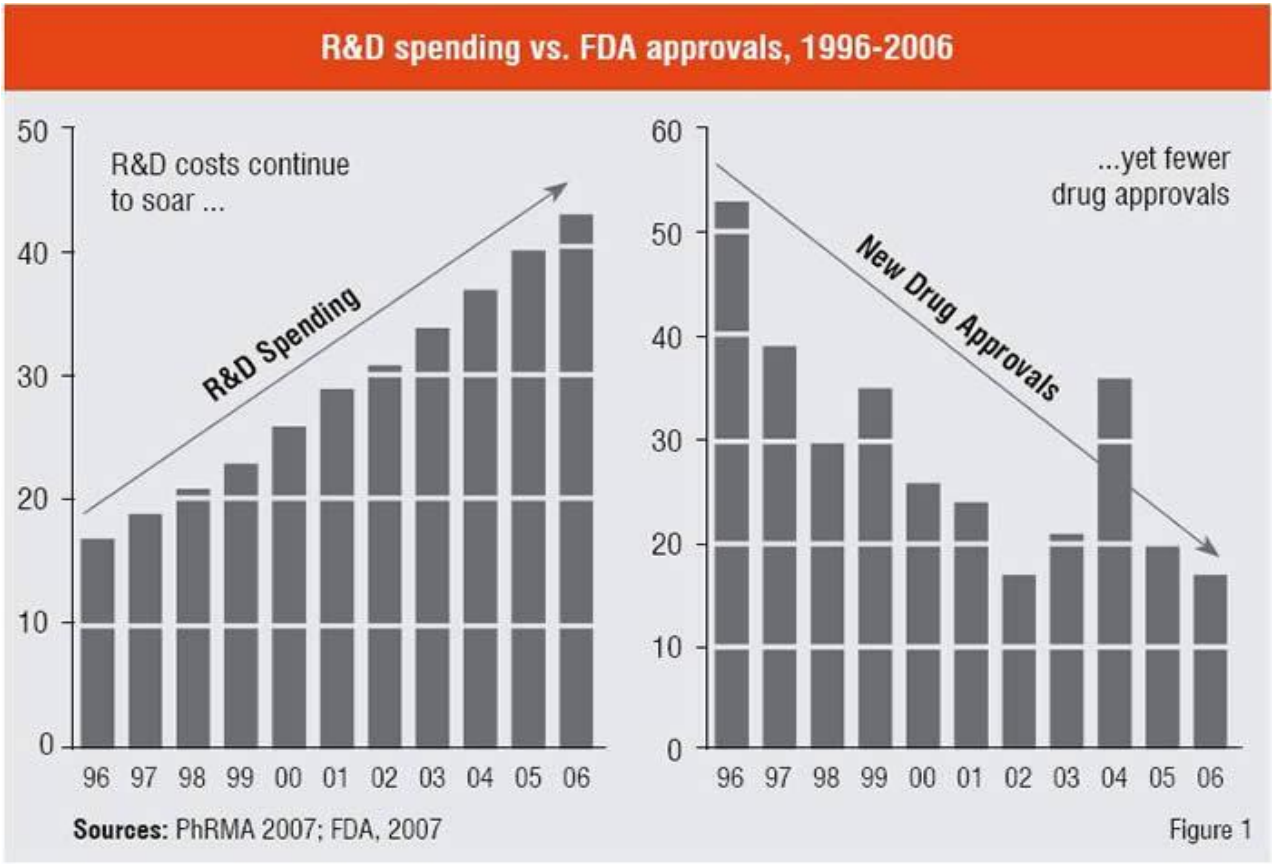
"Yeah, I see him too...But nobody wants to talk about it!"

Where is the design of new chemicals?

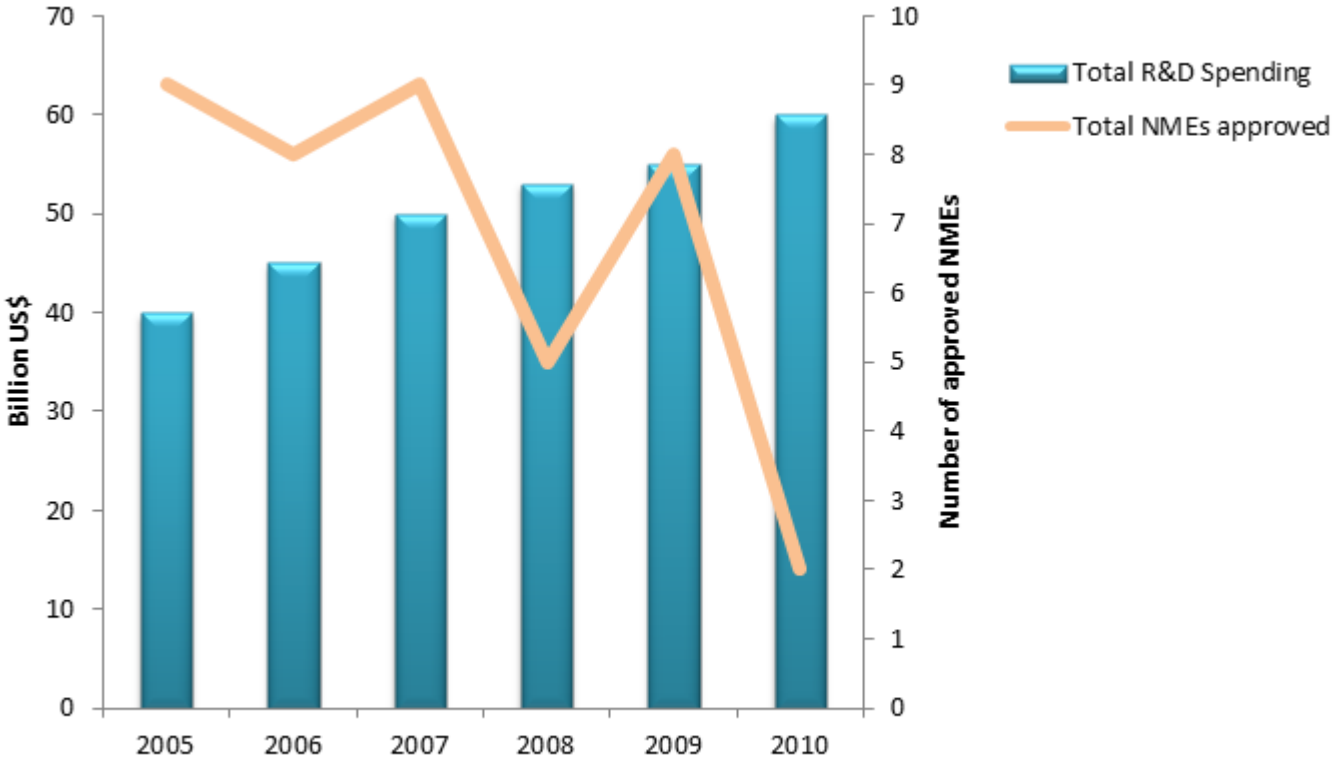
"Many societal challenges are chemical challenges"

G. Whitesides

- health (drugs)
- infrastructure (rust)
- water (desalinate)
- energy (renewable)
- light (OLEDs)



– Infections
– Metabolic syndrome
– Aging
– Cancer
...



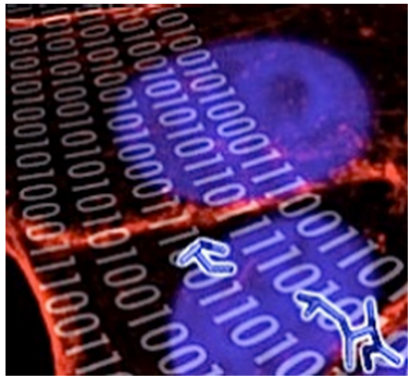
– Infections
– Metabolic syndrome
– Aging
– Cancer
...

www.cadd.ethz.ch/laboratory.html

ETH Zurich → D-CHAB → IPW → Computer-Assisted Drug Design →

How to find us

The Laboratory



The Molecular Design Laboratory (MODLAB) develops and implements new concepts, algorithms and software for rapid identification of bioactive tool compounds and pharmaceutical lead structures.

Head of group:

Prof. Gisbert Schneider →

The molecular design cycle involves multiple scientific disciplines and requires rigorous trans-disciplinary thinking. We employ a broad repertoire of machine-learning methods and bio/cheminformatics techniques for automated hypothesis generation, activity prediction and validation.

About MODLAB

Lab Presentation: **The Computer-Assisted Drug Design Group at ETH Zurich**. *MedChemWatch* (2011) 12:55-57.

Schneider, G. (2012) **From theory to bench experiment by computer-assisted drug design**. *Chimia* 66:120-124.

OPINION

Virtual screening: an endless staircase?

Gisbert Schneider

Abstract | Computational chemistry — in particular, virtual screening — can provide valuable contributions in hit- and lead-compound discovery. Numerous software tools have been developed for this purpose. However, despite the applicability of virtual screening technology being well established, it seems that there are relatively few examples of drug discovery projects in which virtual screening has been the key contributor. Has virtual screening reached its peak? If not, what aspects are limiting its potential at present, and how can significant progress be made in the future?

Nature Reviews (2010)

Virtual screening: an endless staircase?

Gisbert Schneider

Abstract | Computational chemistry — in particular, virtual screening — can provide valuable contributions in hit- and lead-compound discovery. Numerous software tools have been developed for this purpose. However, despite the applicability of virtual screening technology being well established, it seems that there are relatively few examples of drug discovery projects in which virtual screening has been the key contributor. Has virtual screening reached its peak? If not, what aspects are limiting its potential at present, and how can significant progress be made in the future?

Dynamic descriptions of molecules will have to replace our predominantly static view of both targets and ligands[19]. Molecular dynamics simulations can sample conformational ensembles of targets and ligands. However, some of the popular force-field approaches used to describe the energetics of molecular systems might be inadequate for drug design. Furthermore,

Nature Reviews (2010)

OPINION

Virtual screening: an endless staircase?

Gisbert Schneider

Abstract | Computational chemistry — in particular, virtual screening — can provide valuable contributions in hit- and lead-compound discovery. Numerous software tools have been developed for this purpose. However, despite the applicability of virtual screening technology being well established, it seems that there are relatively few examples of drug discovery projects in which virtual screening has been the key contributor. Has virtual screening reached its peak? If not, what aspects are limiting its potential at present, and how can significant progress be made in the future?

Dynamic descriptions of molecules will have to replace our predominantly static view of both targets and ligands[19]. Molecular dynamics simulations can sample conformational ensembles of targets and ligands. However, some of the popular force-field approaches used to describe the energetics of molecular systems might be inadequate for drug design. Furthermore,

Nature Reviews (2010)

If fifty million people say a foolish thing, it is still a foolish thing.

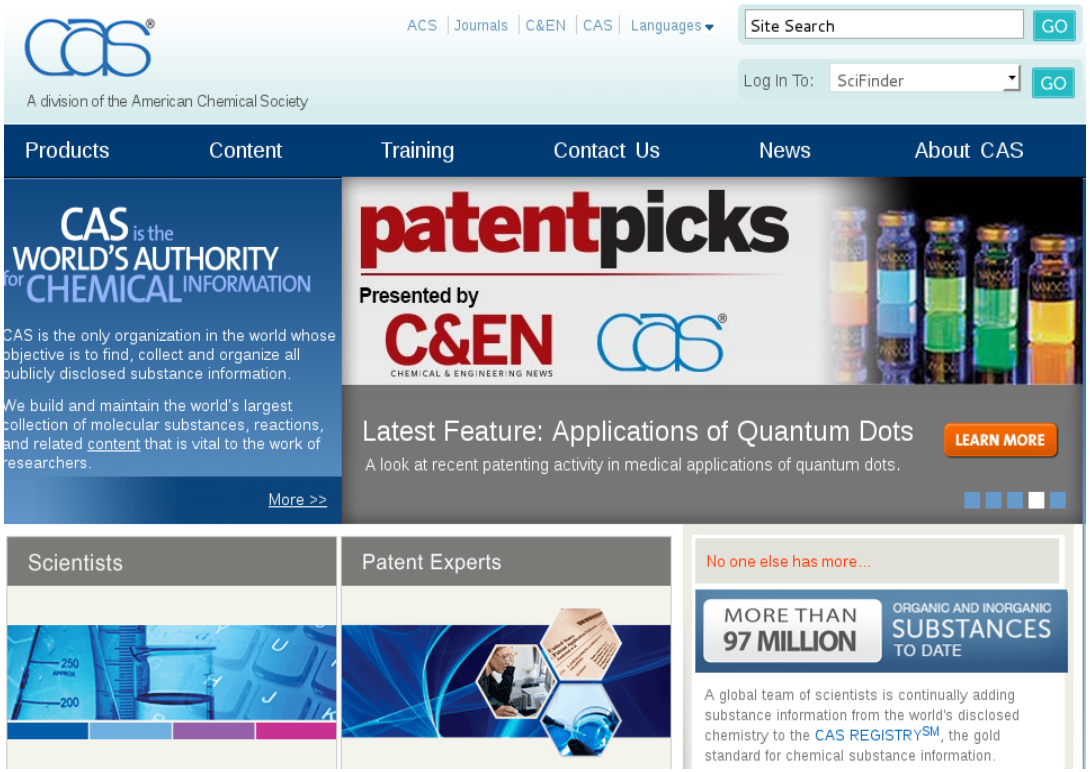
Anatole France

Where is the design of new chemicals?

"Many societal challenges are chemical challenges"

G. Whitesides

- health (drugs)
- infrastructure (rust)
- water (desalinate)
- energy (renewable)
- light (OLEDs)



Question: How many?

American Chemical Society maintains CAS w
~97M substances (alloys, minerals, mixtures, polymers and salts)
~60M sequences (DNA, RNA, proteins)
~10k compounds being added on daily basis

Question: How many?

American Chemical Society maintains CAS w
~97M substances (alloys, minerals, mixtures, polymers and salts)
~60M sequences (DNA, RNA, proteins)
~10k compounds being added on daily basis

But:
Number of (*small organic*) molecules > 10^{60}
[*Nature Insight* on chemical space (2004)]

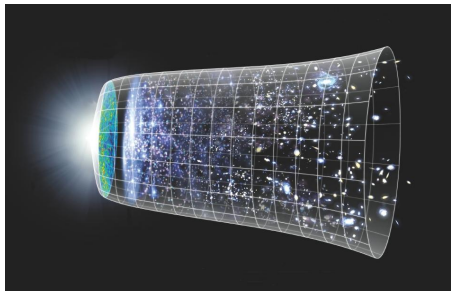
<u>Question:</u> How many?

American Chemical Society maintains CAS w
~97M substances (alloys, minerals, mixtures, polymers and salts)
~60M sequences (DNA, RNA, proteins)
~10k compounds being added on daily basis

<u>But:</u>
Number of (*small organic*) molecules > 10^{60}
[*Nature Insight* on chemical space (2004)]

k years
weeks
years
M years
10 B years

1000000000000000000000000000000000000000000000000000000000



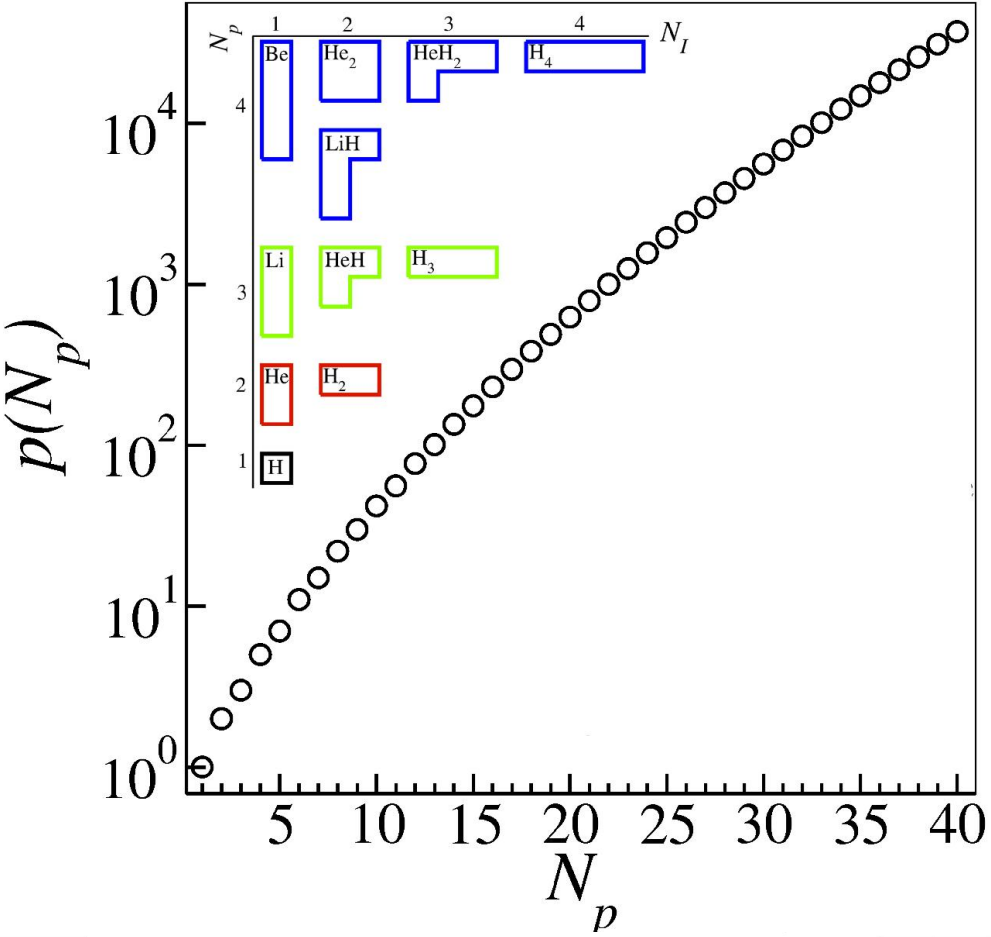
Where is the design of new chemicals?

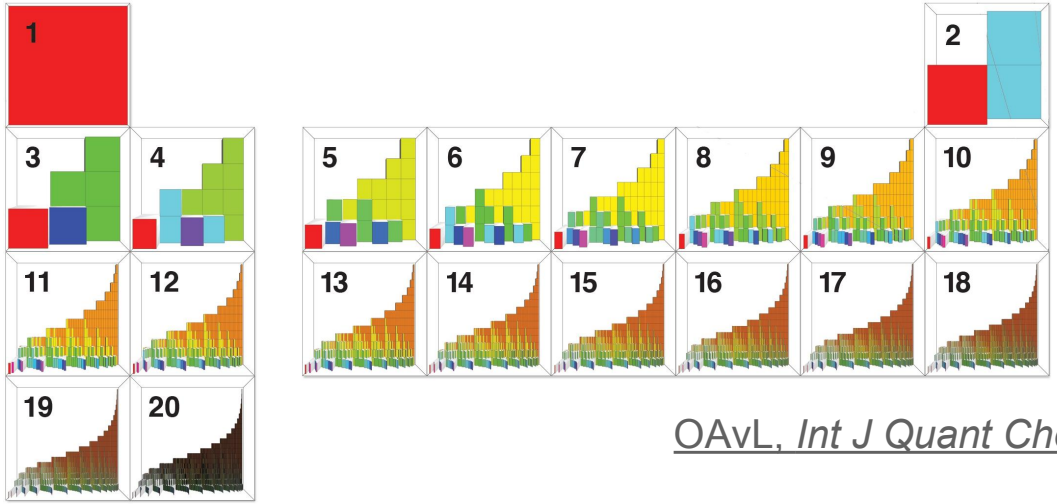
"Many societal challenges are chemical challenges"

G. Whitesides

- health (drugs)
- infrastructure (rust)
- water (desalinate)
- energy (renewable)
- light (OLEDs)

Composition

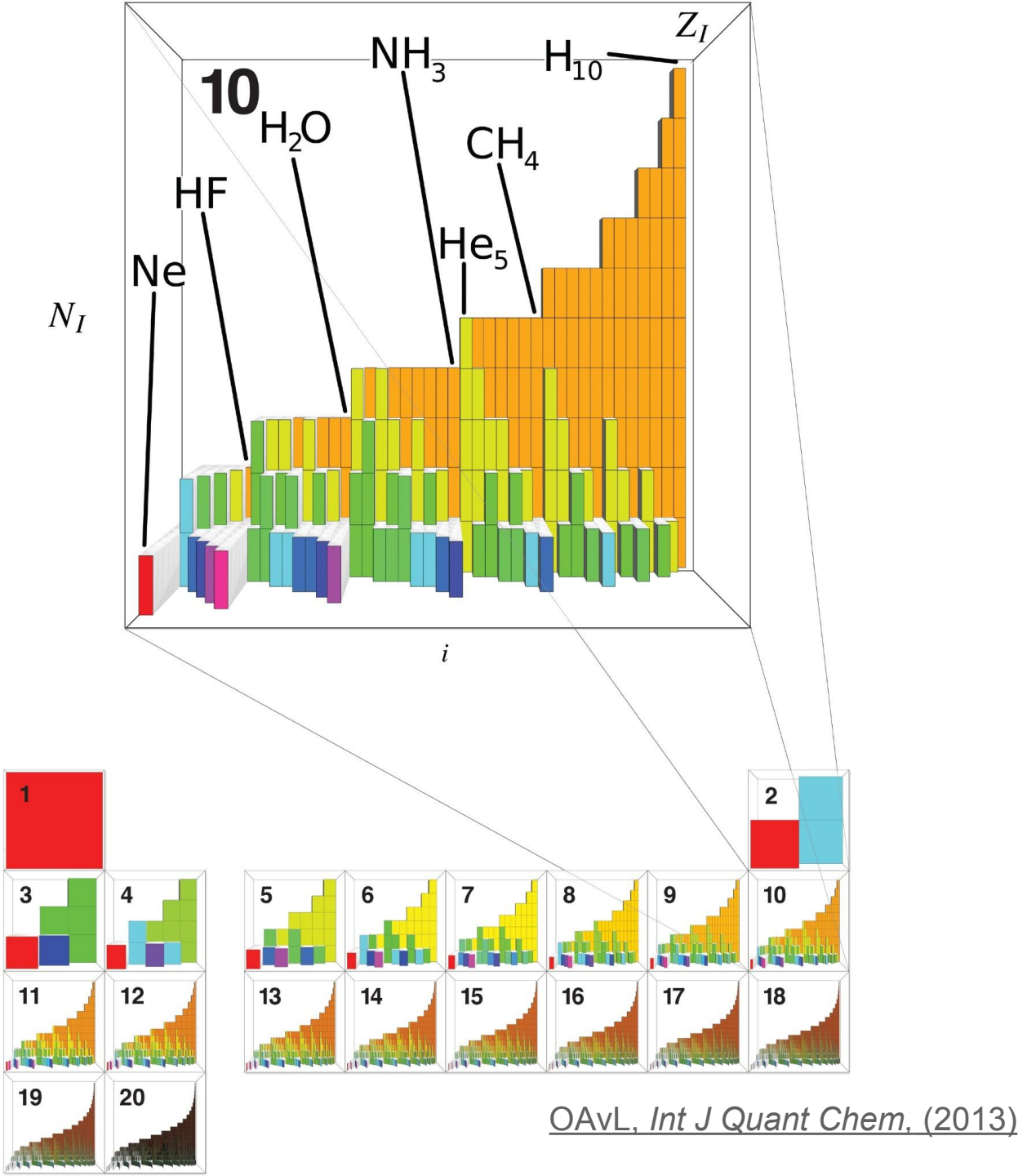




OAvL, *Int J Quant Chem*, (2013)

Composition

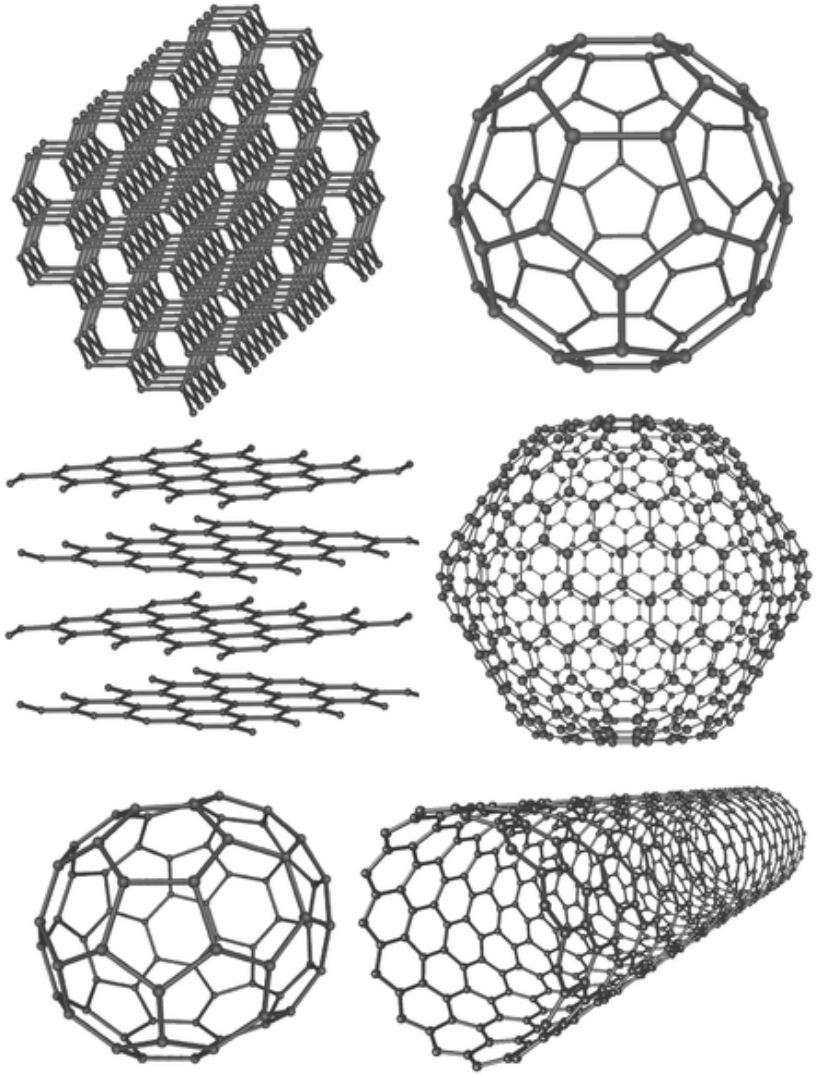
10 protons

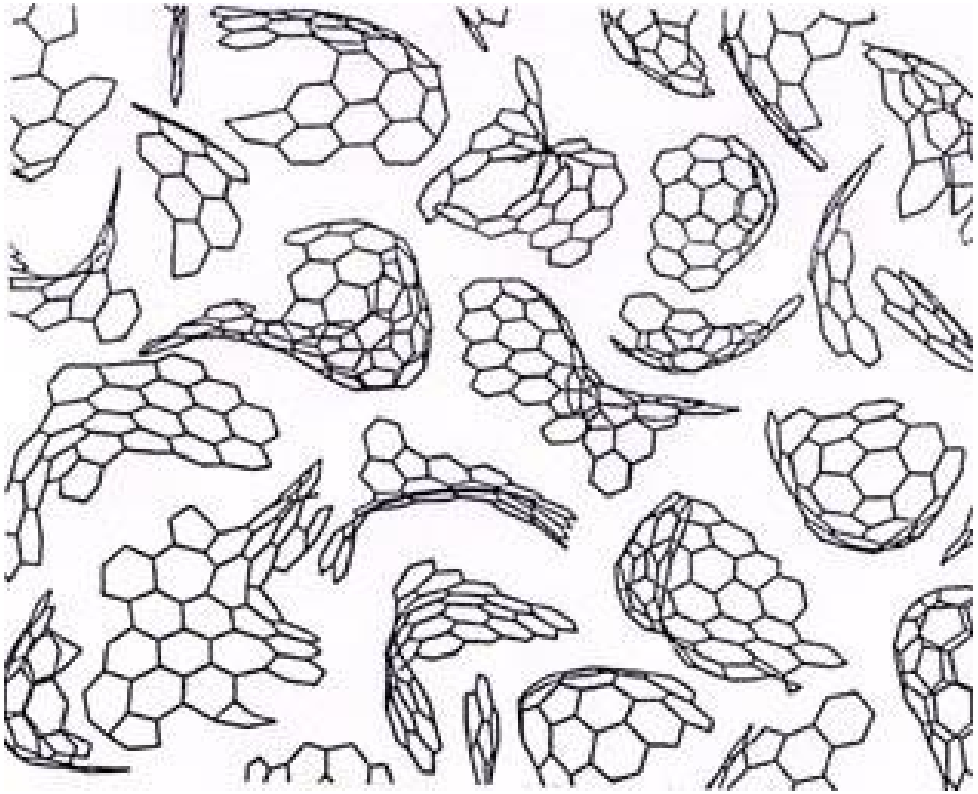


OAvL, *Int J Quant Chem*, (2013)

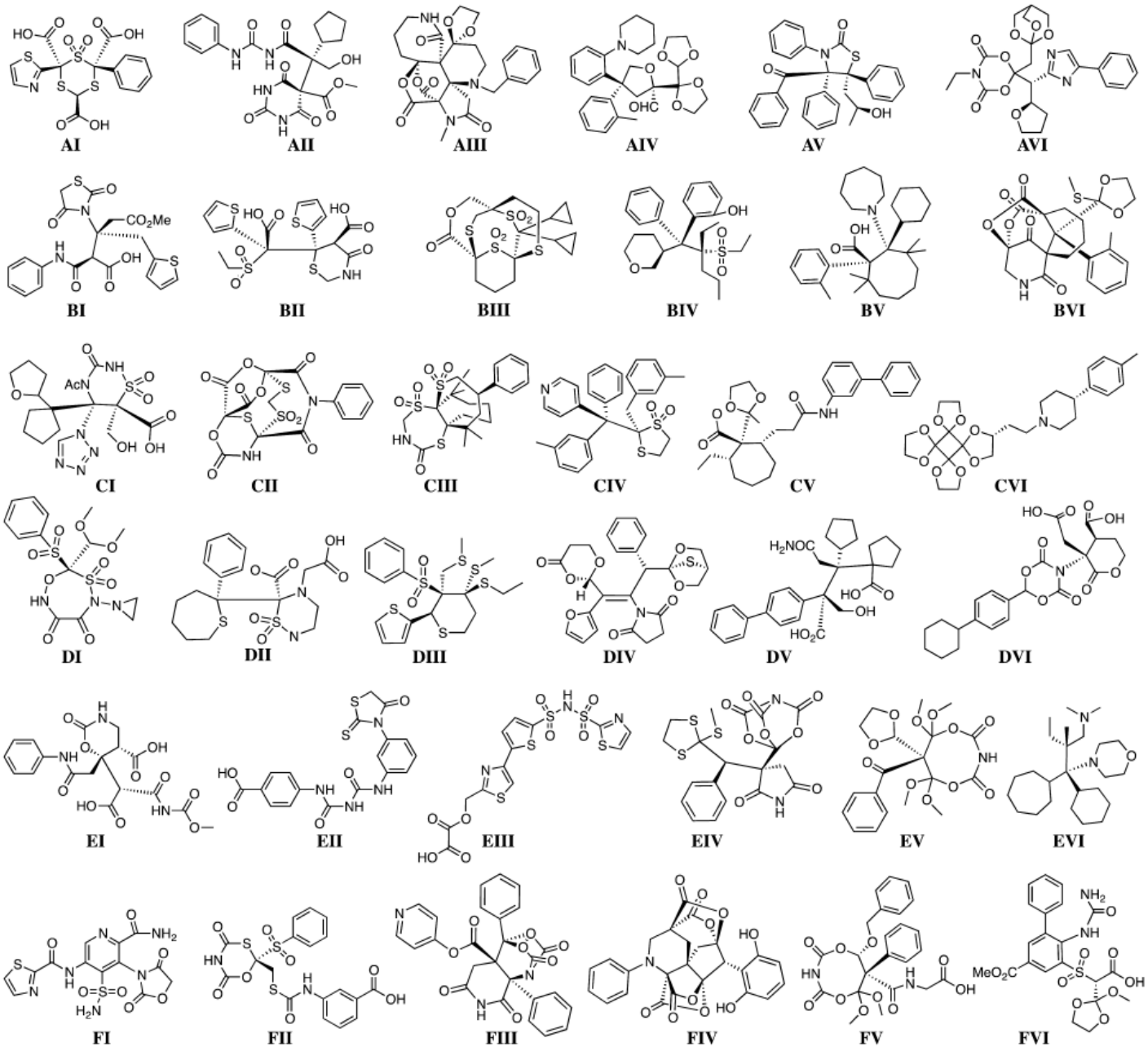
Spatial configuration

Carbon allotropes



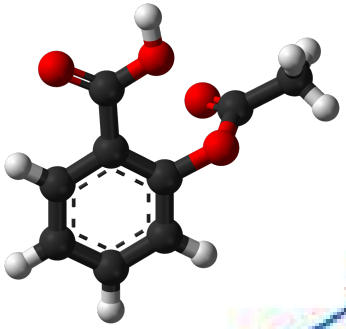


Composition + Configuration



Virshup, Yang, Beratan et al *J Am Chem Soc* (2013)

GDB-*N*: All organic molecules with up to *N* heavy atoms



Fink, Bruggesser, Reymond *ACIE* (2005); Blum, Reymond *JACS* (2009), Ruddigkeit, Reymond, *Chem Inf Model* (2012)

First Principles

$$H(\{Z_I, \mathbf{R}_I\})\Psi(\mathbf{r}) \quad = \quad E\Psi(\mathbf{r})$$

Why first principles?

1. General: Any property
2. Transferable: Any compound and state
3. Rigorous: Guaranteed

First Principles

accuracy (DFT/MD/...)

$$H(\{Z_I, \mathbf{R}_I\})\Psi(\mathbf{r}) \quad = \quad E\Psi(\mathbf{r})$$

Why first principles?

1. General: Any property
2. Transferable: Any compound and state
3. Rigorous: Guaranteed

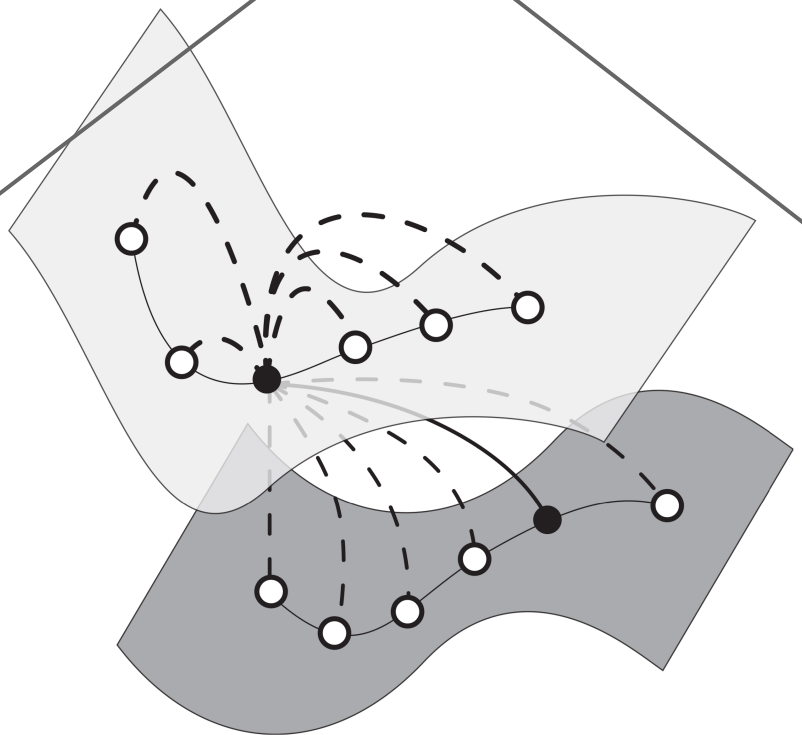
First Principles

accuracy (DFT)

$$H(\{Z_I, \mathbf{R}_I\})\Psi(\mathbf{r}) \quad = \quad E\Psi(\mathbf{r})$$

correlational (inductive)

variational (deductive)

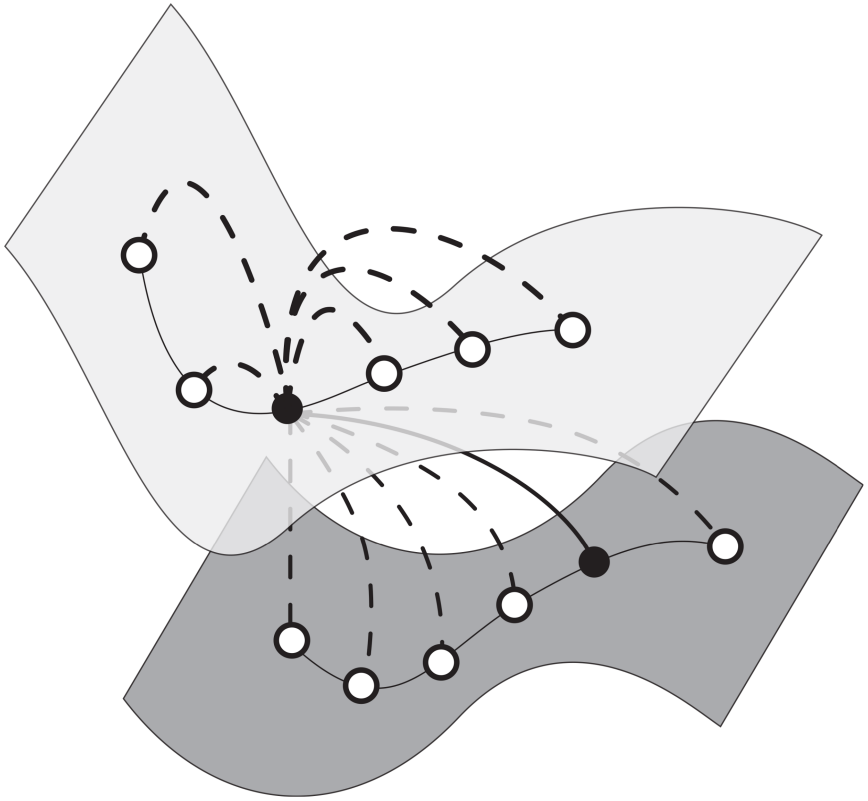


First Principles

$$H(\{Z_I, \mathbf{R}_I\})\Psi(\mathbf{r}) = E\Psi(\mathbf{r})$$

$$H(\{Z_I, \mathbf{R}_I\}) = -\sum_i \nabla_i^2 - \sum_{I,i} \frac{Z_I}{|\mathbf{R}_I - \mathbf{r}_i|} + \sum_{i<j} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|} + \sum_{I<J} \frac{Z_I Z_J}{|\mathbf{R}_I - \mathbf{R}_J|}$$

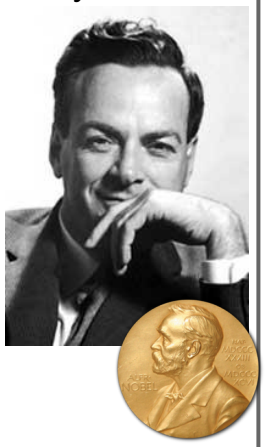
Schrödinger



$$\frac{\partial E[H]}{\partial R_{Ix}} = \left\langle \Psi \left| \frac{\partial H}{\partial R_{Ix}} \right| \Psi \right\rangle$$

$$\frac{\partial E[H]}{\partial Z_I} = \left\langle \Psi \left| \frac{\partial H}{\partial Z_I} \right| \Psi \right\rangle$$

Feynman



First Principles

$$H(\{Z_I, \mathbf{R}_I\})\Psi(\mathbf{r}) = E\Psi(\mathbf{r})$$

$$H(\{Z_I, \mathbf{R}_I\}) = -\sum_i \nabla_i^2 - \sum_{I,i} \frac{Z_I}{|\mathbf{R}_I - \mathbf{r}_i|} + \sum_{i<j} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|} + \sum_{I<J} \frac{Z_I Z_J}{|\mathbf{R}_I - \mathbf{R}_J|}$$

Schrödinger

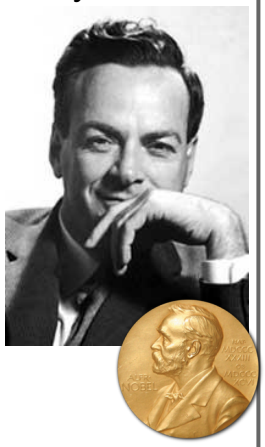
Wilson, *J Phys Chem* (1962); Politzer, Parr *J Phys Chem* (1974); Weigend, Schrodt, Ahlrichs *J Chem Phys* (2004); Beratan, Yang et al *J Am Chem Soc* (2006); Beste et al *J Chem Phys* (2006)

Phys Rev Lett (2005); *J Chem Phys* (2006);

$$\frac{\partial E[H]}{\partial R_{Ix}} = \left\langle \Psi \left| \frac{\partial H}{\partial R_{Ix}} \right| \Psi \right\rangle$$

$$\frac{\partial E[H]}{\partial Z_I} = \left\langle \Psi \left| \frac{\partial H}{\partial Z_I} \right| \Psi \right\rangle$$

Feynman



First Principles

$$H(\{Z_I, \mathbf{R}_I\})\Psi(\mathbf{r}) = E\Psi(\mathbf{r})$$

$$H(\{Z_I, \mathbf{R}_I\}) = -\sum_i \nabla_i^2 - \sum_{I,i} \frac{Z_I}{|\mathbf{R}_I - \mathbf{r}_i|} + \sum_{i<j} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|} + \sum_{I<J} \frac{Z_I Z_J}{|\mathbf{R}_I - \mathbf{R}_J|}$$

Schrödinger

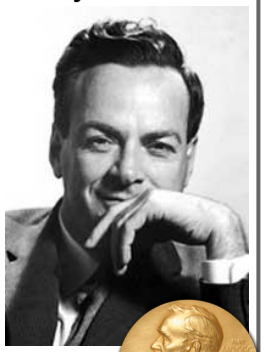
Wilson, *J Phys Chem* (1962); Politzer, Parr *J Phys Chem* (1974); Weigend, Schrodt, Ahlrichs *J Chem Phys* (2004); Beratan, Yang et al *J Am Chem Soc* (2006); Beste et al *J Chem Phys* (2006)

Phys Rev Lett (2005); *J Chem Phys* (2006); *J Chem Phys* (2009); *Int J Quant Chem* (2013); *CHIMIA* (2014)

Feynman

$$\frac{\partial E[H]}{\partial R_{Ix}} = \left\langle \Psi \left| \frac{\partial H}{\partial R_{Ix}} \right| \Psi \right\rangle$$

$$\frac{\partial E[H]}{\partial Z_I} = \left\langle \Psi \left| \frac{\partial H}{\partial Z_I} \right| \Psi \right\rangle$$



$$E(H(\lambda)) = E(H_i + \lambda(H_f - H_i))$$

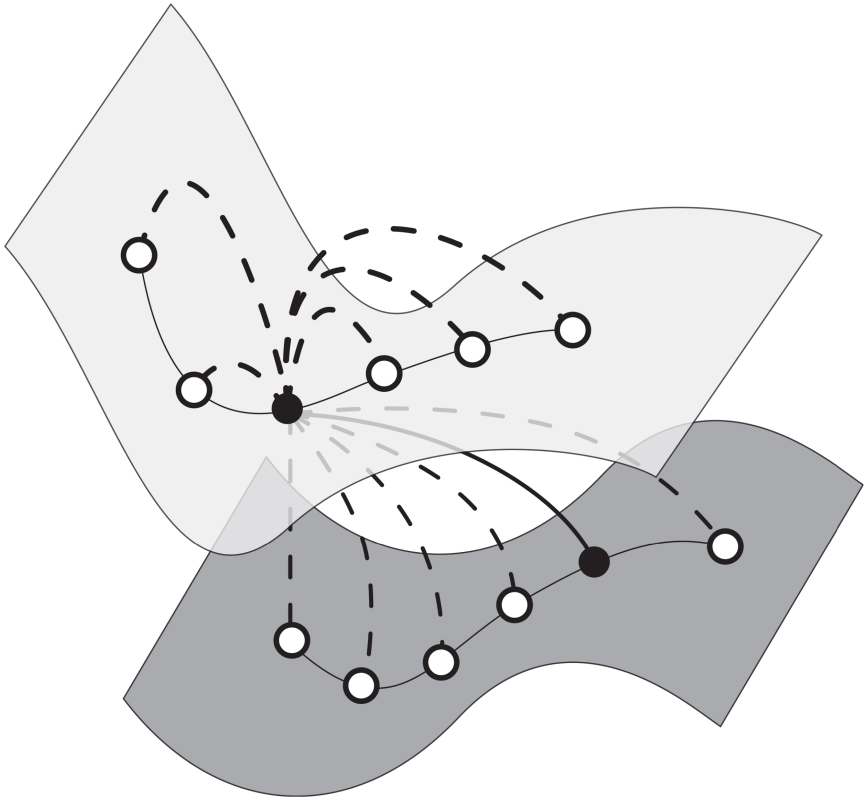
$$\frac{\partial E[H]}{\partial \lambda} = \left\langle \Psi \left| \frac{\partial H(\lambda)}{\partial \lambda} \right| \Psi \right\rangle$$

First Principles

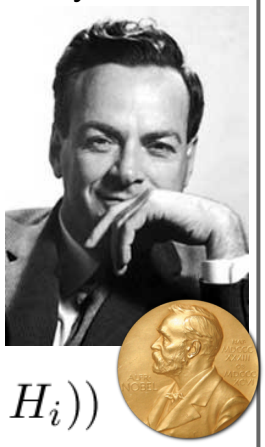
$$H(\{Z_I, \mathbf{R}_I\})\Psi(\mathbf{r}) = E\Psi(\mathbf{r})$$

$$H(\{Z_I, \mathbf{R}_I\}) = -\sum_i \nabla_i^2 - \sum_{I,i} \frac{Z_I}{|\mathbf{R}_I - \mathbf{r}_i|} + \sum_{i<j} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|} + \sum_{I<J} \frac{Z_I Z_J}{|\mathbf{R}_I - \mathbf{R}_J|}$$

Schrödinger



Feynman



$$\frac{\partial E[H]}{\partial R_{Ix}} = \left\langle \Psi \left| \frac{\partial H}{\partial R_{Ix}} \right| \Psi \right\rangle$$

$$\frac{\partial E[H]}{\partial Z_I} = \left\langle \Psi \left| \frac{\partial H}{\partial Z_I} \right| \Psi \right\rangle$$

$$E(H(\lambda)) = E(H_i + \lambda(H_f - H_i))$$

$$\frac{\partial E[H]}{\partial \lambda} = \left\langle \Psi \left| \frac{\partial H(\lambda)}{\partial \lambda} \right| \Psi \right\rangle$$

First Principles

$$H(\{Z_I, \mathbf{R}_I\})\Psi(\mathbf{r}) = E\Psi(\mathbf{r})$$

$$H(\{Z_I, \mathbf{R}_I\}) = -\sum_i \nabla_i^2 - \sum_{I,i} \frac{Z_I}{|\mathbf{R}_I - \mathbf{r}_i|} + \sum_{i<j} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|} + \sum_{I<J} \frac{Z_I Z_J}{|\mathbf{R}_I - \mathbf{R}_J|}$$

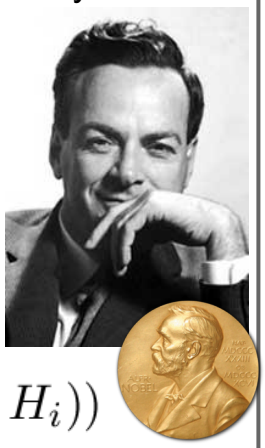
Schrödinger

Machine Learning!?

$$\frac{\partial E[H]}{\partial R_{Ix}} = \left\langle \Psi \left| \frac{\partial H}{\partial R_{Ix}} \right| \Psi \right\rangle$$

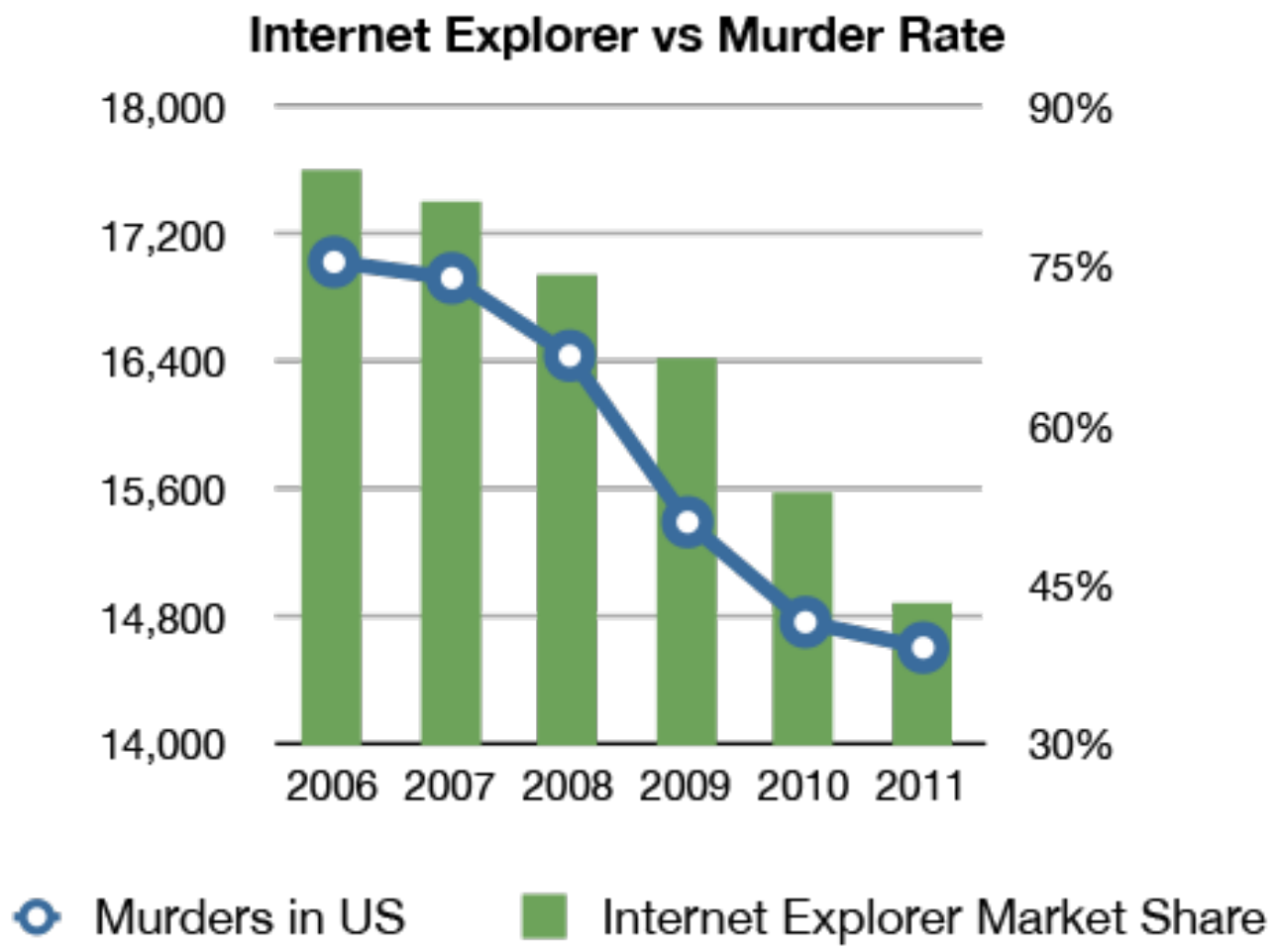
$$\frac{\partial E[H]}{\partial Z_I} = \left\langle \Psi \left| \frac{\partial H}{\partial Z_I} \right| \Psi \right\rangle$$

Feynman



$$E(H(\lambda)) = E(H_i + \lambda(H_f - H_i))$$

$$\frac{\partial E[H]}{\partial \lambda} = \left\langle \Psi \left| \frac{\partial H(\lambda)}{\partial \lambda} \right| \Psi \right\rangle$$

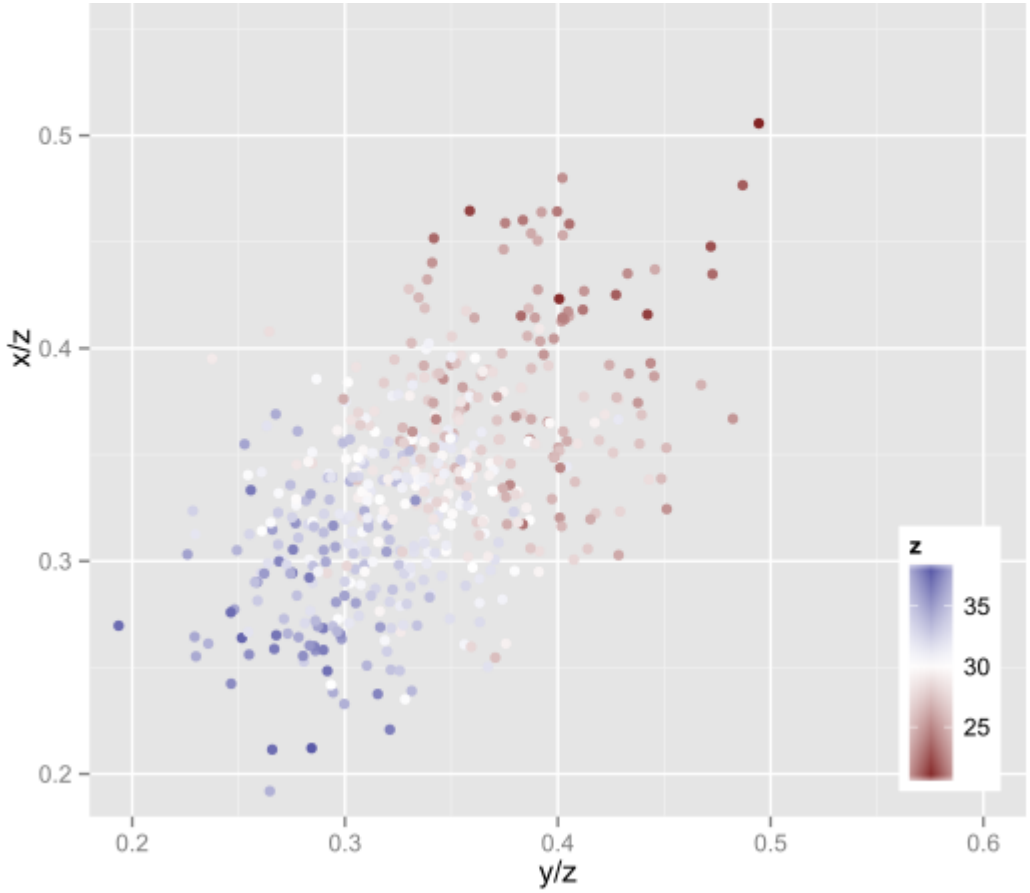

Internet Explorer vs Murder Rate
18,000
17,200
16,400
15,600
14,800
14,000
90%
75%
60%
45%
30%
2006
2007
2008
2009
2010
2011
Murders in US
Internet Explorer Market Share

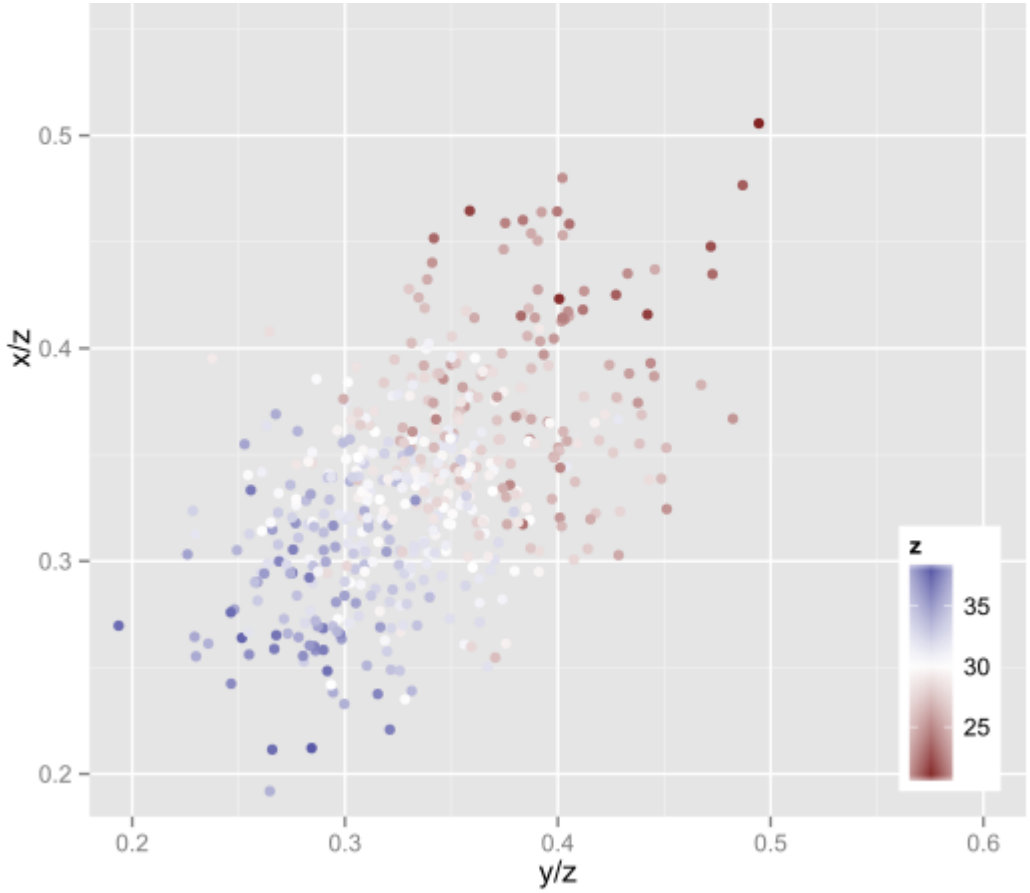
Correlation must *not* be used to infer a causal relationship, however if there is a causal relationship there must be a correlation …
→ Correlation is a necessary but not sufficient condition.

Dangerous: Humans have cognitive bias [“Thinking, Fast and Slow” Tversky and Kahneman, “Fooled by Randomness”, Nassim Taleb]

Correlation can also be due to

1. chance (*any* two variables that change will correlate)
2. a common cause
3. identity relationships
4. spurious correlations


0.5
0.4
0.3
0.2
x/z
0.2
0.3
0.4
0.5
0.6
y/z
z
35
30
25



Spurious correlation for 500 draws of *x,y,z* from

$$x, y \sim N(10, 1)$$
$$z \sim N(30, 9)$$

www.wikipedia.org

First Principles

$$H(\{Z_I, \mathbf{R}_I\})\Psi(\mathbf{r}) = E\Psi(\mathbf{r})$$

$$H(\{Z_I, \mathbf{R}_I\}) = -\sum_i \nabla_i^2 - \sum_{I,i} \frac{Z_I}{|\mathbf{R}_I - \mathbf{r}_i|} + \sum_{i<j} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|} + \sum_{I<J} \frac{Z_I Z_J}{|\mathbf{R}_I - \mathbf{R}_J|}$$

Schrödinger

Vapnik

$$\{Z_I, \mathbf{R}_I\} \overset{H\Psi}{\longmapsto} E$$

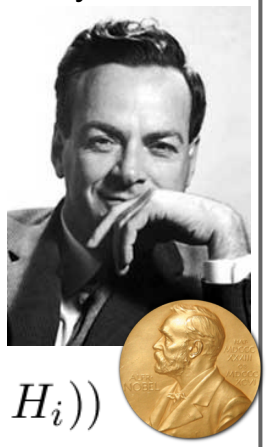
supervised learning

$$\{Z_I, \mathbf{R}_I\} \overset{\text{ML}}{\longmapsto} E$$

$$\frac{\partial E[H]}{\partial R_{Ix}} = \left\langle \Psi \left| \frac{\partial H}{\partial R_{Ix}} \right| \Psi \right\rangle$$

$$\frac{\partial E[H]}{\partial Z_I} = \left\langle \Psi \left| \frac{\partial H}{\partial Z_I} \right| \Psi \right\rangle$$

Feynman



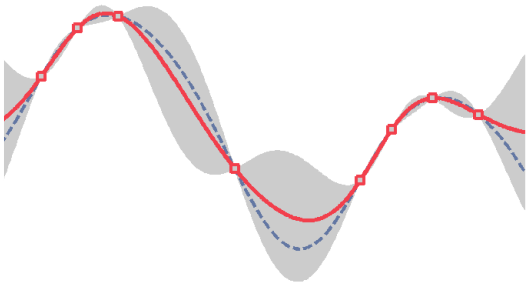
$$E(H(\lambda)) = E(H_i + \lambda(H_f - H_i))$$

$$\frac{\partial E[H]}{\partial \lambda} = \left\langle \Psi \left| \frac{\partial H(\lambda)}{\partial \lambda} \right| \Psi \right\rangle$$

Machine Learning in Chemical Space

$$\{Z_I, \mathbf{R}_I\} \xmapsto{H\Psi} E$$

supervised learning



$$\{Z_I, \mathbf{R}_I\} \xmapsto{\mathrm{ML}} E$$

Data-driven Science

Inductive

1. Assume a law
2. Metric
3. Examples
4. Infer
5. New combination

Fast (ms)
Arbitrary reference
Automatic improvement

Transferable?
Minimally condensed

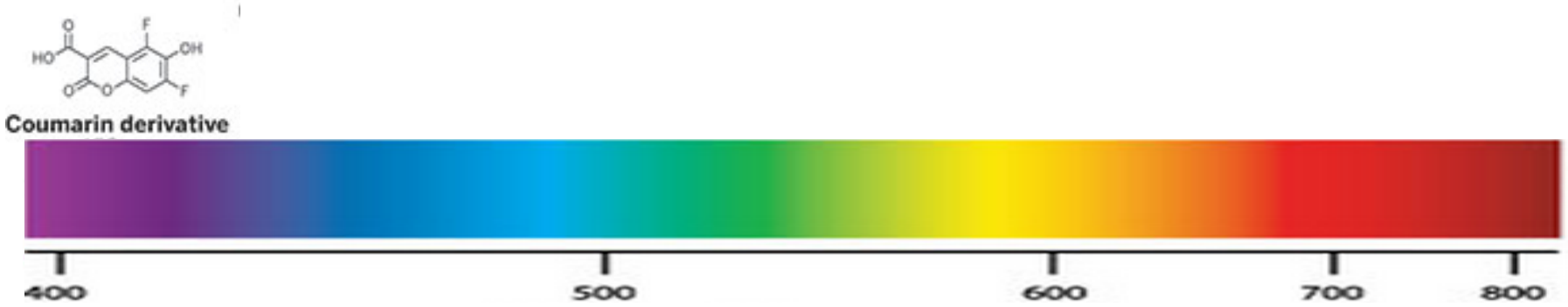
Deductive

1. Assume a law
2. Approximate
3. Solve
4. Predict
5. New regimes

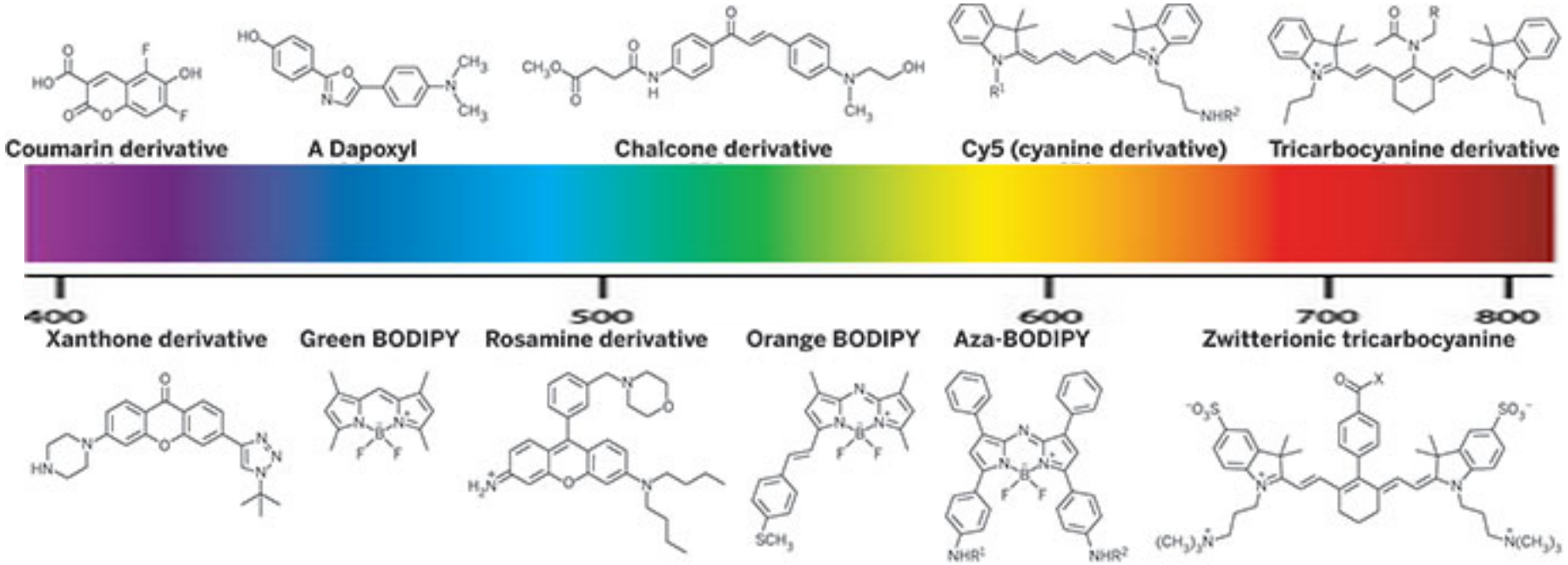
Slow (depending on approx.)
Approximation dependent
Human improvement

Transferable?
Maximally condensed

Configuration + Composition → Chemical Space

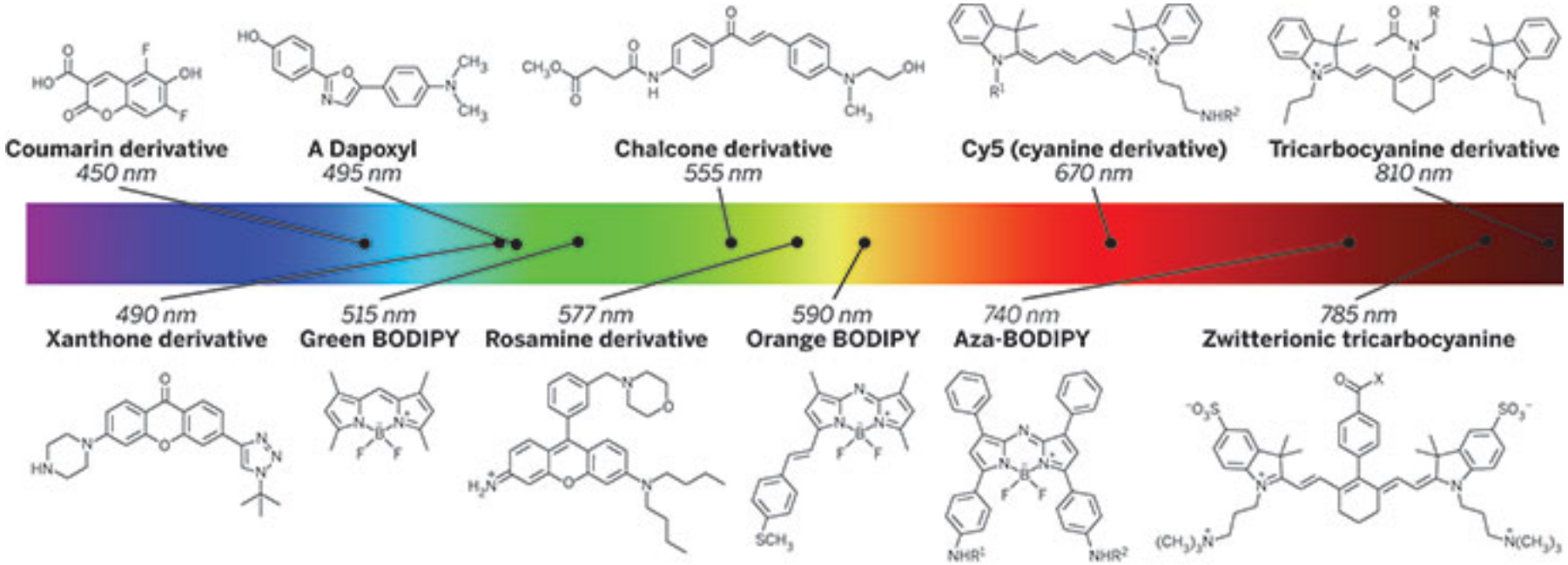


Configuration + Composition → Chemical Space



Young-Tae Chang et al *C&E News* **93 (12) 39-40** (2015)

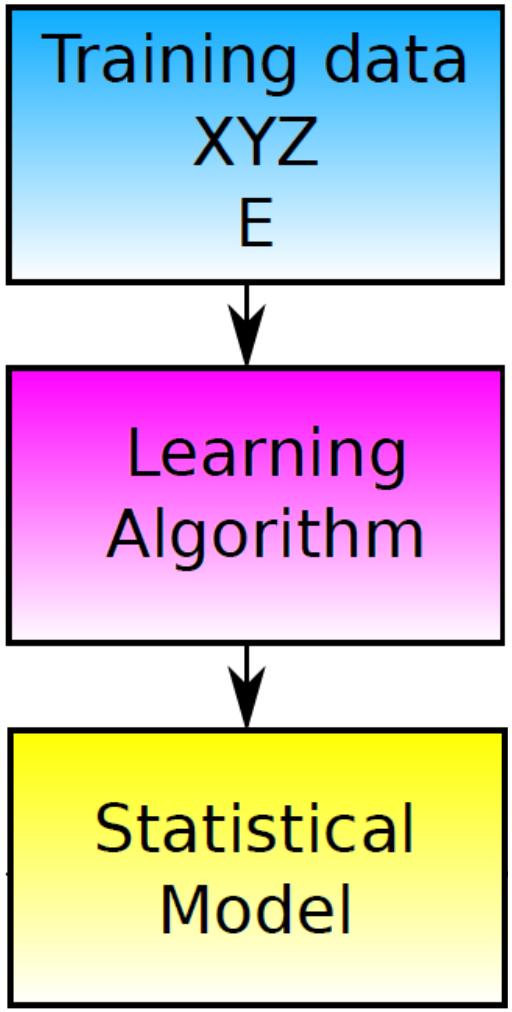
Configuration + Composition → Chemical Space



Young-Tae Chang et al *C&E News* **93 (12) 39-40** (2015)

Machine Learning in Chemical Space

1. Train

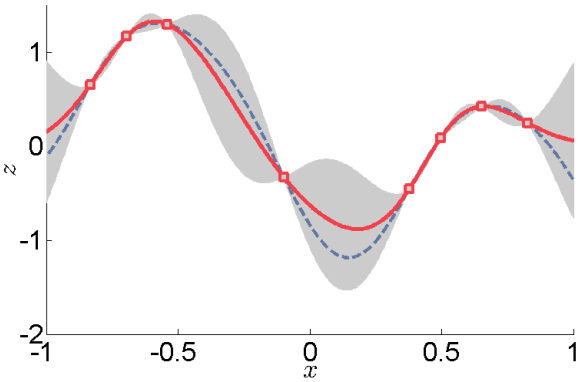


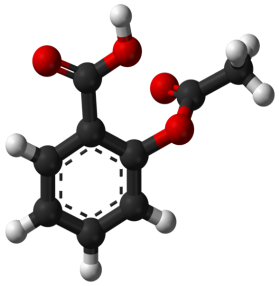
The bigger the data
the better

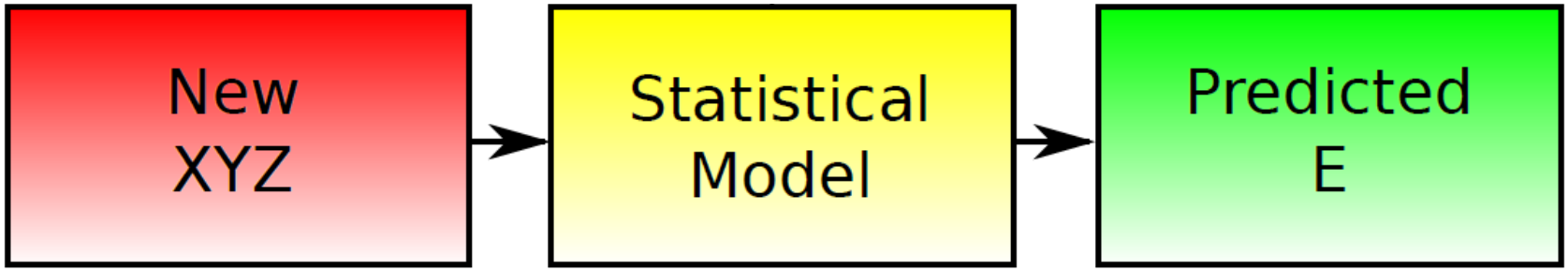
Machine Learning in Chemical Space

1. Train

2. Predict







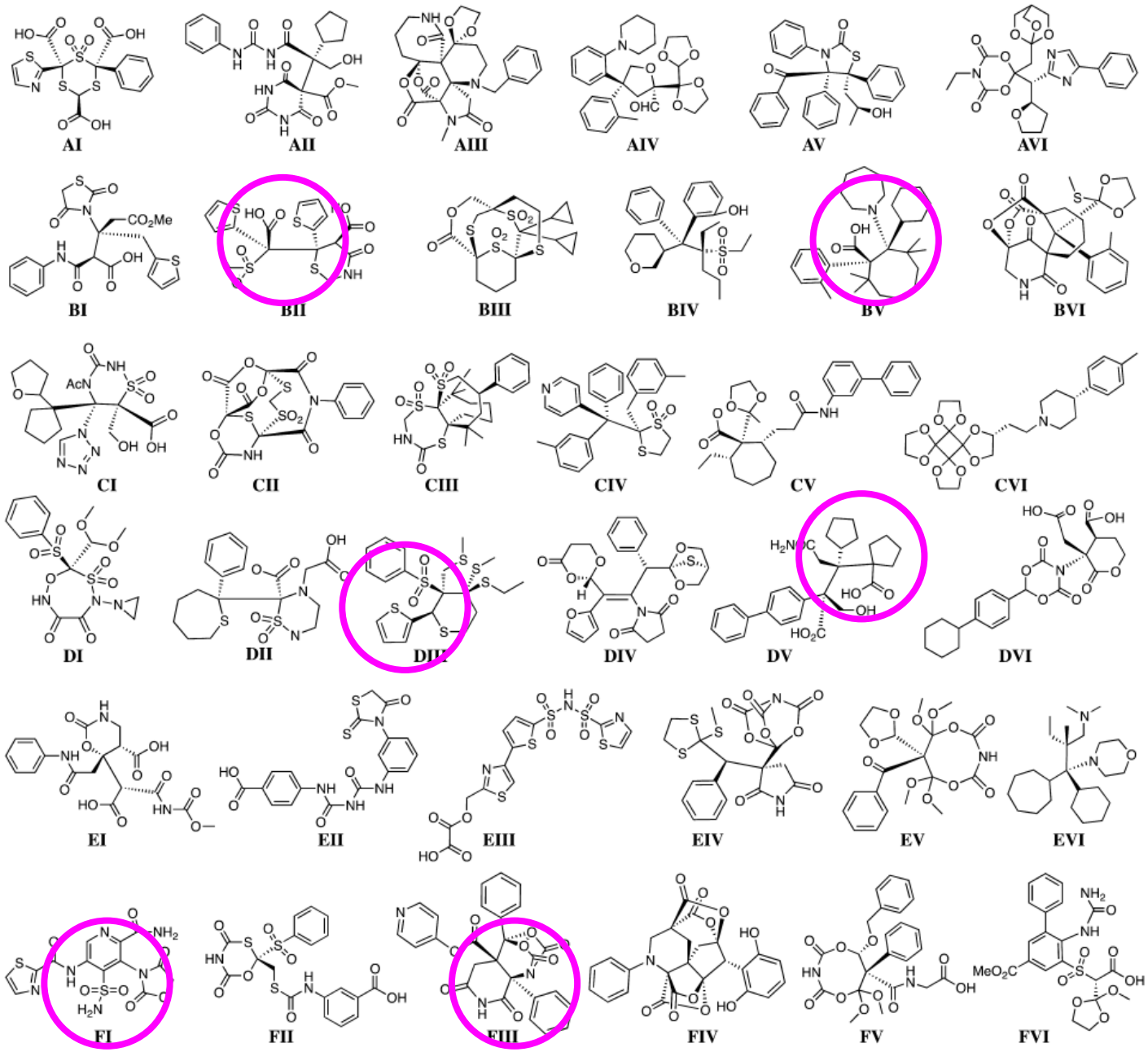
Kernel Ridge Regression

$$E^{est}(\mathbf{M}) = \sum_{i}^{N} \alpha_i k(\mathbf{M}, \mathbf{M}_i)$$

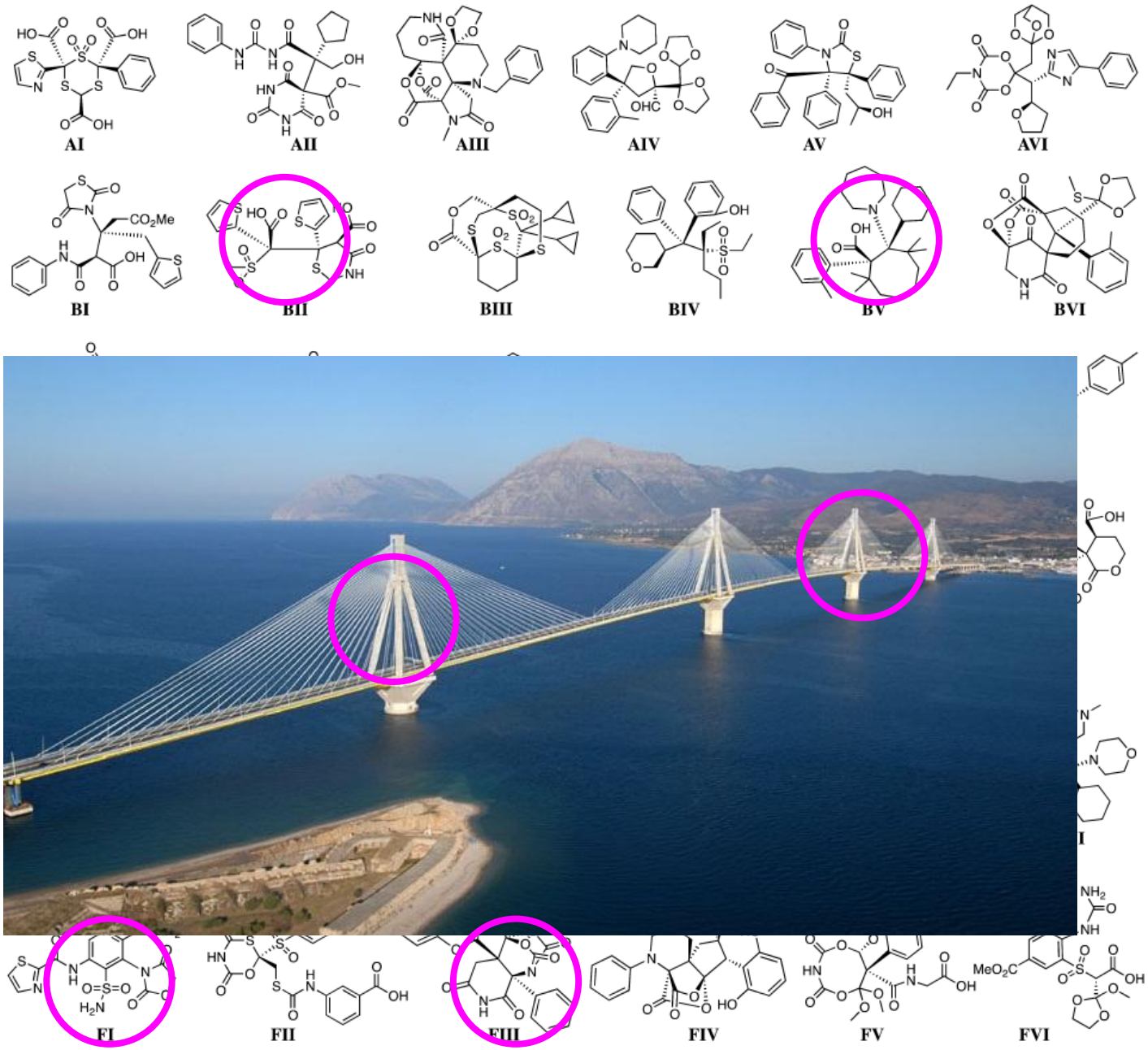
$$\min_{\alpha} = \left(\sum_{i} \left(E^{est}(\mathbf{M}_i) - E_i^{ref} \right)^2 + \lambda \sum_{ij} \alpha_i \alpha_j k(\mathbf{M}_i, \mathbf{M}_j) \right)$$

$$\alpha = (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{E}^{ref}$$ **Solution**

$$k(\mathbf{M}, \mathbf{M}') = \exp\left(-\frac{d(\mathbf{M}, \mathbf{M}')^2}{2\sigma^2} \right)$$



Virshup, Yang, Beratan et al *J Am Chem Soc* (2013)

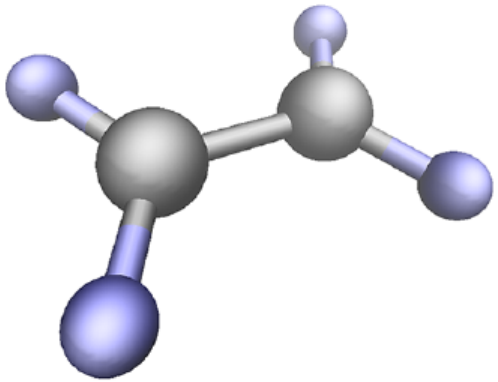


Virshup, Yang, Beratan et al *J Am Chem Soc* (2013)

Model

$$E^{est}(\mathbf{M}) = \sum_i \alpha_i e^{-\frac{d(\mathbf{M},\mathbf{M}_i)^2}{2\sigma^2}}$$

Need to compare
→ What is **M**?



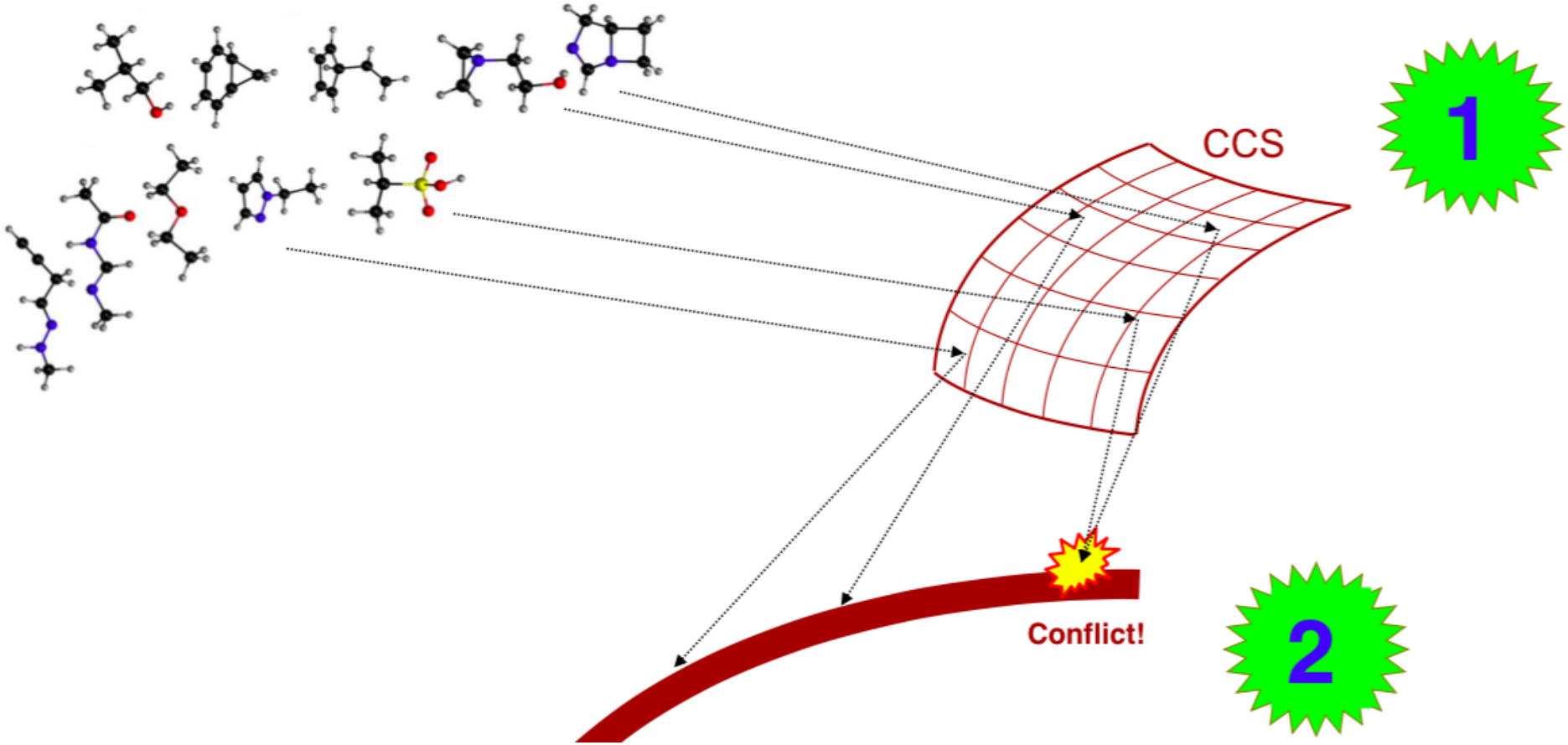
→

Crucial property

- **unique**

Desirable properties

- **translation invariant**
- **rotation invariant**
- **symmetry invariant**
- **index invariant**
- **constant length**
- **continuous**



Slide from Klaus-Robert Mueller

The reason for the uniqueness requirement can be shown by *reductio ad absurdum* in three steps—in analogy to the first Hohenberg–Kohn theorem[45]—for any quantum mechanical observable $\mathcal{O}=\langle\Psi|\hat{O}|\Psi\rangle$. Here, the unperturbed ground-state Hamiltonian H is defined by its external potential, determined by $\{Z_I, \mathbf{R}_I\}$, the set of nuclear charges and coordinates, as well as number of electrons N_e. The variational principle yields the system's many-body wavefunction Ψ for any given H.

OAvL, R. Ramakrishnan, M. Rupp, A. Knoll, *IJQC* (2015), arXiv

i. Let D denote a descriptor that is not unique. Then, two systems $H_1 \neq H_2$ exist that differ in excess of the invariants, but they are mapped to the same descriptor value d, $H_1 \rightarrow d$ and $H_2 \rightarrow d$.

OAvL, R. Ramakrishnan, M. Rupp, A. Knoll, *IJQC* (2015), arXiv

i. Let D denote a descriptor that is not unique. Then, two systems $H_1 \neq H_2$ exist that differ in excess of the invariants, but they are mapped to the same descriptor value d, $H_1 \rightarrow d$ and $H_2 \rightarrow d$.

ii. Because H_1 and H_2 differ by more than their property's invariances, they have different wave-functions, $\Psi_1 \neq \Psi_2$, yielding two different observables, $\mathcal{O}_1 = \langle \Psi_1 | \hat{O} | \Psi_1 \rangle$ and $\mathcal{O}_2 = \langle \Psi_2 | \hat{O} | \Psi_2 \rangle$. Here, we deliberately ignore the obvious exception and special situation of all observables which happen to be degenerate.

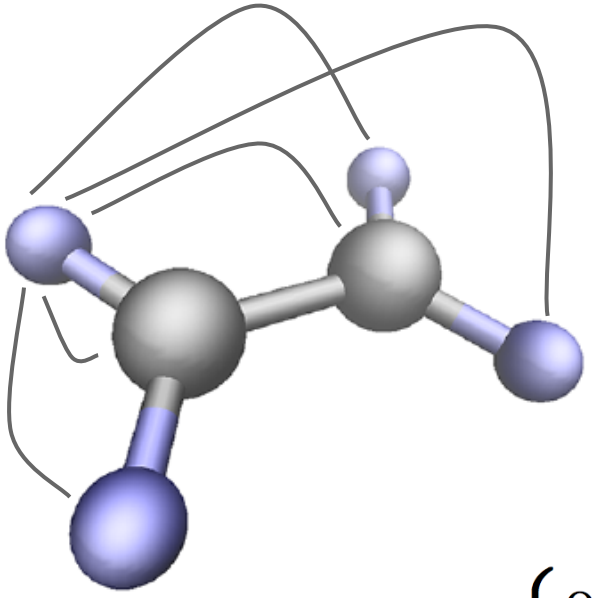
OAvL, R. Ramakrishnan, M. Rupp, A. Knoll, *IJQC* (2015), arXiv

i. Let D denote a descriptor that is not unique. Then, two systems $H_1 \neq H_2$ exist that differ in excess of the invariants, but they are mapped to the same descriptor value d, $H_1 \rightarrow d$ and $H_2 \rightarrow d$.

ii. Because H_1 and H_2 differ by more than their property's invariances, they have different wave-functions, $\Psi_1 \neq \Psi_2$, yielding two different observables, $\mathcal{O}_1 = \langle \Psi_1 | \hat{O} | \Psi_1 \rangle$ and $\mathcal{O}_2 = \langle \Psi_2 | \hat{O} | \Psi_2 \rangle$. Here, we deliberately ignore the obvious exception and special situation of all observables which happen to be degenerate.

iii. A trained statistical model predicts any observable $\mathcal{O}$ solely based on descriptor input d leading to identical predictions $\mathcal{O}_1^{\text{pred}} = \mathcal{O}_2^{\text{pred}}$. In the limit of infinite training data, these predictions will be exact, implying $\mathcal{O}_1 = \mathcal{O}_2$, in contradiction to (ii).

OAvL, R. Ramakrishnan, M. Rupp, A. Knoll, *IJQC* (2015), arXiv



$\longrightarrow$

$$\mathbf{M} =$$

	H	H	C	C	H	H
H	0.5	0.3	2.9	1.5	0.2	0.2
H	0.3	0.5	2.9	1.5	0.2	0.2
C	2.9	2.9	36.9	14.3	1.5	1.5
C	1.5	1.5	14.3	36.9	2.9	2.9
H	0.2	0.2	1.5	2.9	0.5	0.3
H	0.2	0.2	1.5	2.9	0.3	0.5

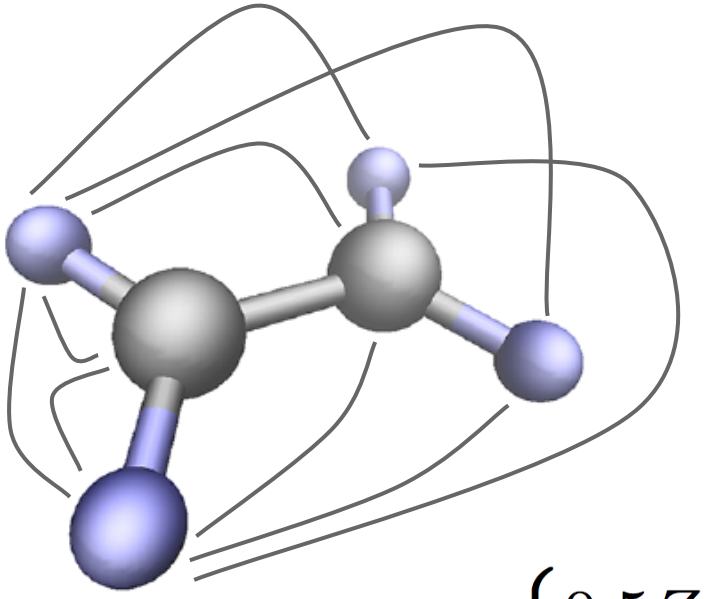
$$M_{IJ} = \begin{cases} 0.5 Z_I^{2.4} & \forall \ I = J, \\ \frac{Z_I Z_J}{|\mathbf{R}_I - \mathbf{R}_J|} & \forall \ I \neq J. \end{cases}$$

Desirable descriptors are

- unique
- translation invariant
- rotation invariant
- symmetry invariant
- index invariant
- constant length
- continuous

Coulomb-matrix

- unique
- translation
- rotation
- symmetry
- fill up w zeros
- sort/diagonalize/permutate
- continuous



$\longrightarrow$

$$\mathbf{M} =$$

	H	H	C	C	H	H
H	0.5	0.3	2.9	1.5	0.2	0.2
H	0.3	0.5	2.9	1.5	0.2	0.2
C	2.9	2.9	36.9	14.3	1.5	1.5
C	1.5	1.5	14.3	36.9	2.9	2.9
H	0.2	0.2	1.5	2.9	0.5	0.3
H	0.2	0.2	1.5	2.9	0.3	0.5

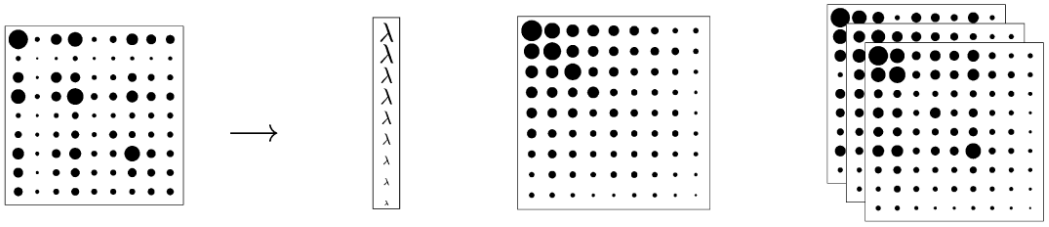
$$M_{IJ} = \begin{cases} 0.5 Z_I^{2.4} & \forall \ I = J, \\ \frac{Z_I Z_J}{|\mathbf{R}_I - \mathbf{R}_J|} & \forall \ I \neq J. \end{cases}$$

Desirable descriptors are

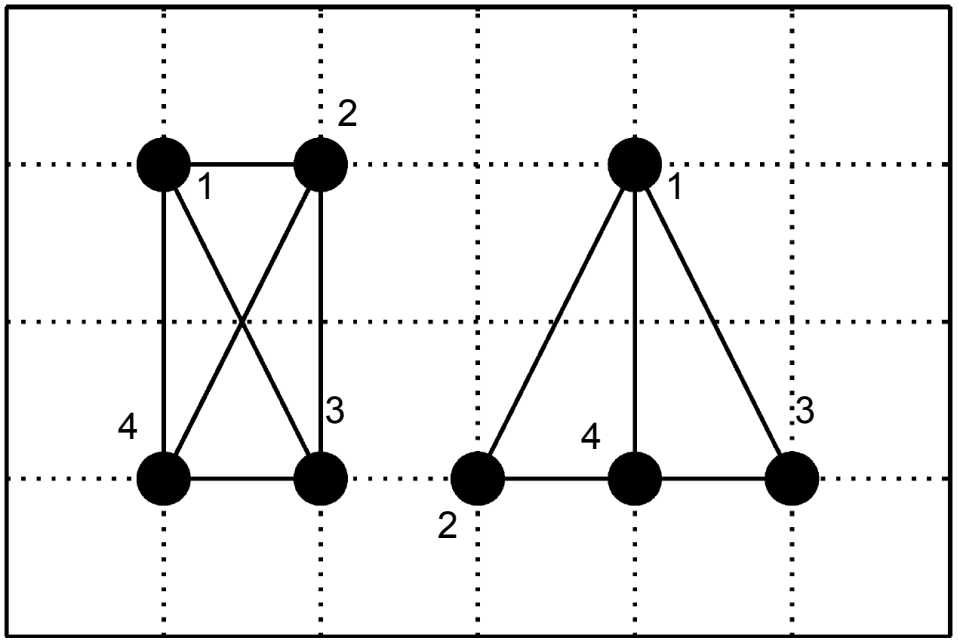
- unique
- translation invariant
- rotation invariant
- symmetry invariant
- index invariant
- constant length
- continuous

Coulomb-matrix

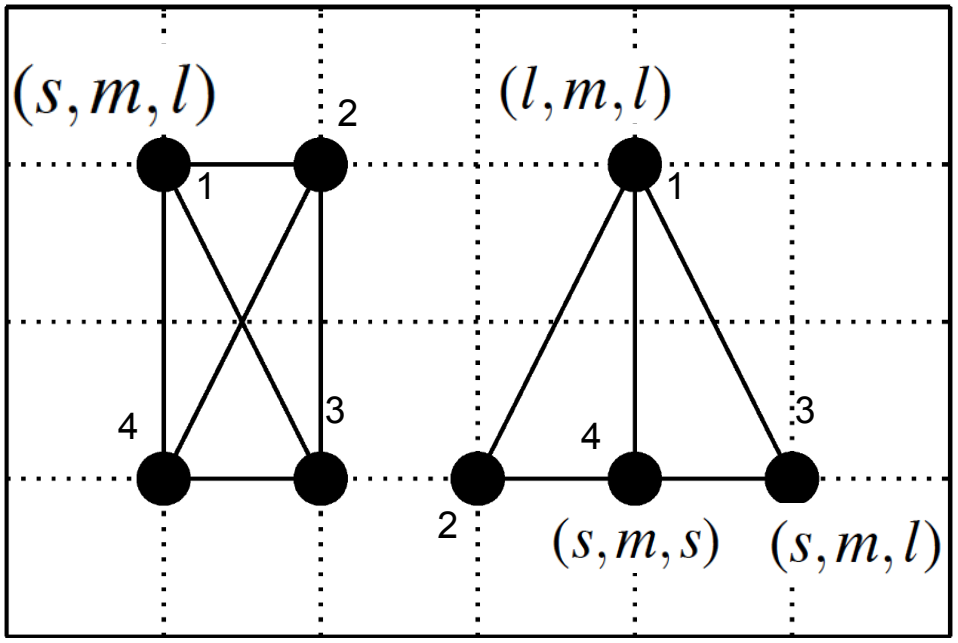
- unique
- translation
- rotation
- symmetry
- fill up w zeros
- sort/diagonalize/permutate
- continuous



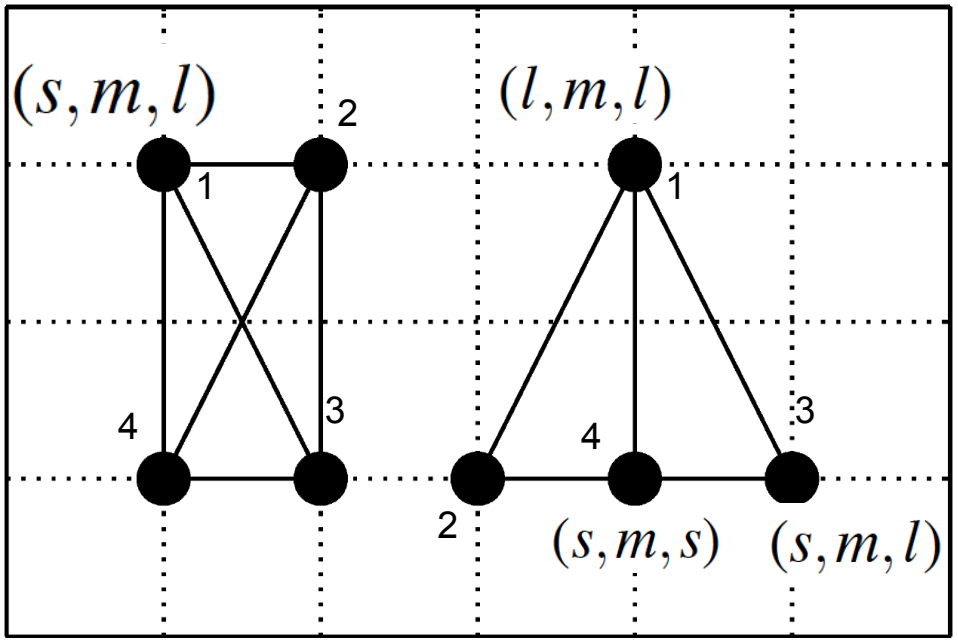
Homometric molecules?



Homometric molecules?

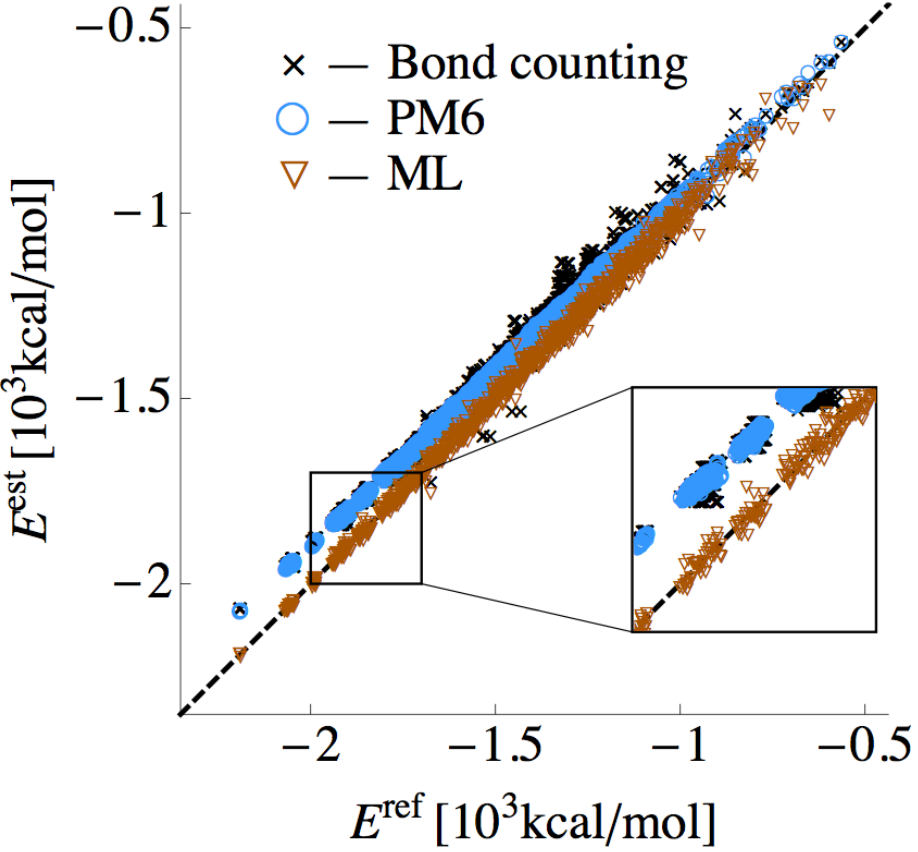


Homometric molecules?



	s	l	m
s		m	l
l	m		s
m	l	s	

	l	l	m
l		m	s
l	m		s
m	s	s	



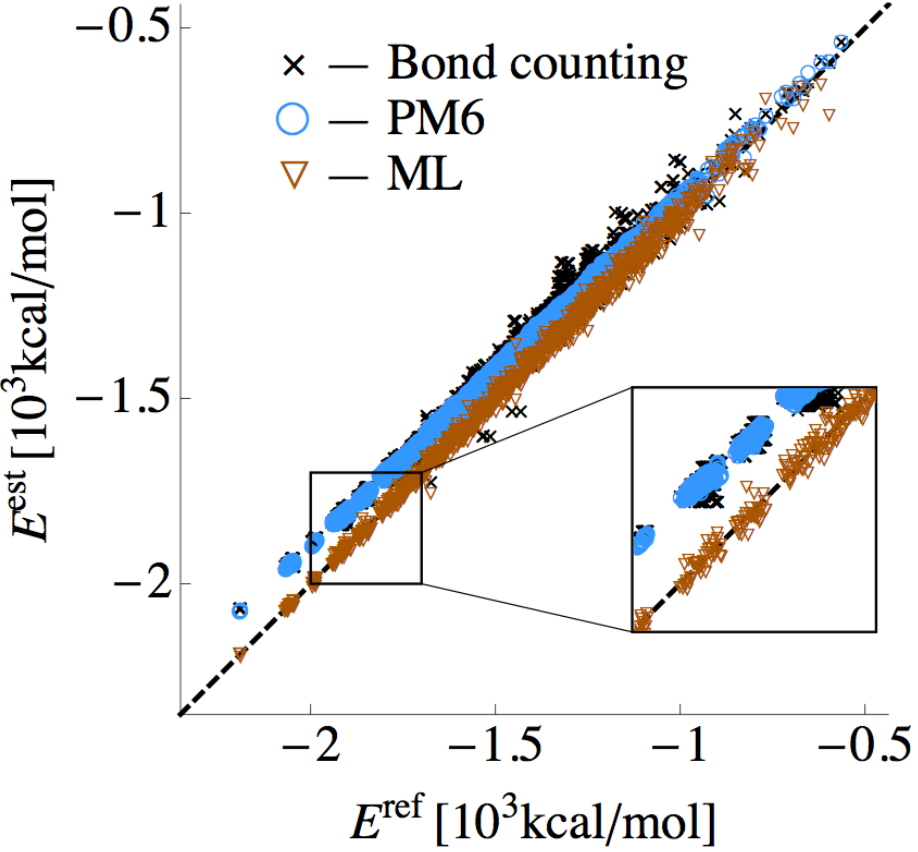
Training for N = 1000 molecules

MAE ~15 kcal/mol

PBE0: ~1000 seconds

ML: ~0.001 seconds

M. Rupp, R. Tkatchenko, K.-R. Muller, OAvL, *PRL* (2012), arXiv



The bigger the data the better

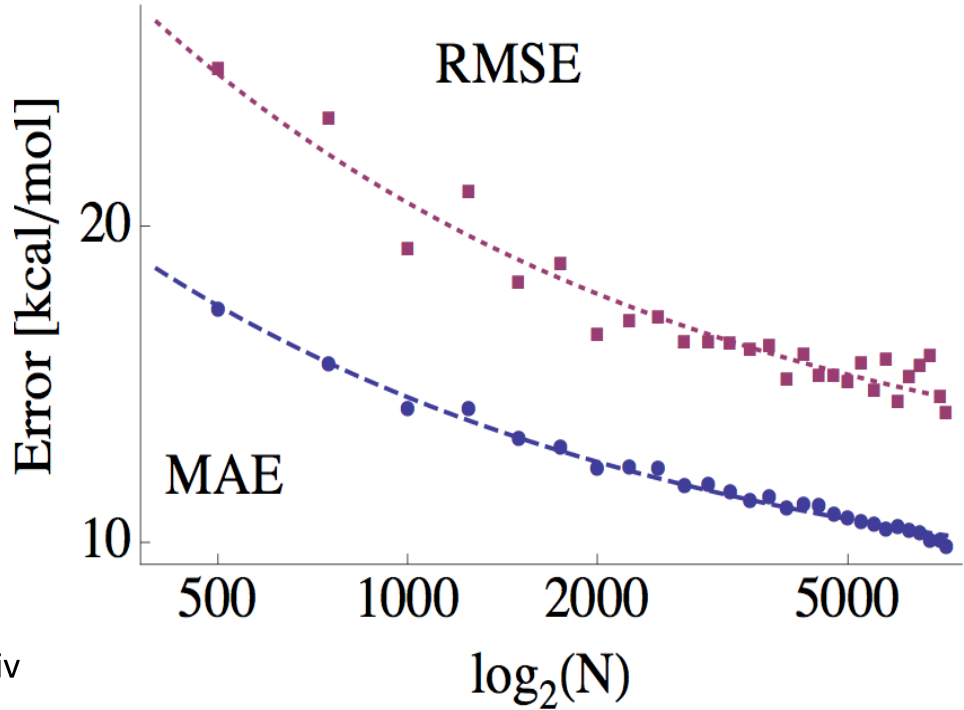
Training for N = 1000 molecules

MAE ~15 kcal/mol

PBE0: ~1000 seconds

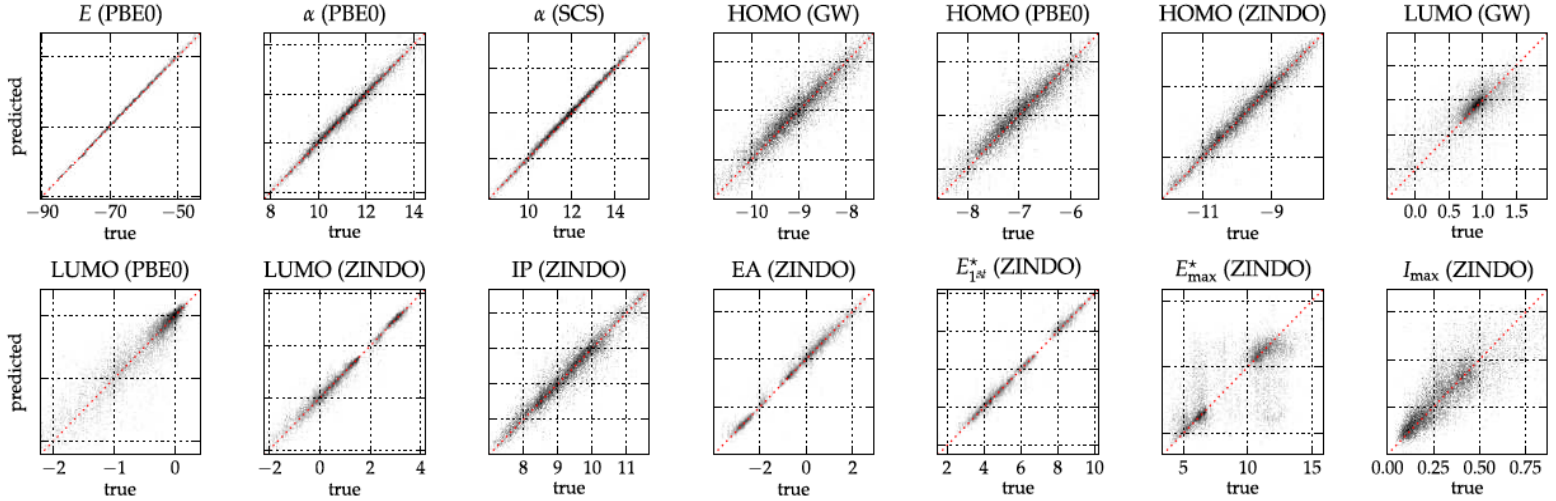
ML: ~0.001 seconds

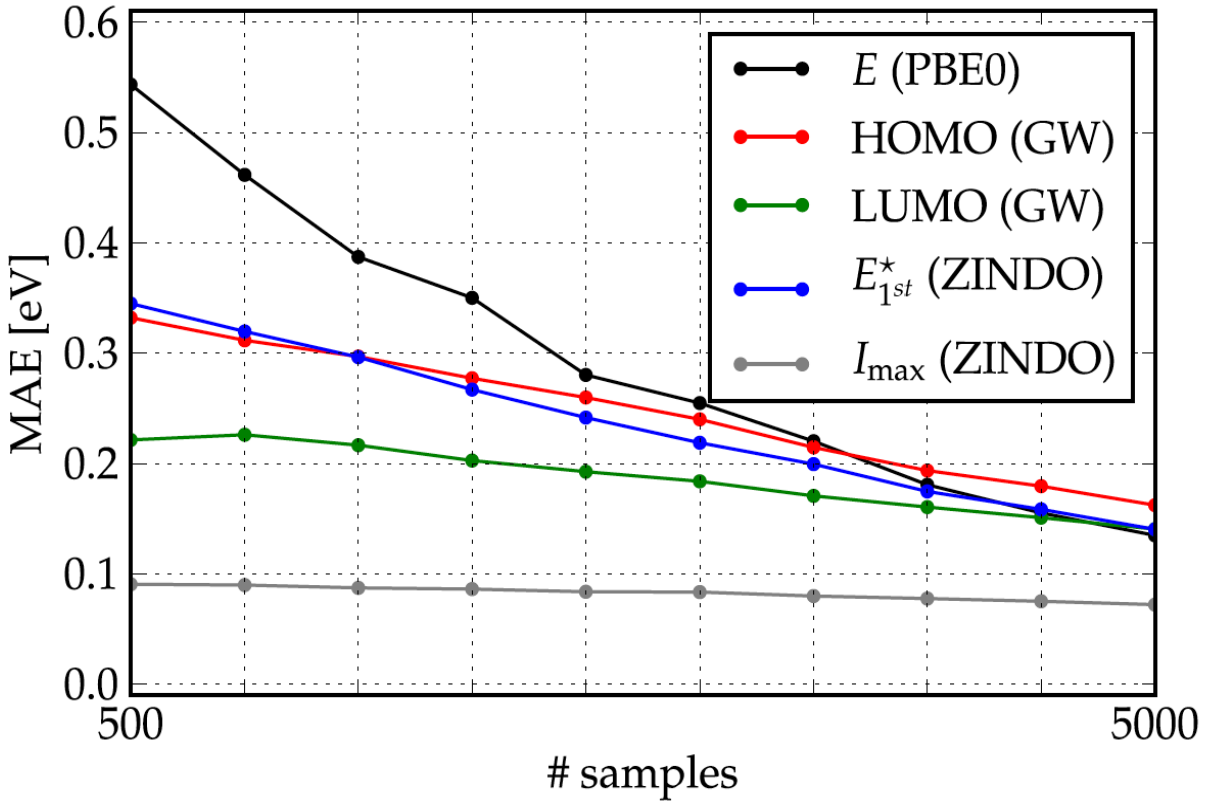
Systematic error decay



M. Rupp, R. Tkatchenko, K.-R. Muller, OAvL, *PRL* (2012), arXiv

http://www.quantum-machine.org/





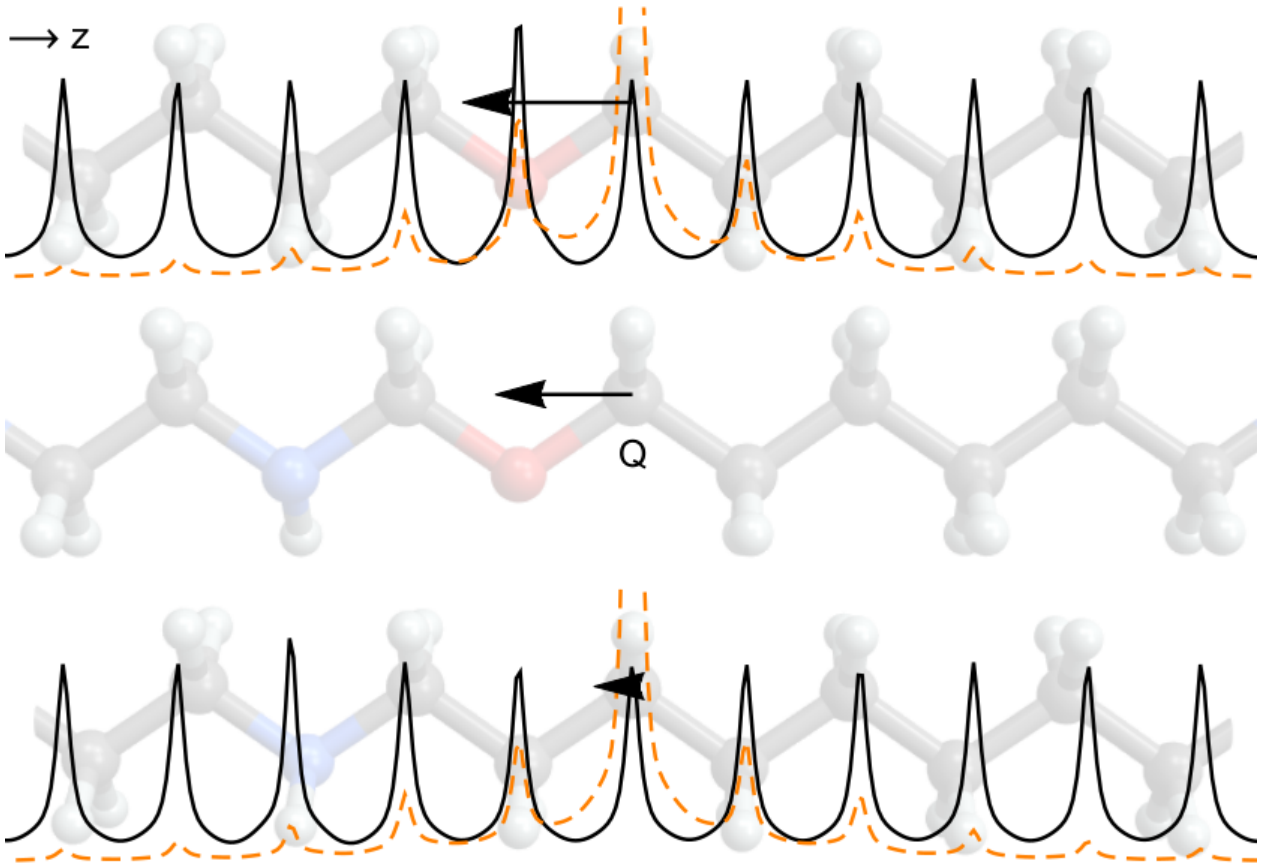
G. Montavon et al, *New J Phys* (2013)

An Atom in Many Molecules

M. Rupp, R. Ramakrishnan, OAvL, *submitted* (2015), arXiv

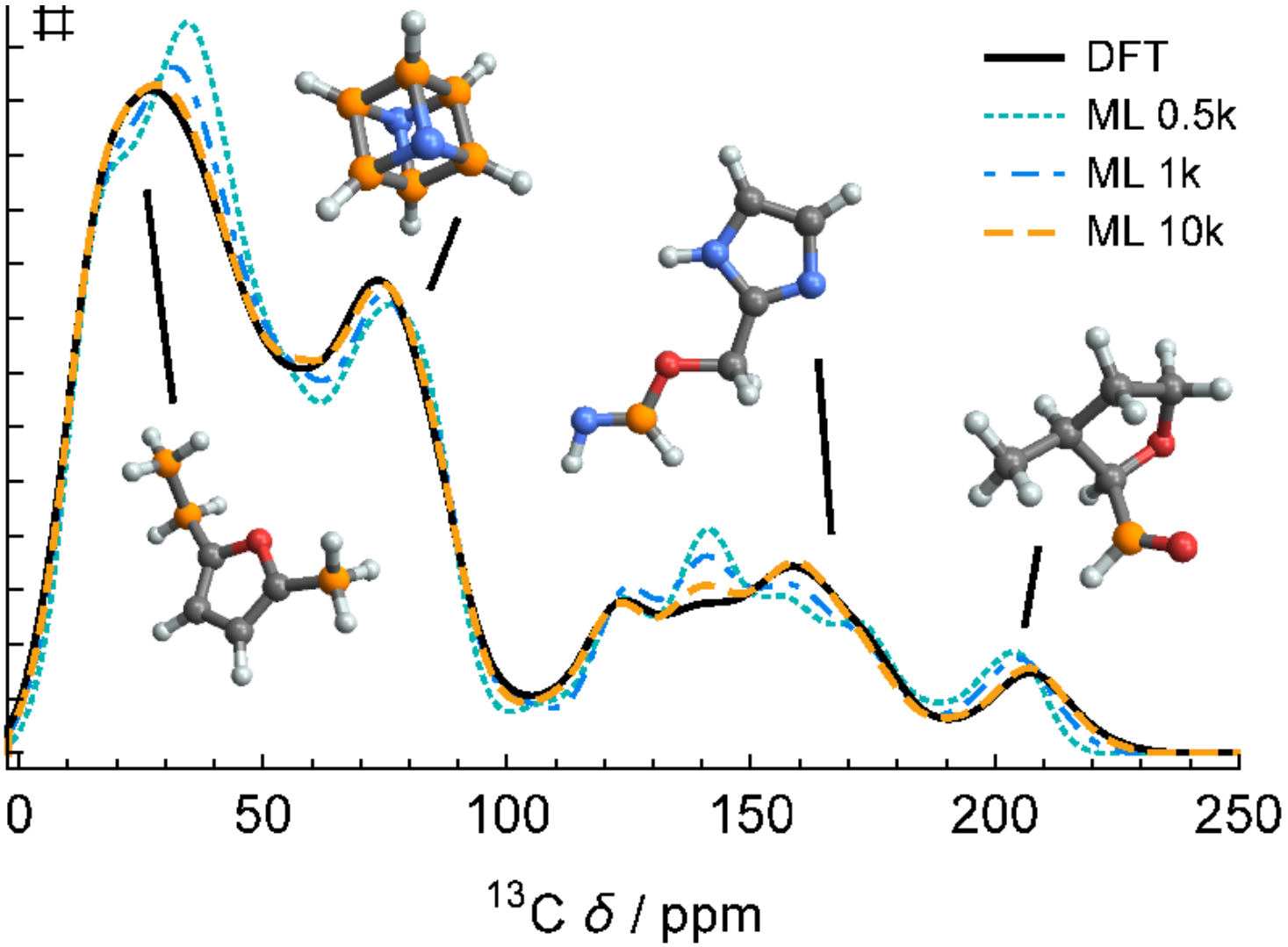
An Atom in Many Molecules

$$\langle \Psi | \partial_{\boldsymbol{R}_Q} \hat{H} | \Psi \rangle = \int \mathrm{d}\boldsymbol{r} \, (\boldsymbol{r} - \boldsymbol{R}_Q) Z_Q \, n(\boldsymbol{r}) / \|\boldsymbol{r} - \boldsymbol{R}_Q\|^3$$



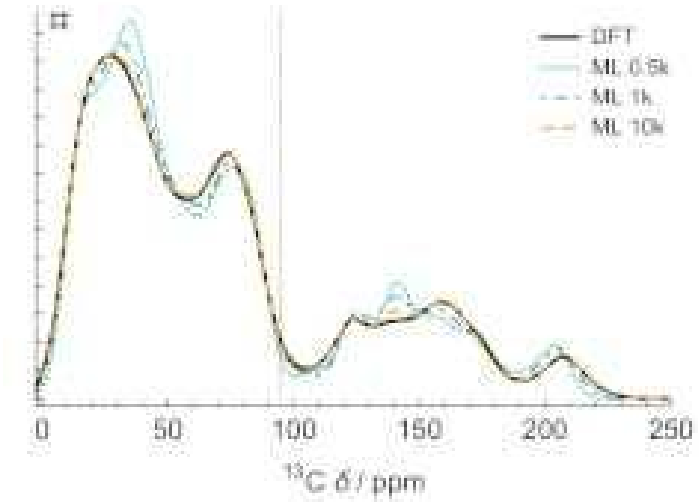
M. Rupp, R. Ramakrishnan, OAvL, *submitted* (2015), arXiv

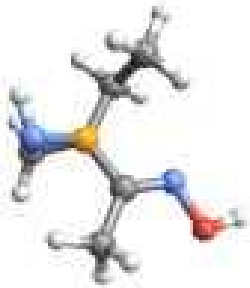
An Atom in Many Molecules



M. Rupp, R. Ramakrishnan, OAvL, *submitted* (2015), arXiv

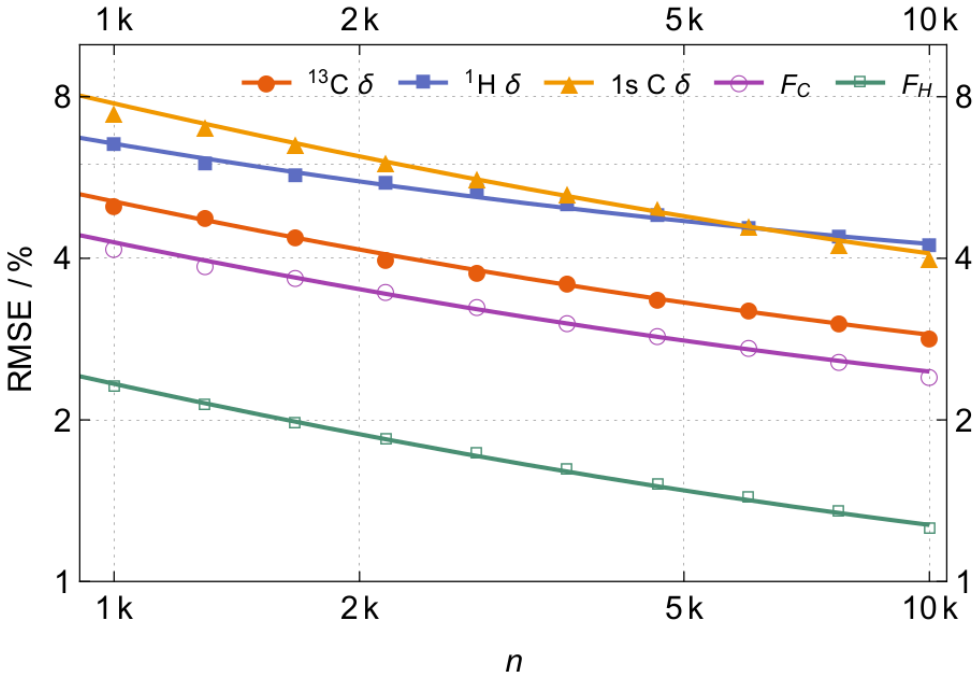
An Atom in Many Molecules





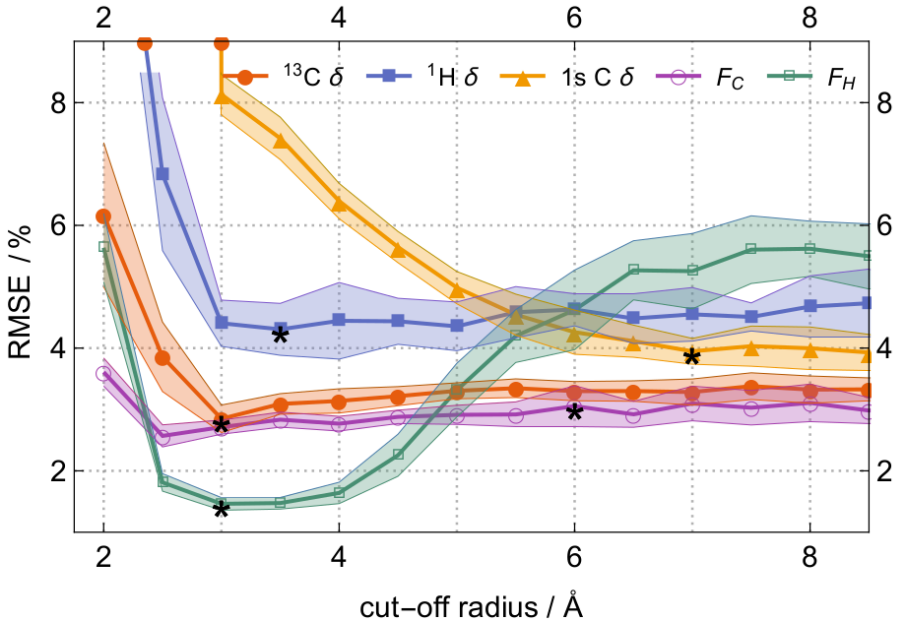
M. Rupp, R. Ramakrishnan, OAvL, *submitted* (2015), arXiv

An Atom in Many Molecules



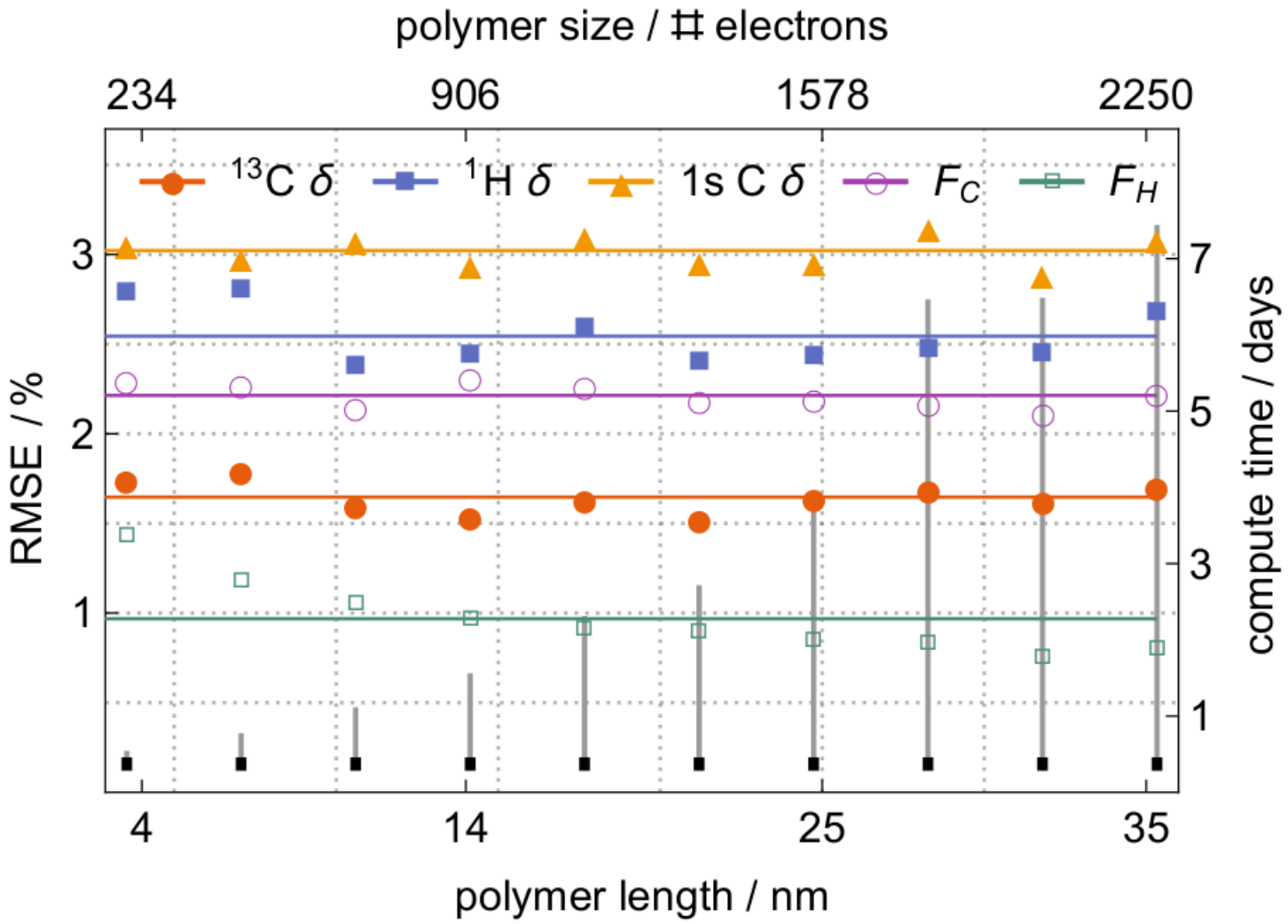
10k atoms from:
-16800 distortions from 168 $C_7H_{10}O_2$ isomers
-9k GDB molecules with 7 to 9 atoms CONF/molecule

For 10k models

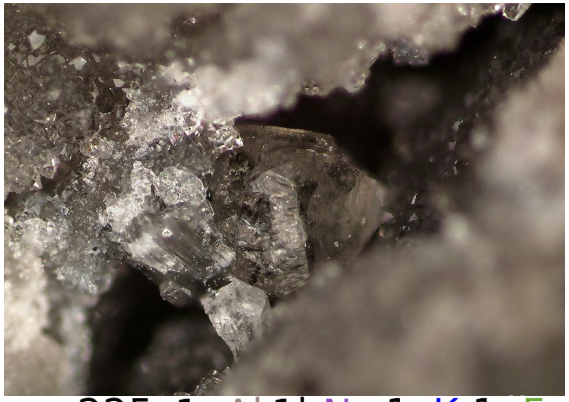


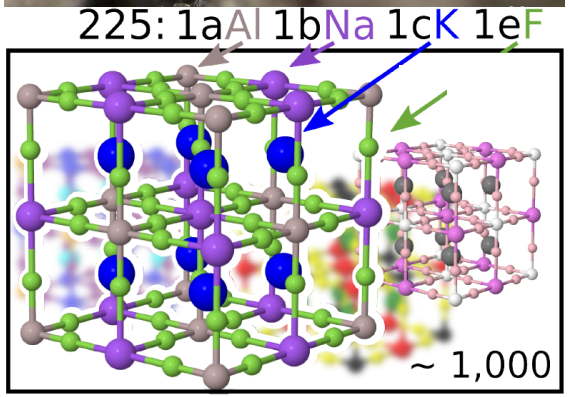
M. Rupp, R. Ramakrishnan, OAvL, *submitted* (2015), arXiv

An Atom in Many Molecules



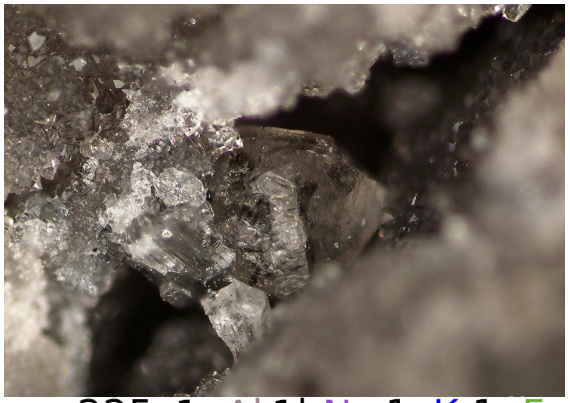
M. Rupp, R. Ramakrishnan, OAvL, *submitted* (2015), arXiv



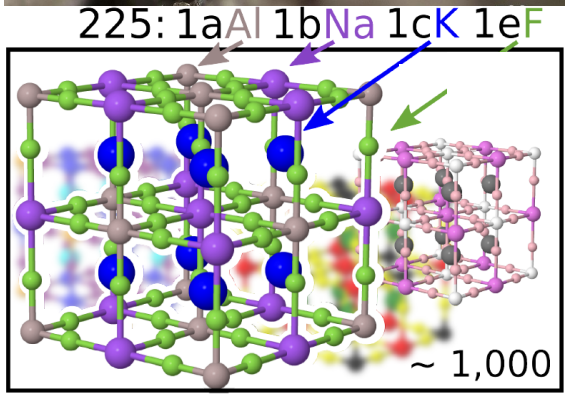


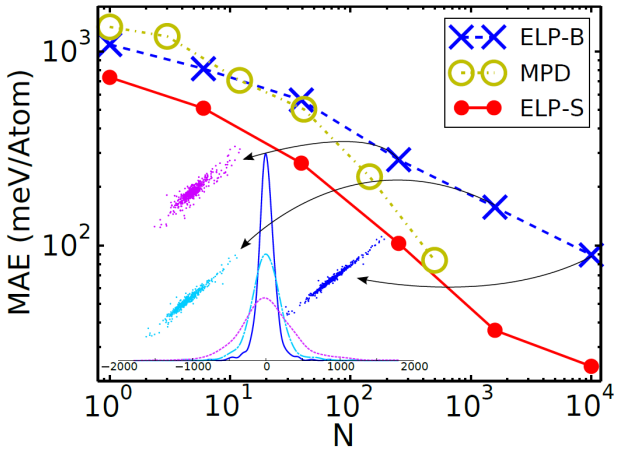
Elpasolite (K_2NaAlF_6-symmetry) is a vitreous, transparent, luster, colorless and soft quaternary crystal in the Fm3m space group which can be found in the Rocky Mountains, Virginia, or the Apennines. It is the most abundant quaternary crystal present in the Inorganic Crystal Structure Database; and some Elpasolites emit light when exposed to ionic radiation, which makes them interesting material candidates for scintillator devices.

Faber et al, *IJQC* (2015), *in preparation* (2015)

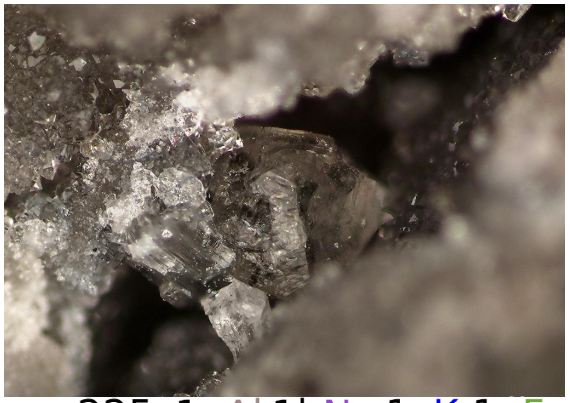


Elpasolite (K_2NaAlF_6-symmetry) is a vitreous, transparent, luster, colorless and soft quaternary crystal in the Fm3m space group which can be found in the Rocky Mountains, Virginia, or the Apennines. It is the most abundant quaternary crystal present in the Inorganic Crystal Structure Database; and some Elpasolites emit light when exposed to ionic radiation, which makes them interesting material candidates for scintillator devices.

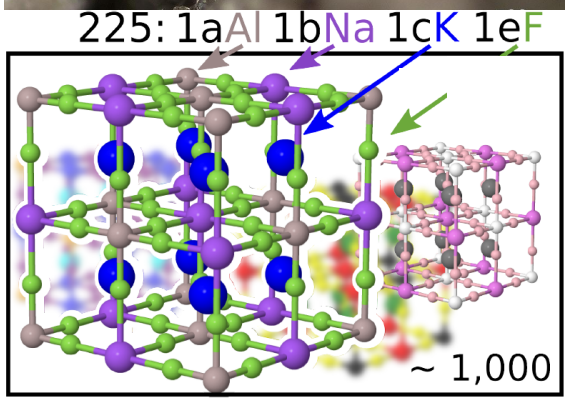


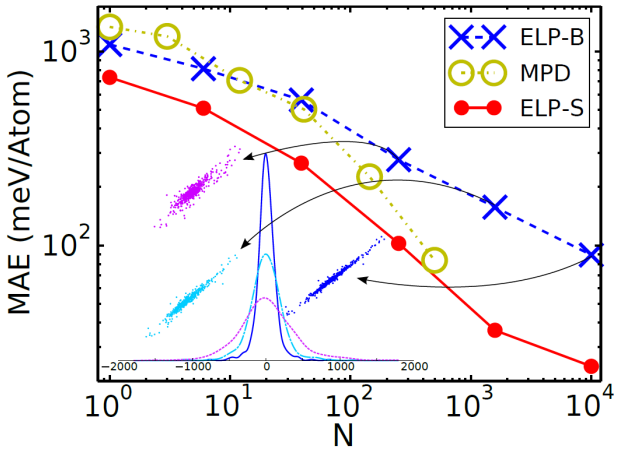


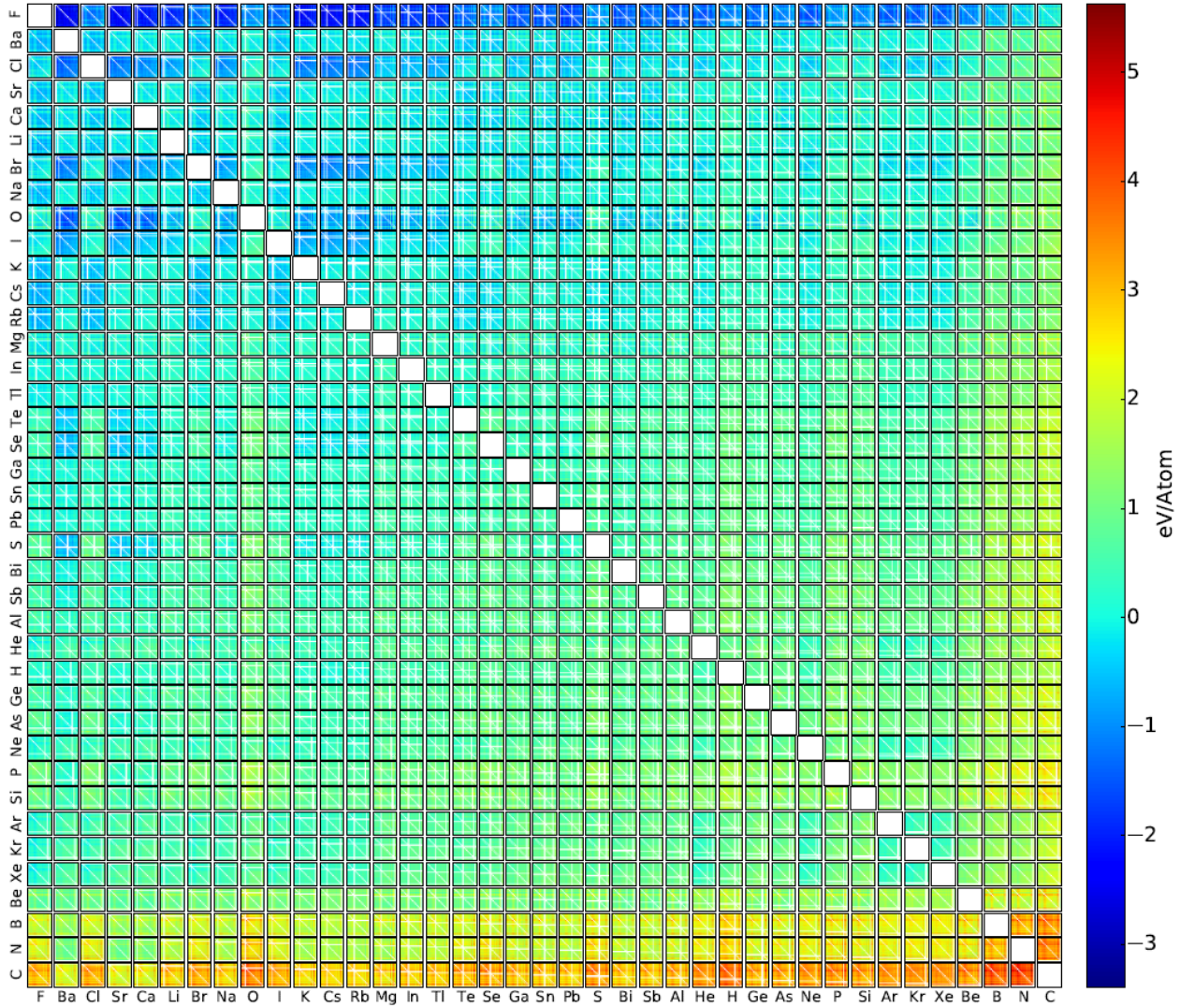
Faber et al, *IJQC* (2015), *in preparation* (2015)



Elpasolite (K_2NaAlF_6-symmetry) is a vitreous, transparent, luster, colorless and soft quaternary crystal in the Fm3m space group which can be found in the Rocky Mountains, Virginia, or the Apennines. It is the most abundant quaternary crystal present in the Inorganic Crystal Structure Database; and some Elpasolites emit light when exposed to ionic radiation, which makes them interesting material candidates for scintillator devices.



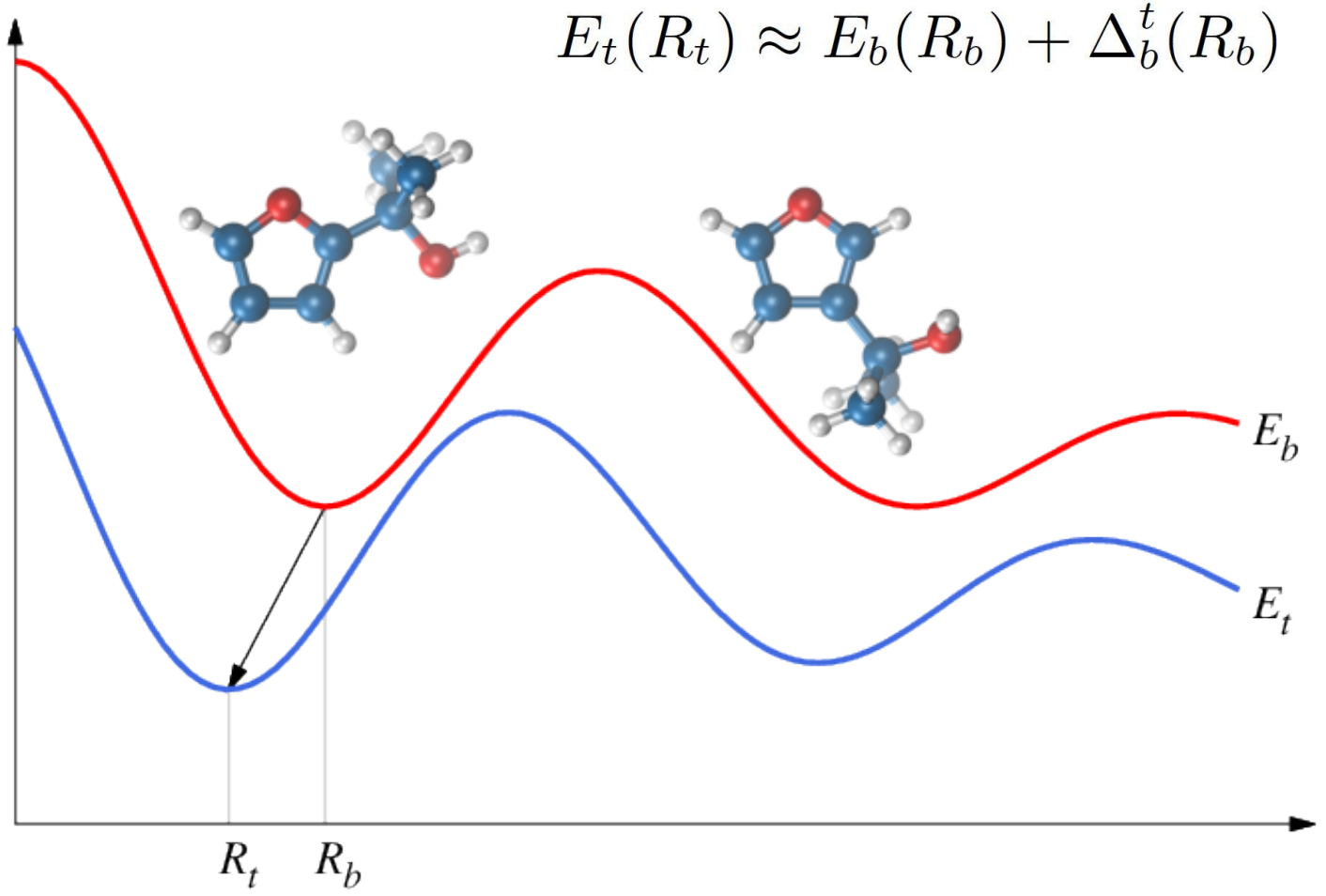




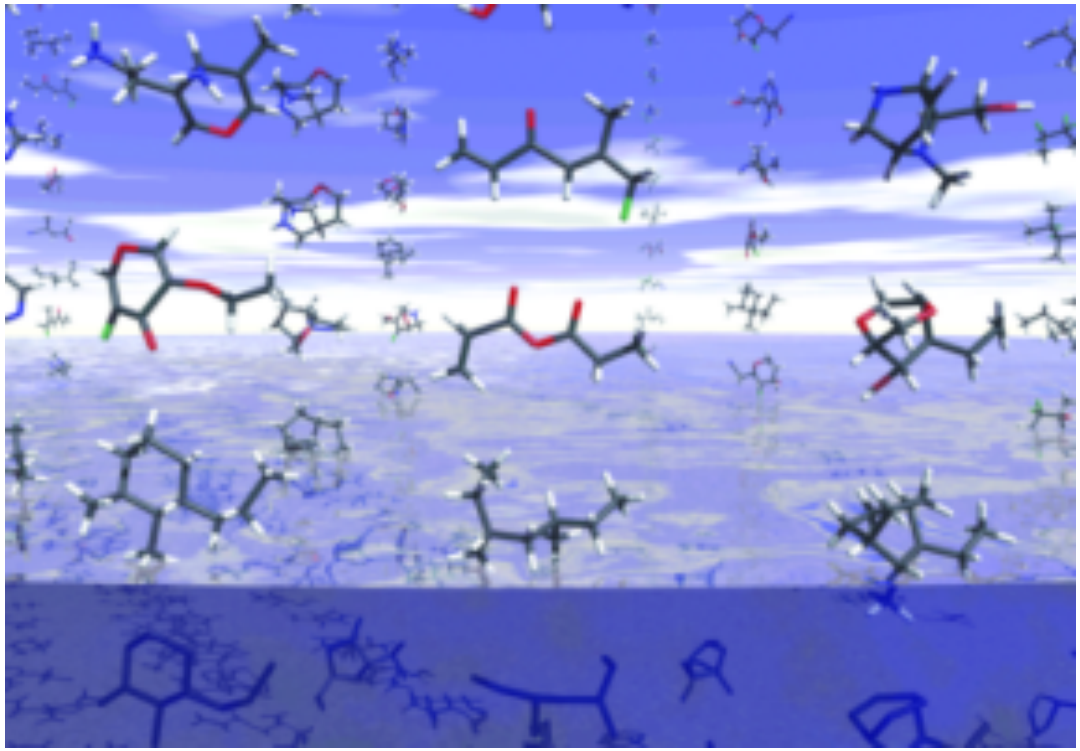
Faber et al, *IJQC* (2015), *in preparation* (2015)

Δ-Machine Learning Approach

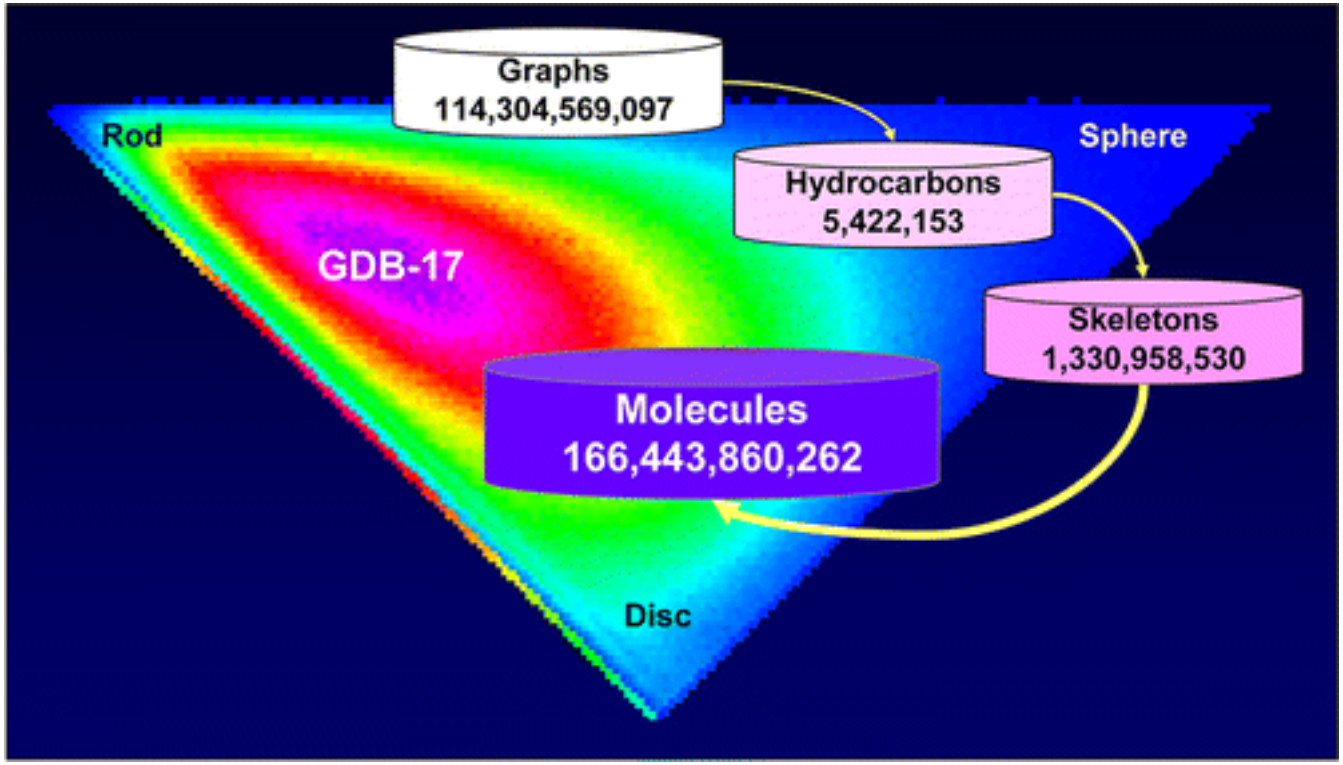
$$E_t(R_t) \approx E_b(R_b) + \Delta_b^t(R_b)$$

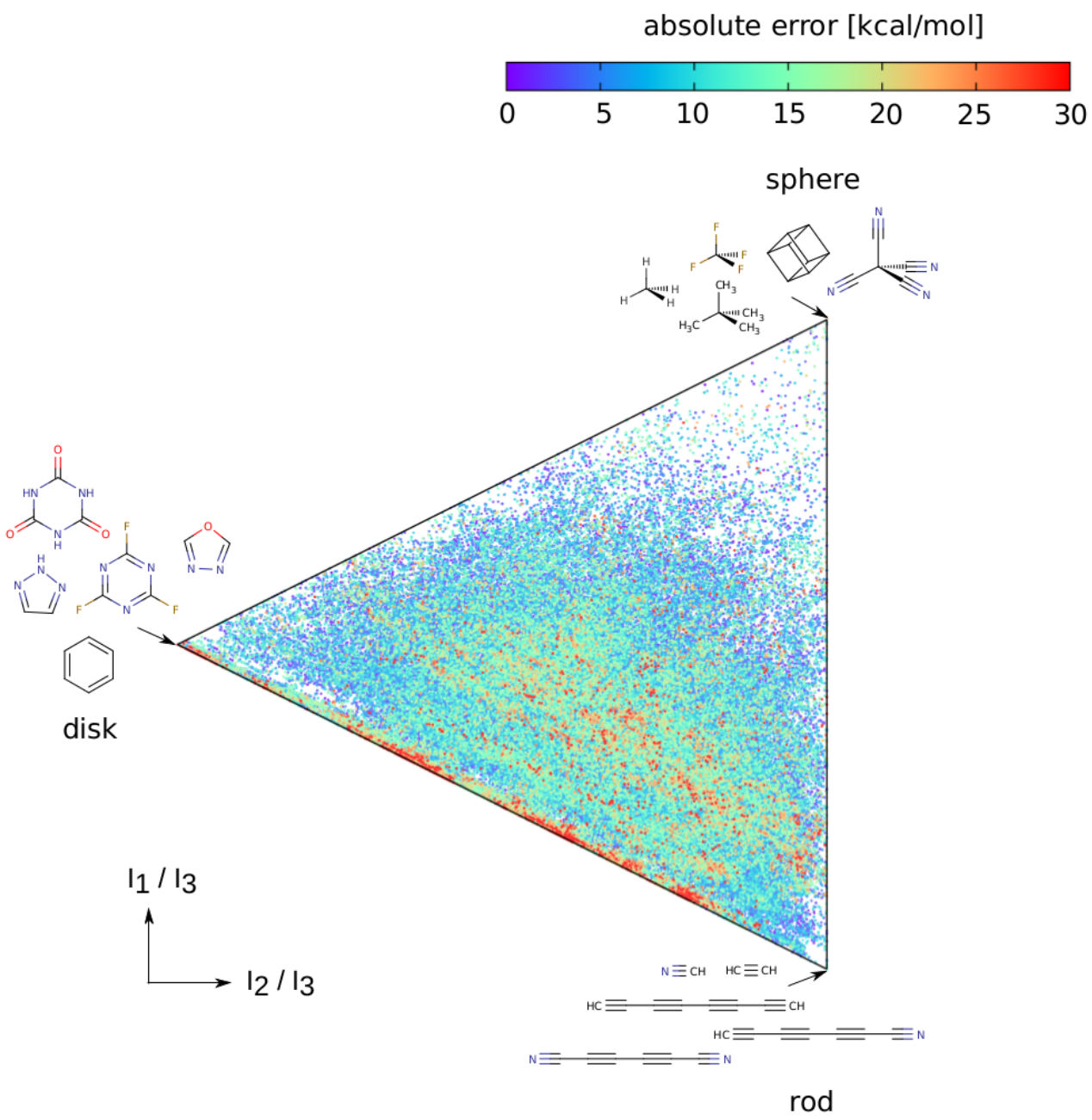


R Ramakrishnan et al *JCTC* (2015)

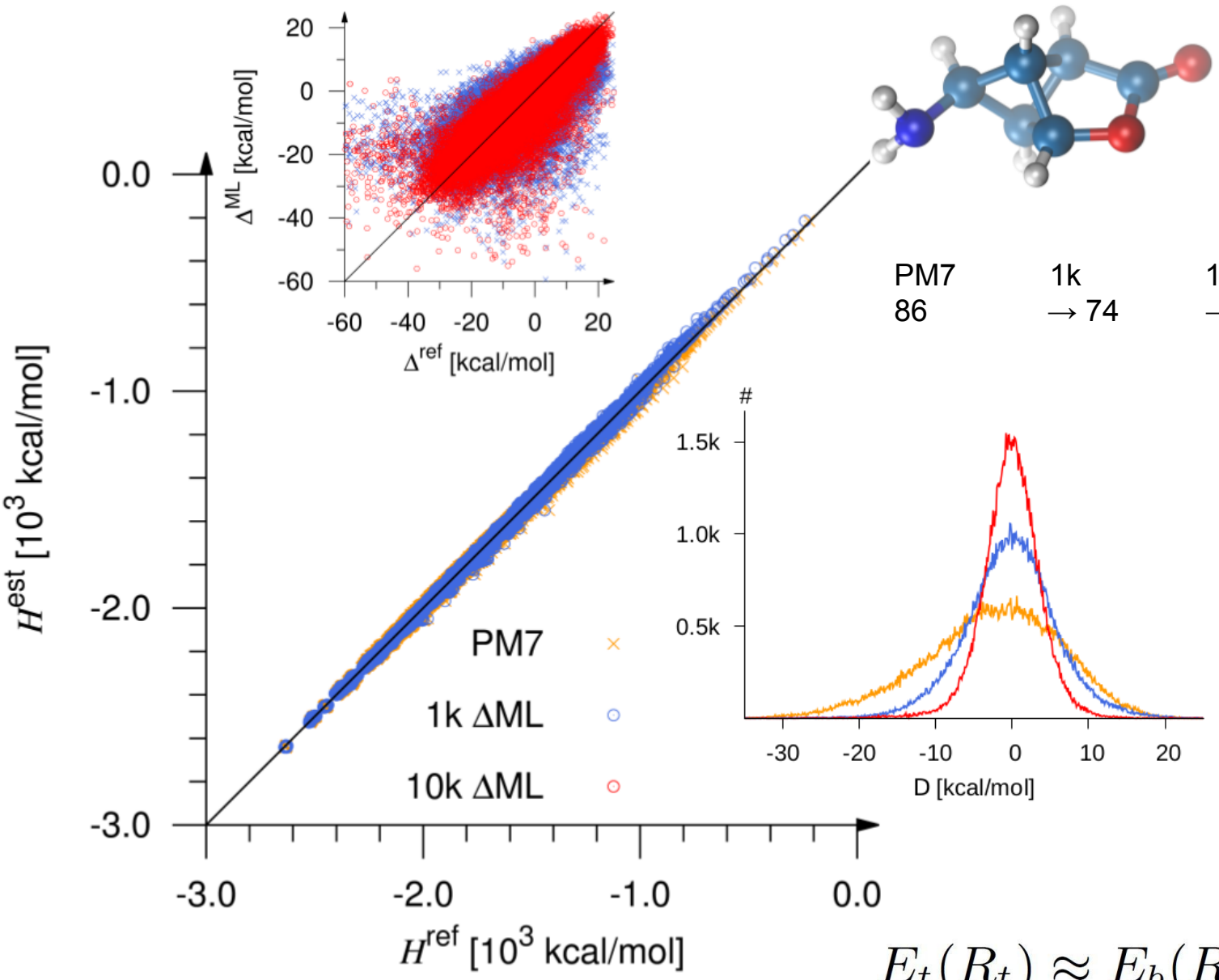


From GDB-17: All of GDB-9 → 134k molecules





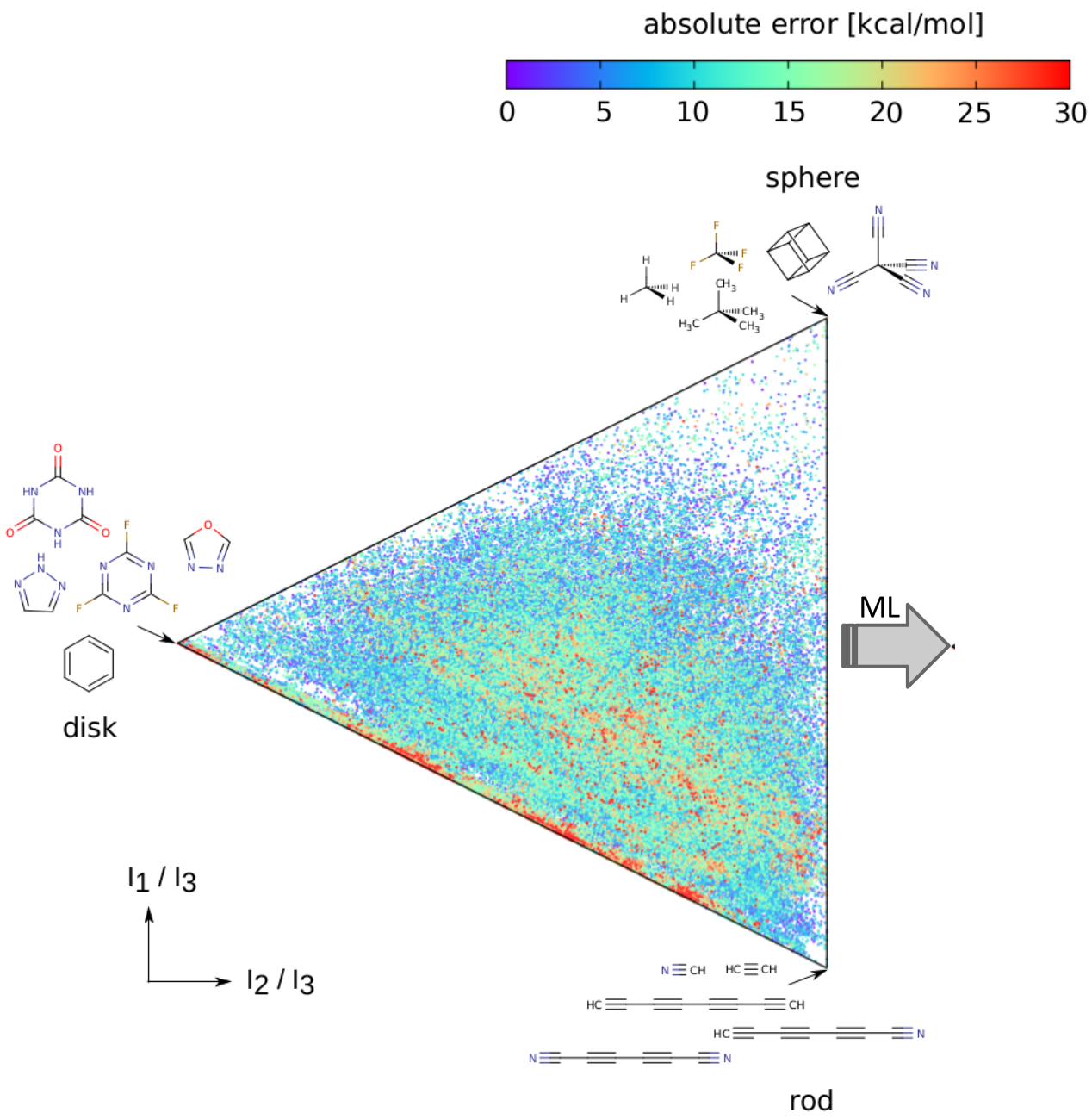
R Ramakrishnan et al *JCTC* (2015)



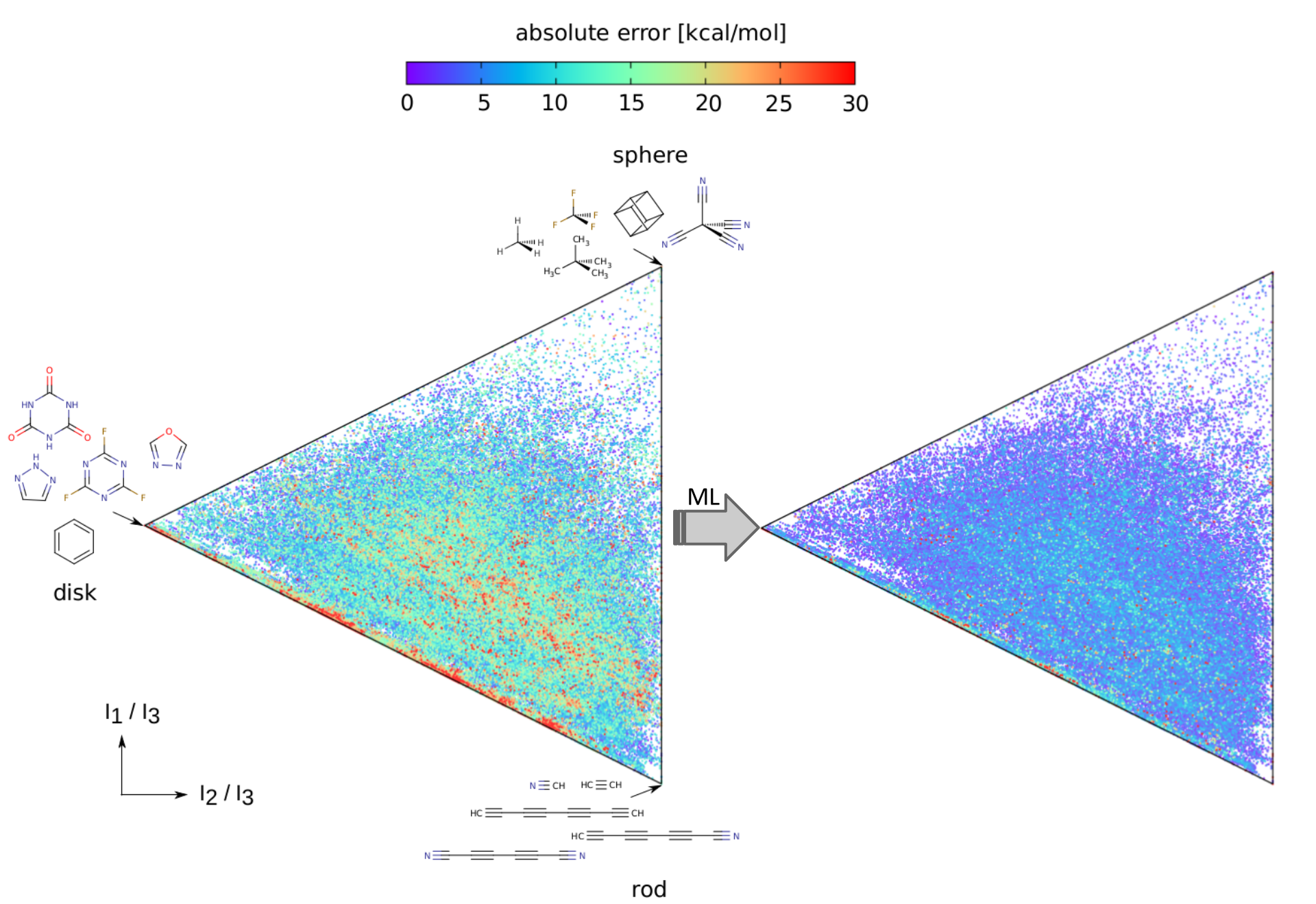
PM7 86 | 1k → 74 | 10k → 58 kcal/mol

$$E_t(R_t) \approx E_b(R_b) + \Delta_b^t(R_b)$$

R Ramakrishnan et al *JCTC* (2015)



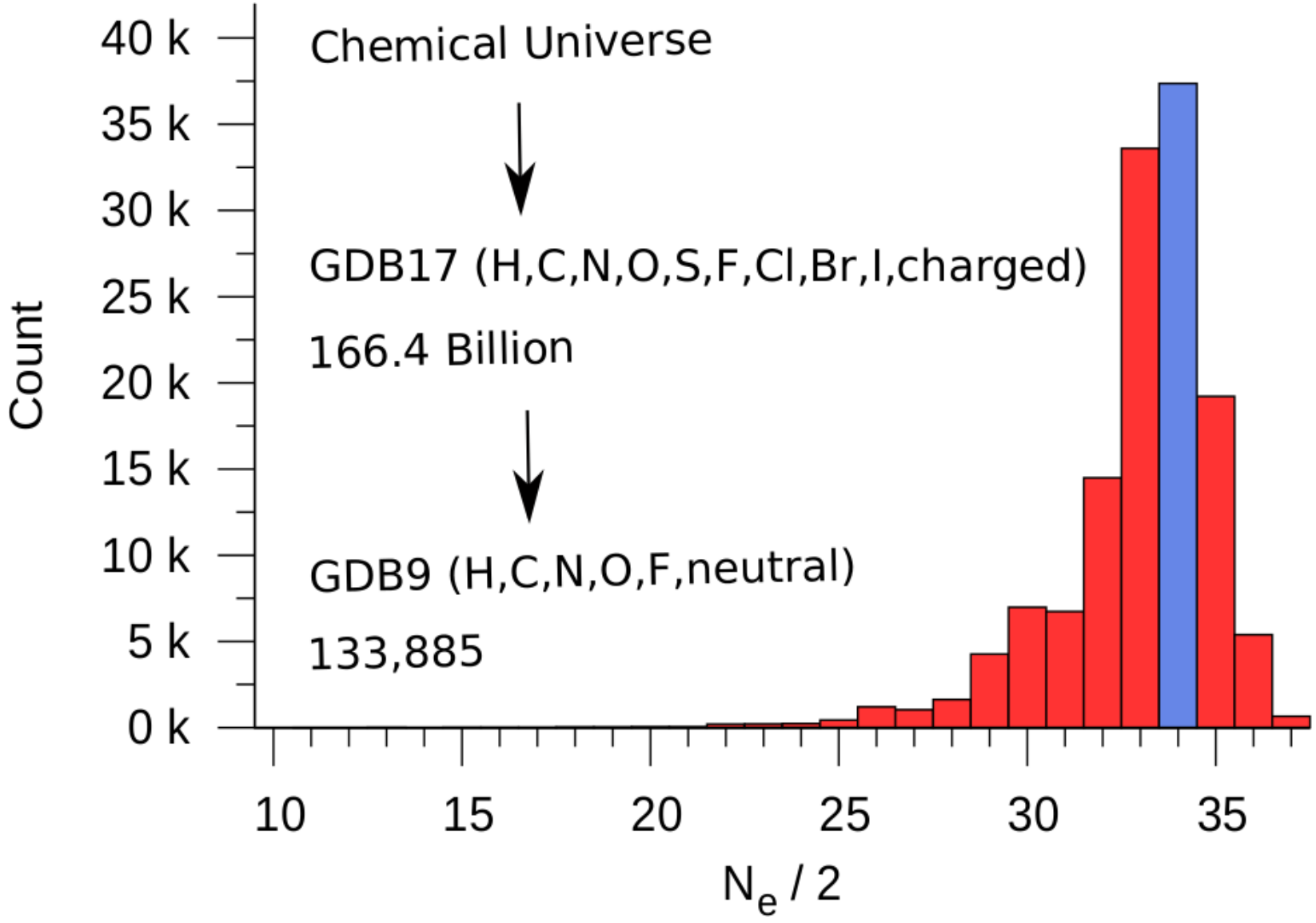
R Ramakrishnan et al *JCTC* (2015)



R Ramakrishnan et al *JCTC* (2015)

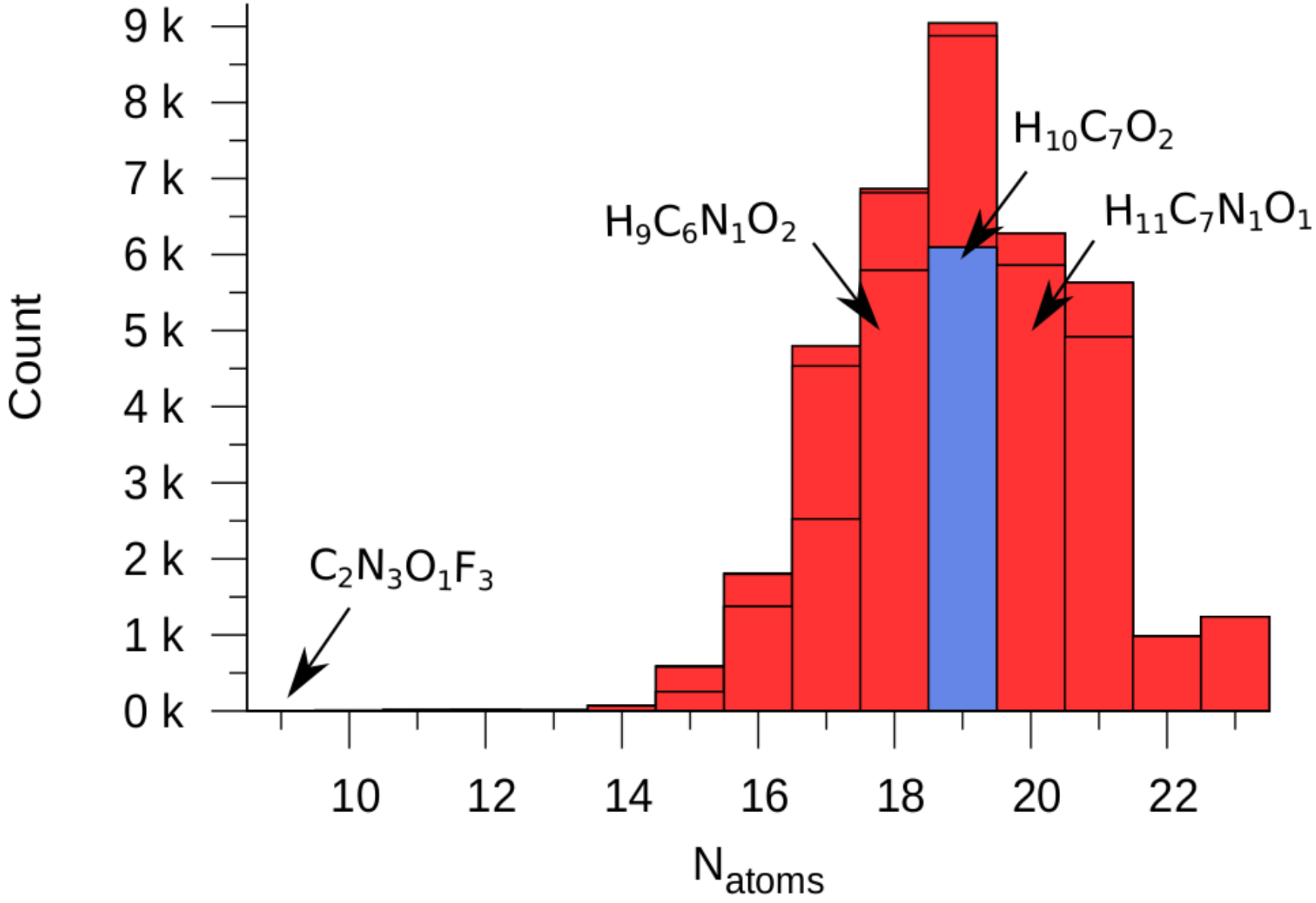
Supervised-learning: Learning energy *and* geometry?

→ **Thermochemistry**



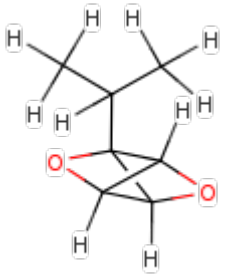
Supervised-learning: Learning energy *and* geometry?

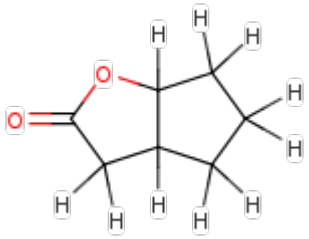
→ **Thermochemistry**

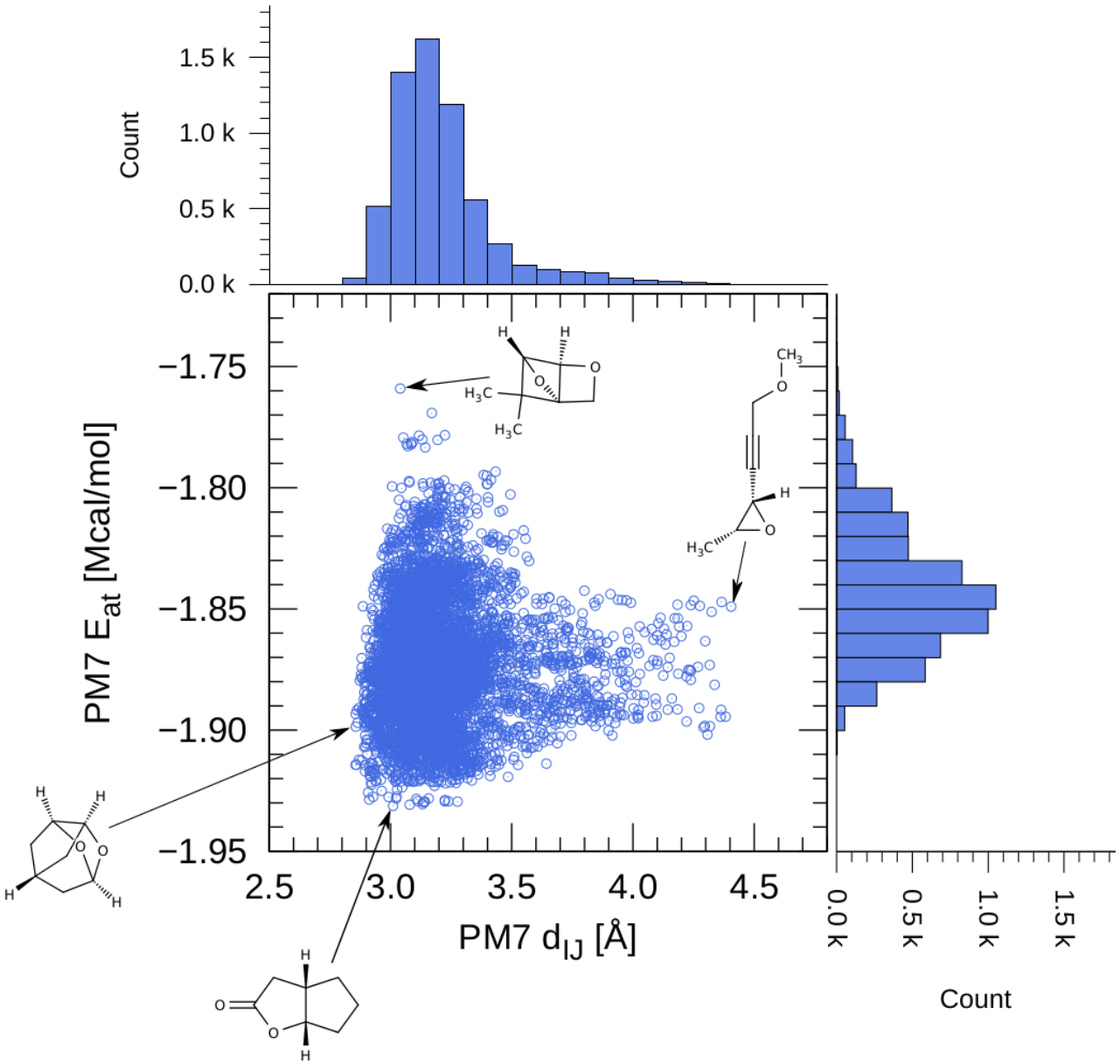


Supervised-learning: Learning energy *and* geometry?

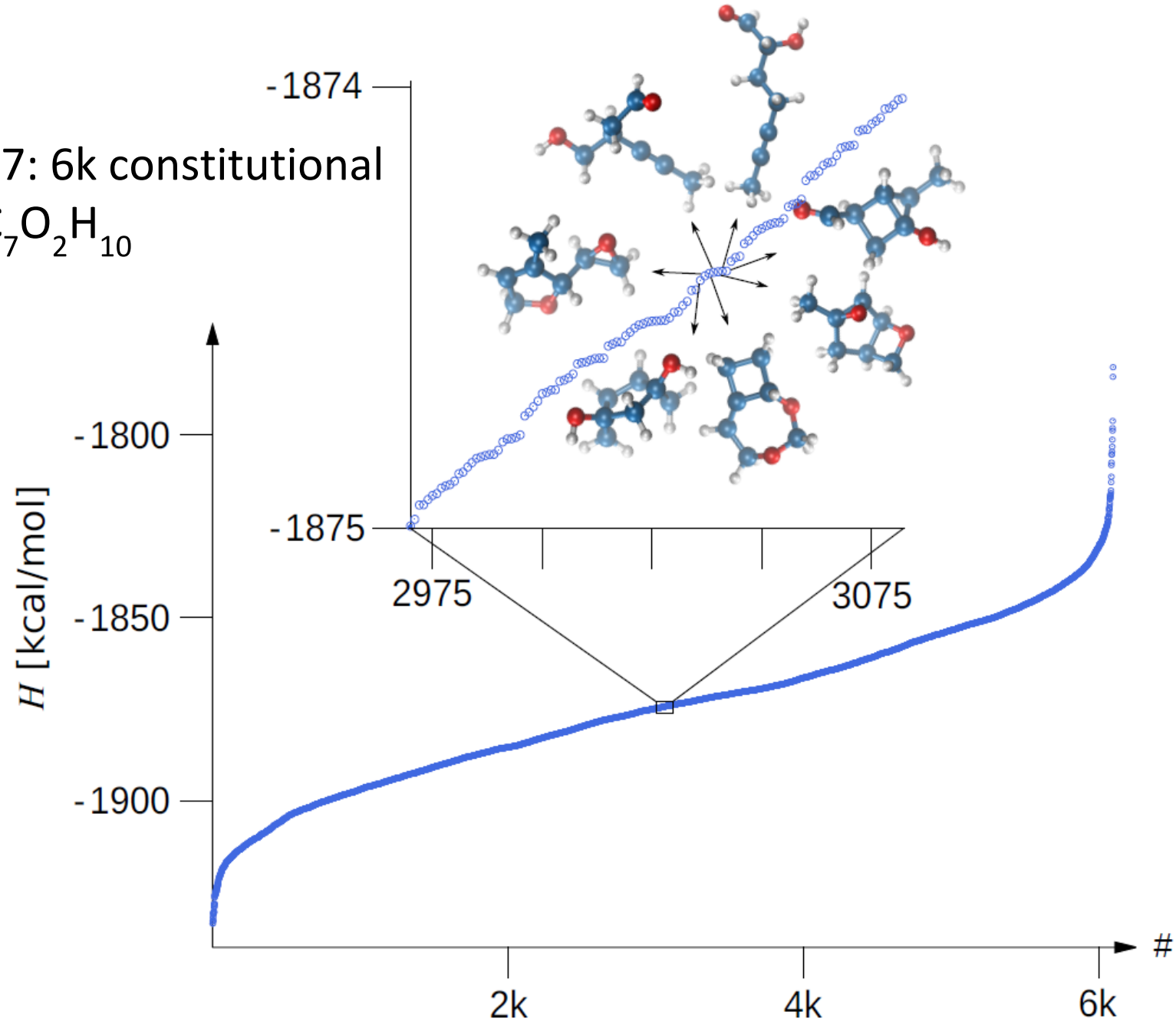
→ **Thermochemistry**







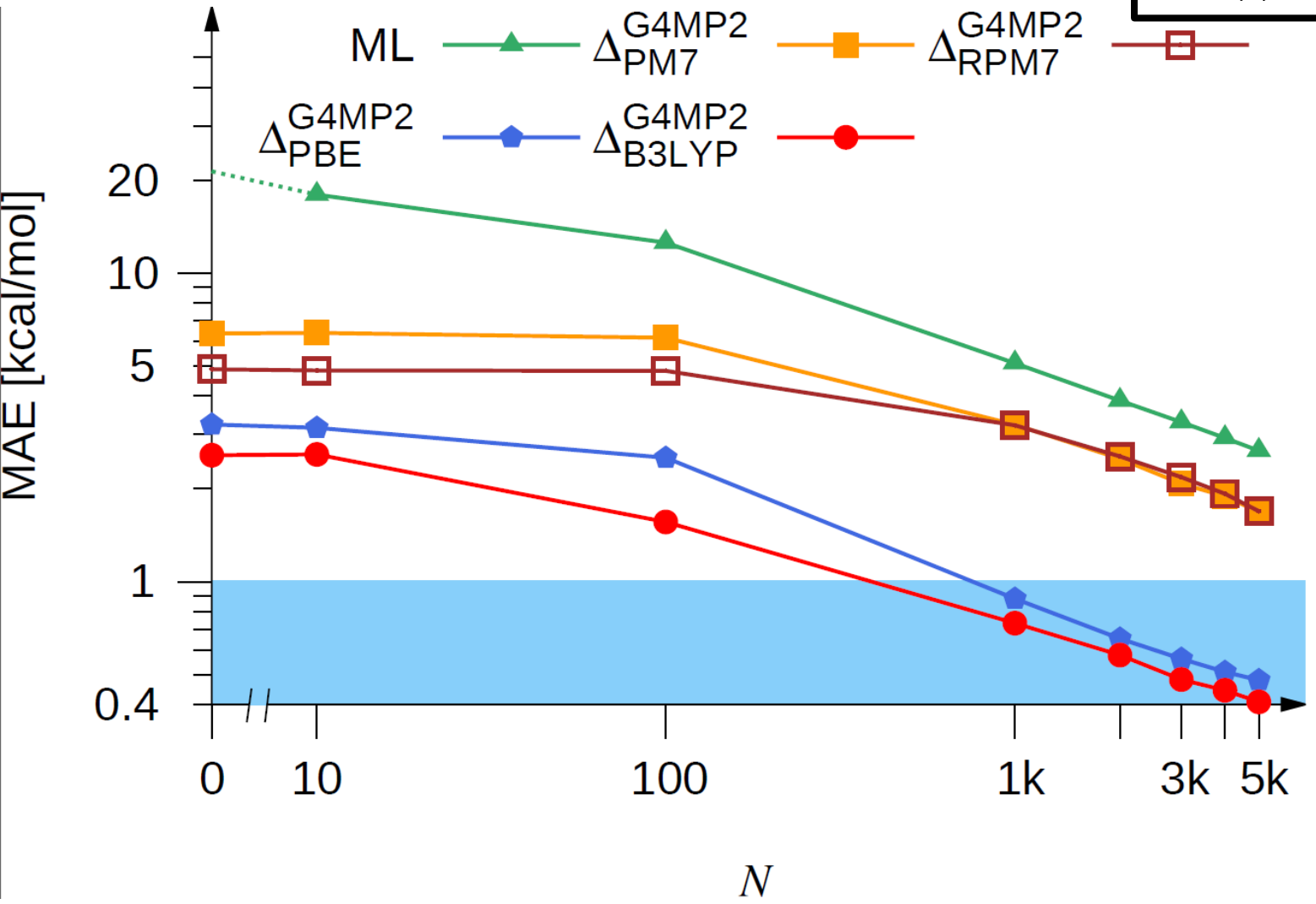
From GDB-17: 6k constitutional isomers of $C_7O_2H_{10}$



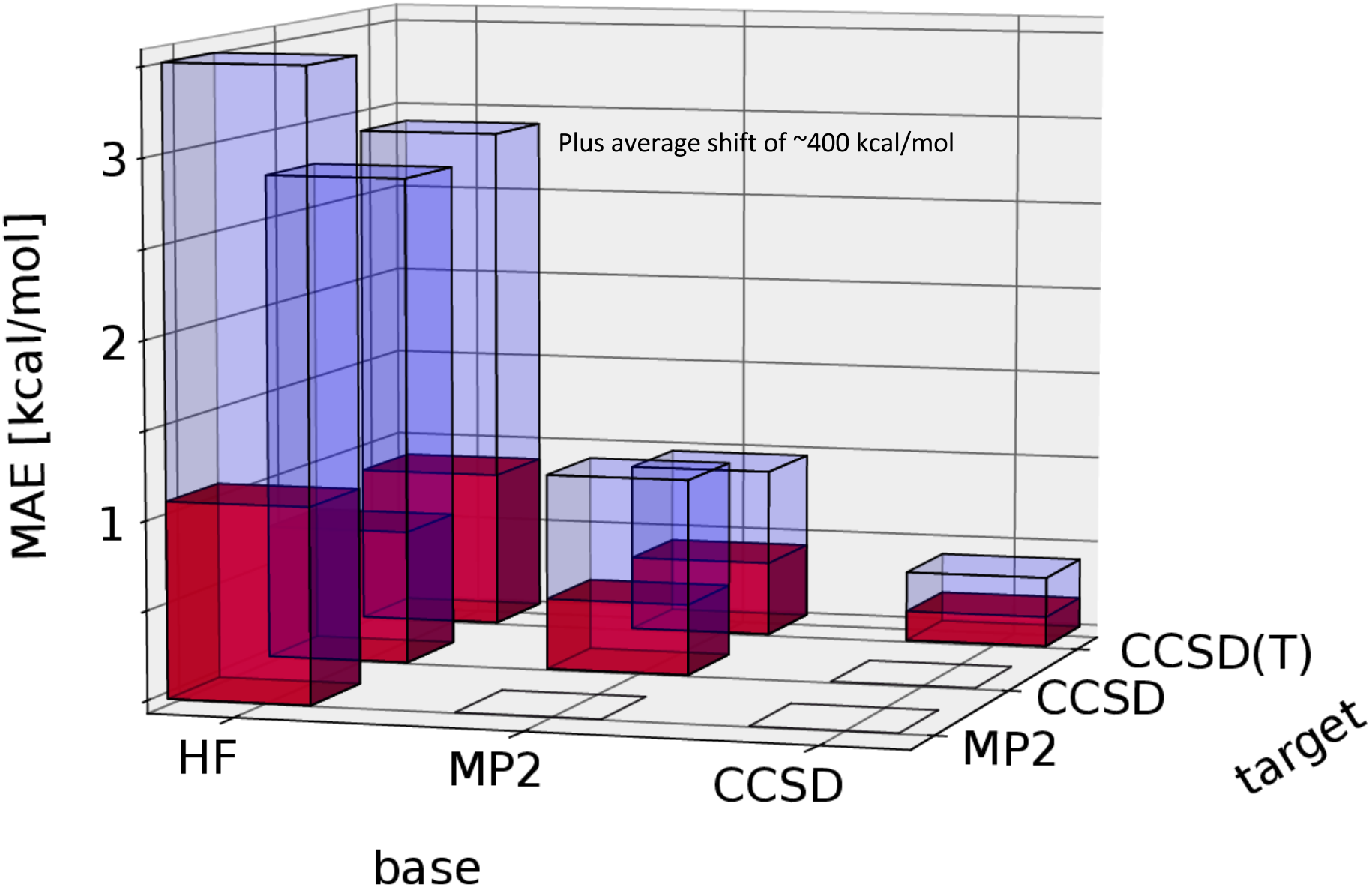
R Ramakrishnan et al *Scientific Data* (2014)

$$E_t(R_t) \approx E_b(R_b) + \Delta_b^t(R_b)$$

b			*t*
ML	(ms)	→	G4MP2 (18 h)
PM7	(s)	→	G4MP2 (18 h)
PBE	(mins)	→	G4MP2 (18 h)
B3LYP	(h)	→	G4MP2 (18 h)



R Ramakrishnan et al *JCTC* (2015)

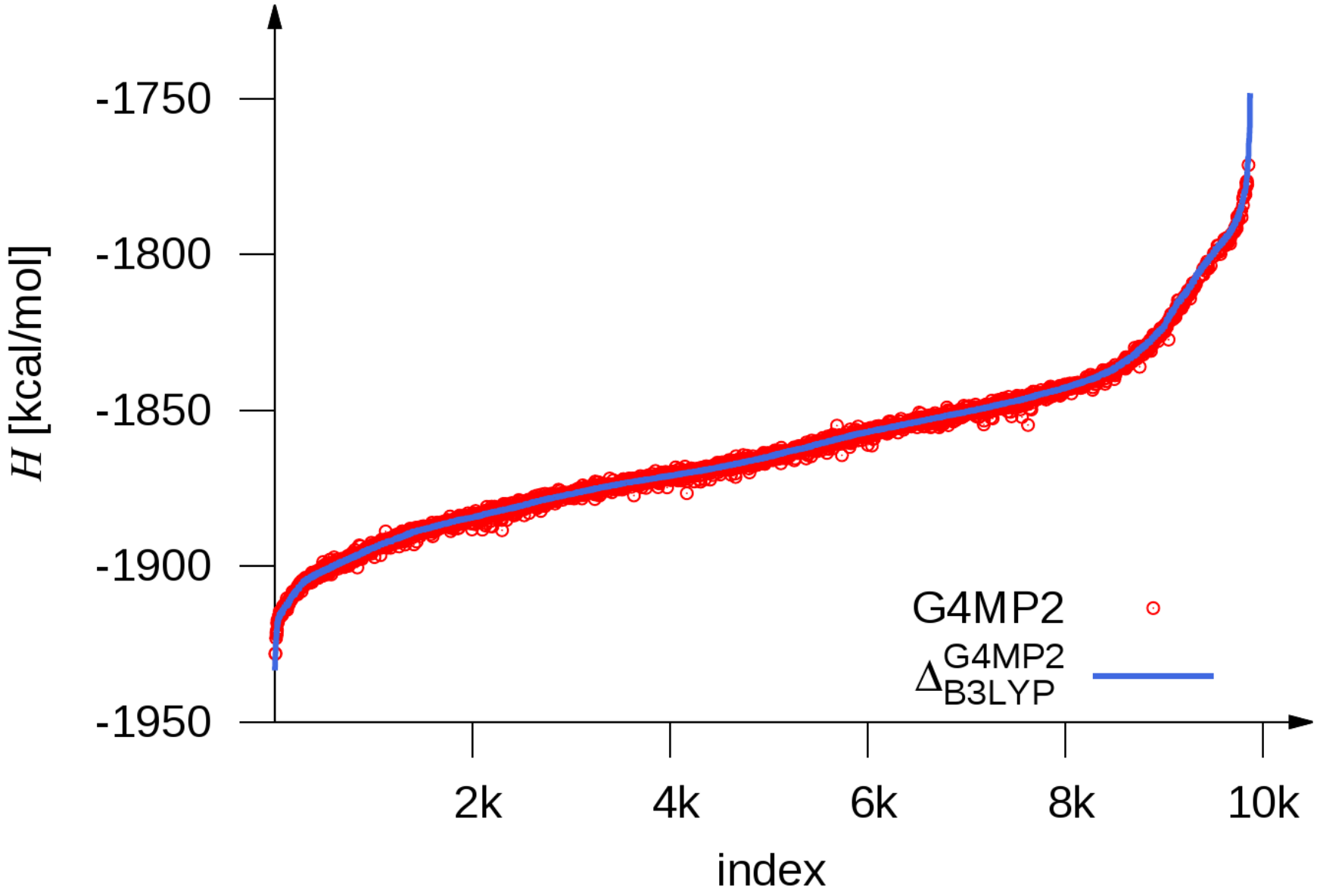


R Ramakrishnan et al *JCTC* (2015)

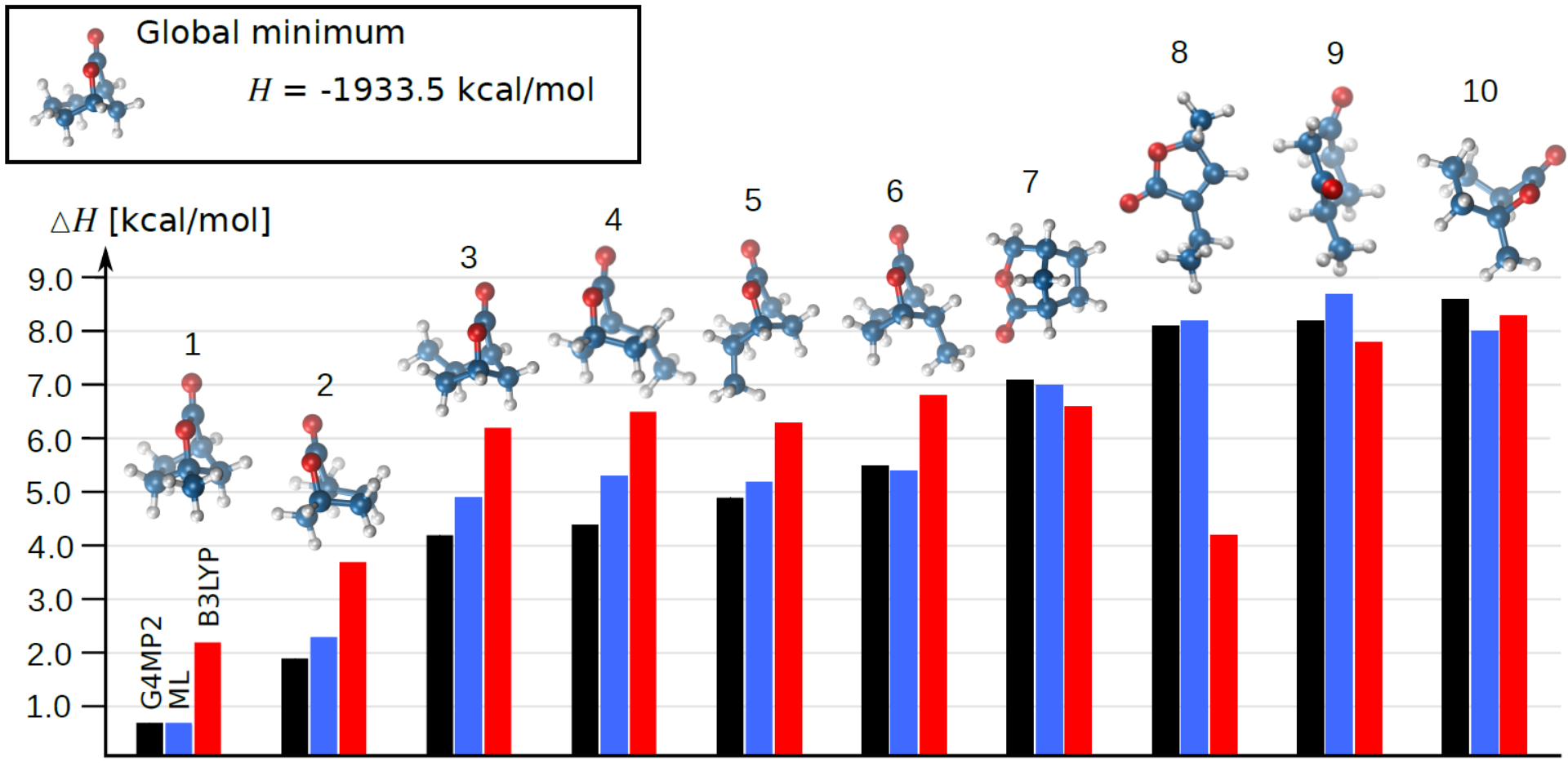
Invert chiral atoms → 10k diastereomers

R Ramakrishnan et al *JCTC* (2015)

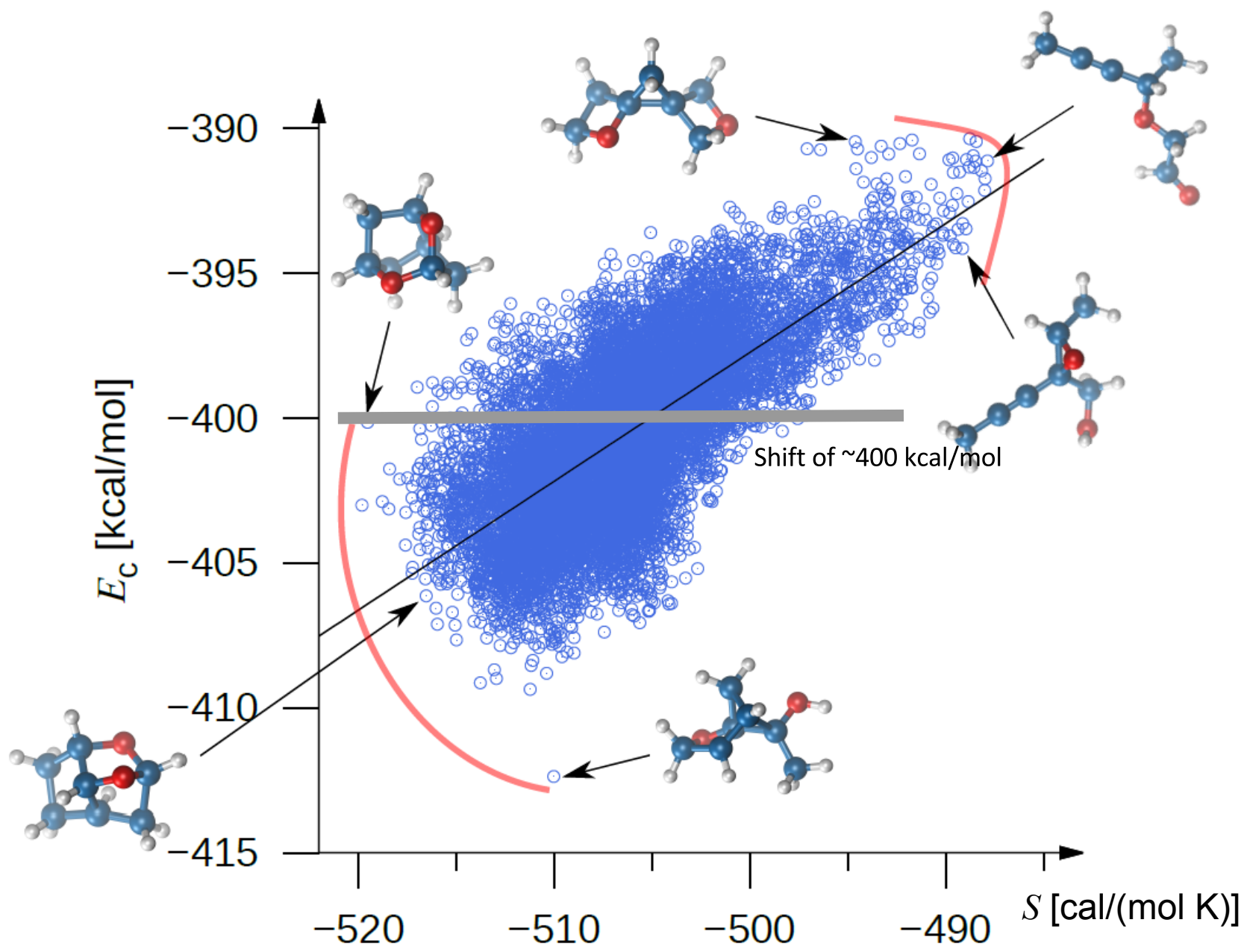
Invert chiral atoms → 10k diastereomers



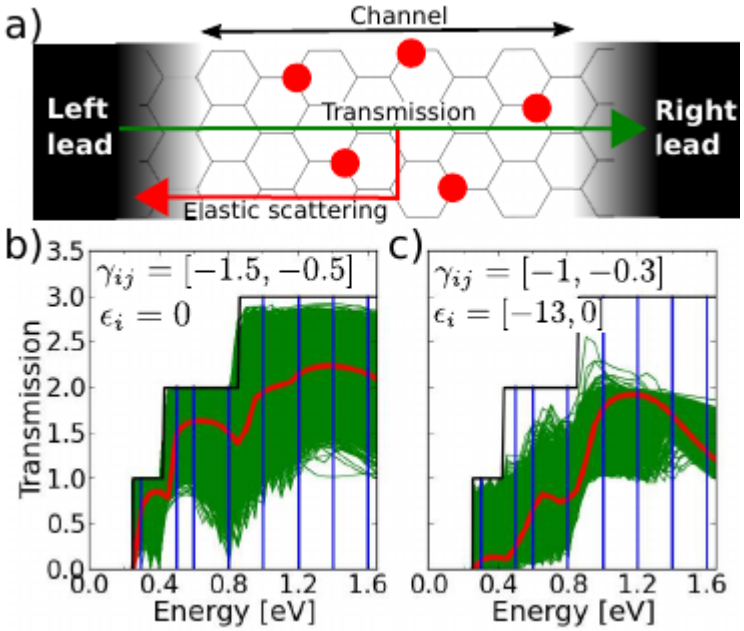
R Ramakrishnan et al *JCTC* (2015)

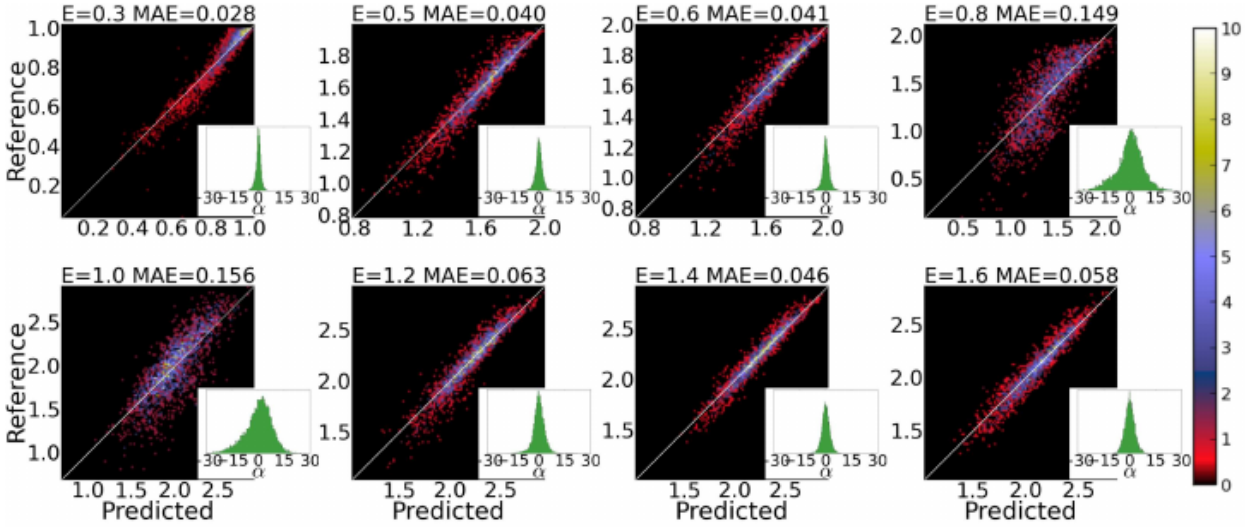


R Ramakrishnan et al *JCTC* (2015)

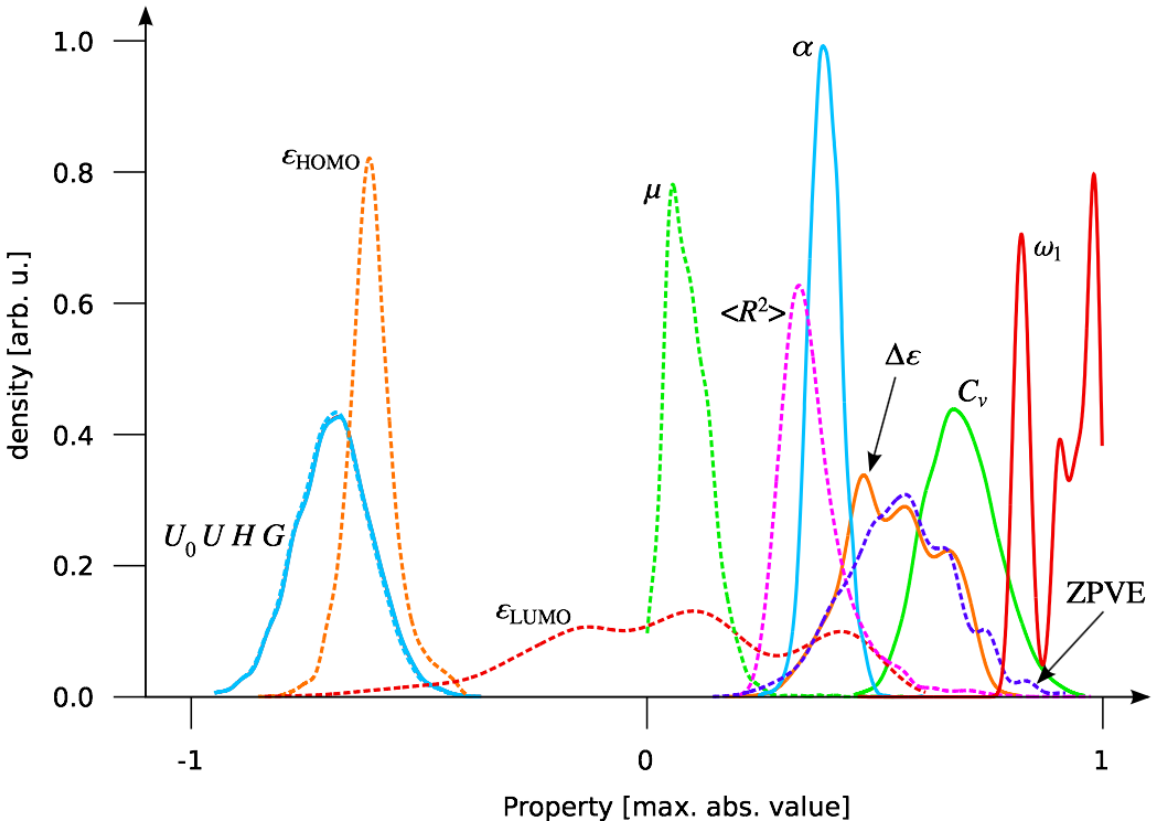


R Ramakrishnan et al *JCTC* (2015)





with A. Lopez-Bezanillal (ANL), accepted in *PRB* (2014), arxiv



Various molecular properties of 134k-*N* organic molecules taken from: Ramakrishnan et al, *Scientific Data* (2014)

Ramakrishnan, OAvL, *CHIMIA* (2015), arXiv

$$p_q = \sum_{t=1}^{N} c_t^p K_{qt}$$

$$\mathbf{c}^p = \left(\mathbf{K} + \lambda \mathbf{I}\right)^{-1} \mathbf{p}^r$$

We set $\lambda = 0$ ...

$$\mathcal{L} = \left(\mathbf{p}^r - \mathbf{K}\mathbf{c}^p\right)^{\mathrm{T}} \left(\mathbf{p}^r - \mathbf{K}\mathbf{c}^p\right) + \lambda \mathbf{c}^{p\,\mathrm{T}} \mathbf{K} \mathbf{c}^p$$

$$\left[\mathbf{c}^{p_1} \mathbf{c}^{p_2} \dots \mathbf{c}^{p_n}\right] = \mathbf{K}^{-1} \left[\mathbf{p}_1^r \mathbf{p}_2^r \dots \mathbf{p}_n^r\right] \quad \Rightarrow \quad \mathbf{C} = \mathbf{K}^{-1} \mathbf{P}^r$$

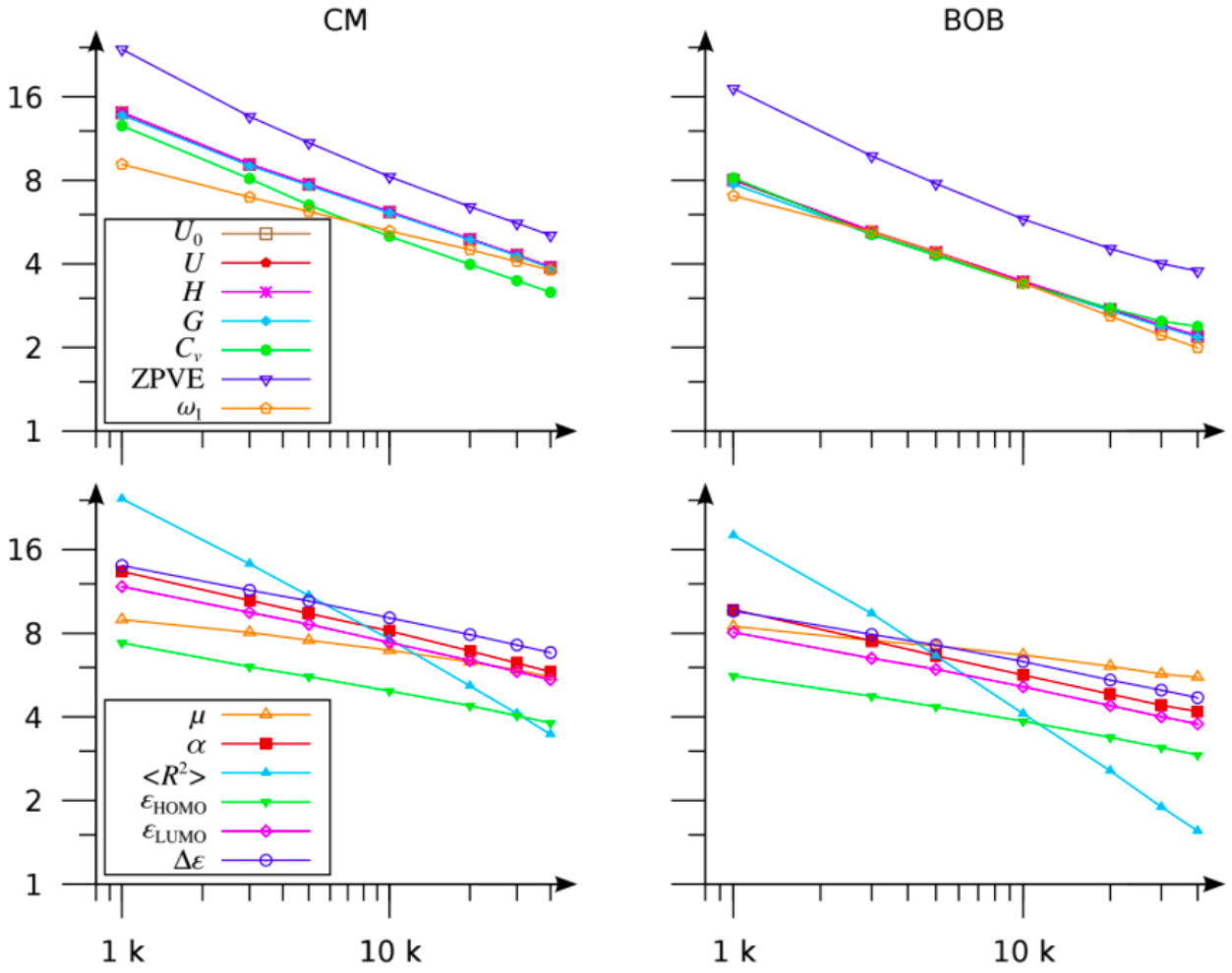
Ramakrishnan, OAvL, *CHIMIA* (2015), arXiv

$$k_{ij} = e^{-D_{ij}/\sigma}$$

$$\frac{1}{2} \leq k_{ij} \leq 1$$

$$\sigma_{\mathrm{opt}} = D_{ij}^{\mathrm{max}} / \log(2)$$

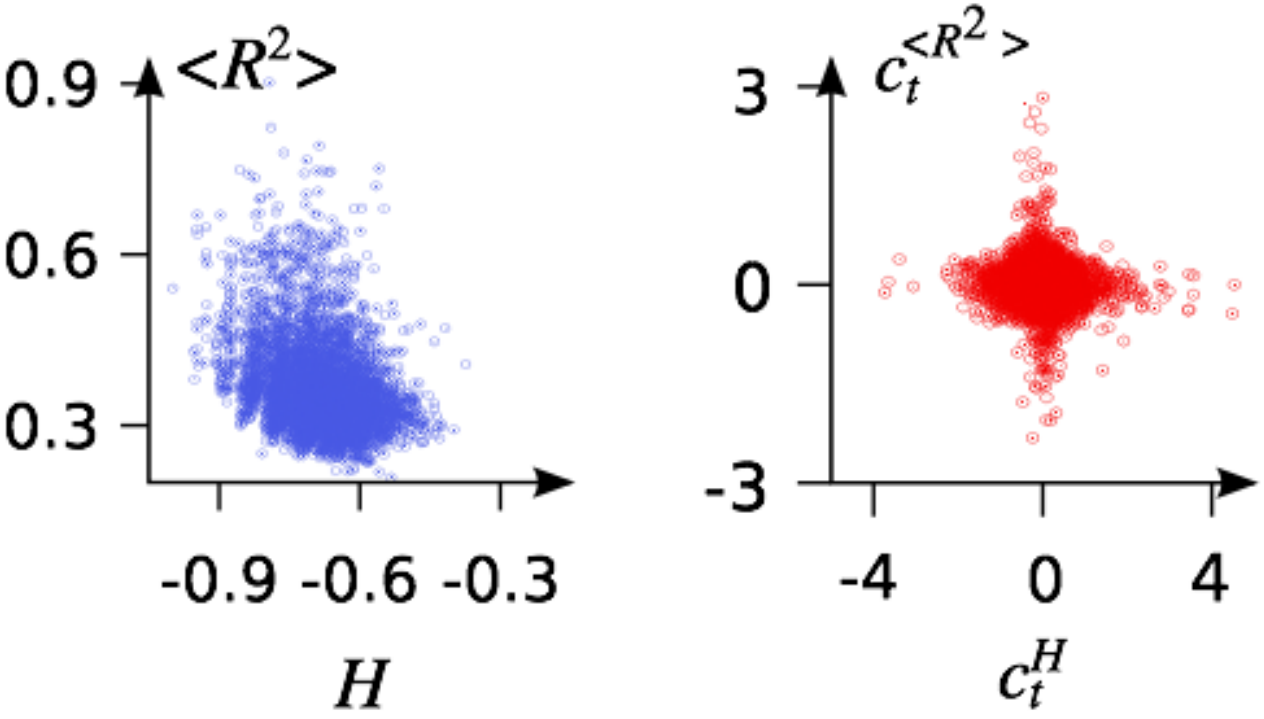
Ramakrishnan, OAvL, *CHIMIA* (2015), arXiv



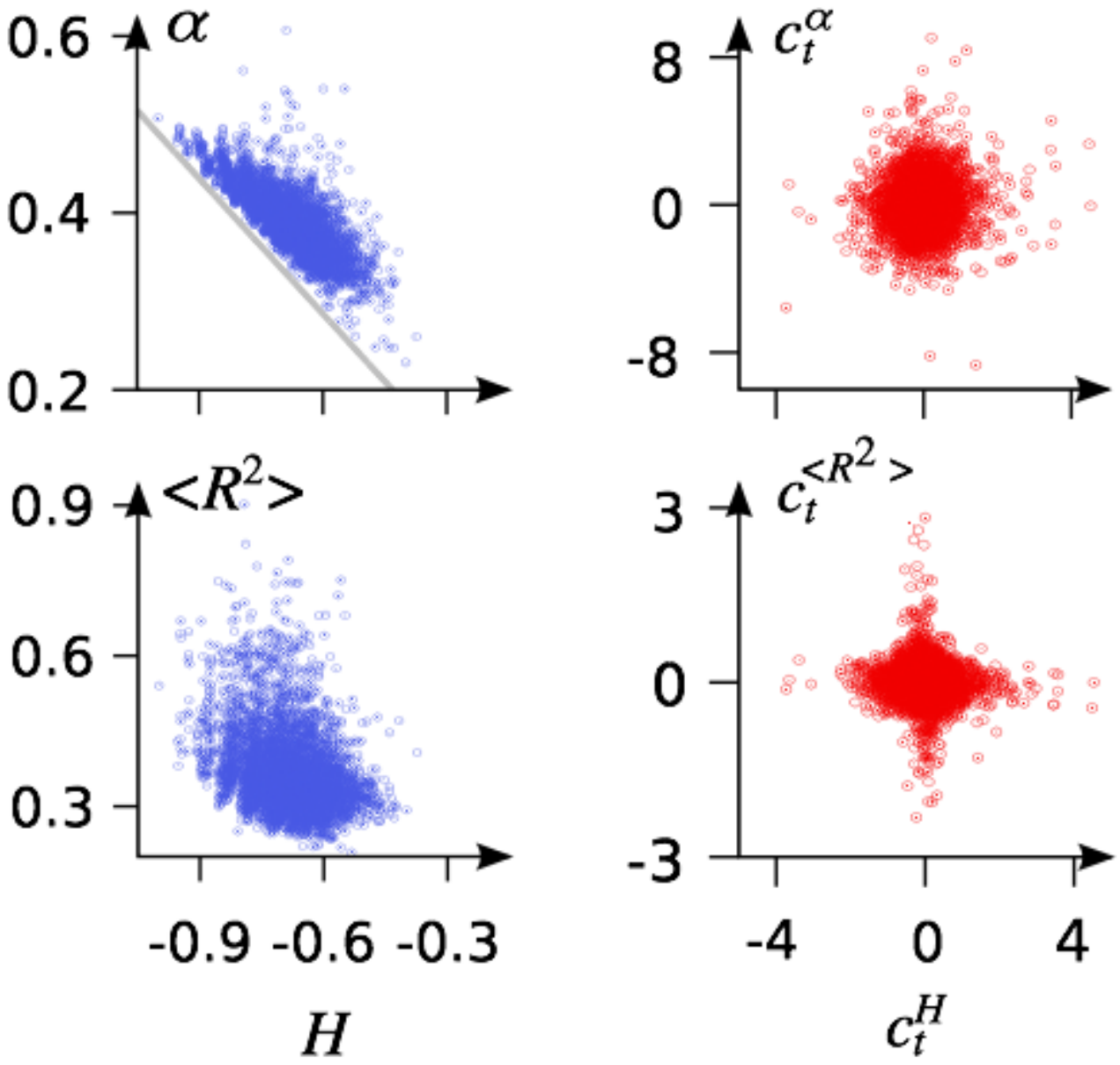
Tested on 134k-*N* organic molecules taken from:
Ramakrishnan et al, *Scientific Data* (2014)

*BOB, Hansen et al, *submitted* (2015)

Ramakrishnan, OAvL, *CHIMIA* (2015), arXiv



Ramakrishnan, OAvL, *CHIMIA* (2015), arXiv



Ramakrishnan, OAvL, *CHIMIA* (2015), arXiv

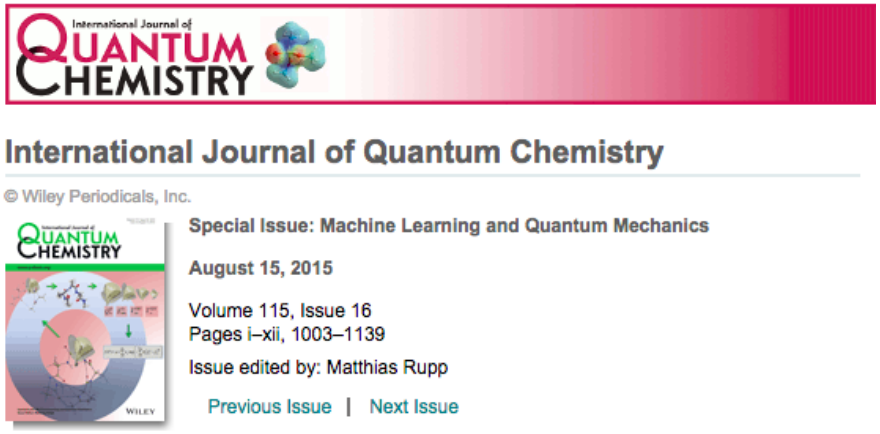
Concluding remarks

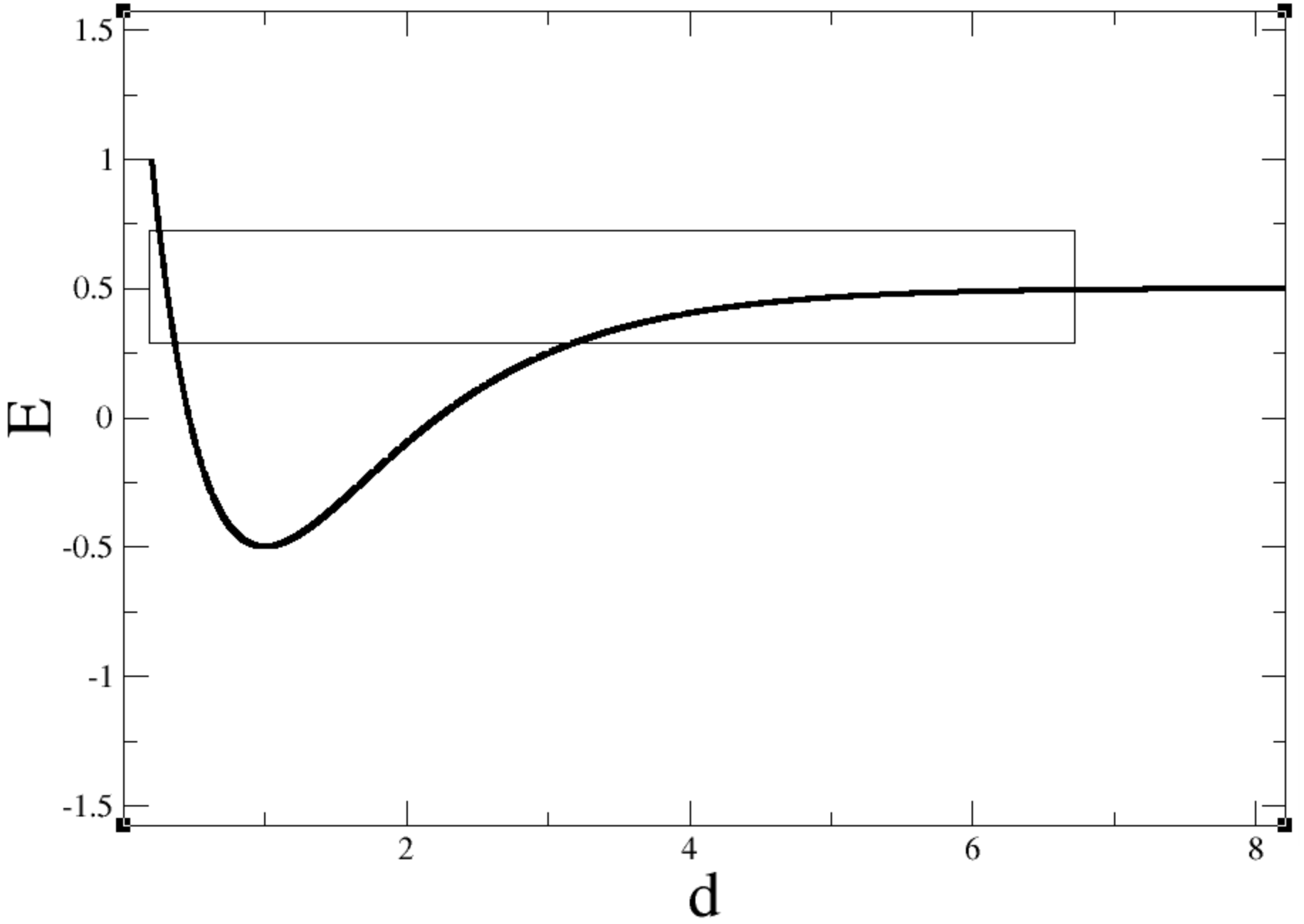
1. ML accuracy depends on
   a. better descriptors
   b. datasets
   c. baseline

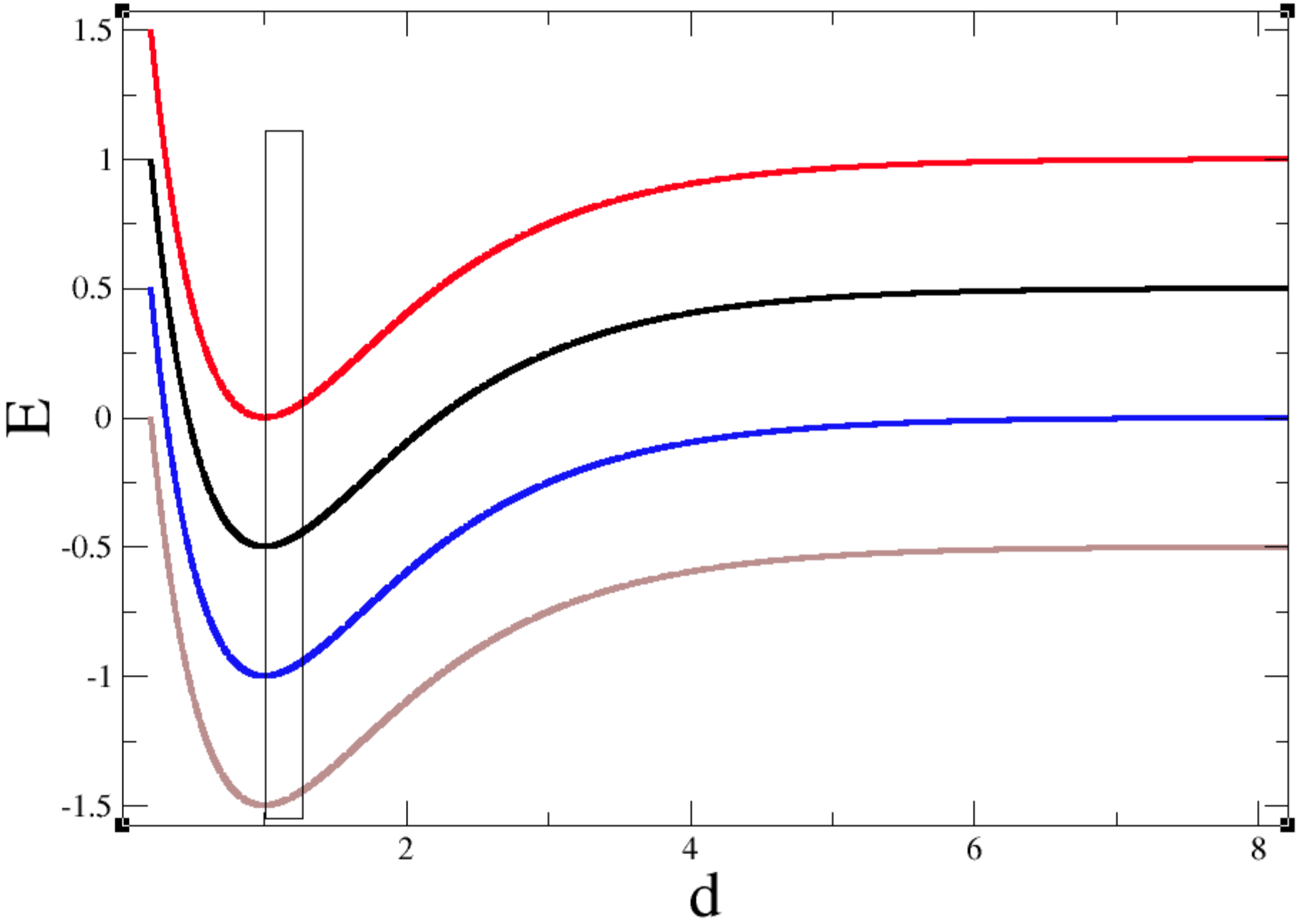
2. Systematic improvement (The bigger the better)

3. Milli-second predictions

4. $\mathbf{K} \sim \Psi$, $\boldsymbol{\alpha} \sim \hat{O}$




1.5
1
0.5
0
-0.5
-1
-1.5
E
2
4
6
8
d


1.5
1
0.5
0
-0.5
-1
-1.5
E
2
4
6
8
d