Quantum Machine Learning

Learning curves, representations, training sets

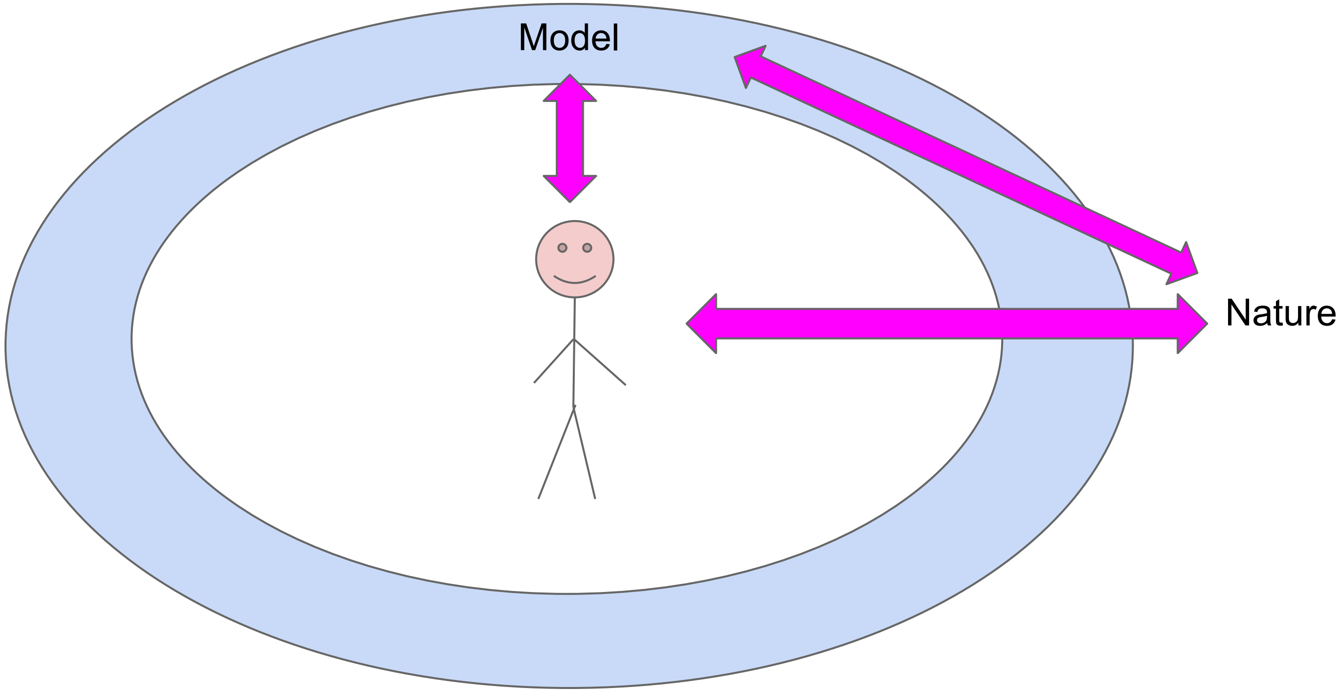
Anatole von Lilienfeld

"Machine Learning, Quantum Mechanics, and Chemical Compound Space"
By Ramakrishnan and von Lilienfeld
published in: *Reviews in Computational Chemistry*
edited by Abby L. Parrill and Kenny B. Lipkowitz
Volume **30**, Chapter 5, pages 225-256 (2017)


UNI BASEL
Research at Google
U.S. DEPARTMENT OF ENERGY
INCITE
LEADERSHIP COMPUTING
MARVEL
NATIONAL CENTRE OF COMPETENCE IN RESEARCH
Institute for Pure & Applied Mathematics
IPAM
University of California, Los Angeles
FNSNF
FONDS NATIONAL SUISSE
SCHWEIZERISCHER NATIONALFONDS
FONDO NAZIONALE SVIZZERO
SWISS NATIONAL SCIENCE FOUNDATION
AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
UNITED STATES AIR FORCE

What’s more important?

The right skill set or the right mind set?

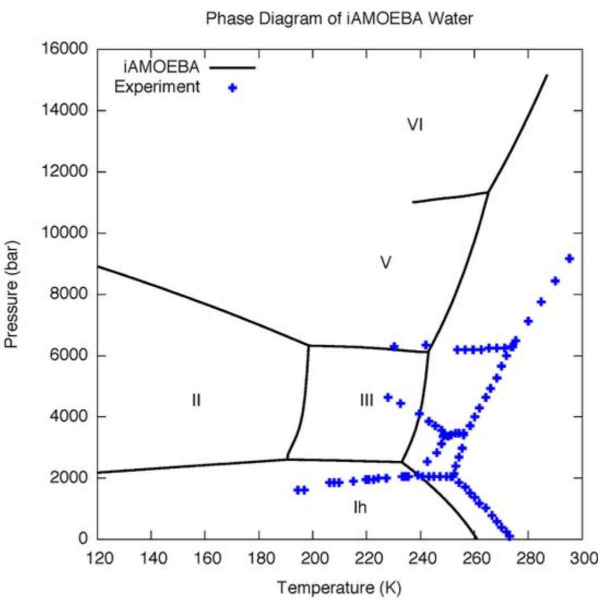


How do we do physics?

1. Guess a law
2. Build a model
3. Predict an outcome
4. If it does not compare to experiment it’s wrong

Richard

Why do we do (chemical) physics?

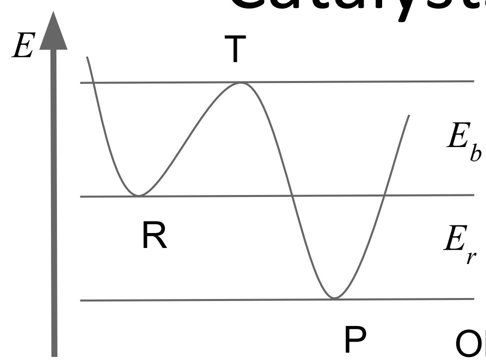


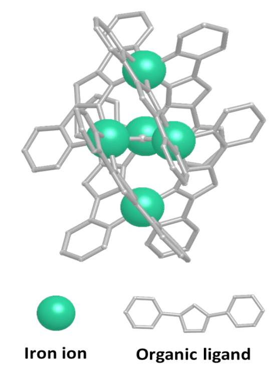
Pande et al, *J. Phys. Chem B* (2013)

Understanding!!!

Structure

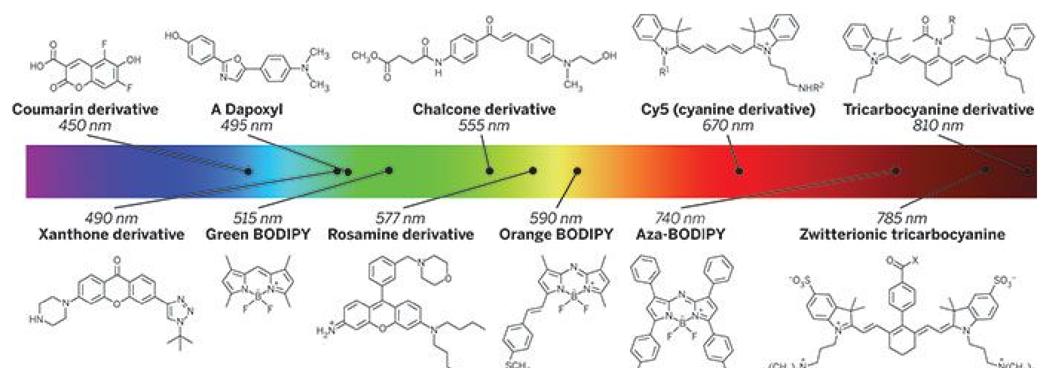
Catalysts



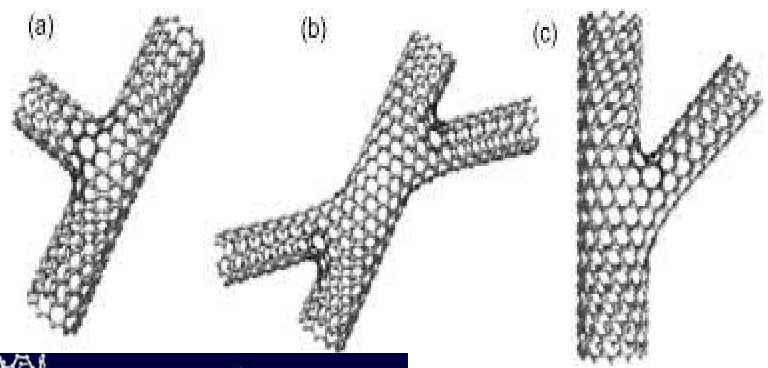


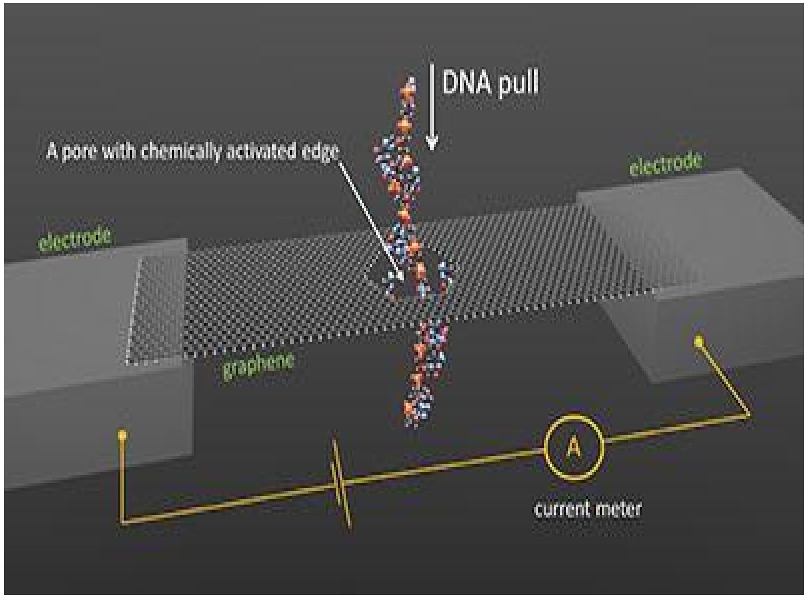
Okamura et al, *Nature* (2016)

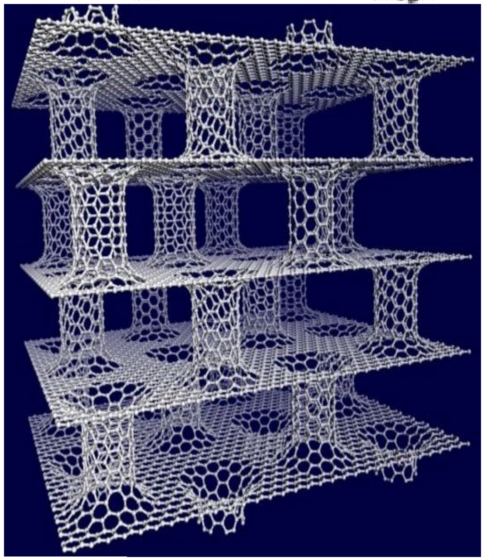
OLEDs

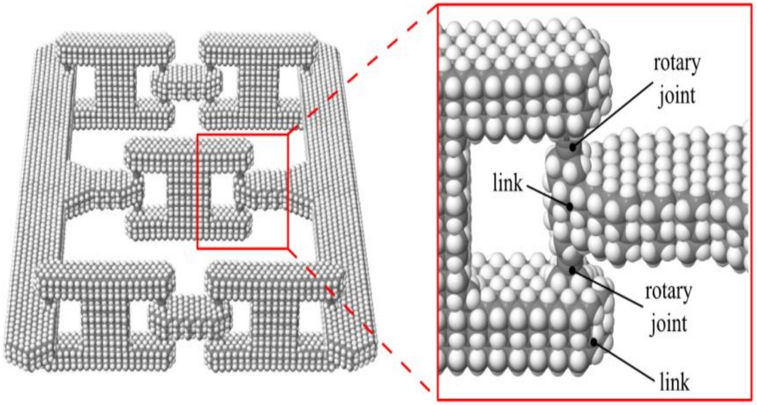


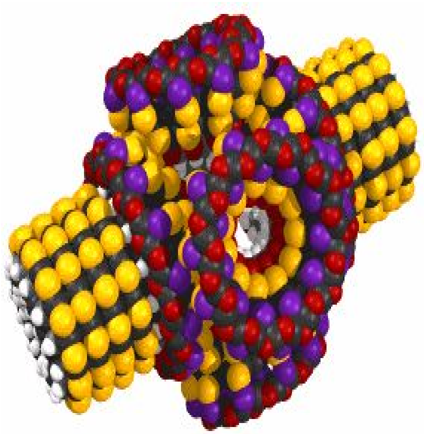
Complex Integrated Nanosystems





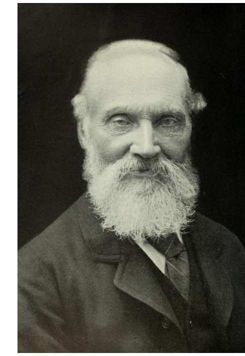






Theory to understand chemistry … to help design experiments

→ predictions that can be falsified

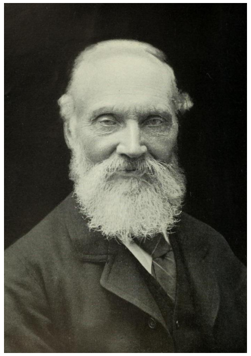


1. “ … *when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind.*” Lord Kelvin

→

Theory to understand chemistry … to help design experiments

→ predictions that can be falsified



1. " … *when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind.*" Lord Kelvin

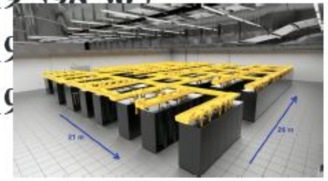
2. " … *we only understand molecules once we predict properties with quantitative accuracy*" M. Quack, ETHZ (2000)

→

TABLE I. History of the accurate calculations of the ground state of helium atom.

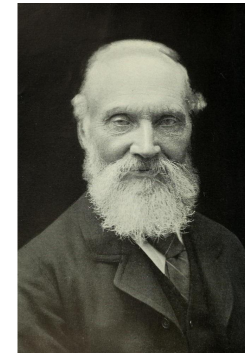
Mercedes Euklid Mod. 8

Year	Ref.	Type	Energy (a.u.)
1929	Hylleraas (Ref. 2)	Hylleraas (three terms)	**−2.902** 43
1957	Kinoshita (Ref. 6)	Kinoshita type	**−2.903 72**2 5
1966	Frankowski and Pekeris (Ref. 7)	Logarithm	**−2.903 724 377 03**2 6
1994	Thakkar and Koga (Ref. 8)	Half-integer	**−2.903 724 377 034 11**4 4
1998	Goldman (Ref. 9)	Polynomial	−**2.903 724 377 034 119 59**4
1999	Drake (Ref. 10)	Double exponent	**−2.903 724 377 034 119 59**6
2002	Sims and Hagstrom (Ref. 12)	Hylleraas-CI	**−2.903 724 377 034 119 598 2**9 99
2002	Drake *et al.* (Ref. 11)	Triple exponent	**−2.903 724 377 034 119 598 305**
2002	Korobov (Ref. 13)	Slater geminal	**−2.903 724 377 034 119**
2006	Schwartz (Ref. 15)	Logarithm (ln(*s*))	**−2.903 724 377 034 119** **194 404 44**0 049 5
2007	This work	ICI (new logarithm)	**−2.903 724 377 034 119 598 311 159 245 194 404 446 696 9**05 37



Theory to understand chemistry … to help design experiments

→ predictions that can be falsified



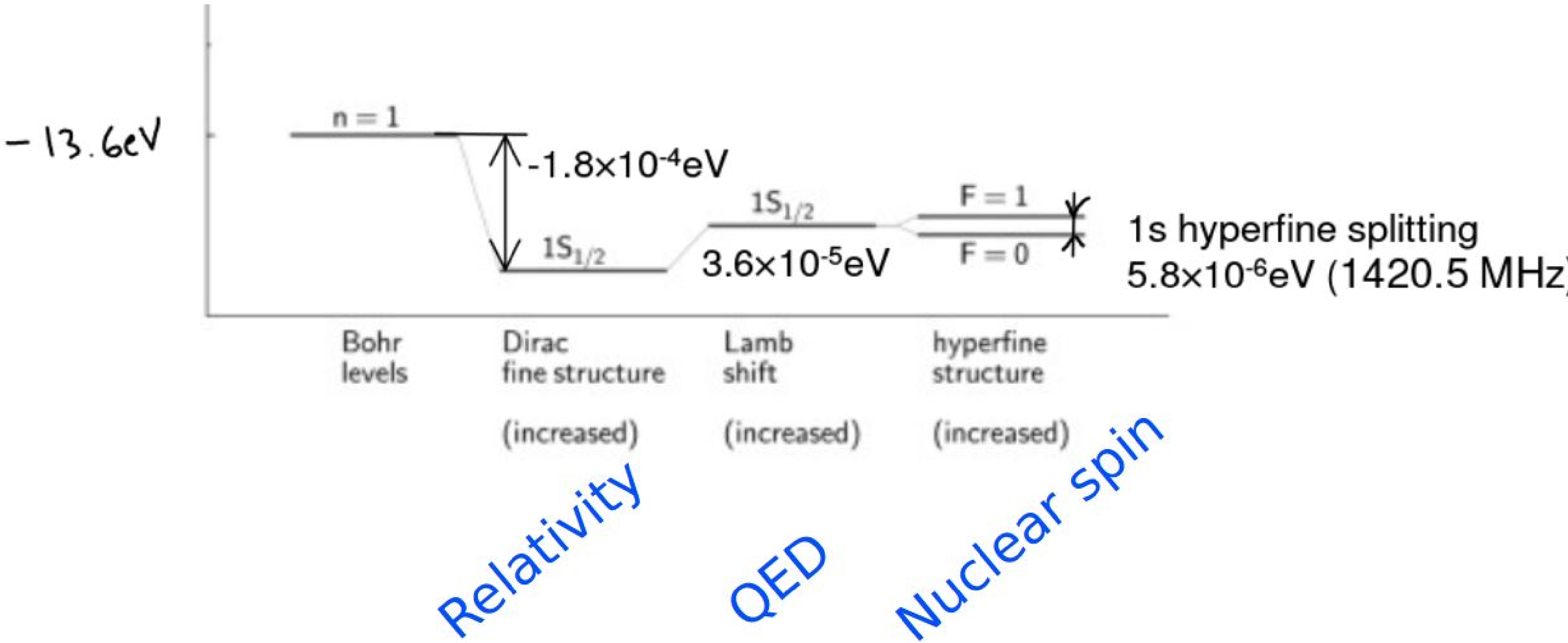
1. " … *when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind.*" Lord Kelvin

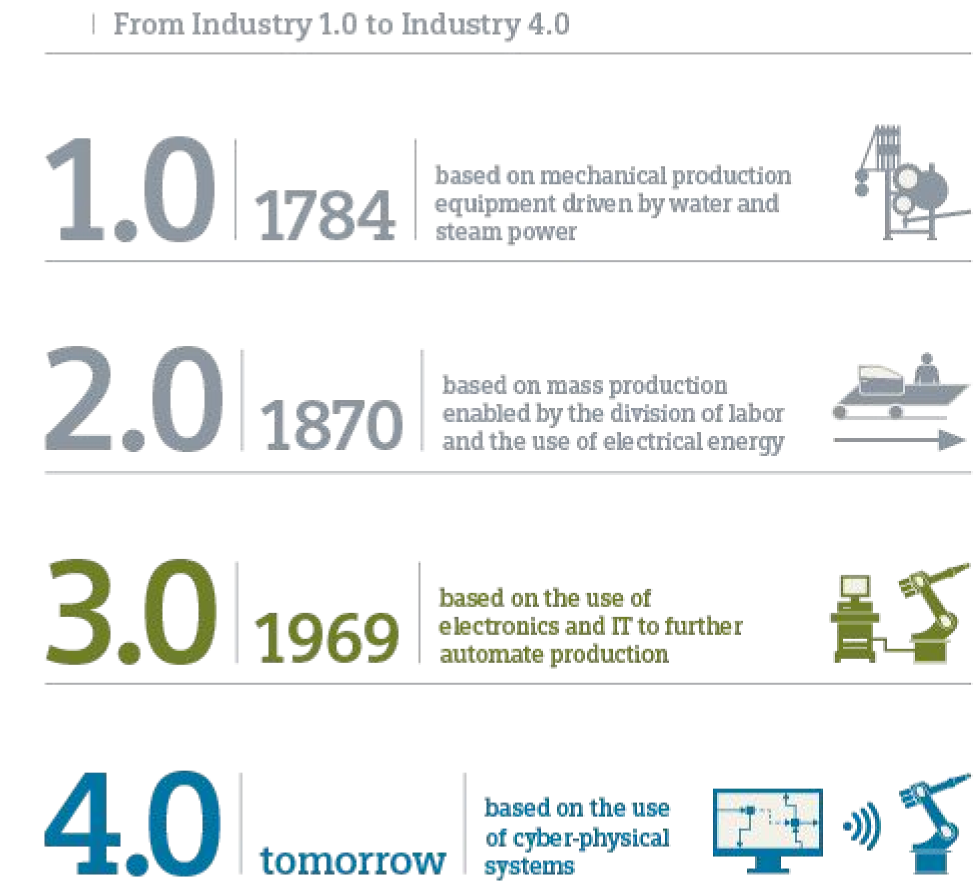
2. " … *we only understand molecules once we predict properties with quantitative accuracy*" M. Quack, ETHZ (2000)

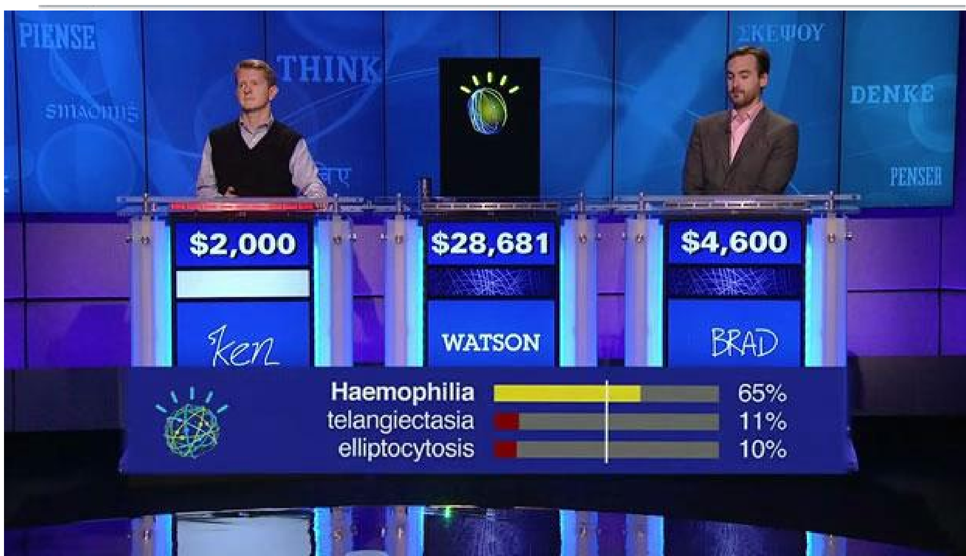
3. *" … It is nice to know that the computer understands the problem. But I would like to understand it too."* E. Wigner

→ compare to experiment (arbiter)

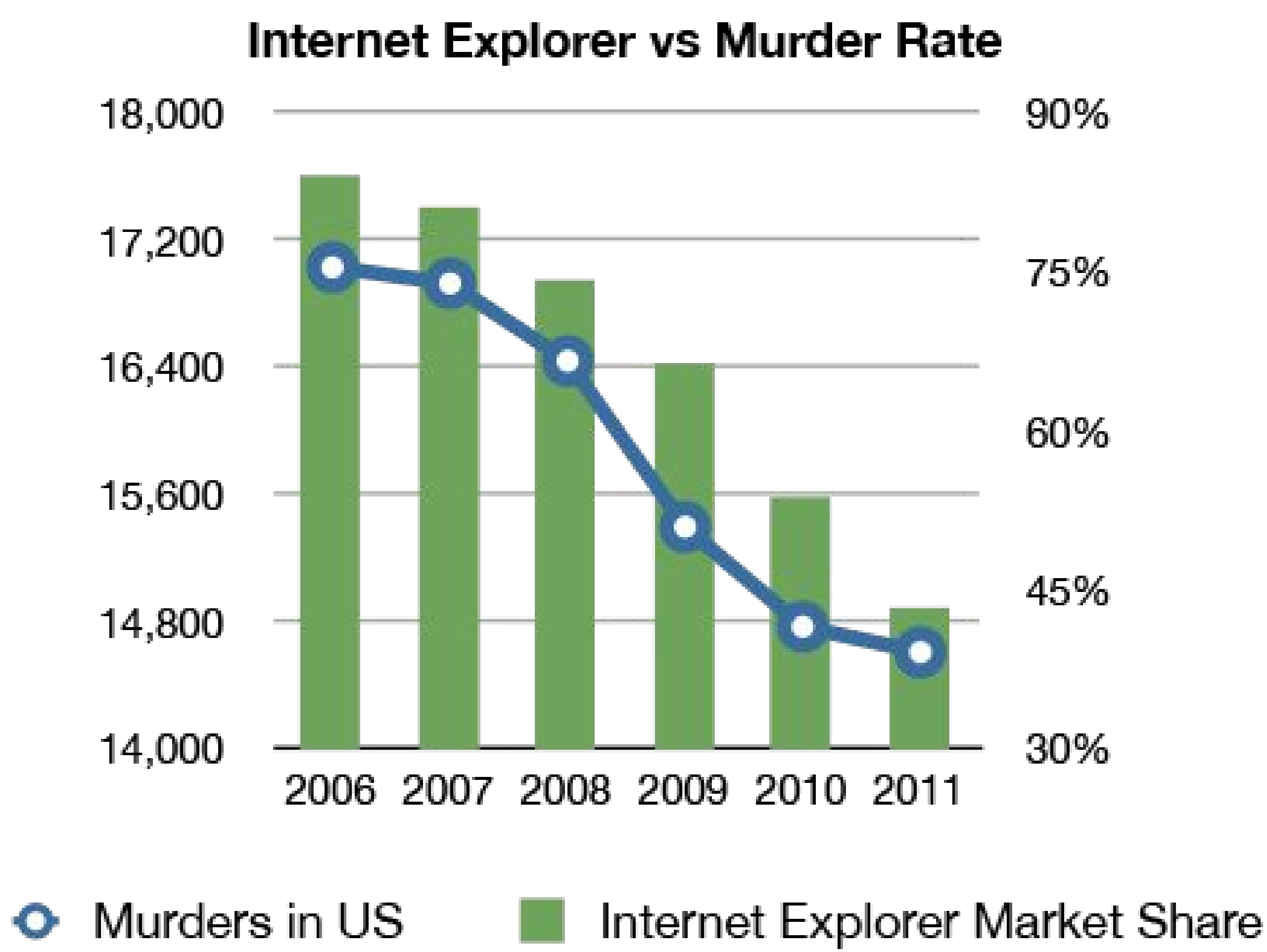
Energy of Hydrogen atom



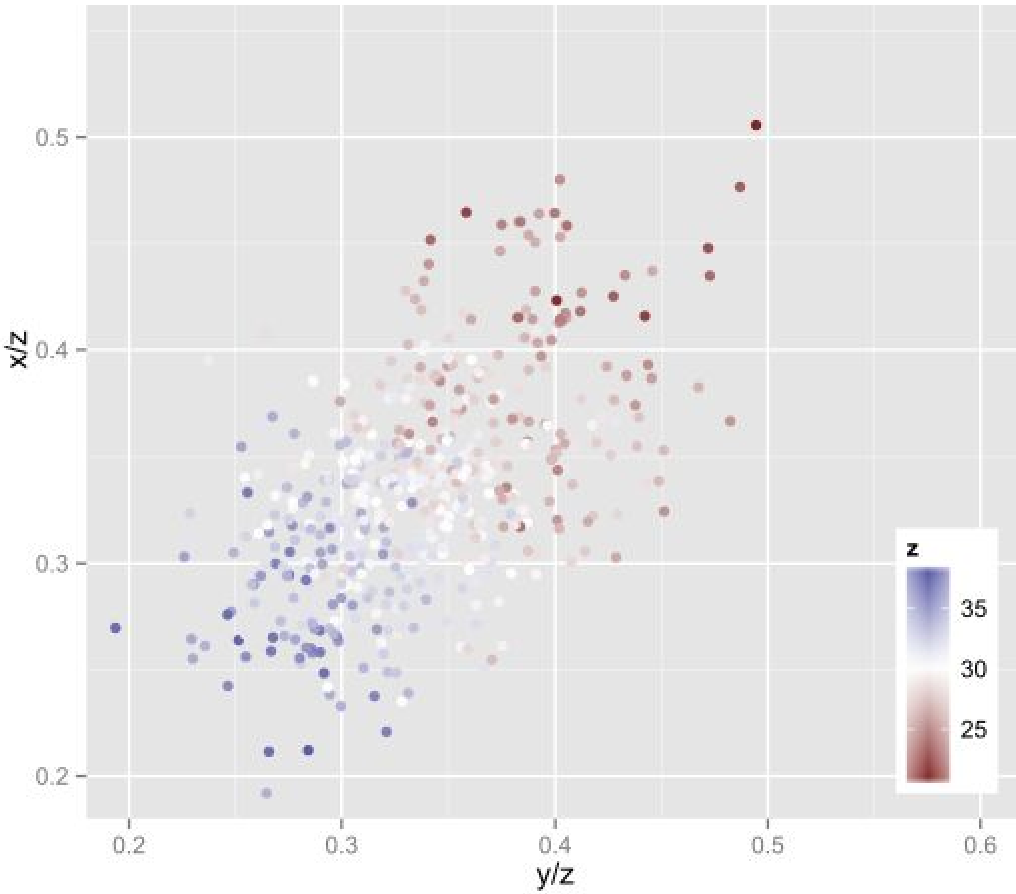

From Industry 1.0 to Industry 4.0
1.0
1784
based on mechanical production equipment driven by water and steam power
2.0
1870
based on mass production enabled by the division of labor and the use of electrical energy
3.0
1969
based on the use of electronics and IT to further automate production
4.0
tomorrow
based on the use of cyber-physical systems

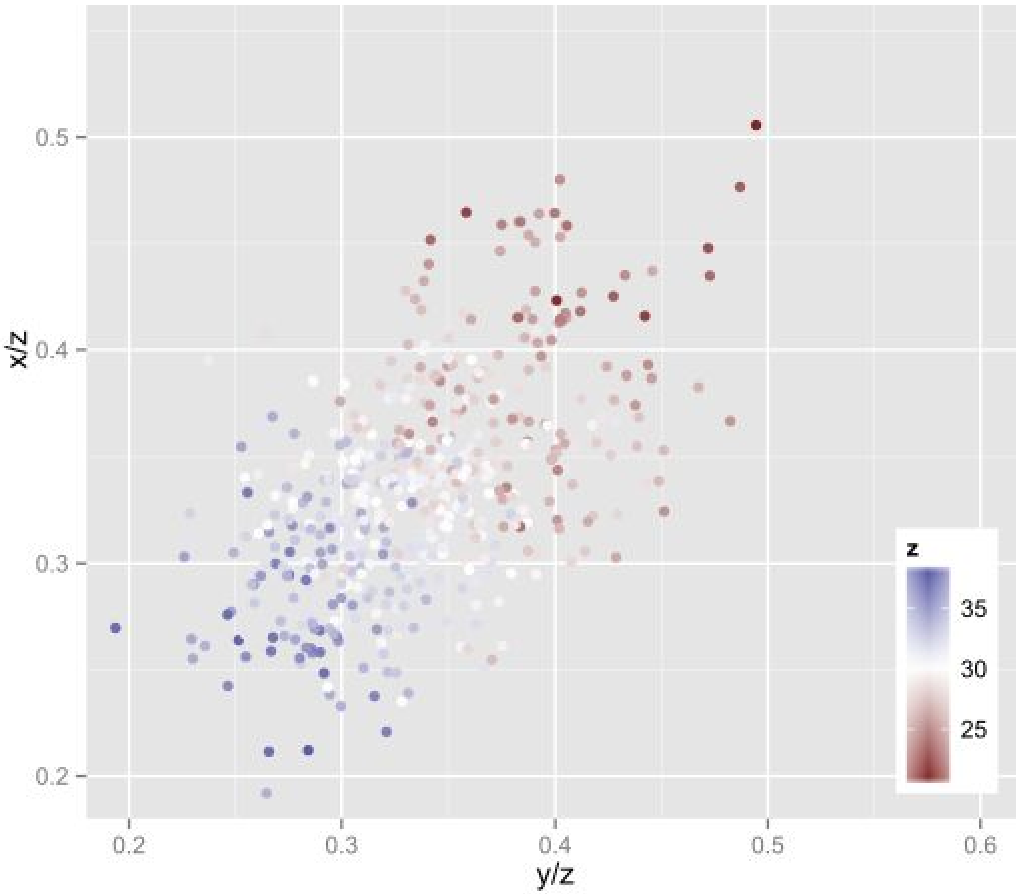
???

PIENSE
THINK
DENKE
PENSER
$2,000
$28,681
$4,600
Ken
WATSON
BRAD
Haemophilia 65%
telangiectasia 11%
elliptocytosis 10%

Big Data



More at
http://www.tylervigen.com/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



Example of spurious correlation for 500 draws of *x,y,z* with respective means of 10,10,30 and standard deviations 1, 1, and 9. From `wikipedia`

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

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]

Spurious correlation can also be due to

1. chance (*anything* which varies simultaneously will correlate)
2. a common cause
3. identity relationships

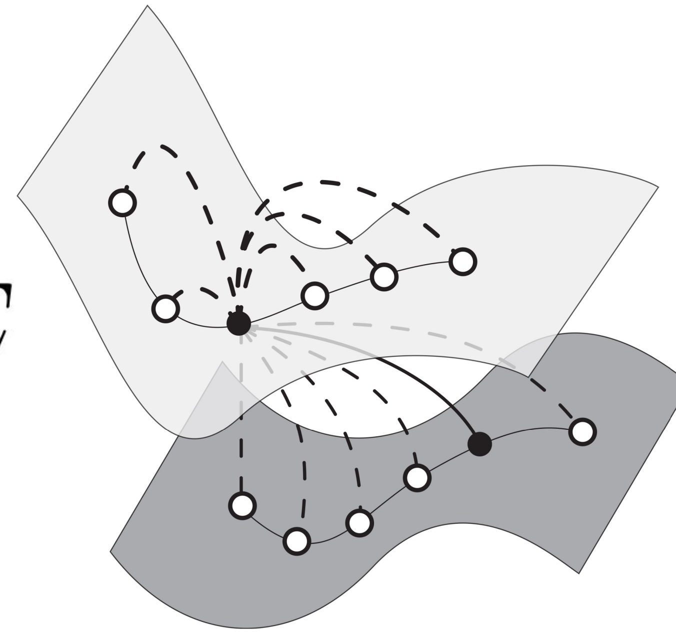
correlations (inductive) vs. **law** (deductive)

Erwin

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

correlations (inductive) vs. **law** (deductive)

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



Chang, OAvL, *CHIMIA* (2014)

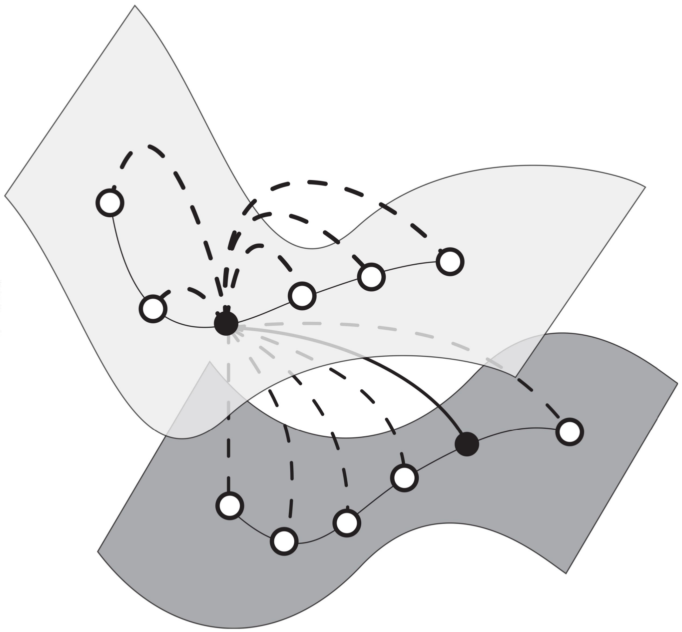
Erwin

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

correlations (inductive) vs. law (deductive)

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

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

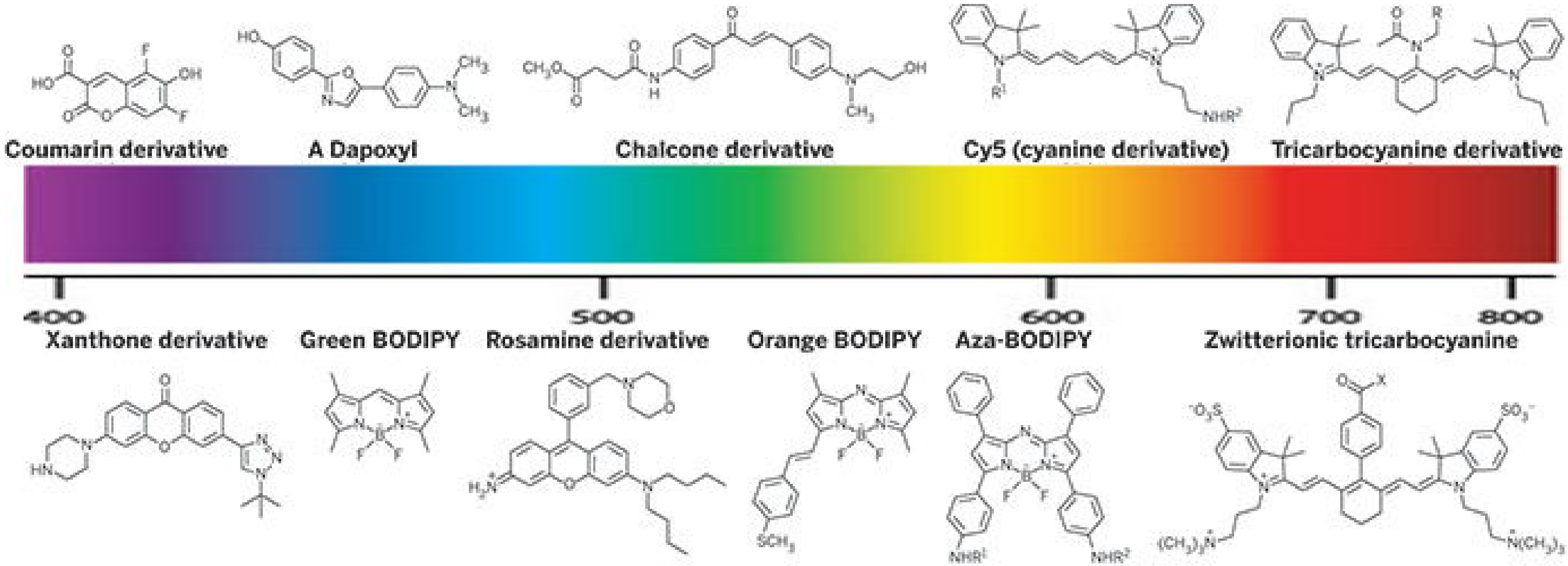


Chang, OAvL, *CHIMIA* (2014)

Erwin

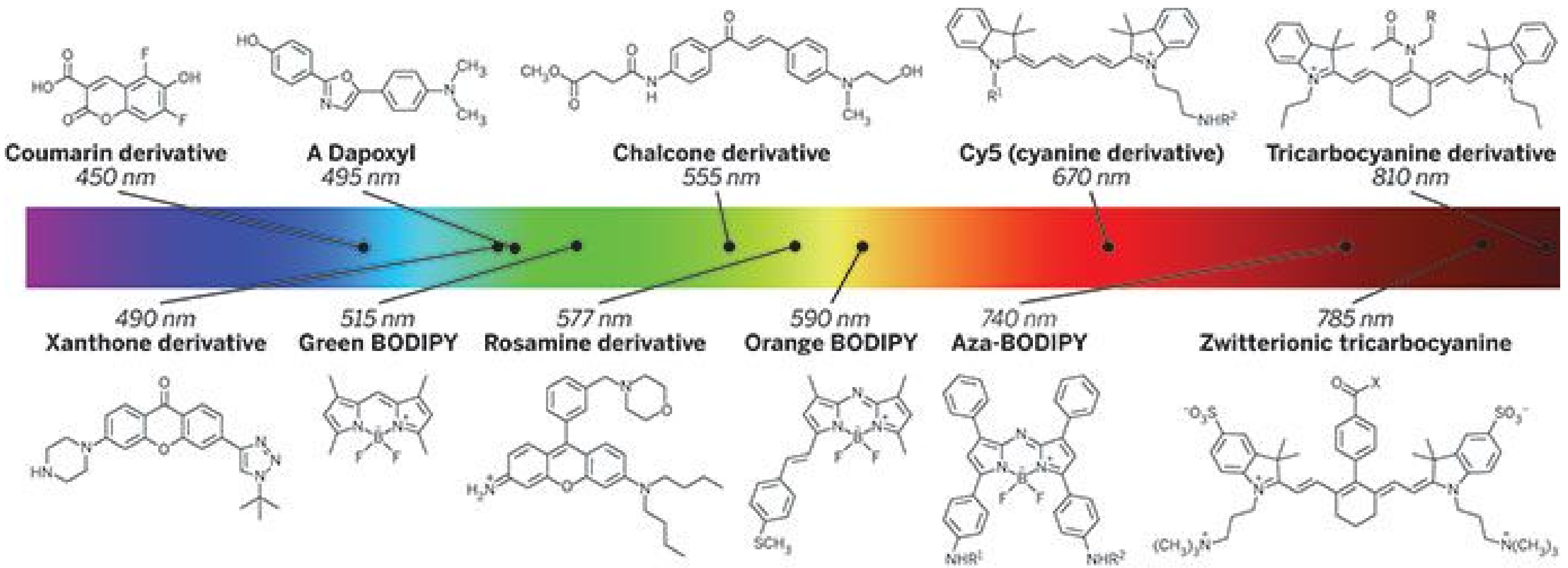
$$H(\{Z_I, \mathbf{R}_I\})\Psi(\mathbf{r}) = E\Psi(\mathbf{r})$$

Configuration + Composition → Chemical Space



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

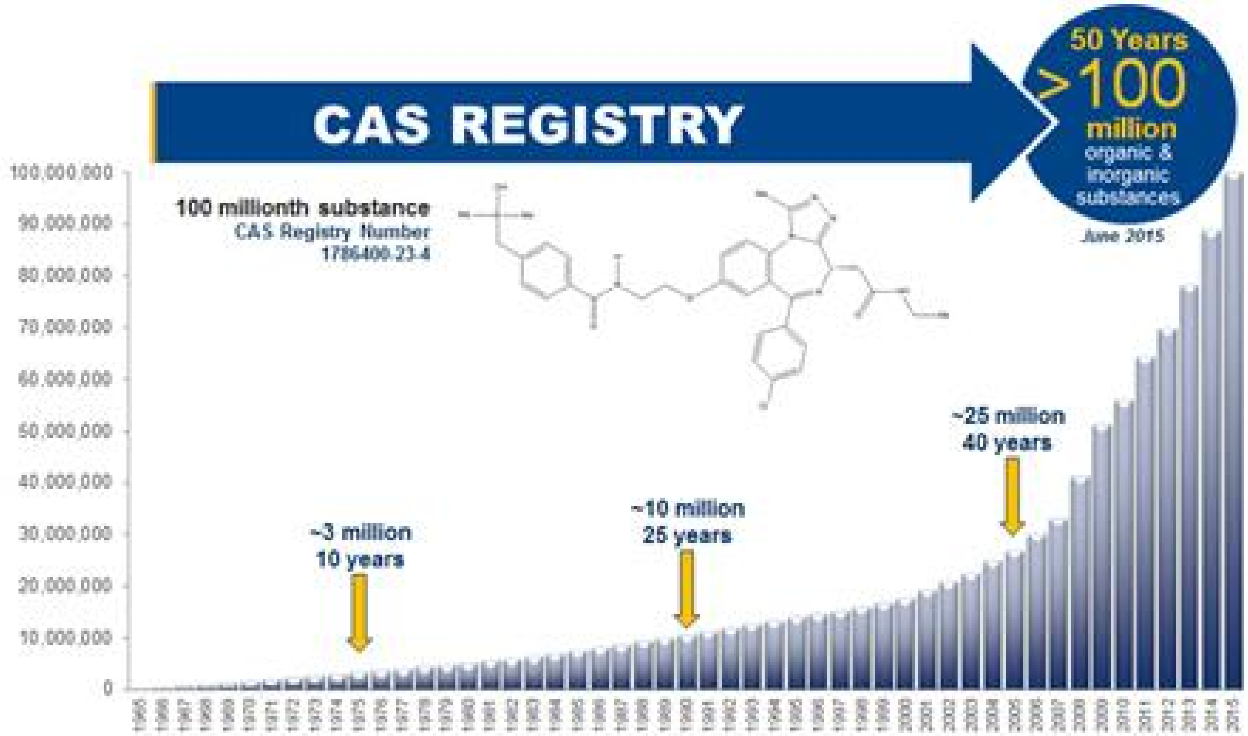
Configuration + Composition → Chemical Space



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

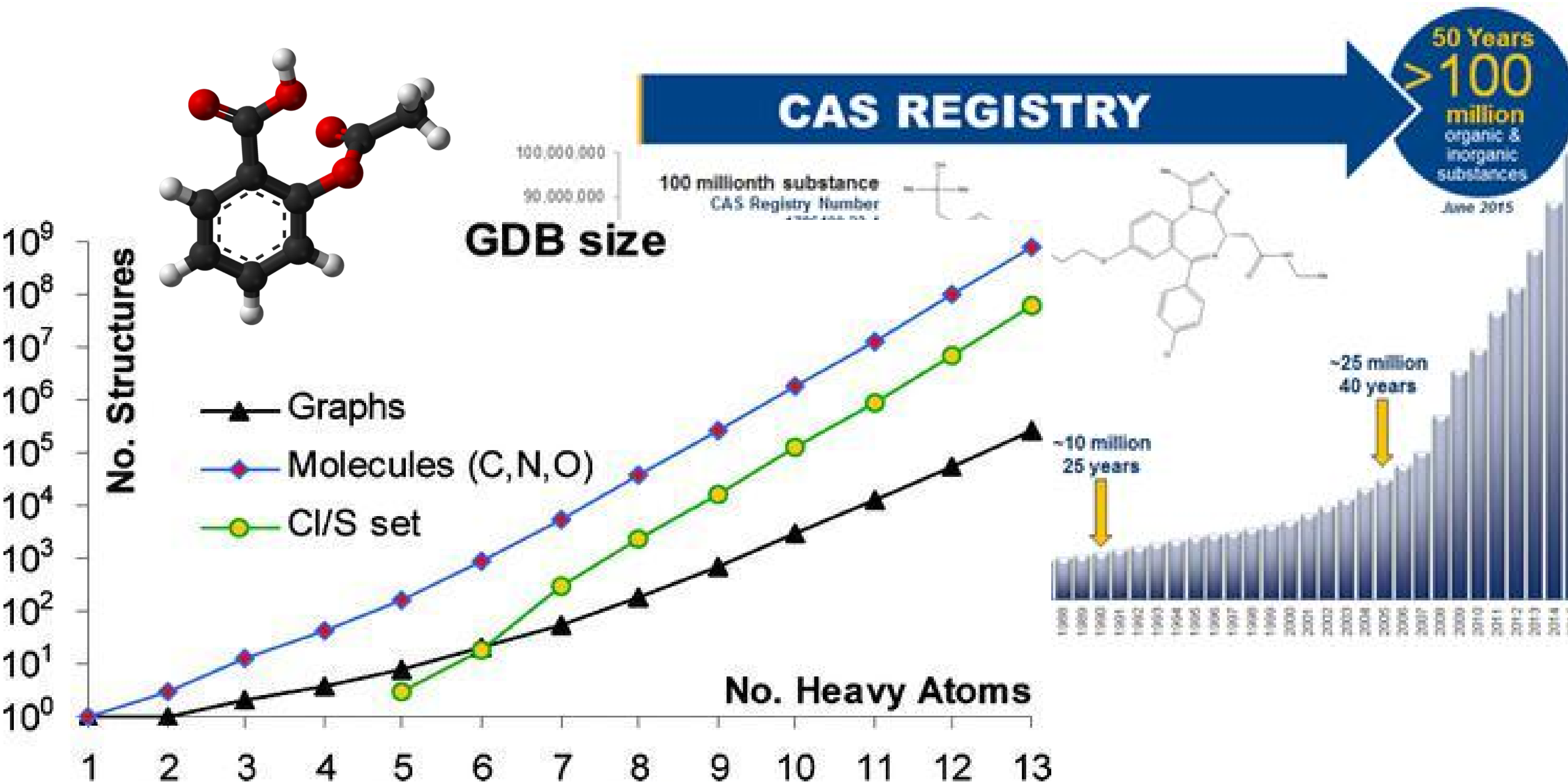
How many are possible?

Differ in composition and constitution (no conformational isomers)
~120 M
~15 k are being added on a daily basis



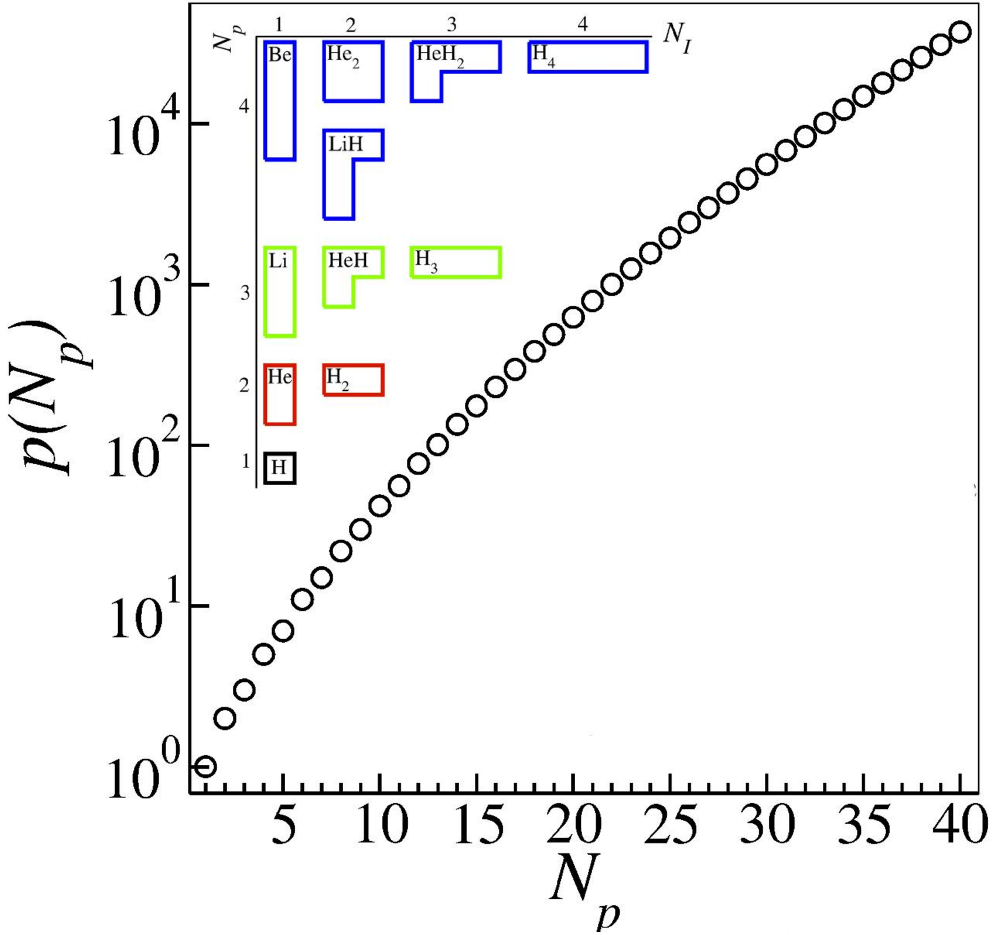
``The greatest shortcoming of the human race is our inability to understand the exponential function''

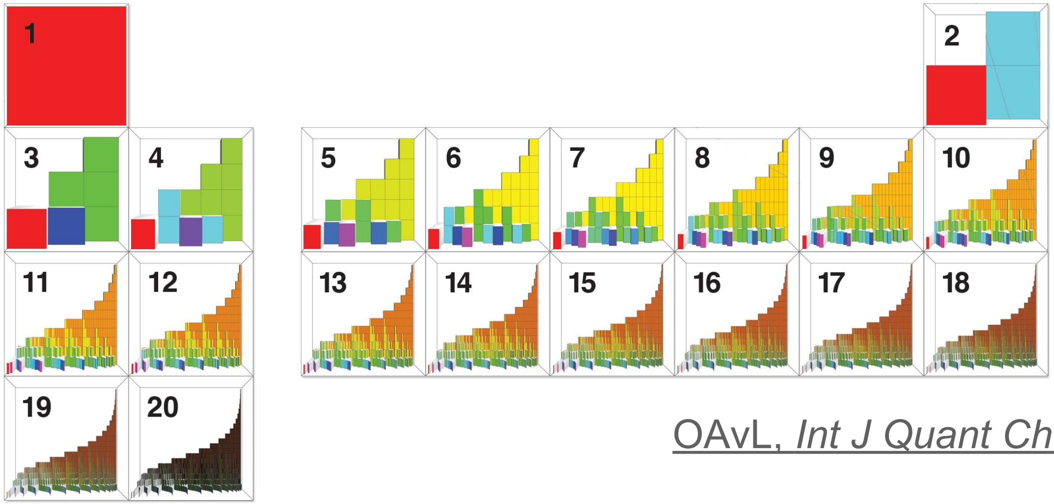
Al Bartlett, U of Colorado Boulder



J.-L. Reymond and coworkers, *J Am Chem Soc* (2009) and ff

Composition

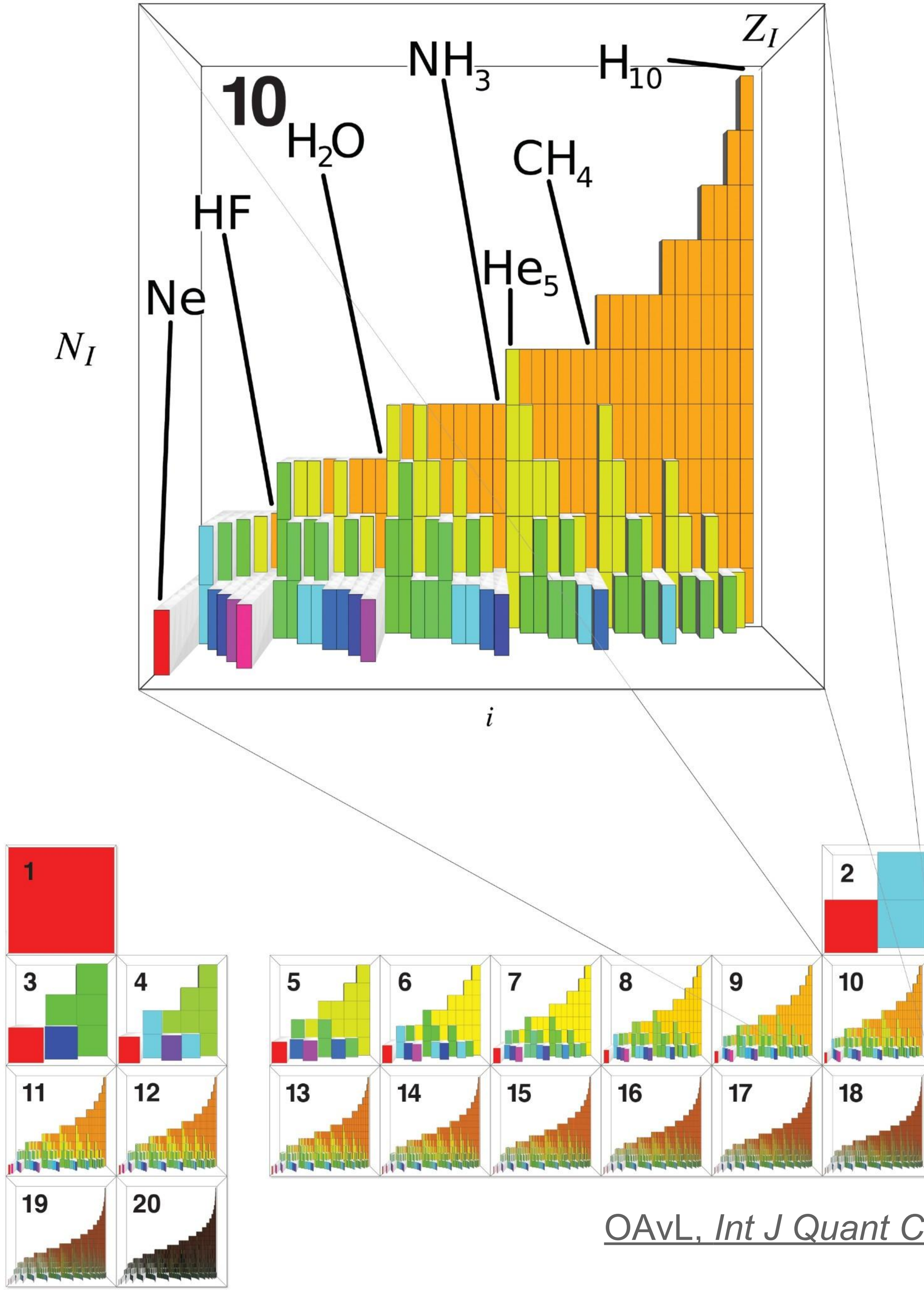




OAvL, *Int J Quant Chem*, (2013)

Composition

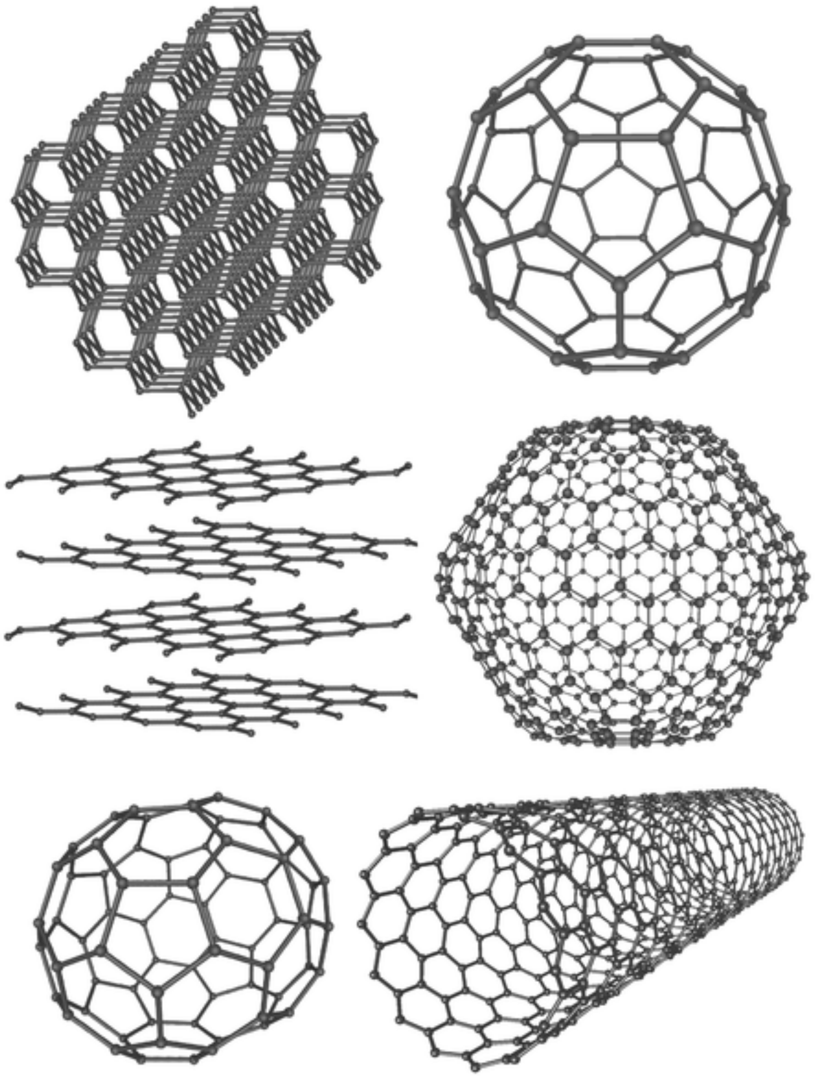
10 protons

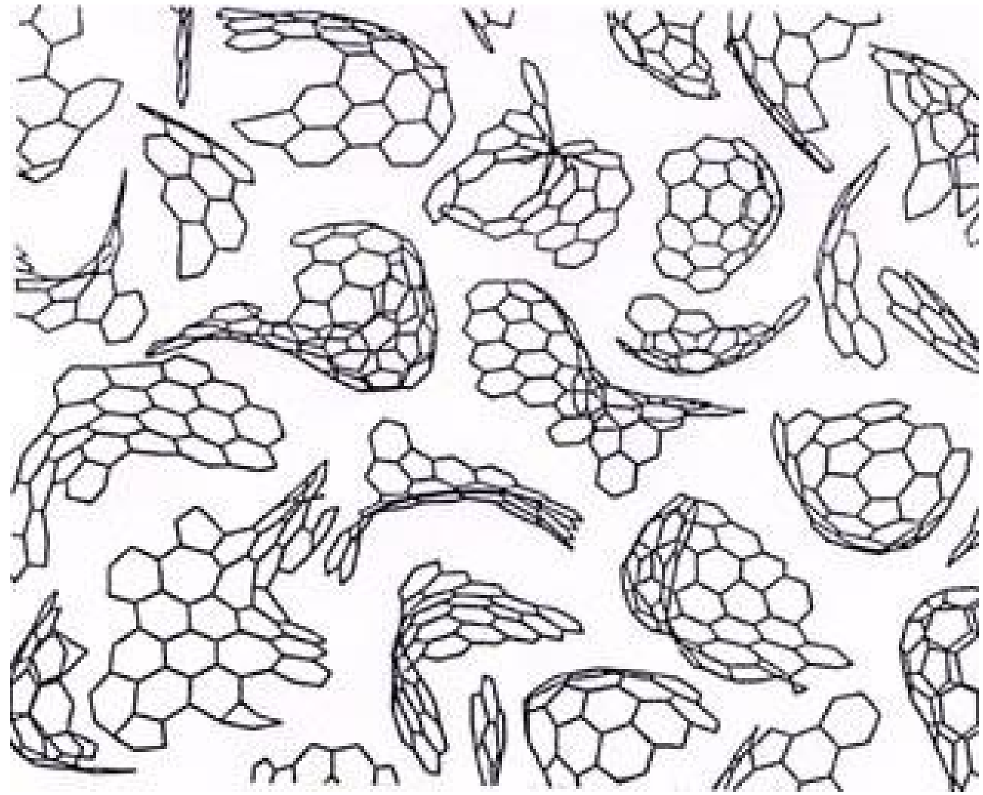


OAvL, *Int J Quant Chem*, (2013)

Spatial configuration

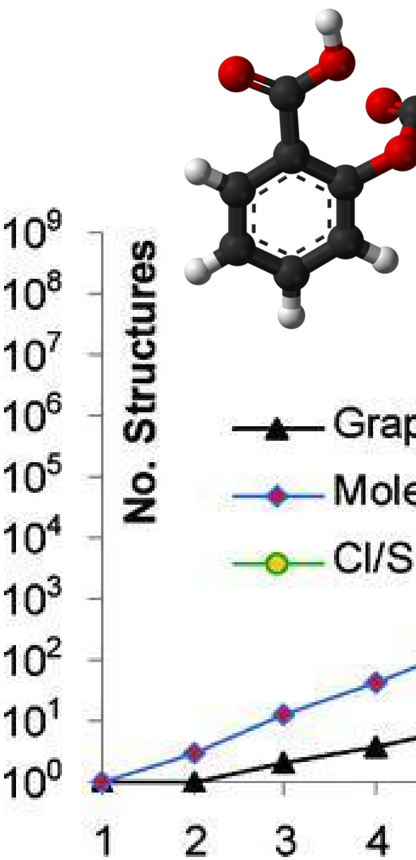
Carbon allotropes





``The great   **s our**
inability to   **n”**

Al Bartlett, U c

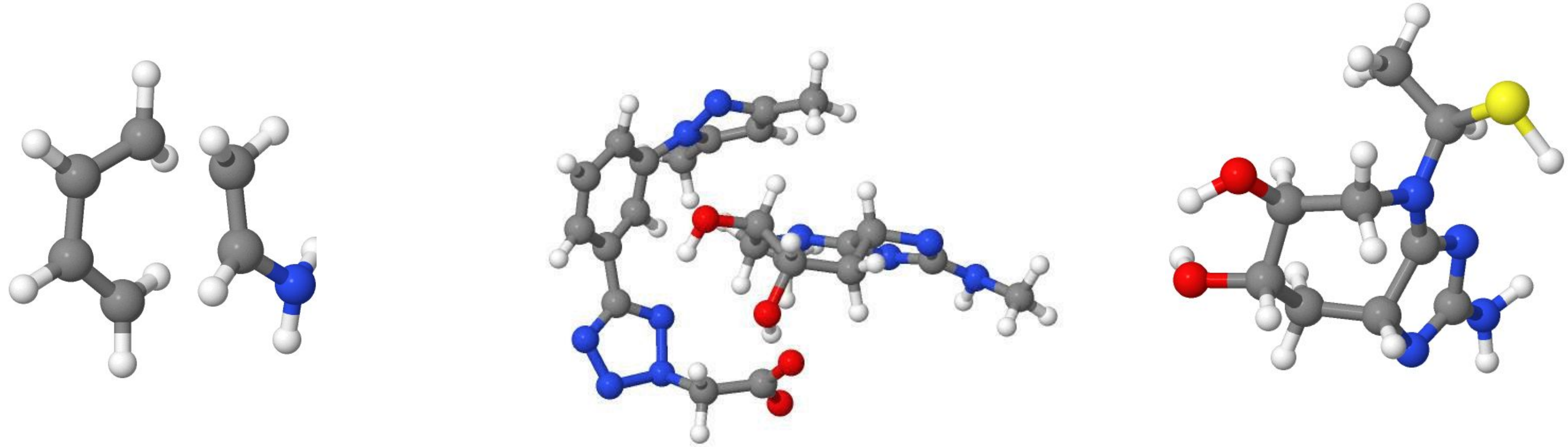


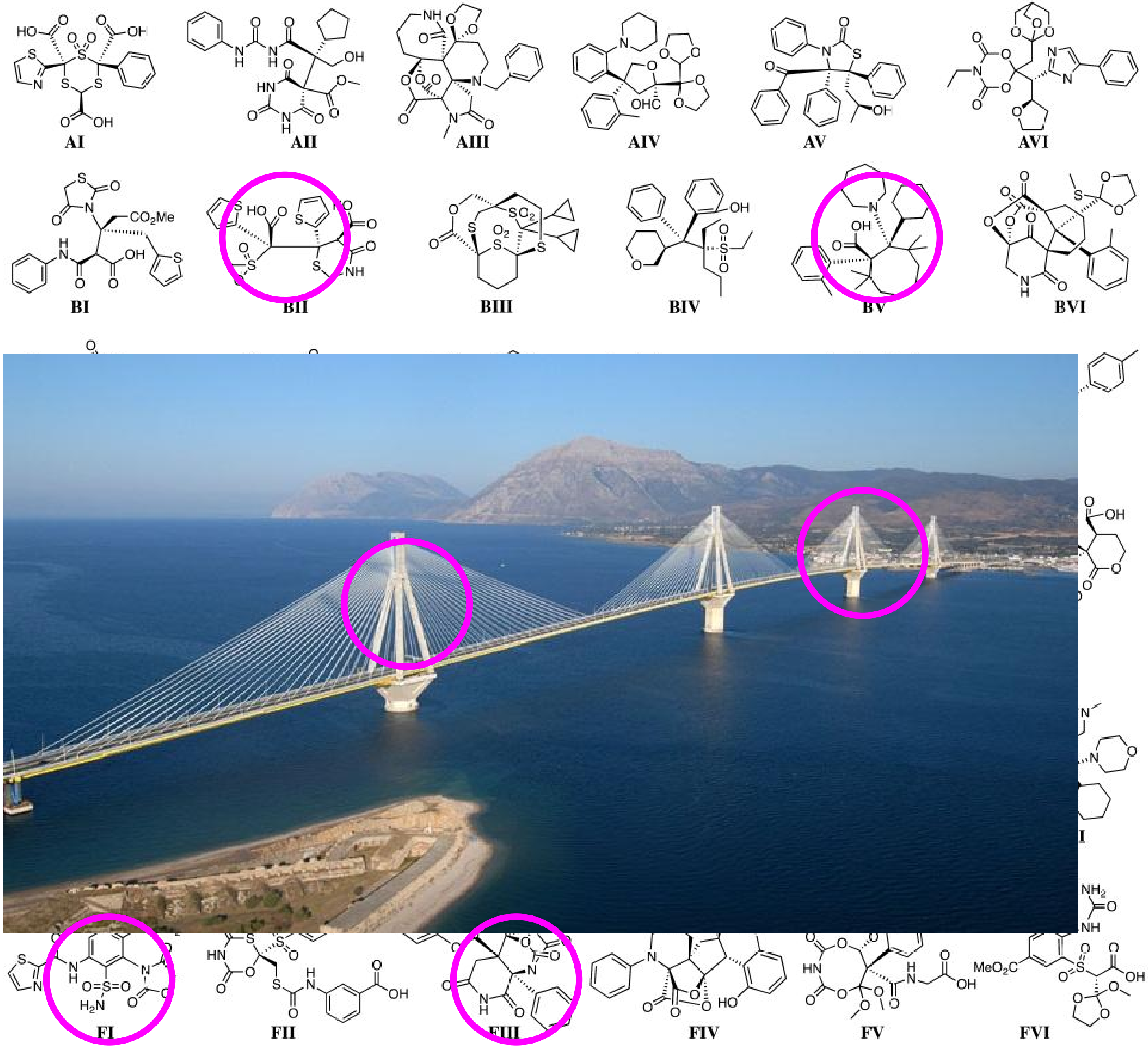
J.-L. Reymond and coworkers, *J Am Chem Soc* (2009) and ff

Conclusions

1. Instantaneous QM quality predictions
2. Learning curves reveal quality of ML model
3. Representations
4. Data sets

-70.160669 96288144_10





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

Kernel Ridge Regression

Kernel

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

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

Regression

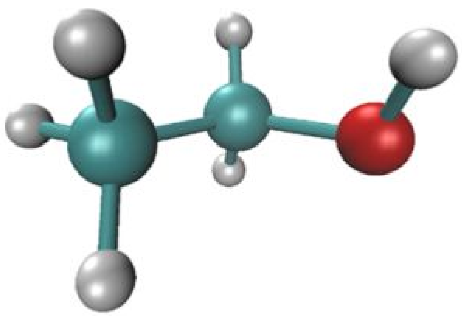
$$\min_\alpha \quad \left(\sum_i \left(E^{est}(\mathbf{M}_i) - E_i^{ref}\right)^2 + \lambda \sum_{ij} \alpha_i \alpha_j k(\mathbf{M}_i, \mathbf{M}_j)\right)$$

Solution

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

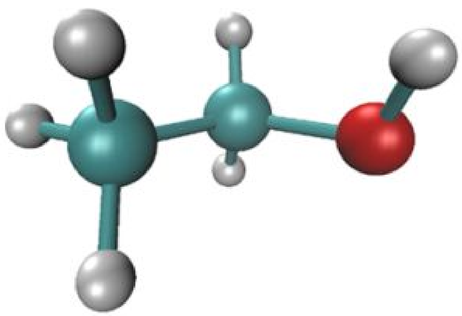
From molecule to representation

Molecule



From molecule to representation

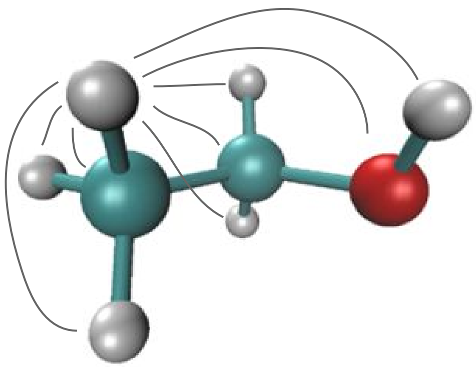
Molecule



From molecule to Coulomb matrix (CM) to Bag of Bonds (BOB)

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

Molecule

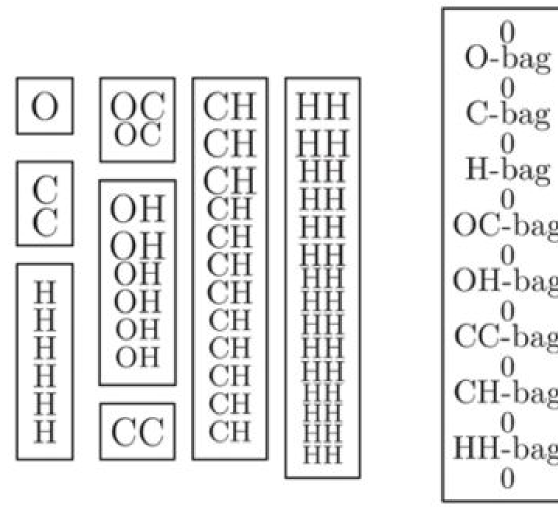


Coulomb matrix (CM)

	O	C	C	H	H	H	H	H	H
O	O	OC	OC	OH	OH	OH	OH	OH	OH
C	OC	C	CC	CH	CH	CH	CH	CH	CH
C	OC	CC	C	CH	CH	CH	CH	CH	CH
H	OH	CH	CH	H	HH	HH	HH	HH	HH
H	OH	CH	CH	HH	H	HH	HH	HH	HH
H	OH	CH	CH	HH	HH	H	HH	HH	HH
H	OH	CH	CH	HH	HH	HH	H	HH	HH
H	OH	CH	CH	HH	HH	HH	HH	H	HH
H	OH	CH	CH	HH	HH	HH	HH	HH	H

Rupp et al, *Phys Rev Lett* (2012)

Bag of Bonds (BoB)



Hansen et al, *J Phys Chem Lett* (2015)

- Unique but overcomplete
- Invariances (Tra&Rot)
- Compact
- Physical meaning
- Fast
- Simple metrics are not smooth if sorted

- Not unique (homometricity)
- Invariant (Tra&Rot)
- Compact
- Physical meaning
- Fast
- Simple metrics are smooth

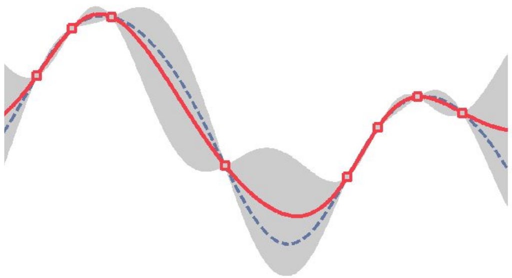
Big Data
N

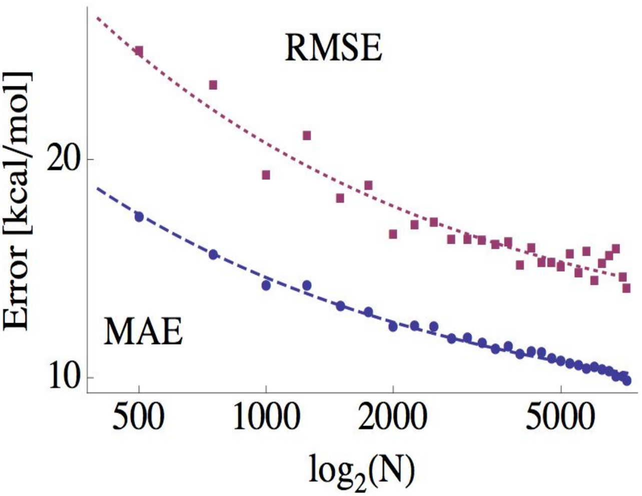
↓

Train

↓

Model(N)





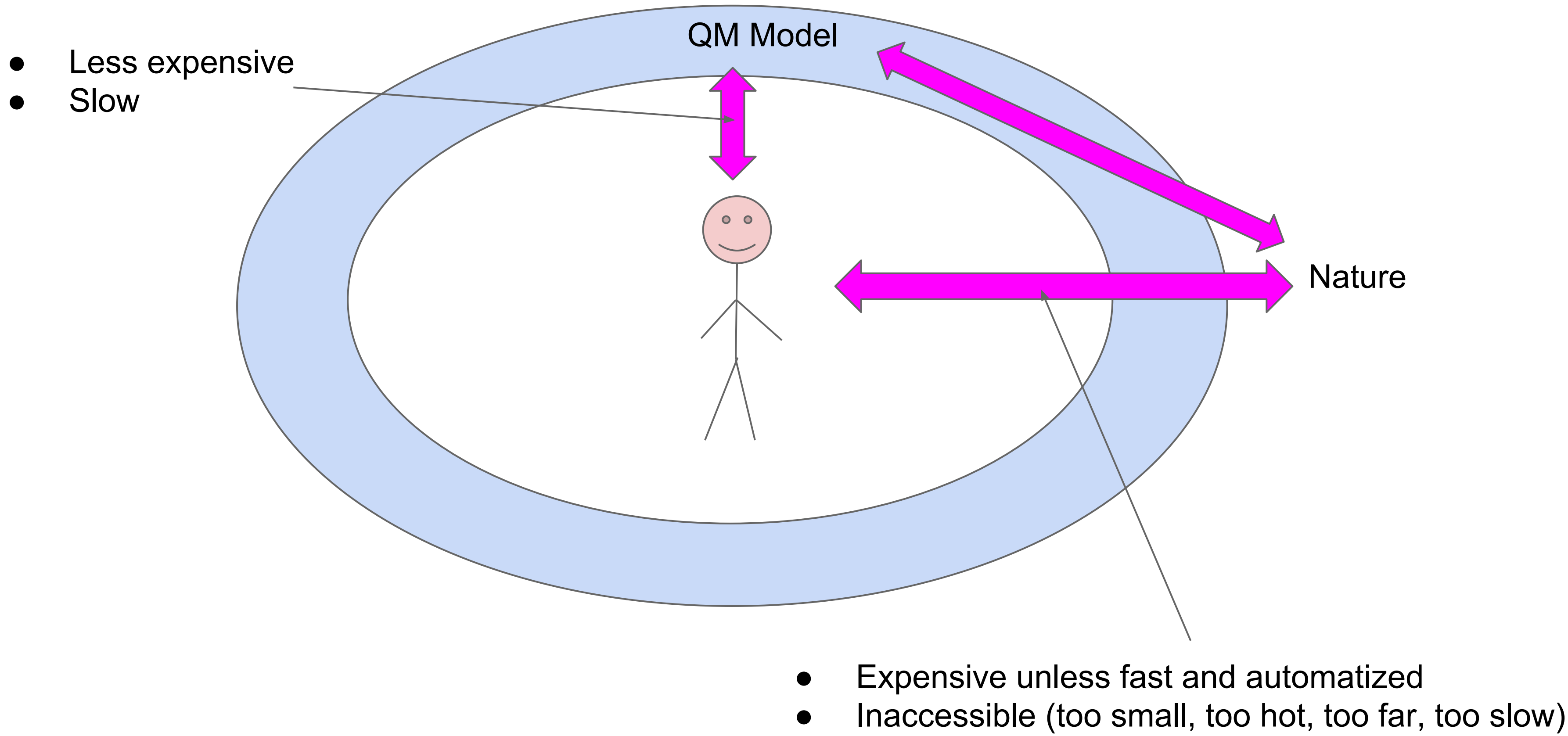
Rupp et al, *Phys Rev Lett* (2012)

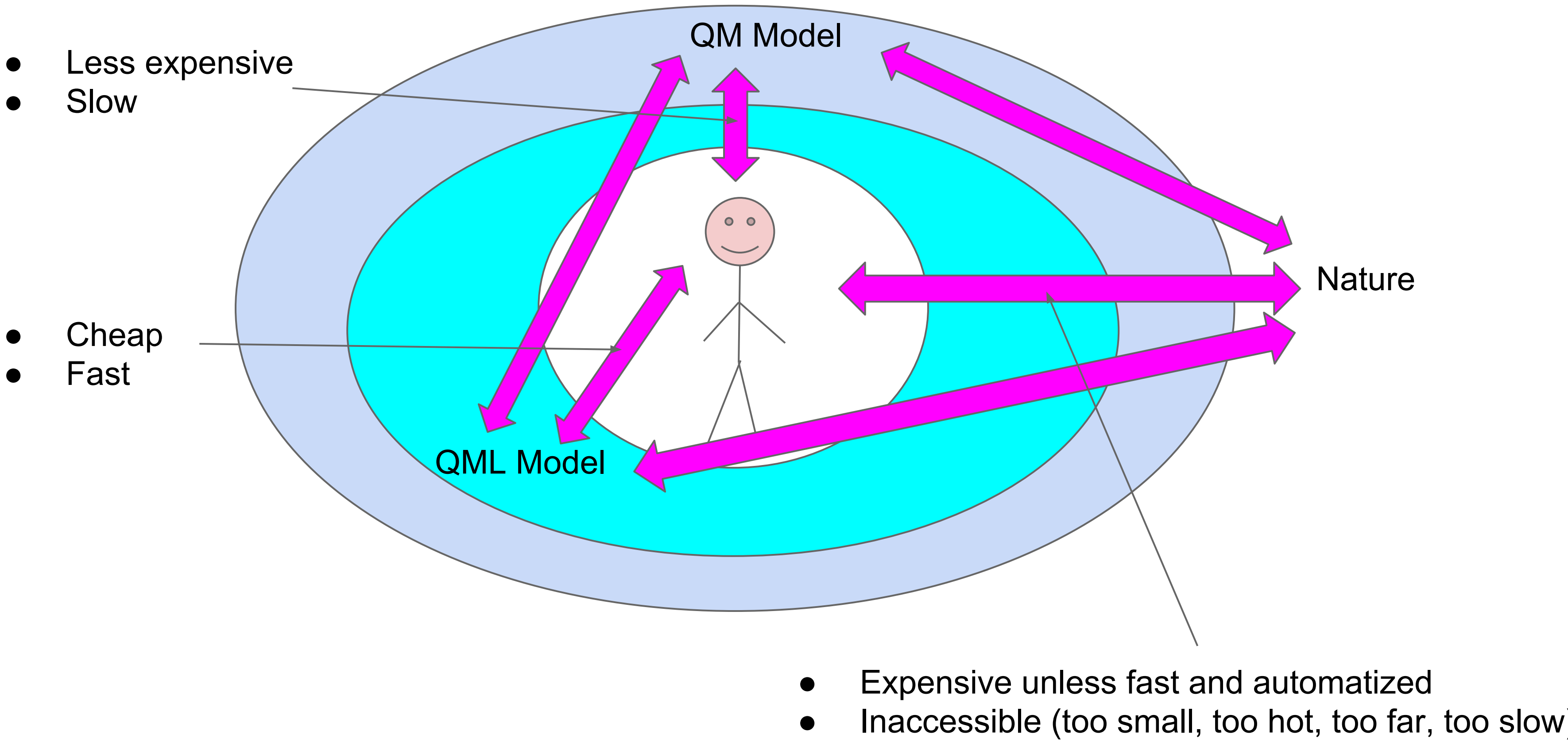
QM: ~1000 seconds
ML: ~milli seconds

Big Data
N

Train

Query → Model(N) → Predict


QM Model
Nature
Less expensive
Slow
Expensive unless fast and automatized
Inaccessible (too small, too hot, too far, too slow)



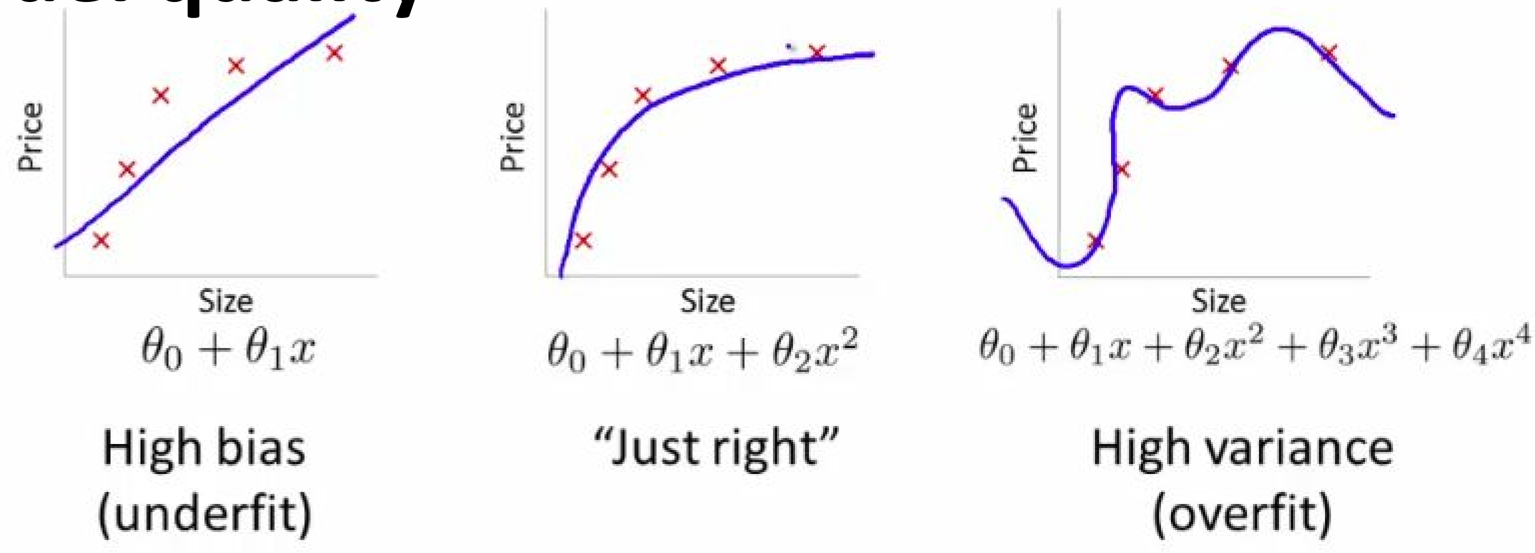
Interactive virtual experimenting possible in real time

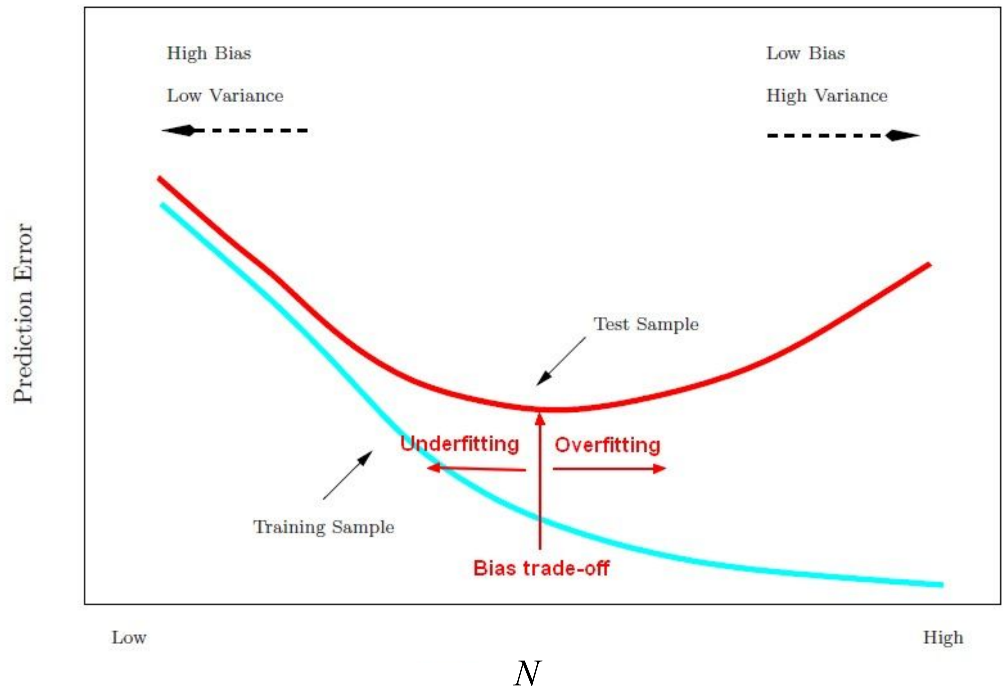
1. Molecular, materials, biochemical design problems
2. Discover new trends/relationships/rules/fill gaps
3. Enhance teaching, communication, and outreach

Conclusions

1. Instantaneous QM quality predictions
2. Learning curves reveal quality of ML model
3. Representations
4. Data sets

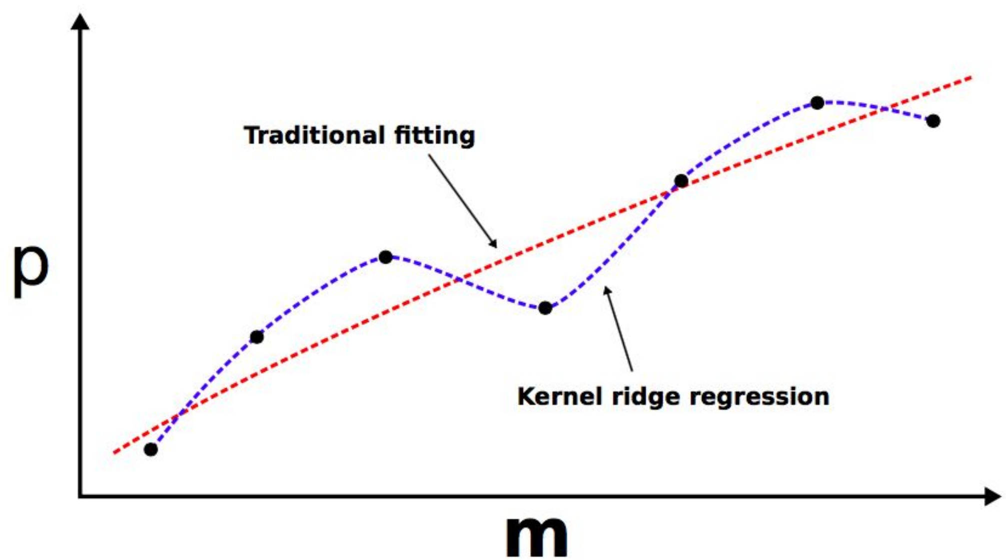
Model quality



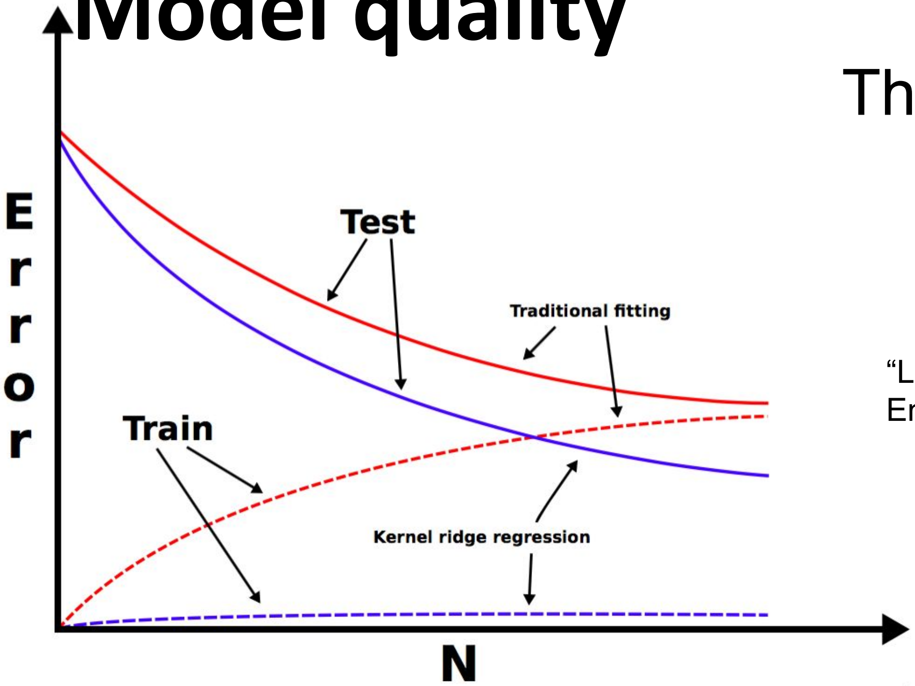


Model quality

The bigger the data the better …

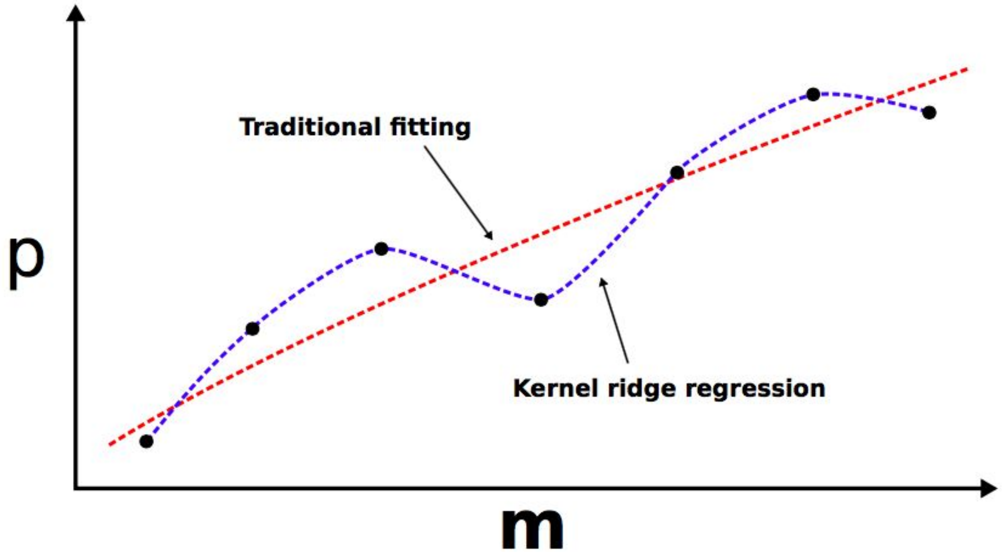


Model quality



The bigger the data the better …

"Learning curves in machine learning" Claudia Perlich, Encyclopedia of Machine Learning (Springer, 2011) pp. 577–580.

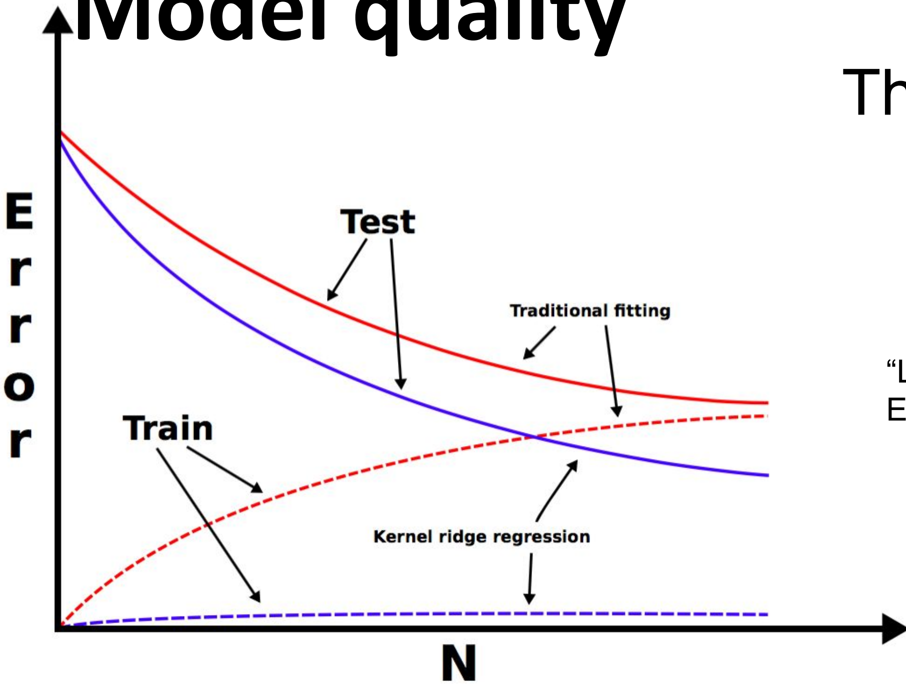


$$\text{Error} \sim a/N^b$$

K.-R. Mueller et al, *Neural Comput* (1996)

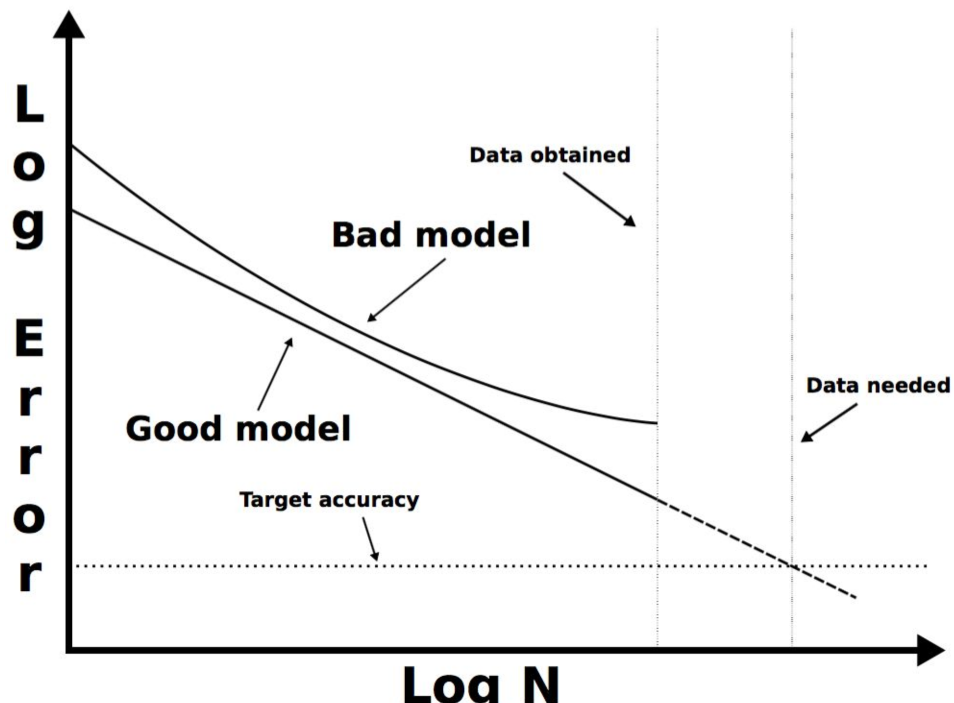
$$\rightarrow \log(\text{Error}) = \log(a) - b\log(N)$$

Model quality



The bigger the data the better …

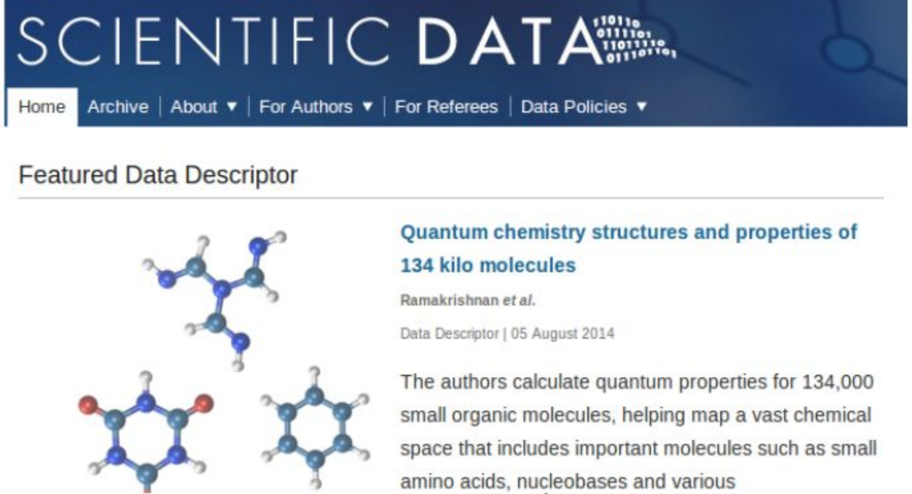
"Learning curves in machine learning" Claudia Perlich, Encyclopedia of Machine Learning (Springer, 2011) pp. 577–580.

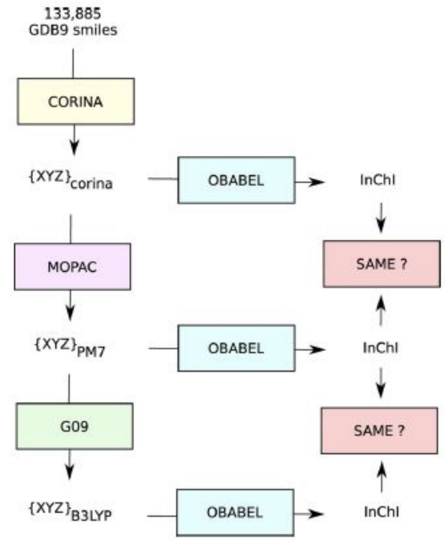


$$\text{Error} \sim a/N^b$$

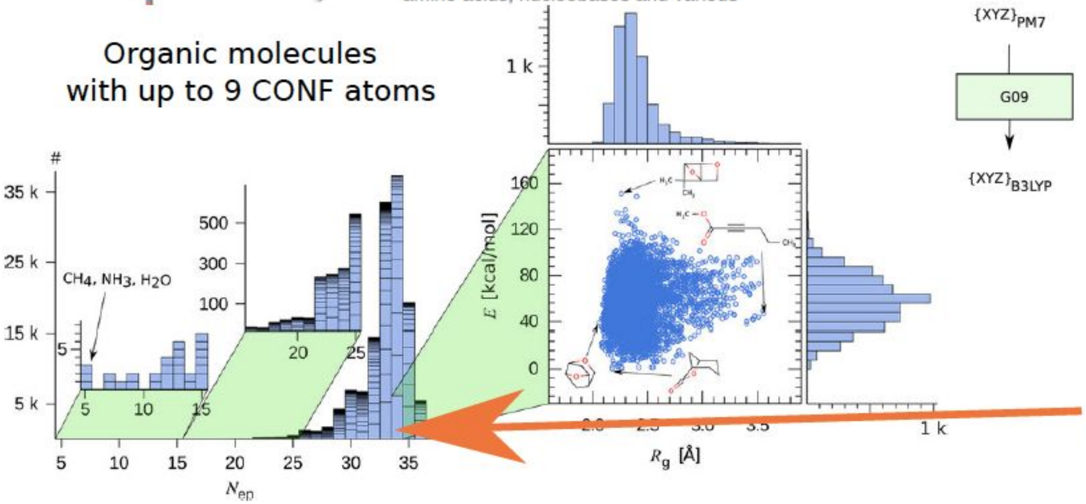
K.-R. Mueller et al, *Neural Comput* (1996)

$$\rightarrow \log(\text{Error}) = \log(a) - b\log(N)$$



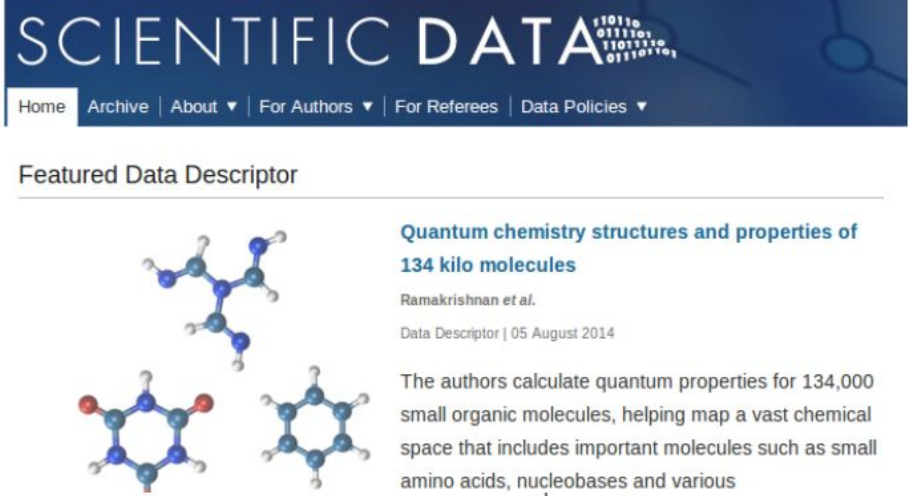


Organic molecules
with up to 9 CONF atoms



6k $C_7O_2H_{10}$
constitutional isomers

Ramakrishnan et al, *Scientific Data* (2014)


SCIENTIFIC DATA

Home | Archive | About ▼ | For Authors ▼ | For Referees | Data Policies ▼

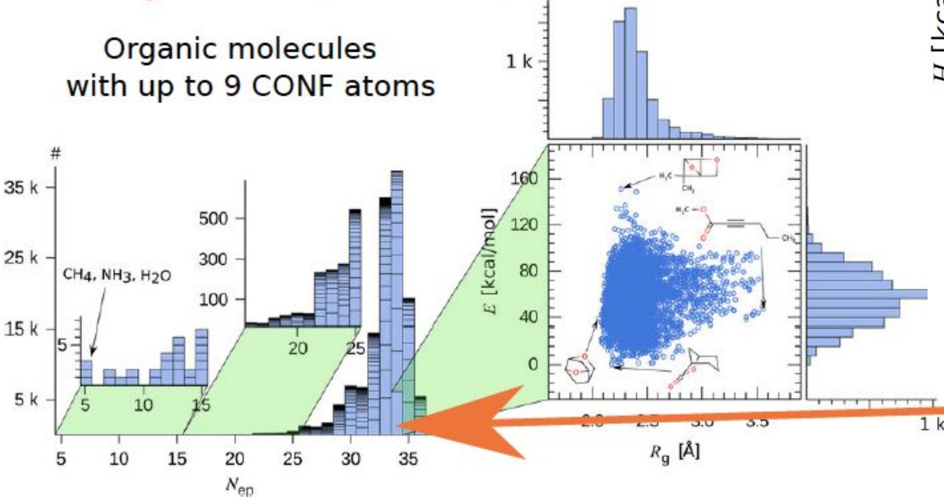
Featured Data Descriptor

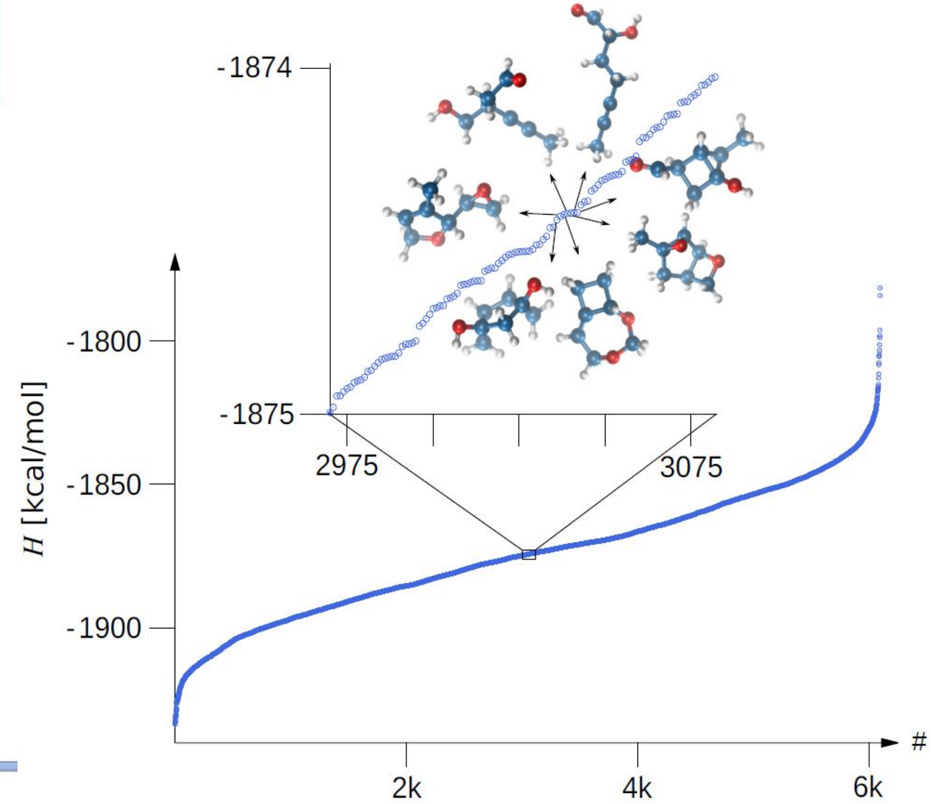
Quantum chemistry structures and properties of 134 kilo molecules

Ramakrishnan *et al.*

Data Descriptor | 05 August 2014

The authors calculate quantum properties for 134,000 small organic molecules, helping map a vast chemical space that includes important molecules such as small amino acids, nucleobases and various

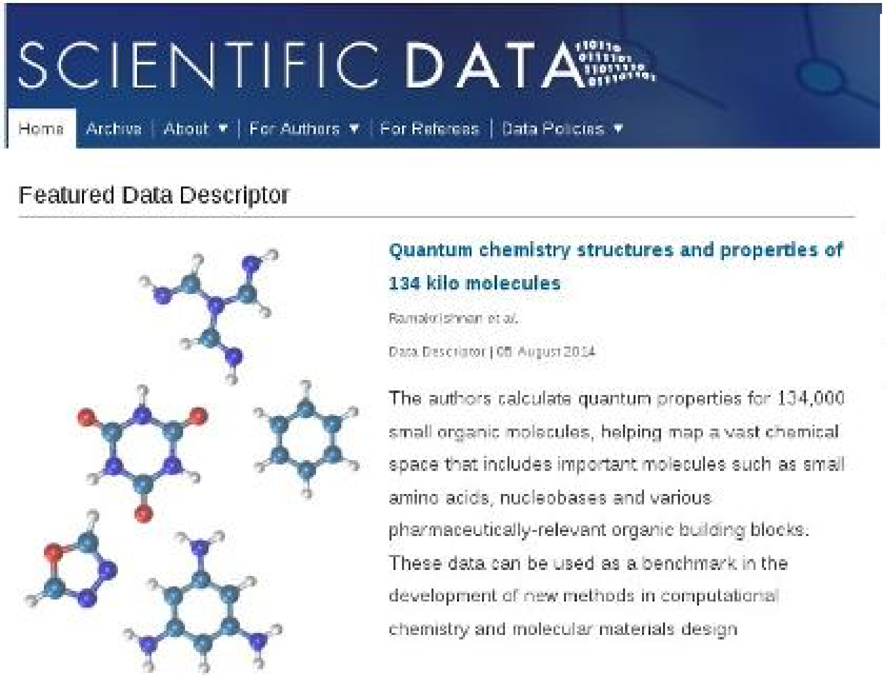


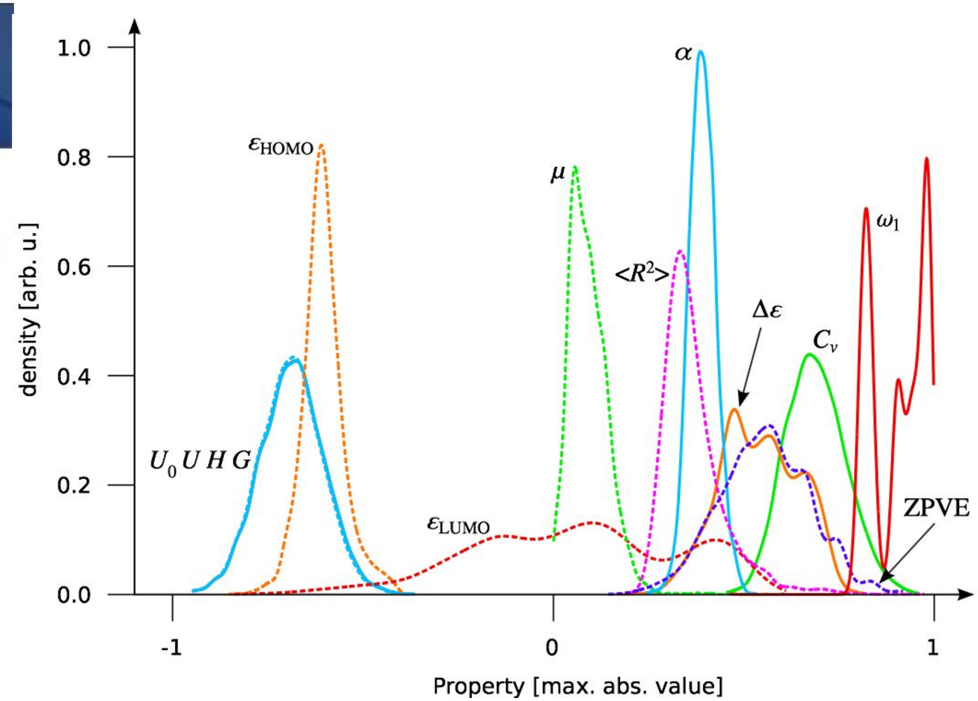


6k $C_7O_2H_{10}$
constitutional isomers

``Enumeration surpasses imagination''
J.-L. Reymond

Ramakrishnan et al, *Scientific Data* (2014)





Ramakrishnan et al, *Scientific Data* (2014)

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

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

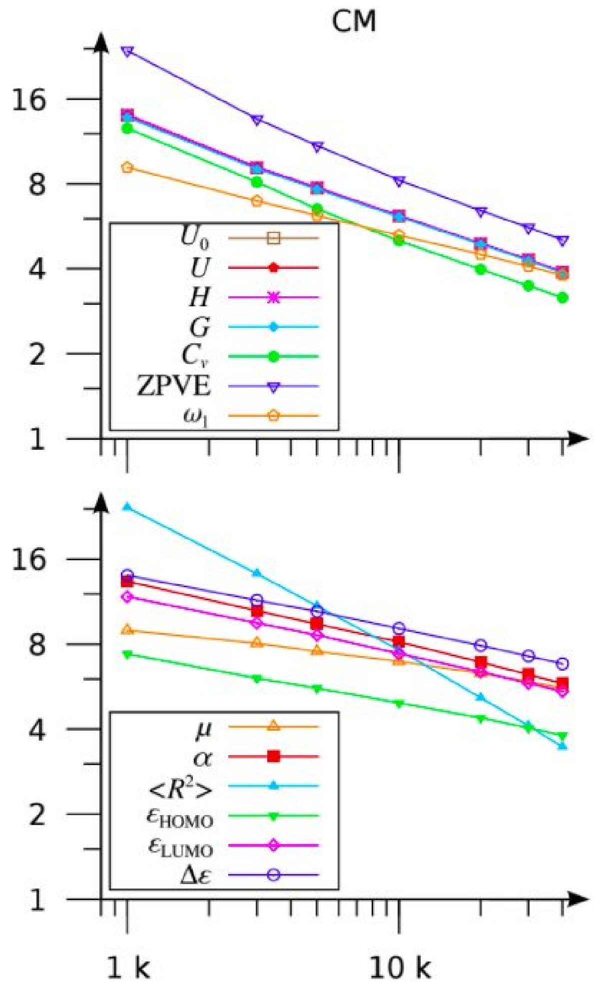
We set $\lambda = 0$ ...

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

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

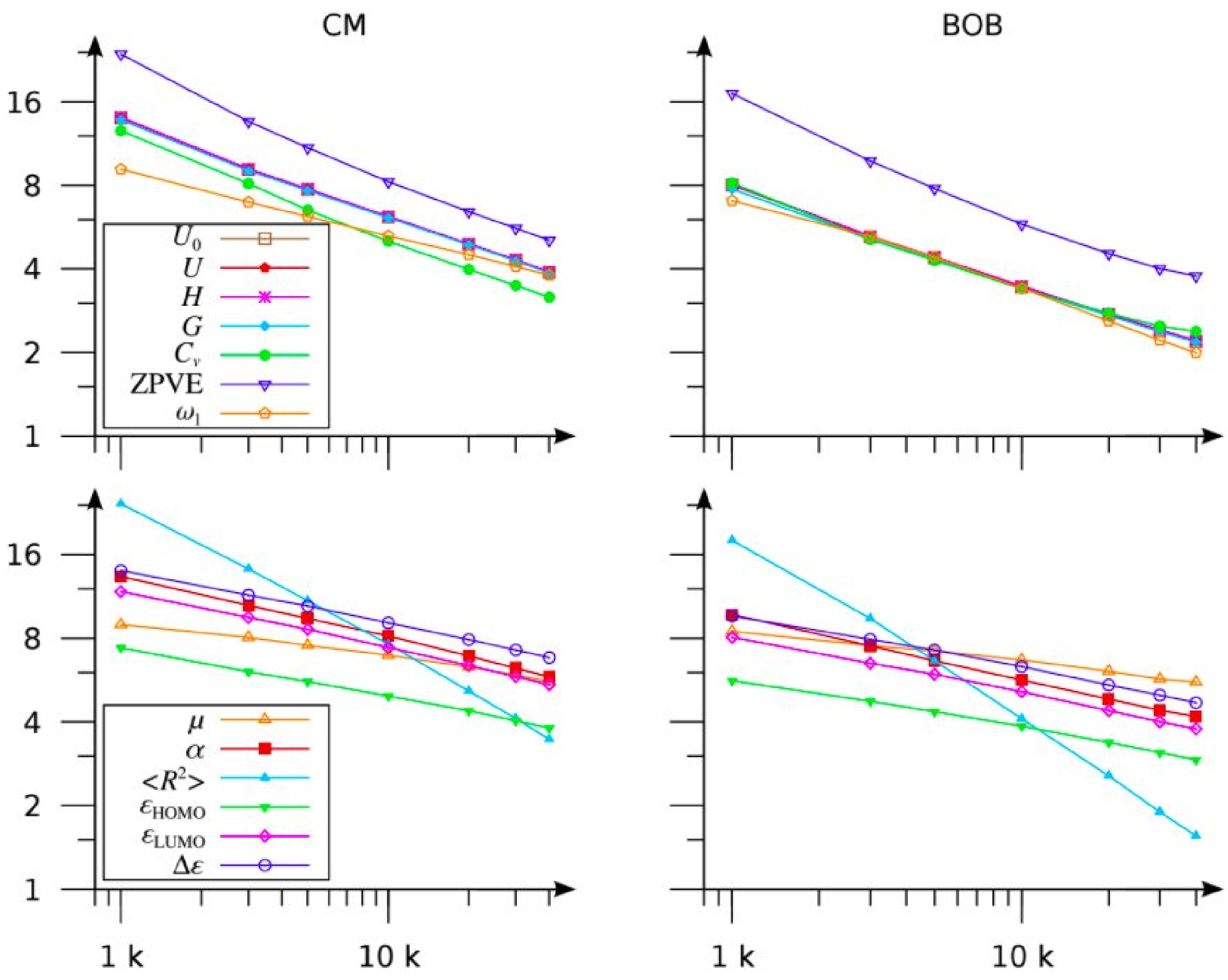
$$k_{ij} = e^{-D_{ij}/\sigma}$$
$$\frac{1}{2} \leq k_{ij} \leq 1$$
$$\sigma_{\mathrm{opt}} = D_{ij}^{\max} / \log(2)$$

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



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

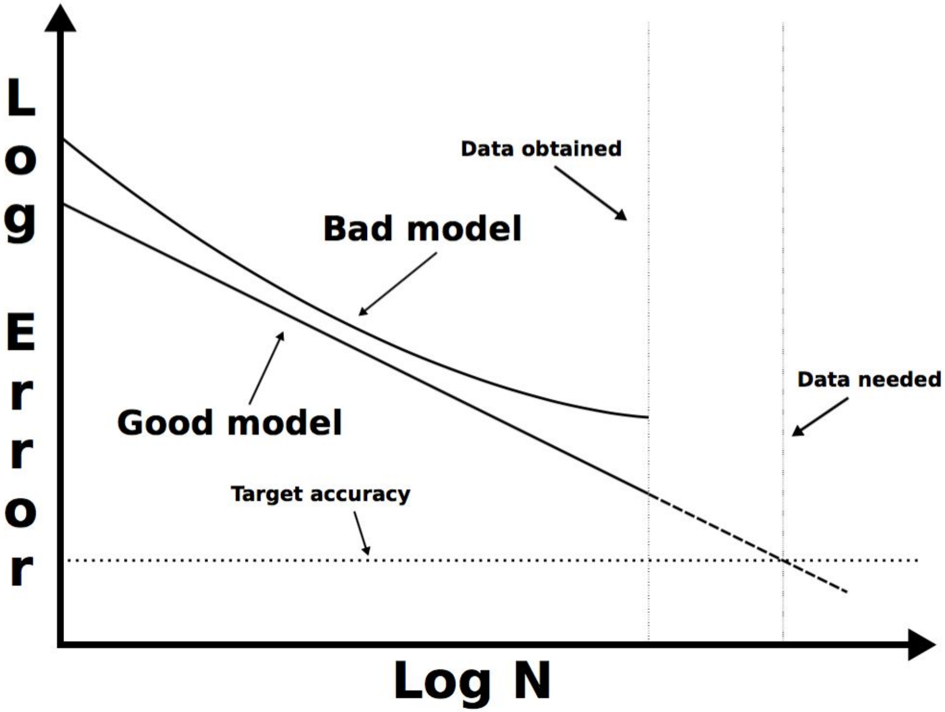
Ramakrishnan, OAvL, *CHIMIA* (2015), arXiv

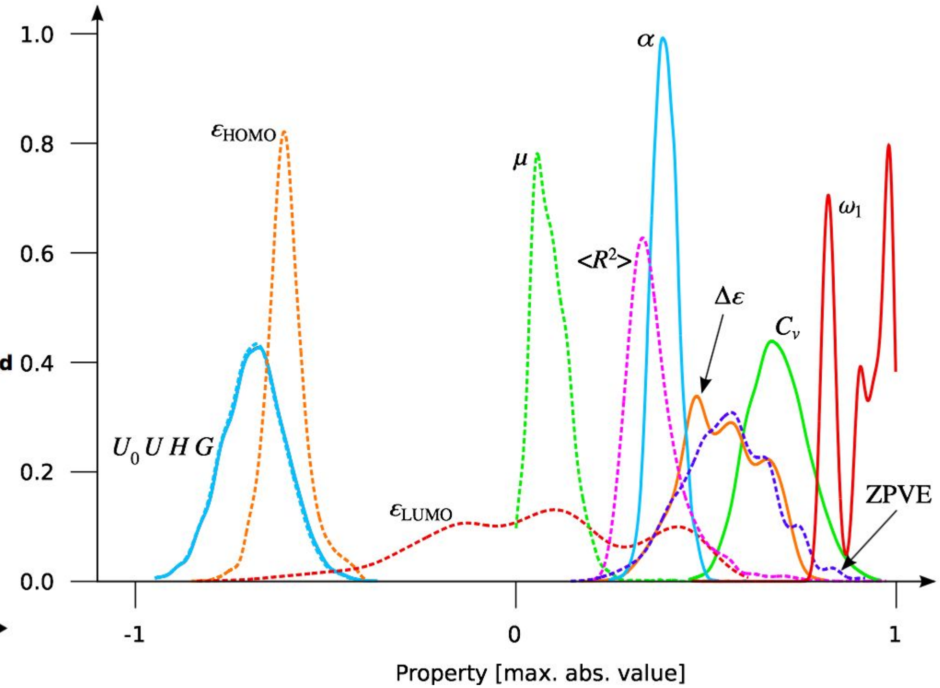


Tested on 134k-*N* organic molecules taken from:
Ramakrishnan et al, *Scientific Data* (2014)

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

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





$$\boxed{\begin{aligned} P^{\text{est}}(\mathbf{M}) &= \textstyle\sum_i \alpha_i k(\mathbf{M}, \mathbf{M}_i) \\ \vec{\alpha} &= \mathbf{K}^{-1} \vec{P}^{\text{ref}} \end{aligned}}$$

$$\sigma = \max\{|\mathbf{d}_i - \mathbf{d}_j|\} / \log(2)$$

$$E_q = \langle \Psi_q | \hat{H} | \Psi_q \rangle$$

$$O_q = \langle \Psi_q | \hat{O} | \Psi_q \rangle$$

Error ~ a/N^b

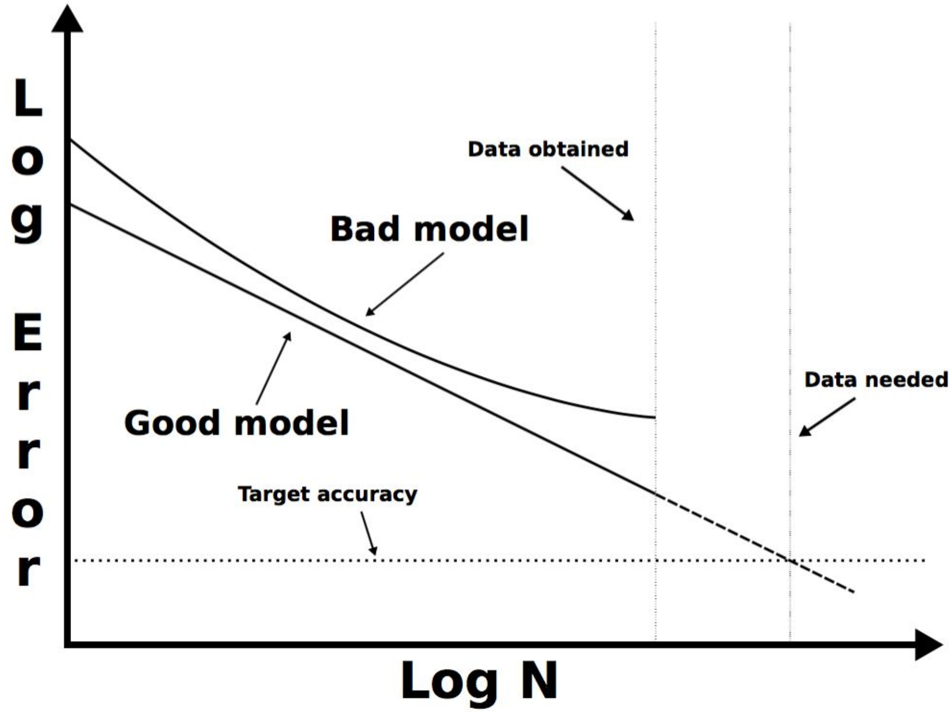
K.-R. Mueller et al, *Neural Comput* (1996)

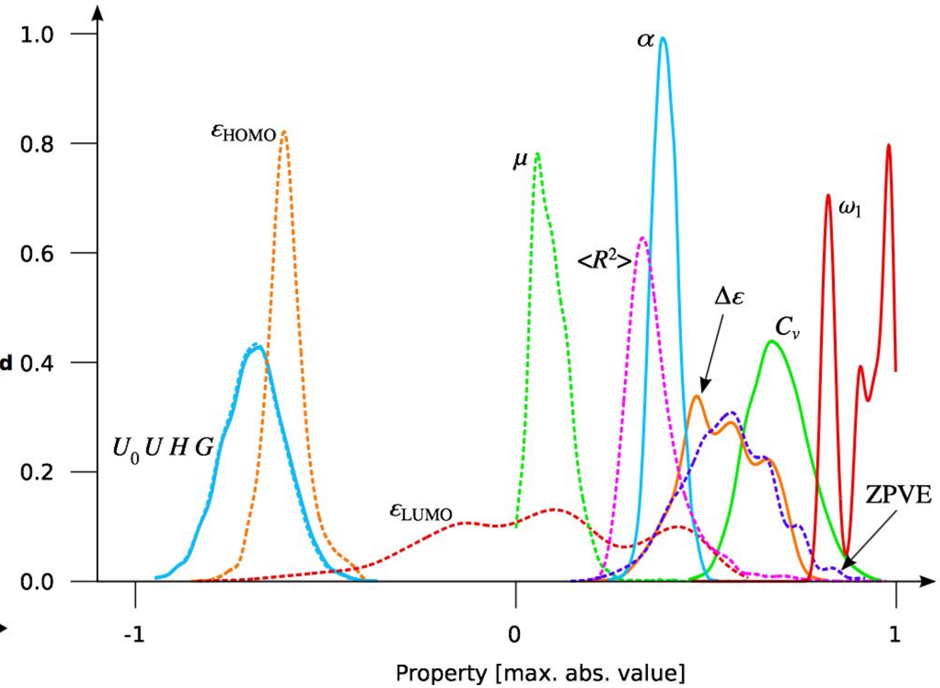
→ log(Error) = log(a) - $b$log(N)

K ~ Ψ

α ~ Ô

Ramakrishnan, OAvL, *CHIMIA* (2015)





$$\boxed{\begin{array}{c} P^{\mathrm{est}}(\mathbf{M}) = \sum_i \alpha_i k(\mathbf{M}, \mathbf{M}_i) \\ \vec{\alpha} = \mathbf{K}^{-1} \vec{P}^{\mathrm{ref}} \end{array}}$$

$$\sigma = \max\{|\mathbf{d}_i - \mathbf{d}_j|\} / \log(2)$$

Error ~ $a/(N')^b$, e.g. $N' = N/x$

K.-R. Mueller et al, *Neural Comput* (1996)

$\rightarrow \log(\mathrm{Error}) = \log(a) + 3x - b\log(N)$

Possible reasons for large a and $b \rightarrow 0$

1. No cause and effect relationship (spurious)

2. Bad representation (no physics/uniqueness …)

3. Bad data (noisy/not representative/ …)

4. Bad regressor: Underfitting (too rigid)/Overfitting (``crazy'' assumptions)/Unconverged (Less coefficients than data points)

5. High dimensionality and curvature

More ways to be wrong than right

Conclusions

1. Instantaneous QM quality predictions
2. Learning curves reveal quality of ML model
3. Representations
4. Data sets

Kernel Ridge Regression

Kernel

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

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

Regression

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

Solution

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

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, $H_1 \to d$ and $H_2 \to d$.

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 ingore 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).

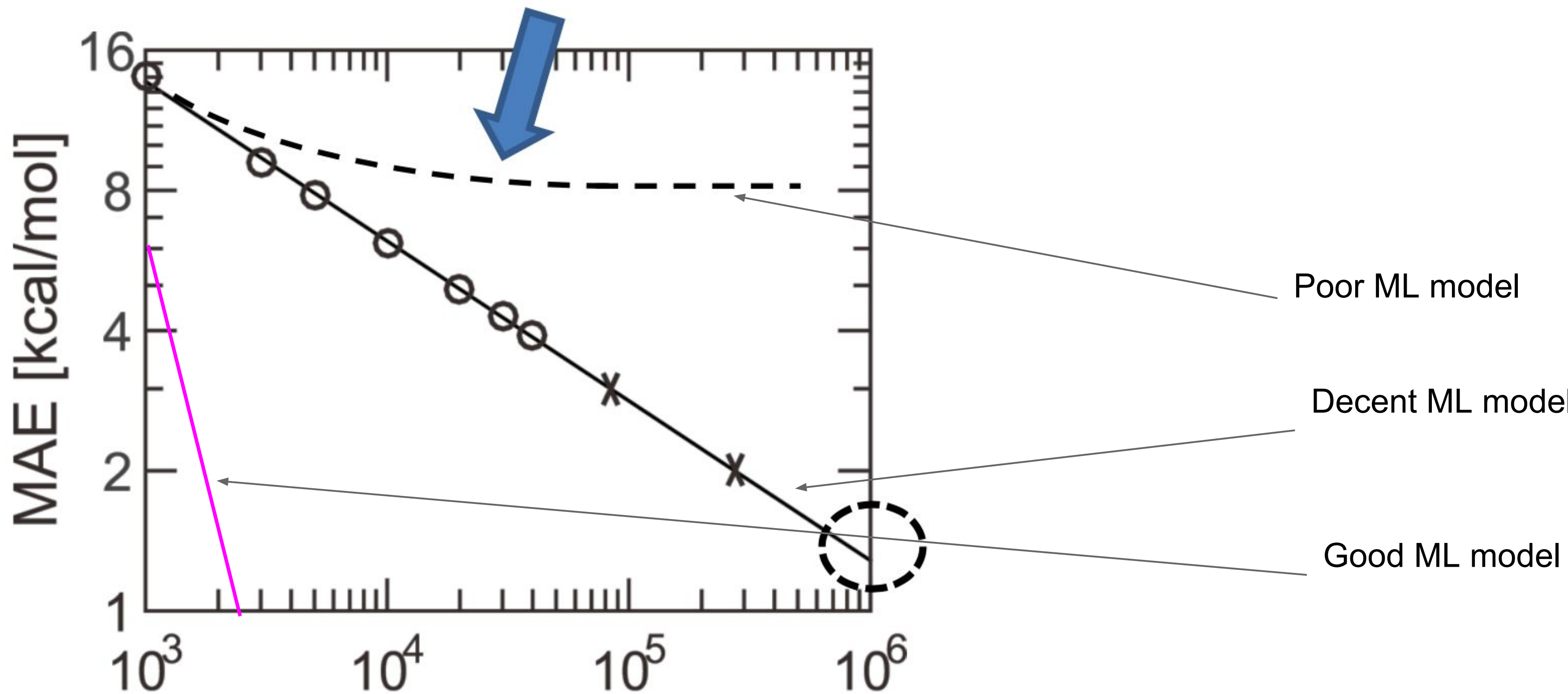
lack of uniqueness → absurd results → noise in training

OAvL et al, *IJQC* (2013)

$$\log(\mathrm{Error}) = a - b\log(N)$$

uniqueness

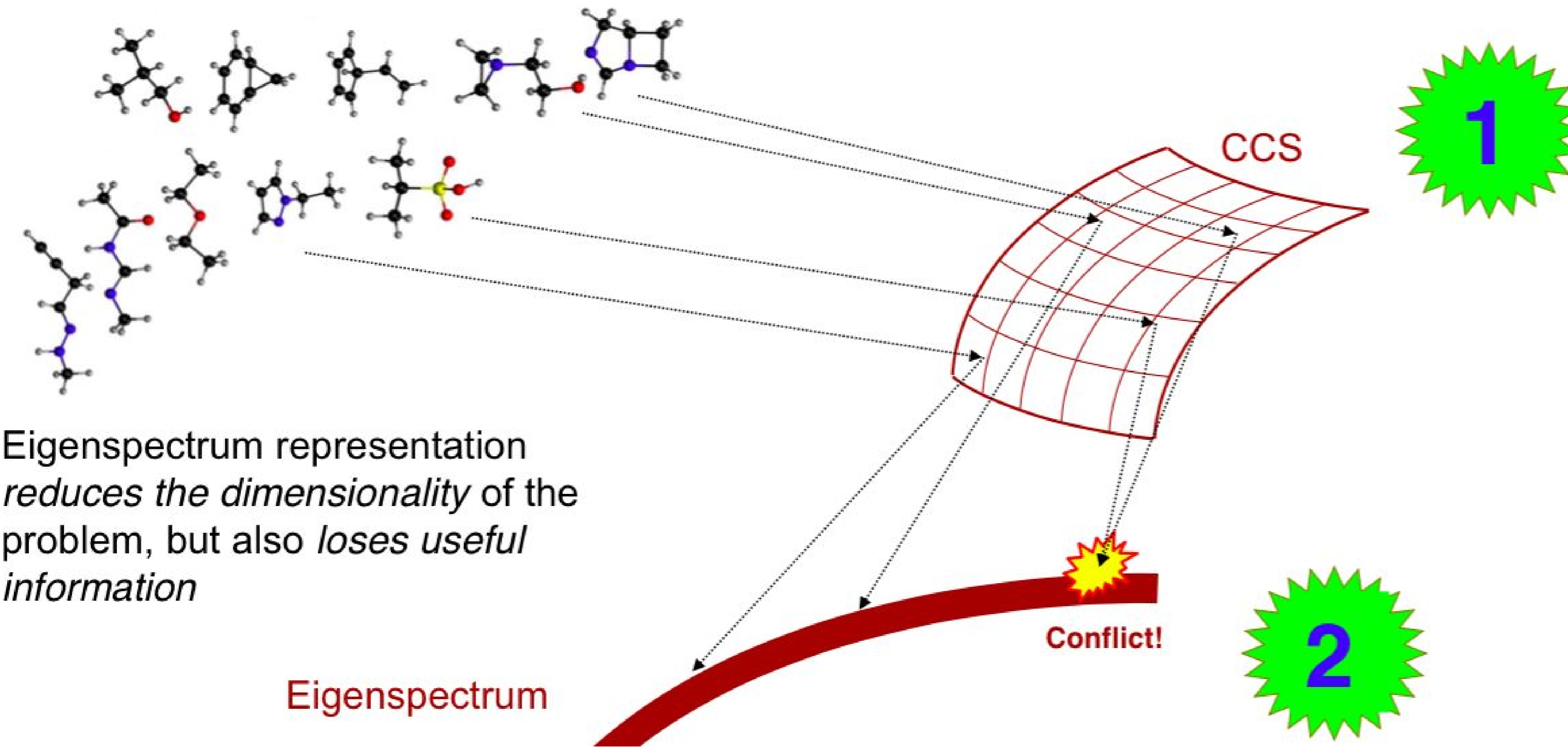
lack of uniqueness



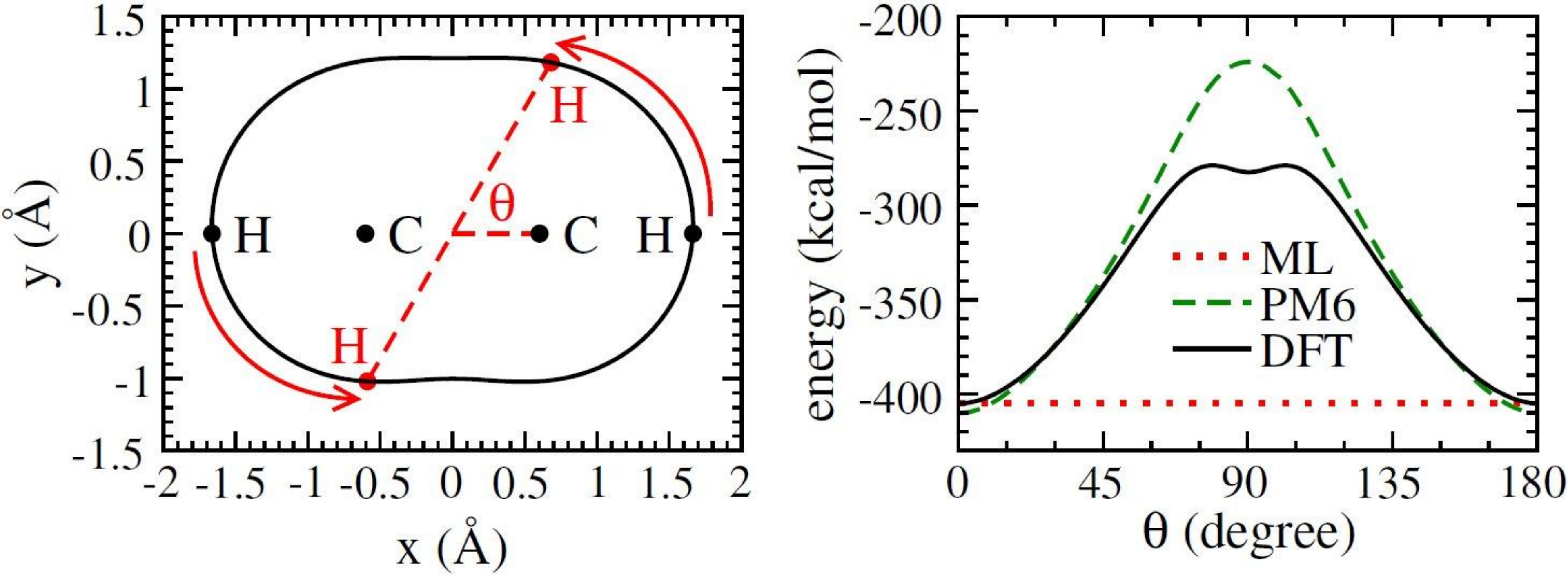
lack of uniqueness → absurd results → noise in training

OAvL et al, *IJQC* (2013)

Huang, OAvL, *J Chem Phys Comm* (2016) arxiv.org/abs/1608.06194



Eigenspectrum representation *reduces the dimensionality* of the problem, but also *loses useful information*



J. E. Moussa, *Phys Rev Lett* (2012)

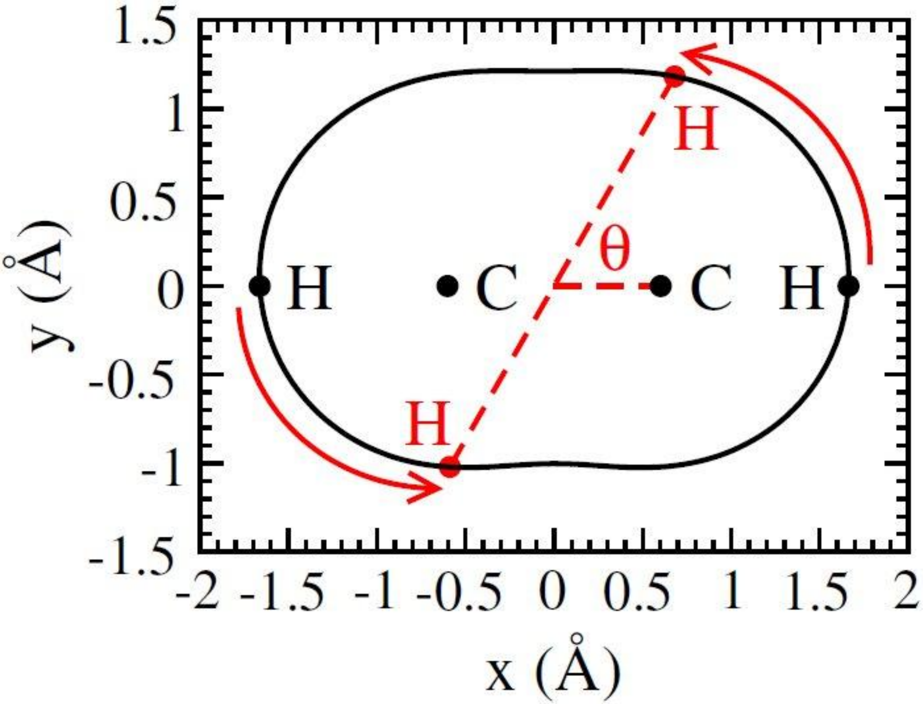
$$M_{IJ} = \begin{cases} 0.5 Z_I^{2.4} & \forall \ I = J, \\ \frac{Z_I Z_J}{|\mathbf{R}_I - \mathbf{R}_J|} & \forall \ I \neq J. \end{cases}$$

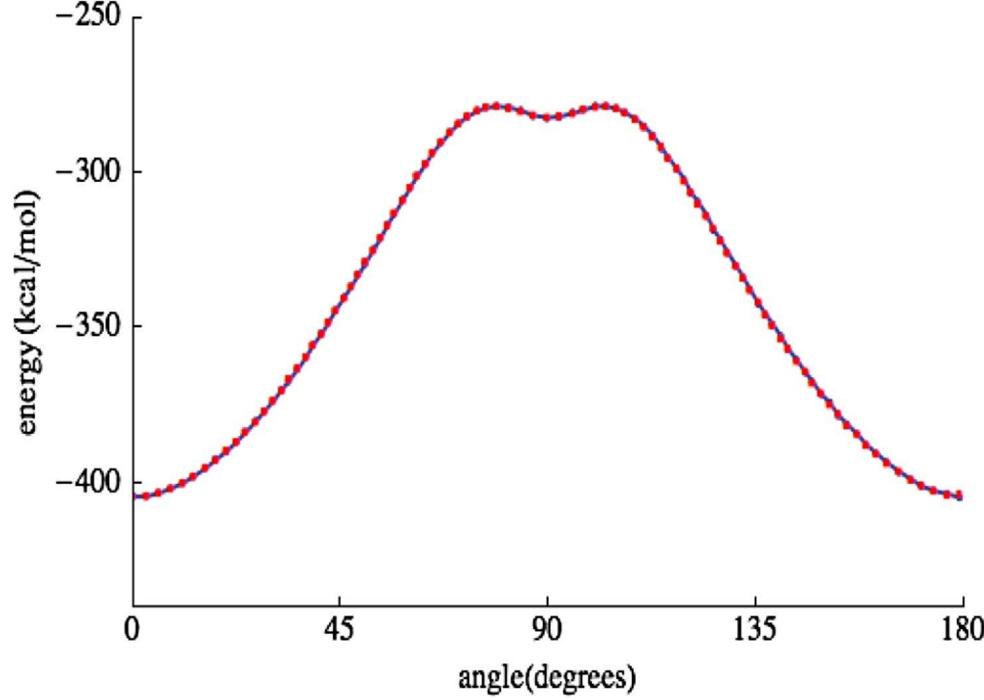
N = 4
-> 3*N-6 = 6 degrees of freedom

Coulomb-matrix

- unique
- translation
- rotation
- symmetry
- diagonalize
- fill up w zeros

???





J. E. Moussa, *Phys Rev Lett* (2012)

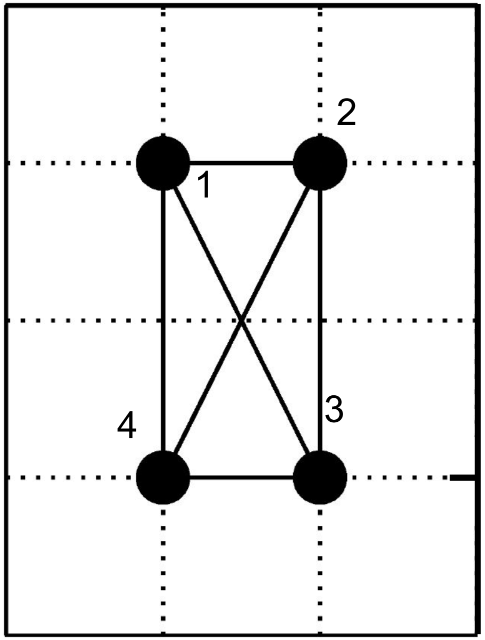
$$M_{IJ} = \begin{cases} 0.5 Z_I^{2.4} & \forall \; I = J, \\ \frac{Z_I Z_J}{|\mathbf{R}_I - \mathbf{R}_J|} & \forall \; I \neq J. \end{cases}$$

Coulomb-matrix

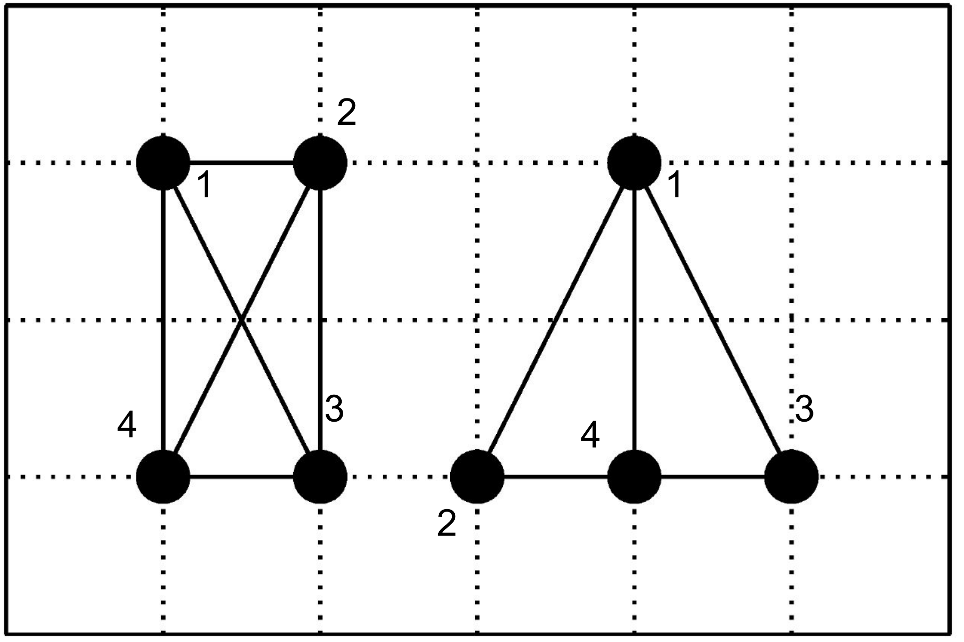
- unique
- translation
- rotation
- symmetry
- ~~diagonalize~~ sort
- fill up w zeros

N = 4
-> 3*N-6 = 6 degrees of freedom

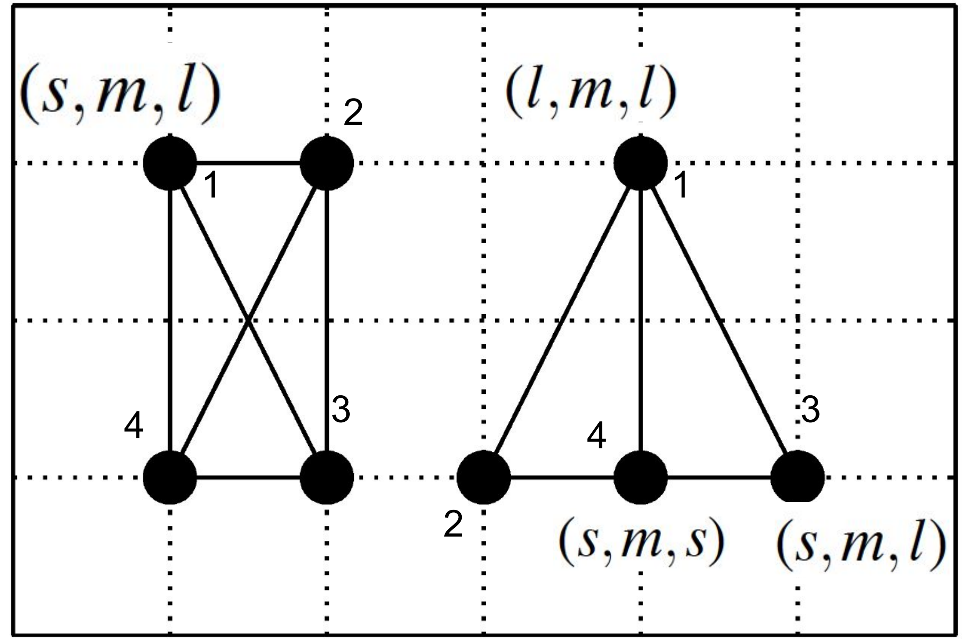
Homometric molecules?



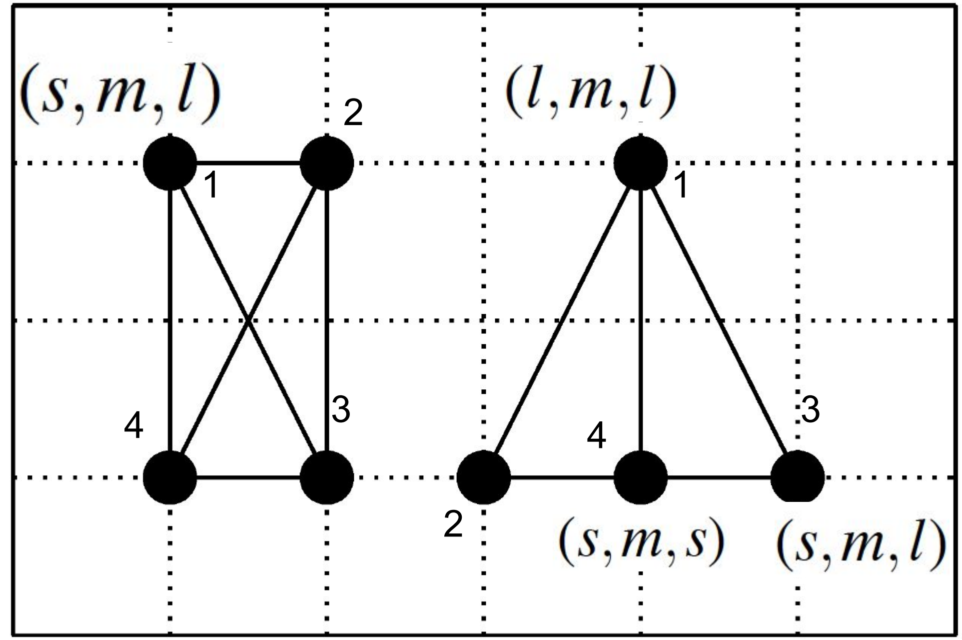
Homometric molecules?



Homometric molecules?



Homometric molecules?



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

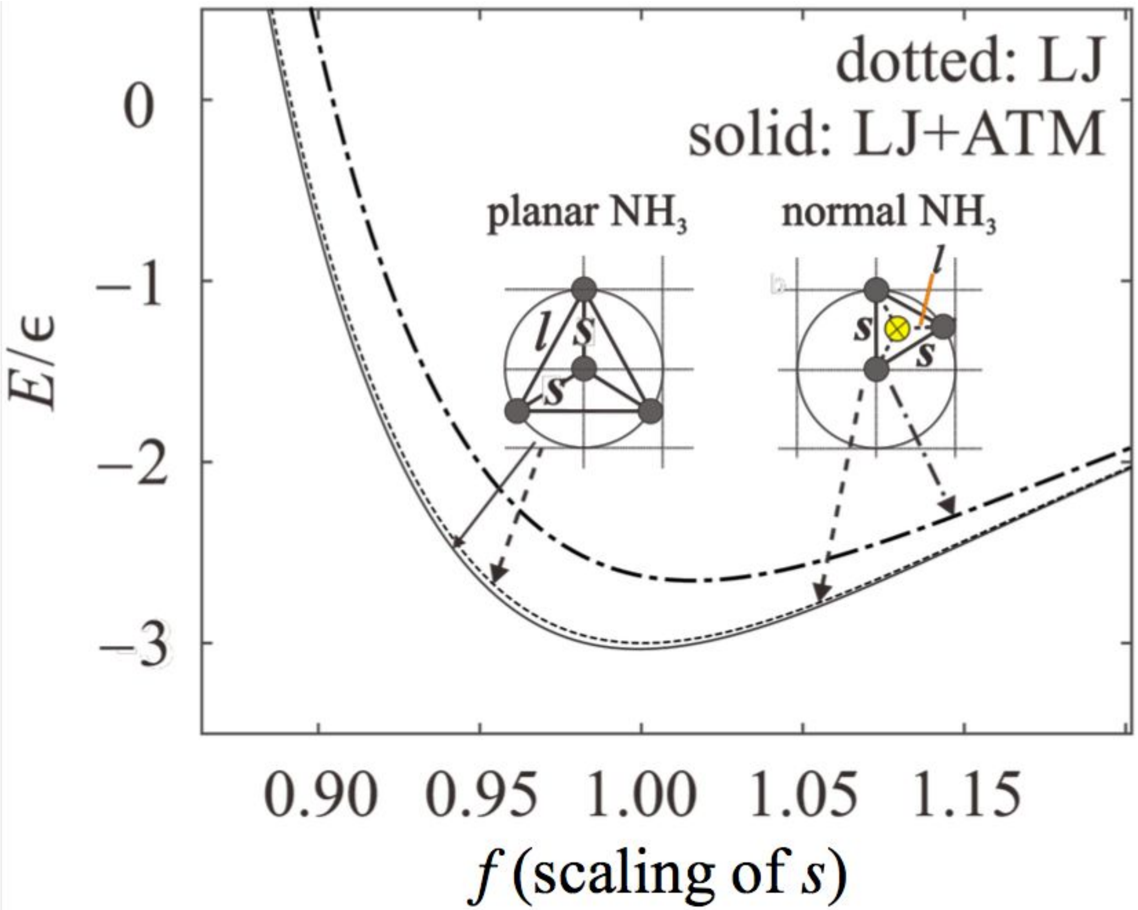
	s	l	m
s		m	l
l	m		s
m	l	s	

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

Learning curves

$$\log(\text{Error}) = a - b\log(N)$$

uniqueness



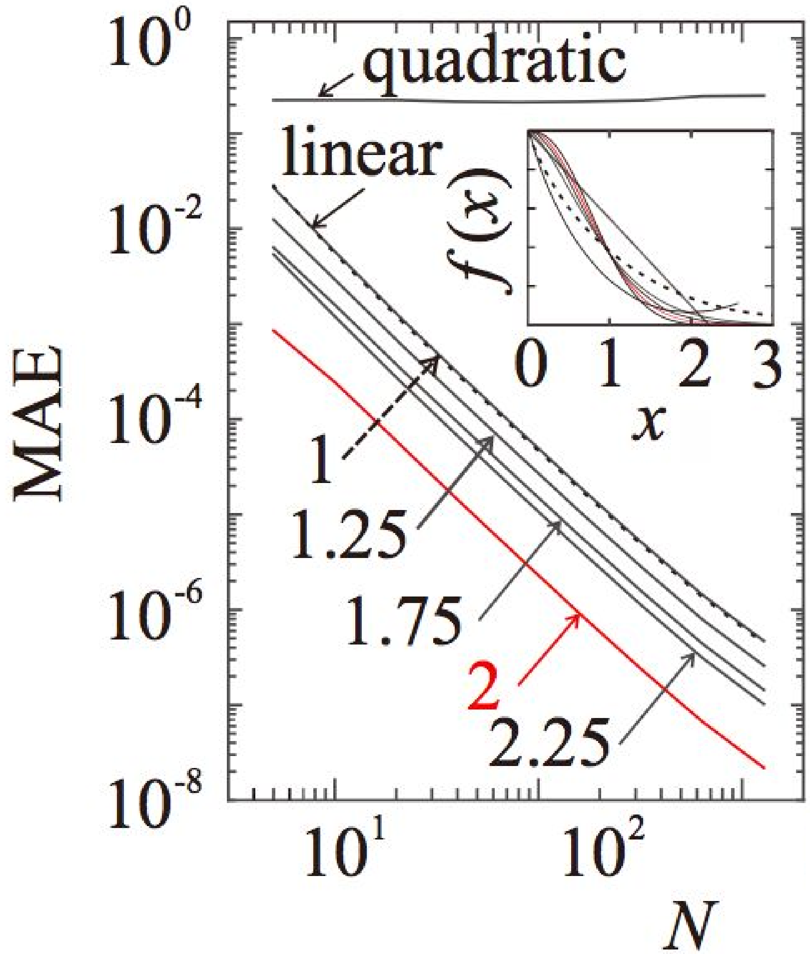
LJ: Lennard-Jones 2-body vdW potential
ATM: Axilrod-Teller-Muto 3-body vdW potential

Huang, OAvL, *J Chem Phys Comm* (2016) arxiv.org/abs/1608.06194

$$\log(\mathrm{Error}) = a - b\log(N)$$

target
similarity

$$f^{\mathrm{est}}(x) = \sum \alpha_i k(\underbrace{ax_i + b}_{\mathrm{M}_i}, \underbrace{ax + b}_{\mathrm{M}})$$



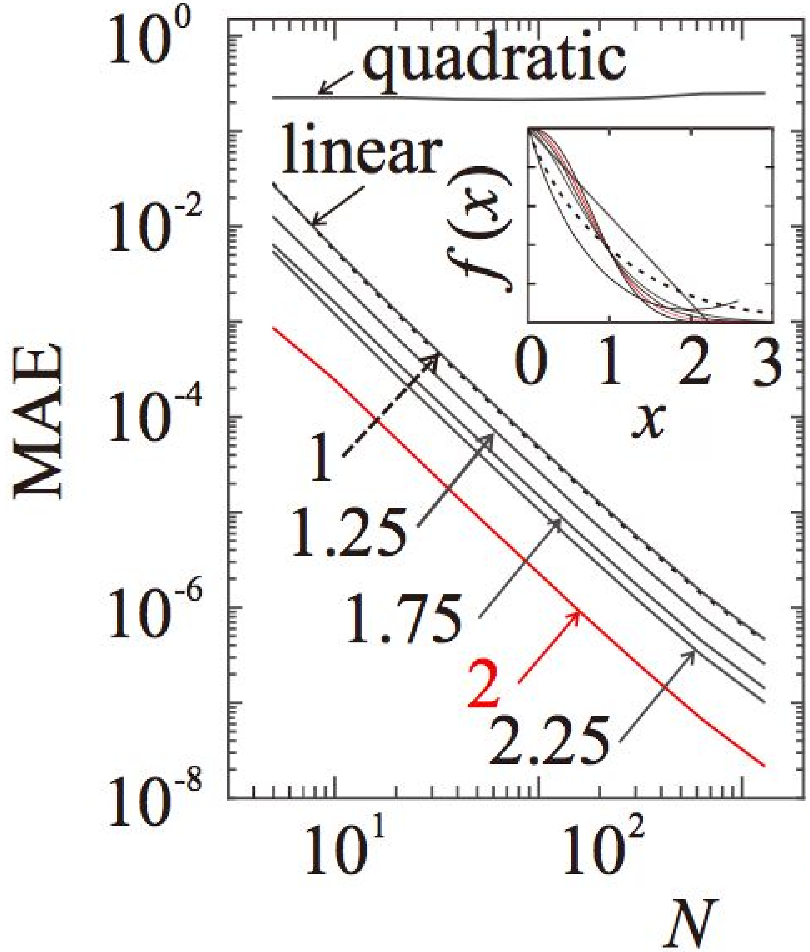
Huang, OAvL, *J Chem Phys Comm* (2016) arxiv.org/abs/1608.06194

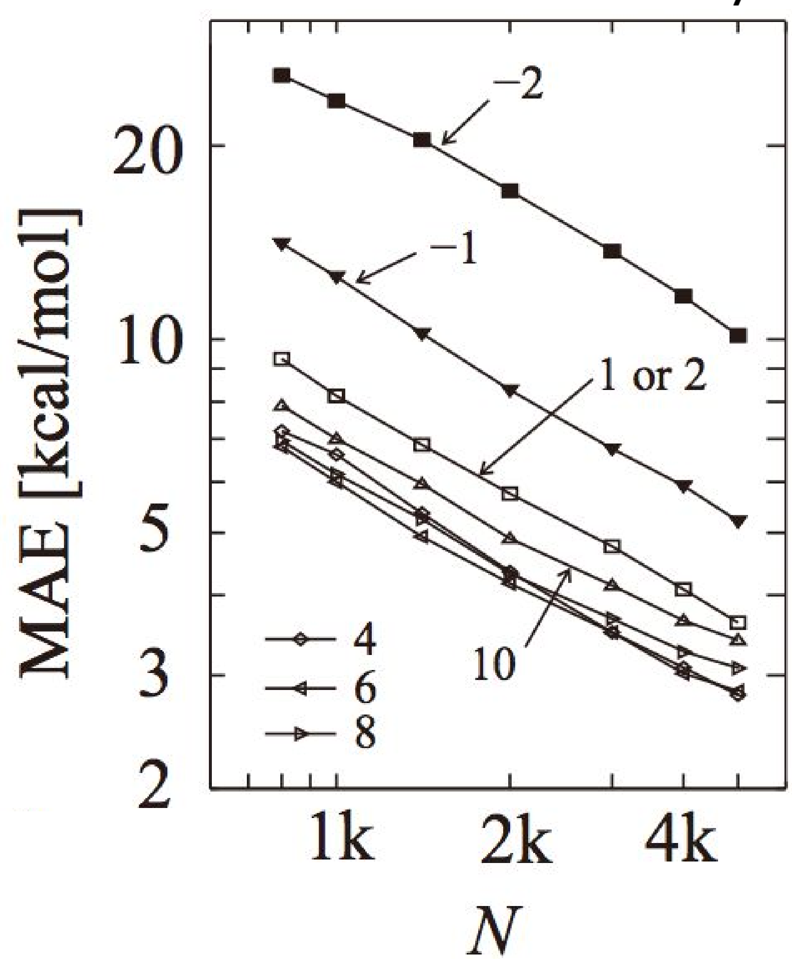
$$\log(\text{Error}) = a - b\log(N)$$

target similarity

$$\mathrm{CM}^{(n)}_{IJ} = \frac{Z_I Z_J}{R^n_{IJ}}$$

$$f^{\text{est}}(x) = \sum \alpha_i k(\underbrace{ax_i+b}_{\mathrm{M}_i}, \underbrace{ax+b}_{\mathrm{M}})$$



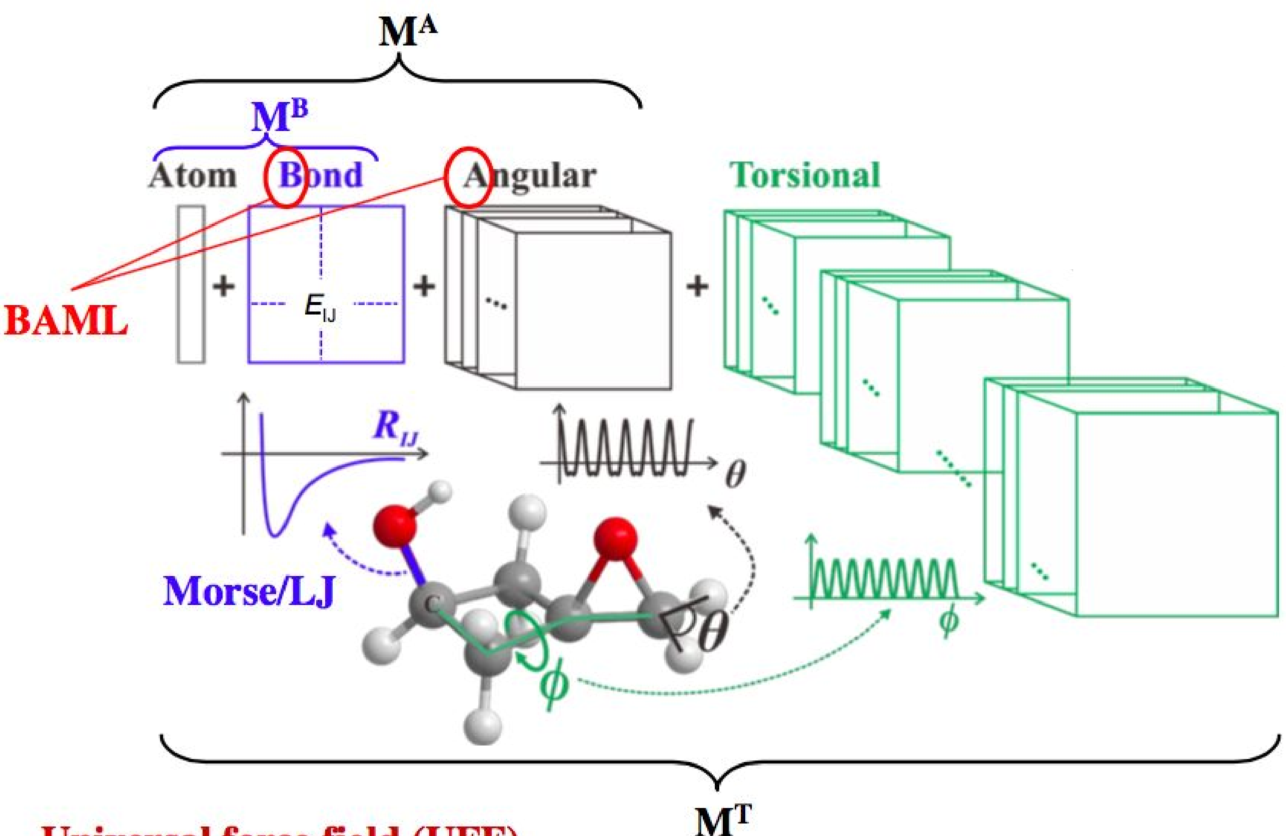


Huang, OAvL, *J Chem Phys Comm* (2016) arxiv.org/abs/1608.06194

BAML

Approach: best M is unique AND good model

bags of UFF contributions



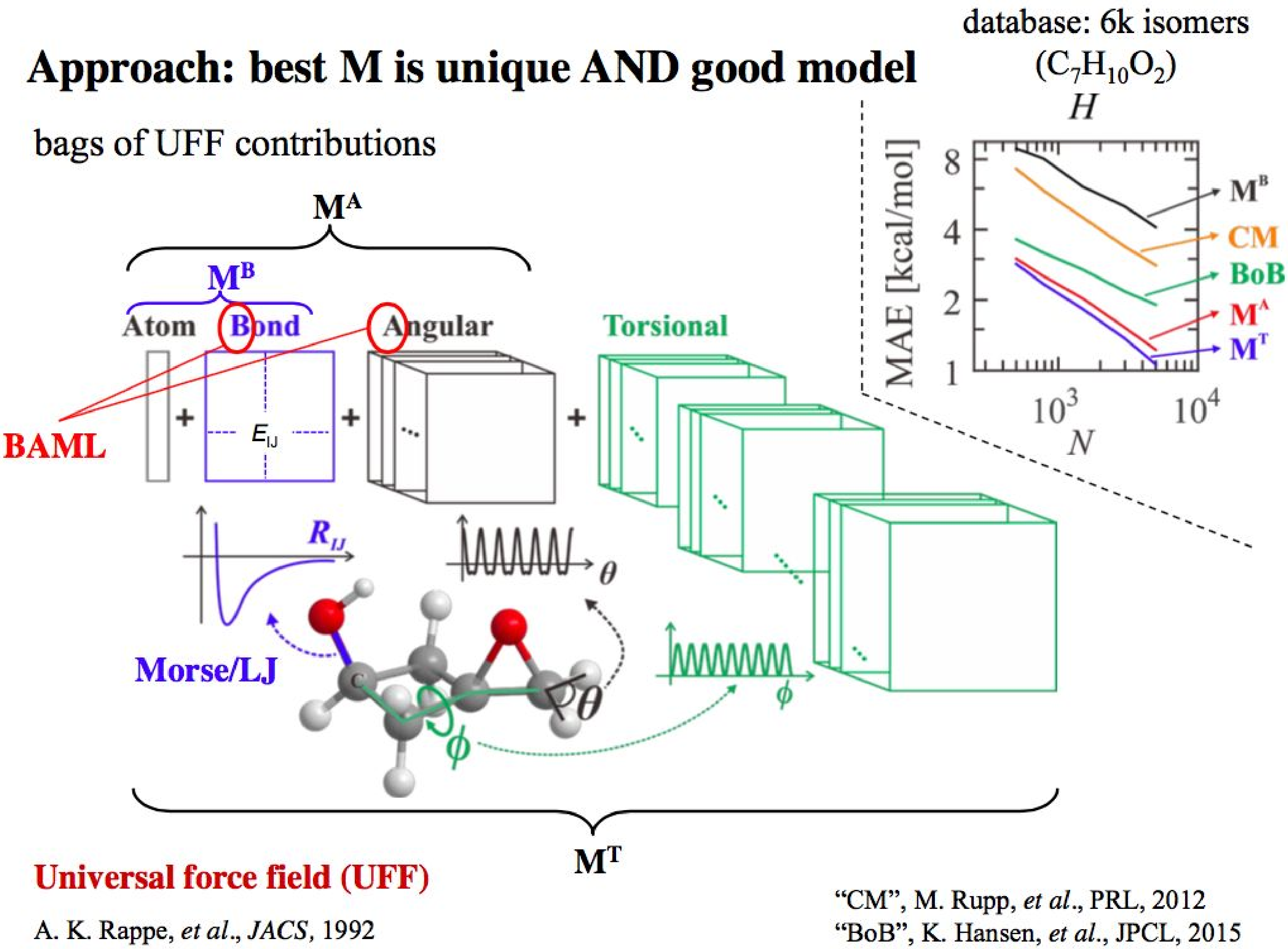
Universal force field (UFF)

A. K. Rappe, *et al*., *JACS*, 1992

"CM", M. Rupp, *et al*., PRL, 2012
"BoB", K. Hansen, *et al*., JPCL, 2015

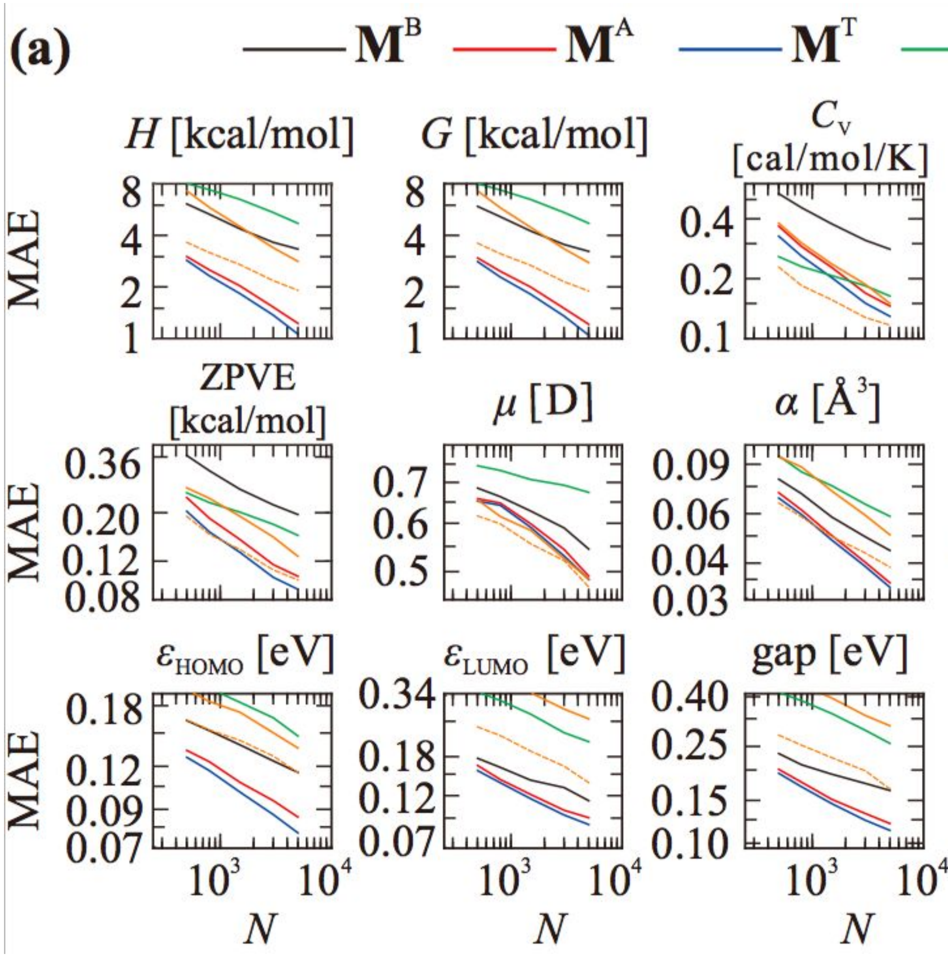
Huang, OAvL, *J Chem Phys Comm* (2016) arxiv.org/abs/1608.06194

BAML



Huang, OAvL, *J Chem Phys Comm* (2016) arxiv.org/abs/1608.06194

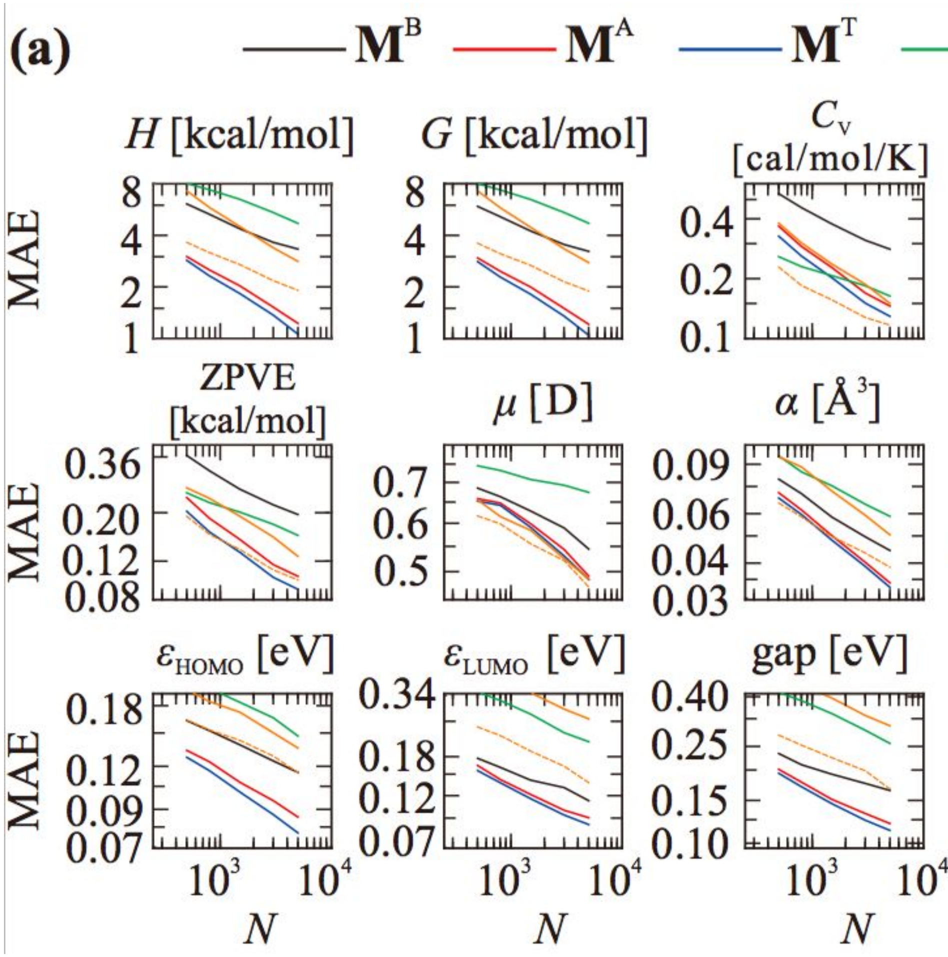
BAML



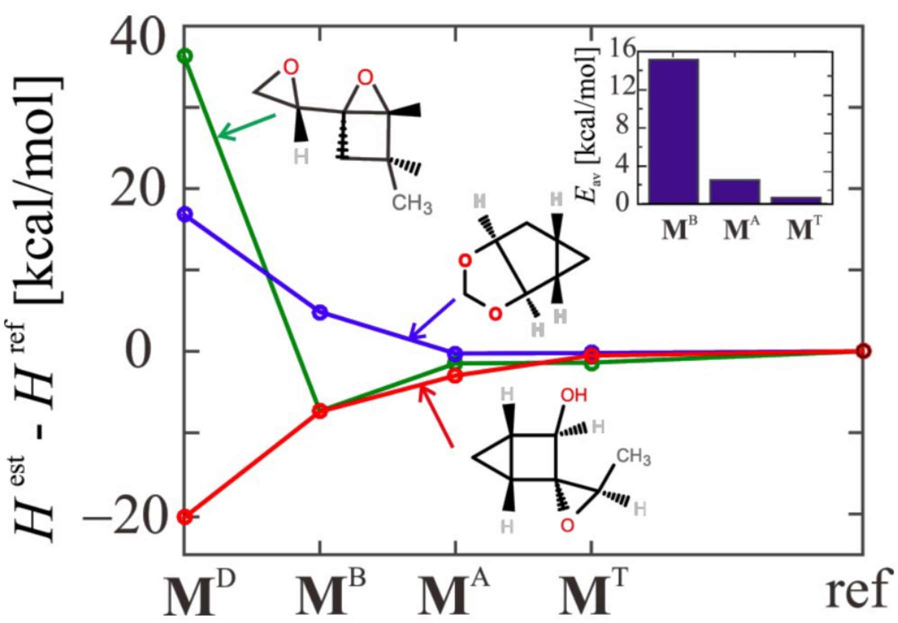
6k constitutional isomers of $C_7O_2H_{10}$

Huang, OAvL, *J Chem Phys Comm* (2016) arxiv.org/abs/1608.06194

BAML



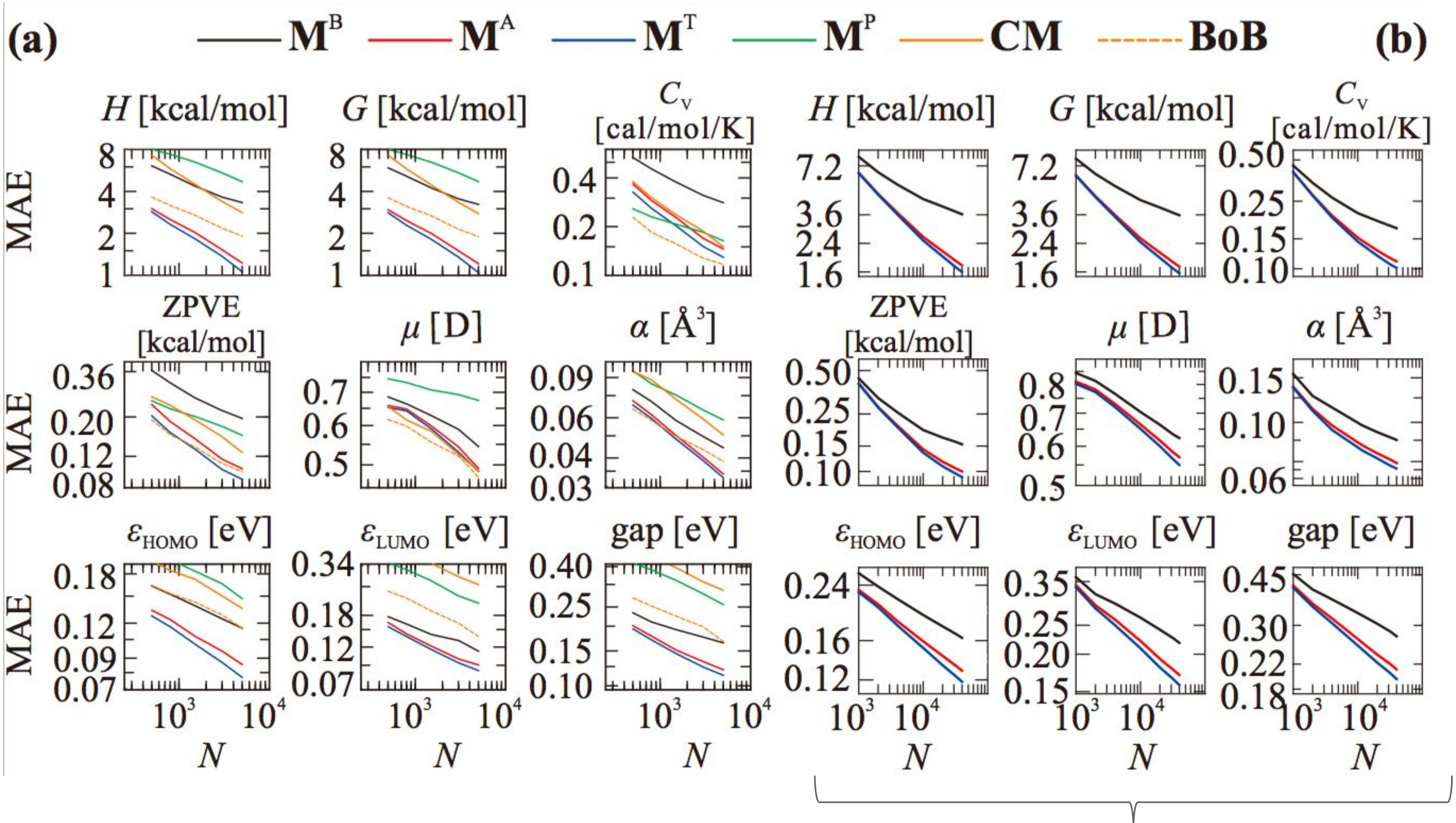
3 outliers



6k constitutional isomers of $C_7O_2H_{10}$

Huang, OAvL, *J Chem Phys Comm* (2016) arxiv.org/abs/1608.06194

BAML



6k constitutional isomers of $C_7O_2H_{10}$

QM9 (134k molecules)

Huang, OAvL, *J Chem Phys Comm* (2016) arxiv.org/abs/1608.06194

BAML

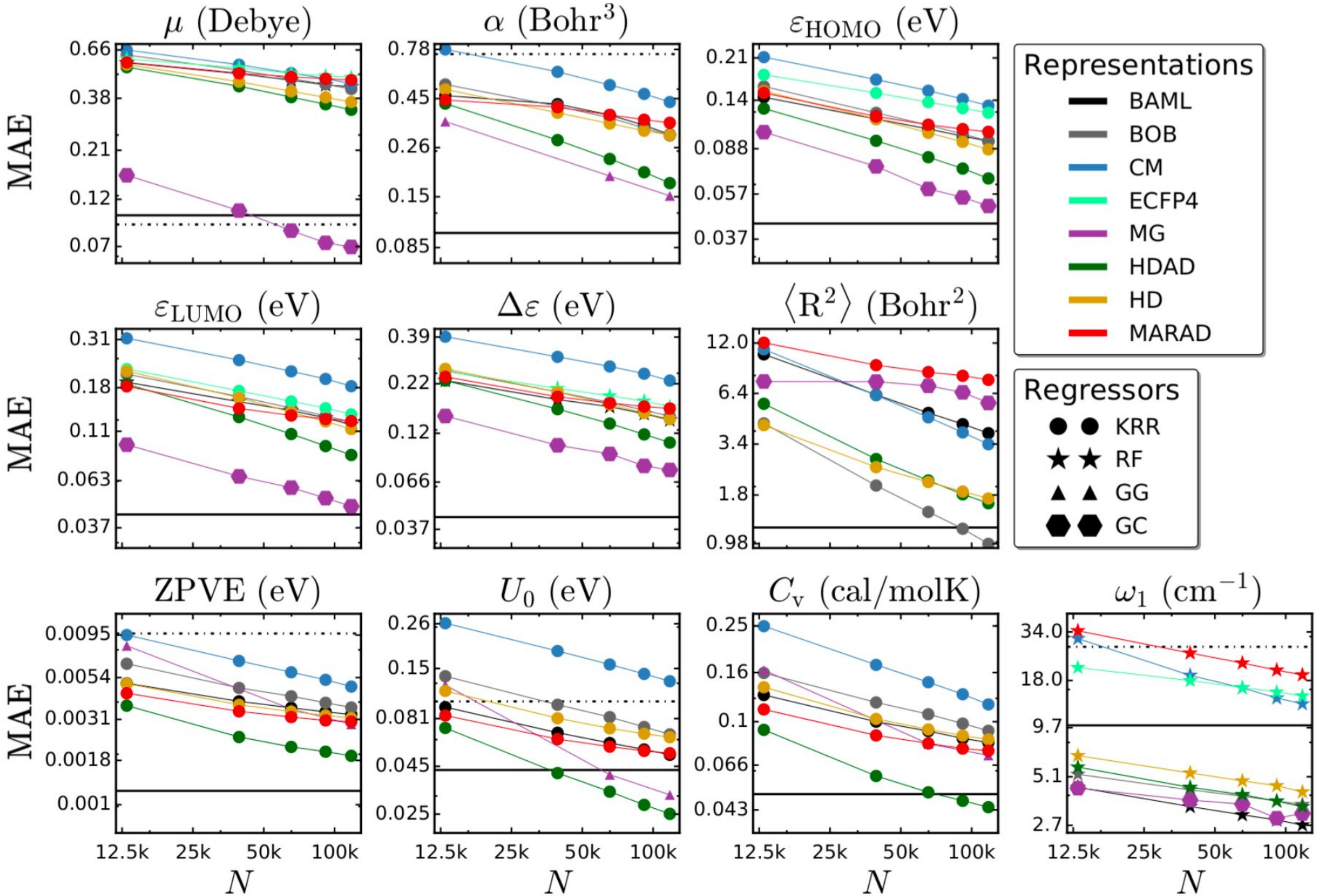
QM7b database (size: 7211)

MAE (5k out-of-sample)

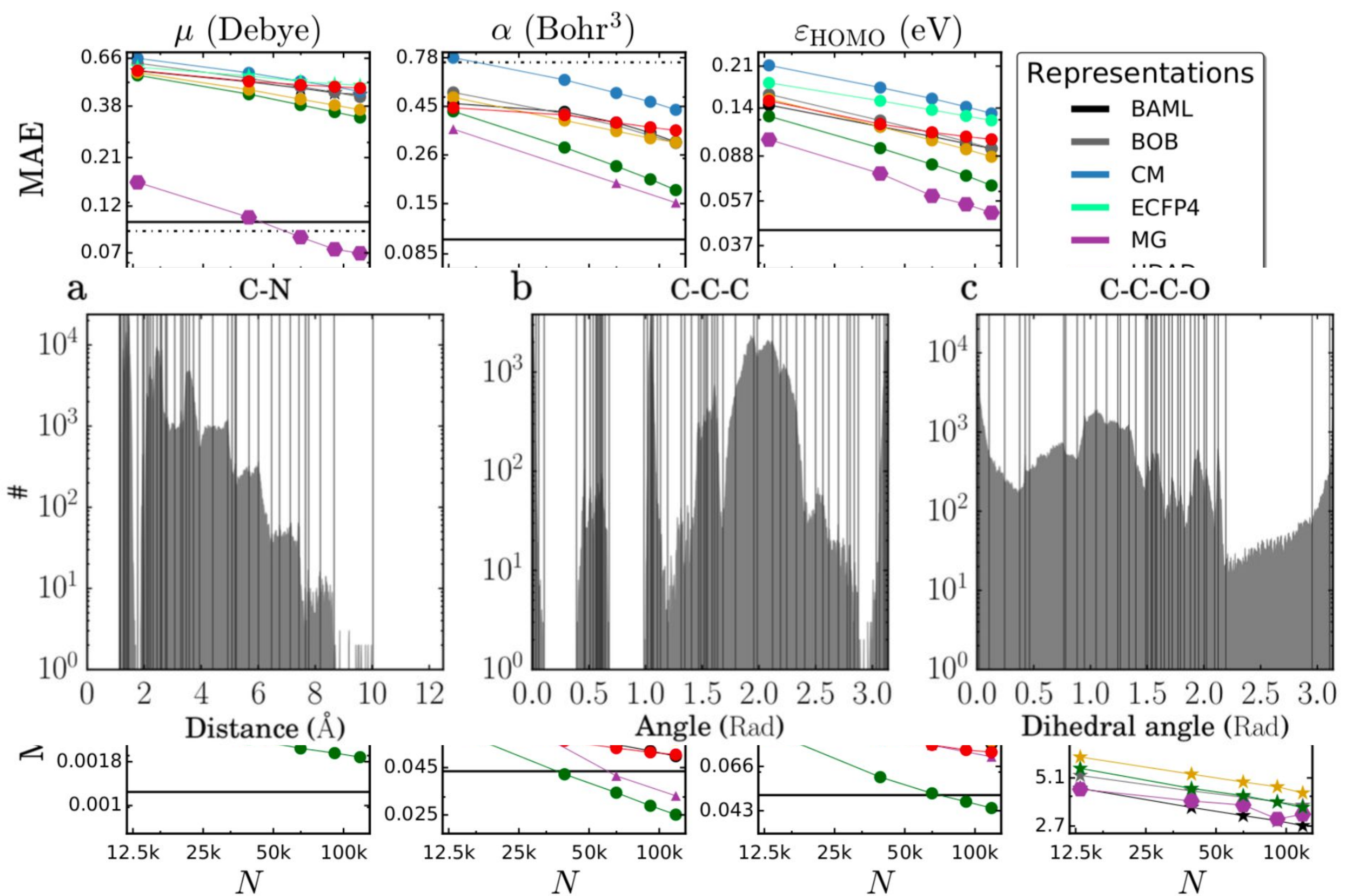
	BAML	BoB	SOAP[a]	CM[b]	accuracy[b]
E (PBE0)/eV	0.05	0.08	**0.04**	0.16	0.15, 0.23, 0.09-0.22
α (PBE0)/ Å^3	0.07	0.09	**0.05**	0.11	0.05-0.27, 0.04-0.14
HOMO (GW)/eV	**0.10**	0.15	0.12	0.16	-
LUMO (GW)/eV	**0.11**	0.16	0.12	0.16	-
IP (ZINDO)/eV	**0.15**	0.20	0.19	0.17	0.20, 0.15
EA (ZINDO)/eV	**0.07**	0.17	0.13	0.11	0.16, 0.11
E_{1st}* (ZINDO)/eV	**0.13**	0.21	0.18	**0.13**	0.18, 0.21

[a] S. De, *et al.*, *PCCP*, 2016
[b] G. Montavon, *et al.*, *NJP*, 2013

Huang, OAvL, *J Chem Phys Comm* (2016) arxiv.org/abs/1608.06194



"Prediction errors of molecular machine learning models lower than hybrid DFT errors", F. A. Faber, L. Hutchison, B. Huang, J. Gilmer, S. S. Schoenholz, G. E. Dahl, O. Vinyals, S. Kearnes, P. F. Riley, OAvL *arxiv.org/abs/1702.05532*

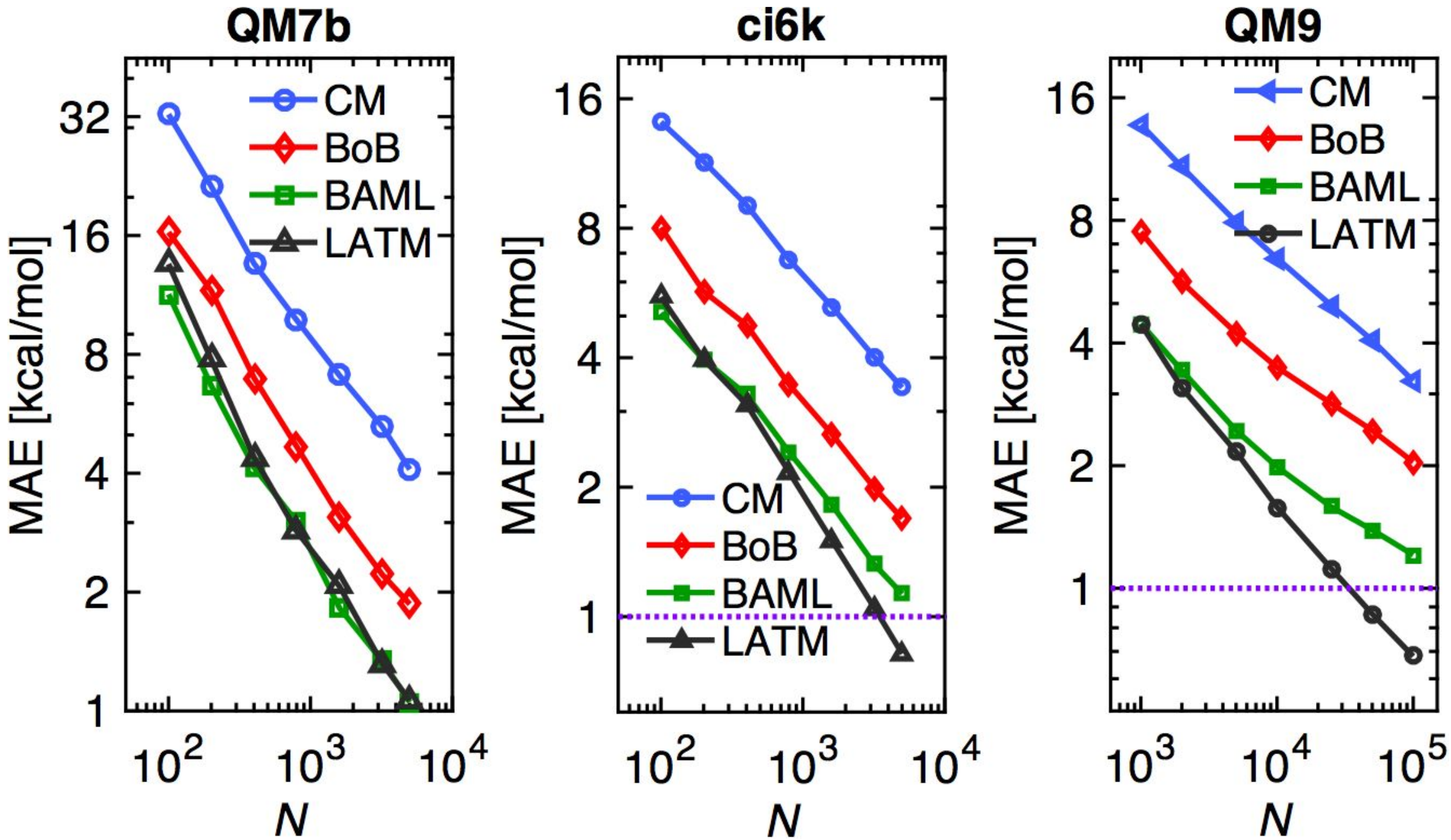


"Prediction errors of molecular machine learning models lower than hybrid DFT errors", F. A. Faber, L. Hutchison, B. Huang, J. Gilmer, S. S. Schoenholz, G. E. Dahl, O. Vinyals, S. Kearnes, P. F. Riley, OAvL *arxiv.org/abs/1702.05532*

LATM

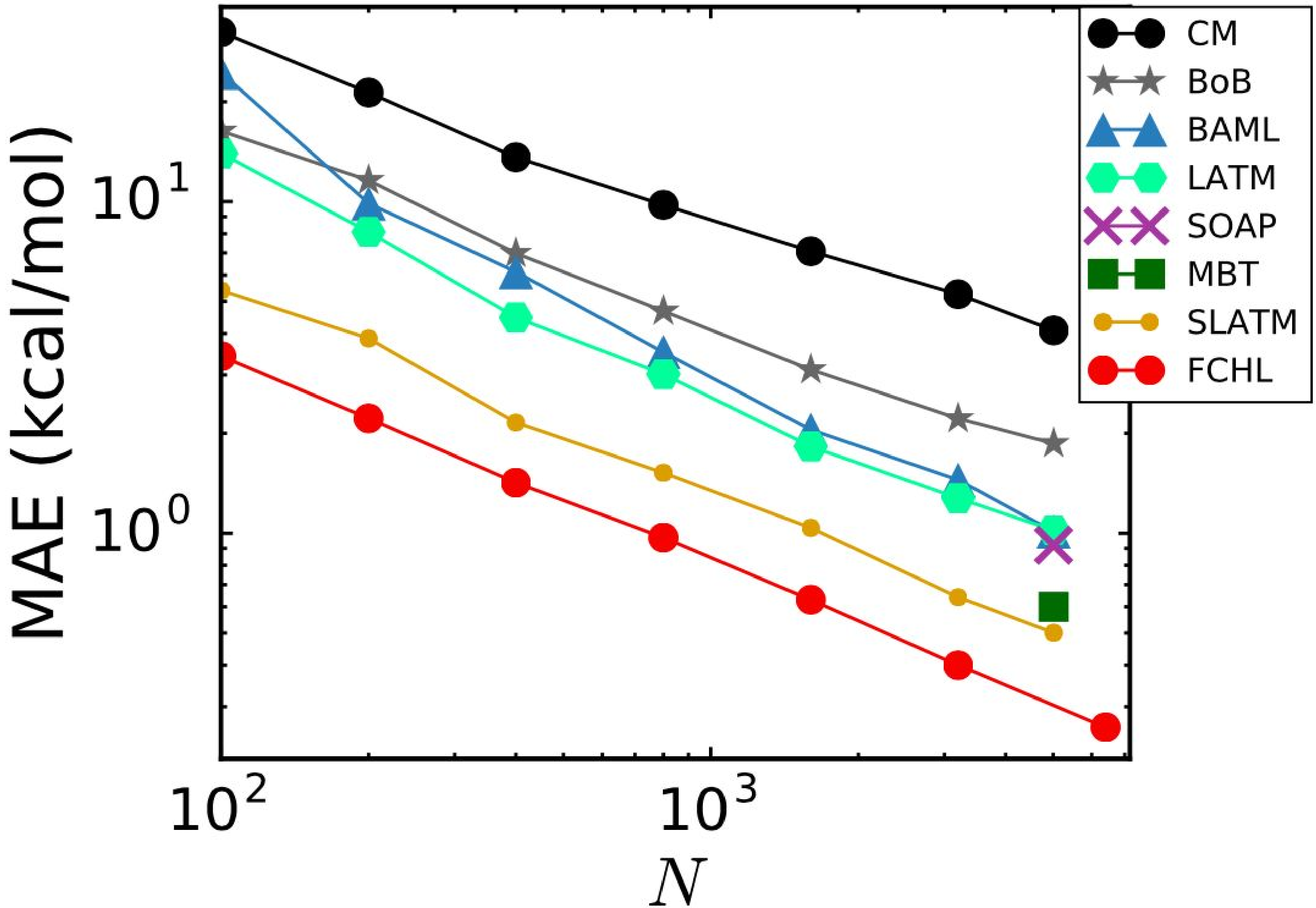
Atoms + **L**ondon + **A**xilrod-**T**eller-**M**uto **(LATM)**

$$E^{(2)}(\mathbf{R_I}, \mathbf{R_J}) = -\frac{C_{6_{IJ}}}{R_{IJ}^6} \qquad E^{(3)}(\mathbf{R_I}, \mathbf{R_J}, \mathbf{R_K}) = C_{9_{IJK}} \frac{3\cos[\phi_I]\cos[\phi_J]\cos[\phi_K] + 1}{R_{IJ}^3 R_{IK}^3 R_{JK}^3}$$



Huang, OAvL, *to be submitted* (2017)

Current performance on QM9



Faber, Christensen, Huang, OAvL, *to be submitted* (2017)

Representation leading to low a and b

1. Unique
2. Similar to target
3. Efficient
4. …

More ways to be wrong than right

Conclusions

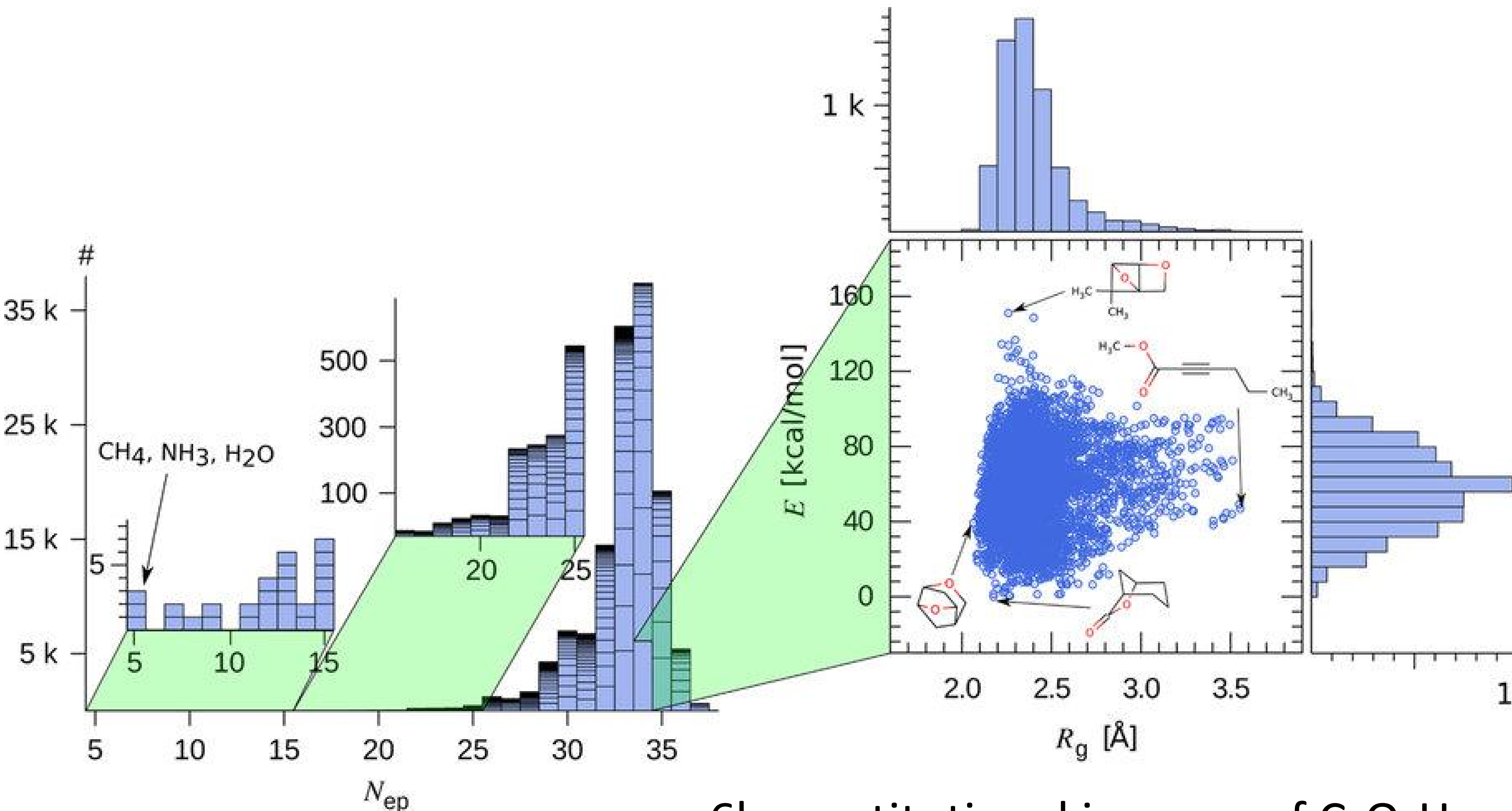
1. Instantaneous QM quality predictions

2. Learning curves reveal quality of ML model

3. Representations

4. Data sets

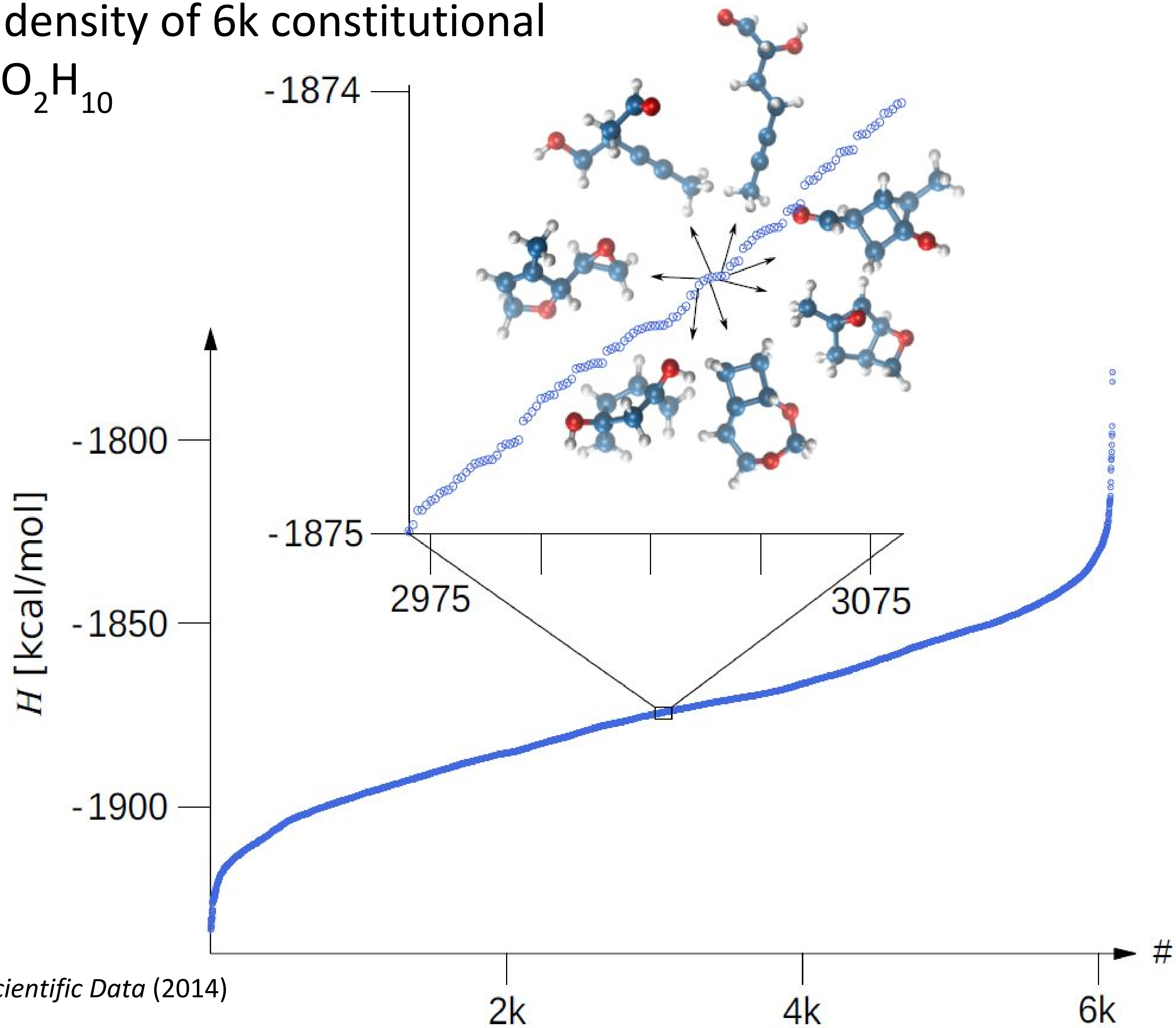
Data: Smallest 134k organic molecules in GDB



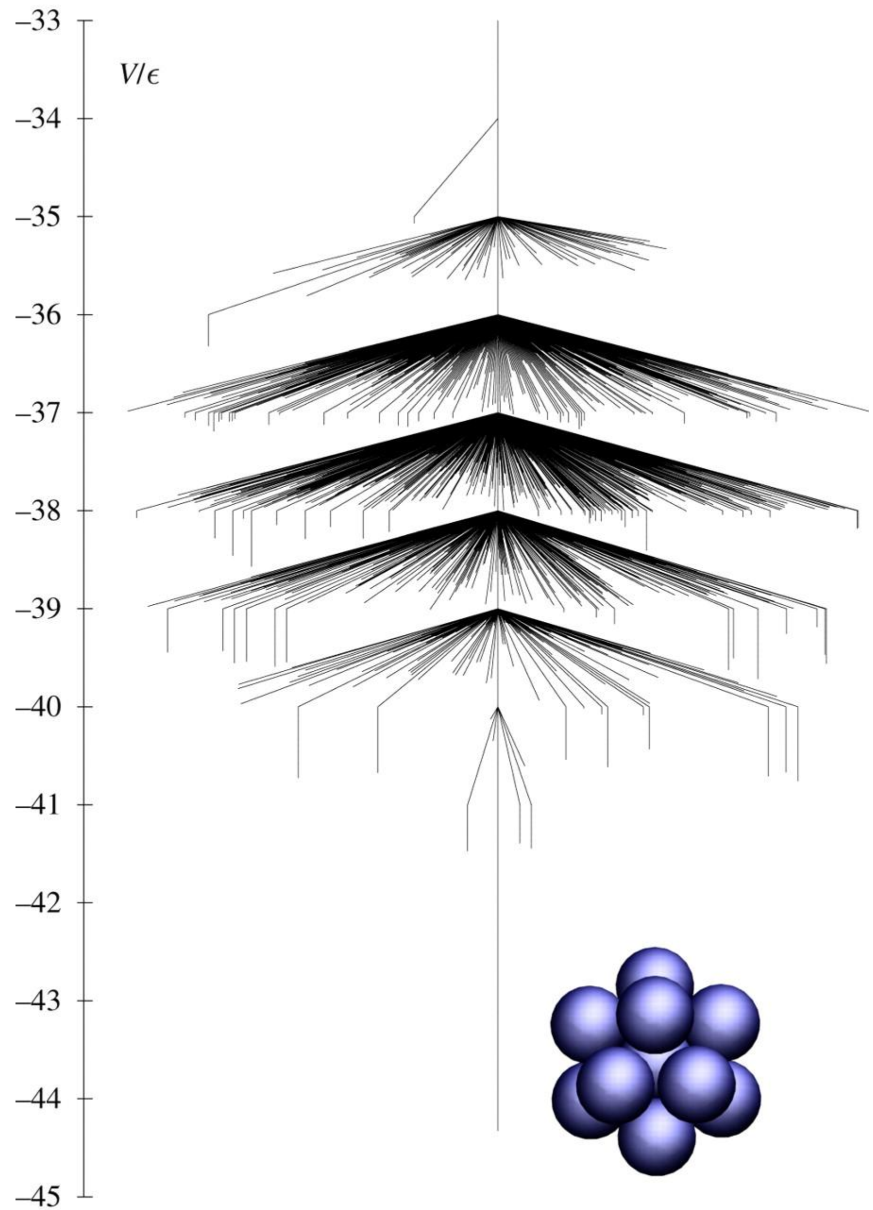
6k constitutional isomers of $C_7O_2H_{10}$

R Ramakrishnan et al *Scientific Data* (2014)

Data: Energy density of 6k constitutional isomers of $C_7O_2H_{10}$



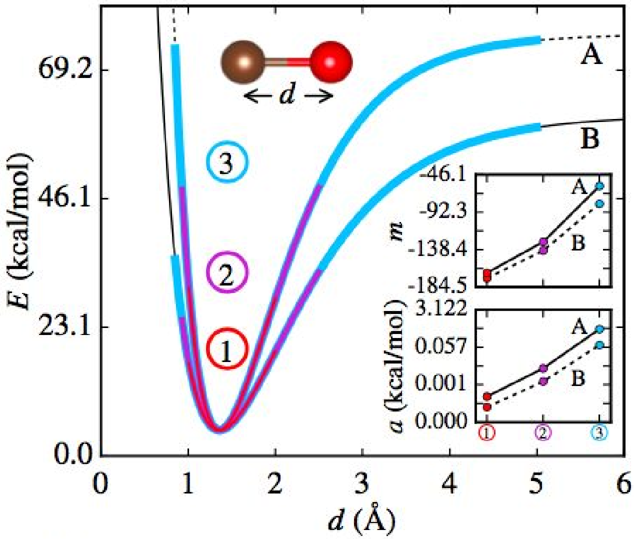
R Ramakrishnan et al *Scientific Data* (2014)

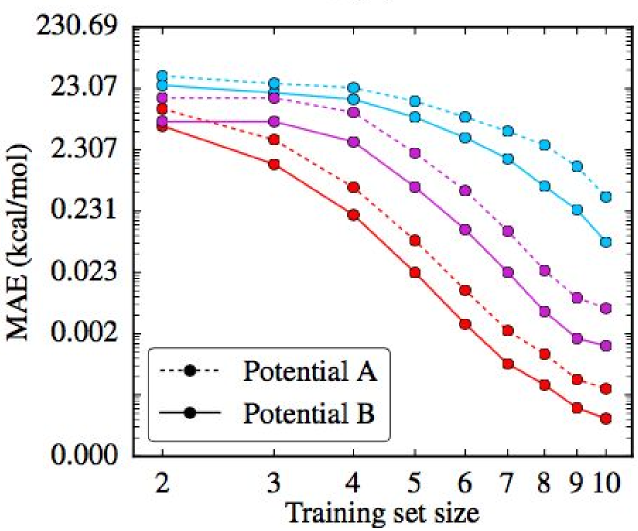


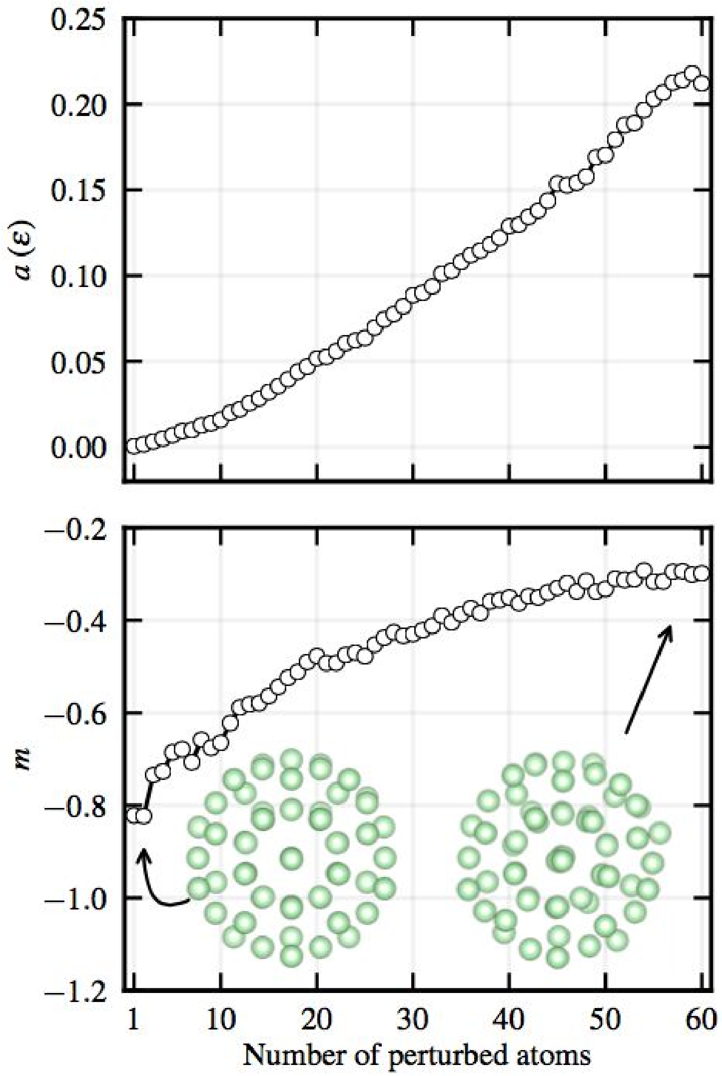
D J Wales *Philosophical Transactions A* (2012)

Temperature

$$\log(\epsilon) = \log(a) + m\log(N)$$





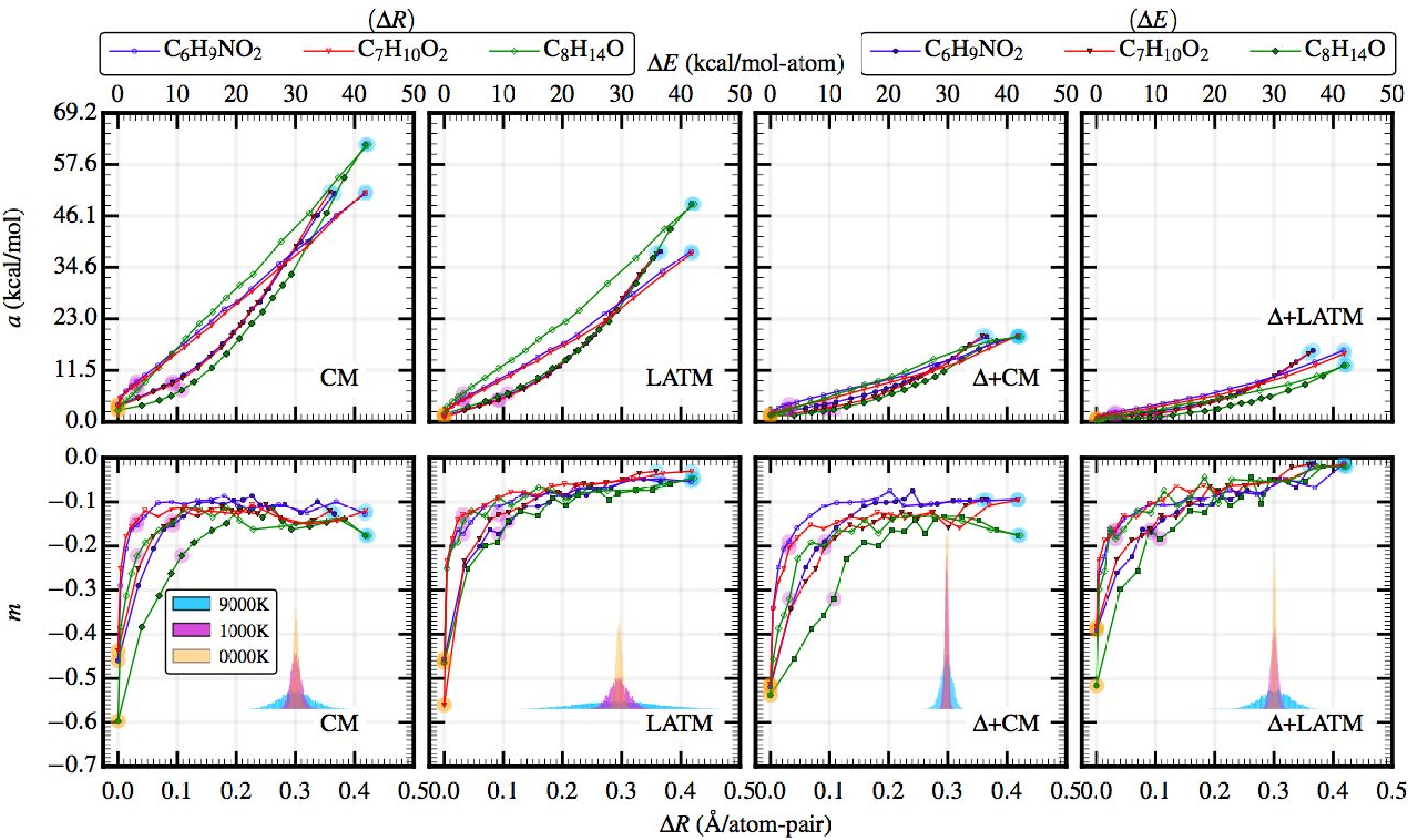


Ar_{60} LJ cluster

R. Sarmiento-Perez, F. Faber, B. Huang, OAvL, *to be submitted* (2017)

Temperature

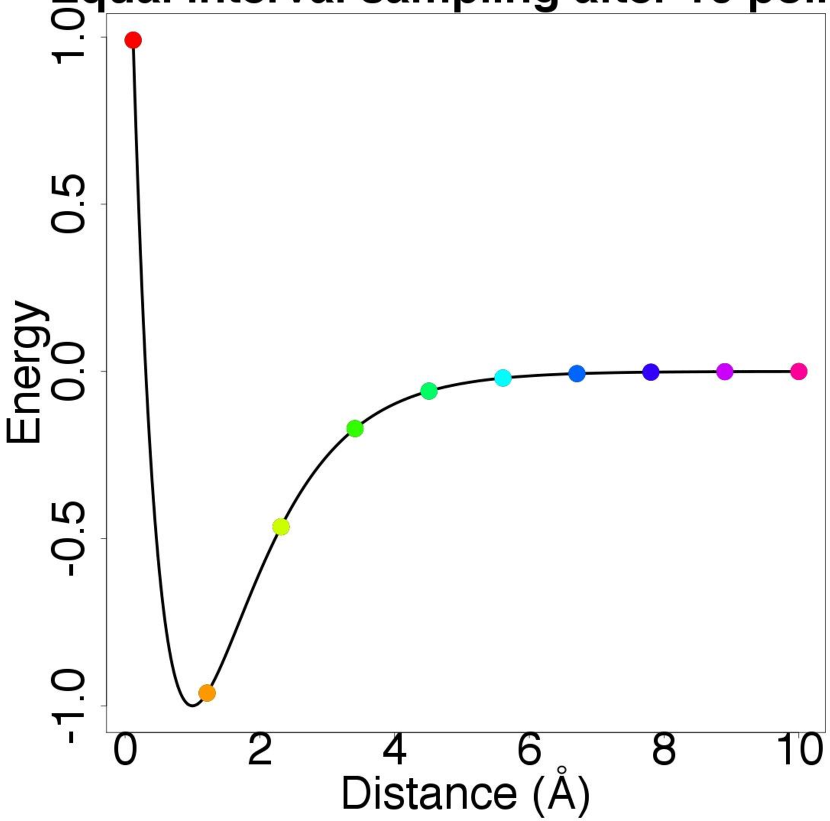
$$\log(\epsilon) = \log(a) + m\log(N)$$

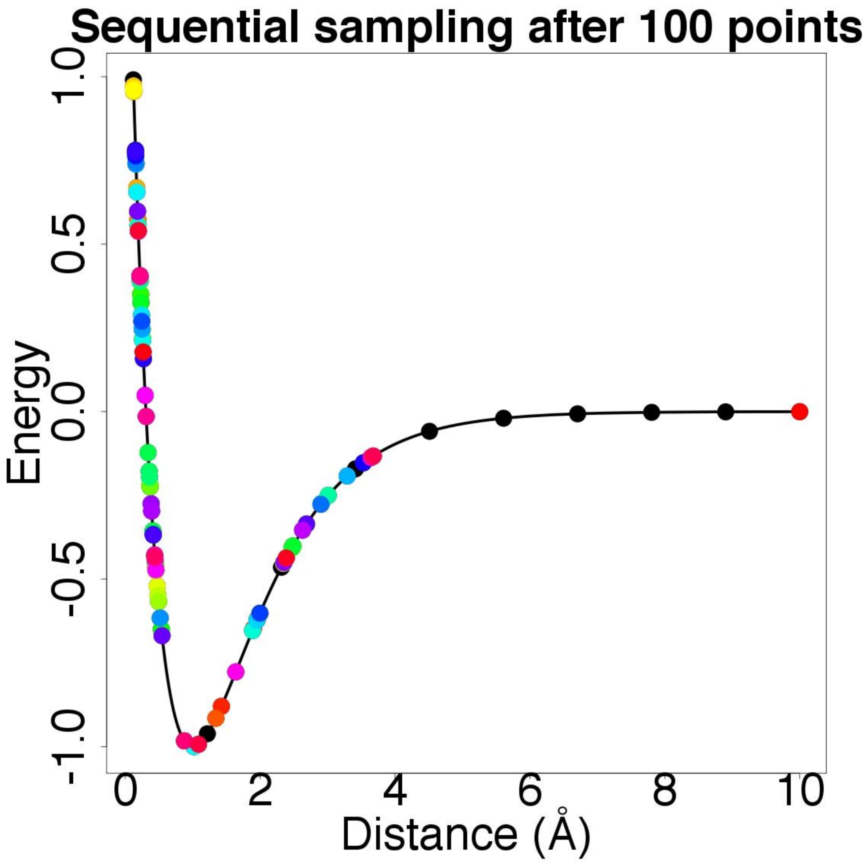


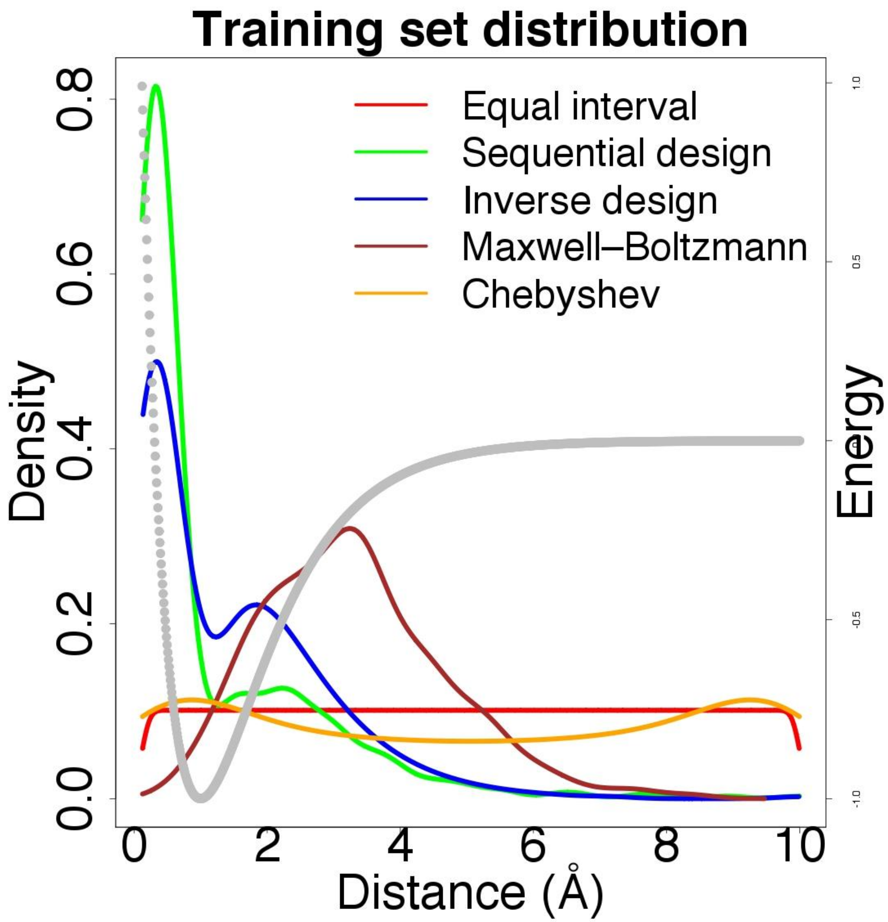
R. Sarmiento-Perez, F. Faber, B. Huang, OAvL, *to be submitted* (2017)

Selection bias

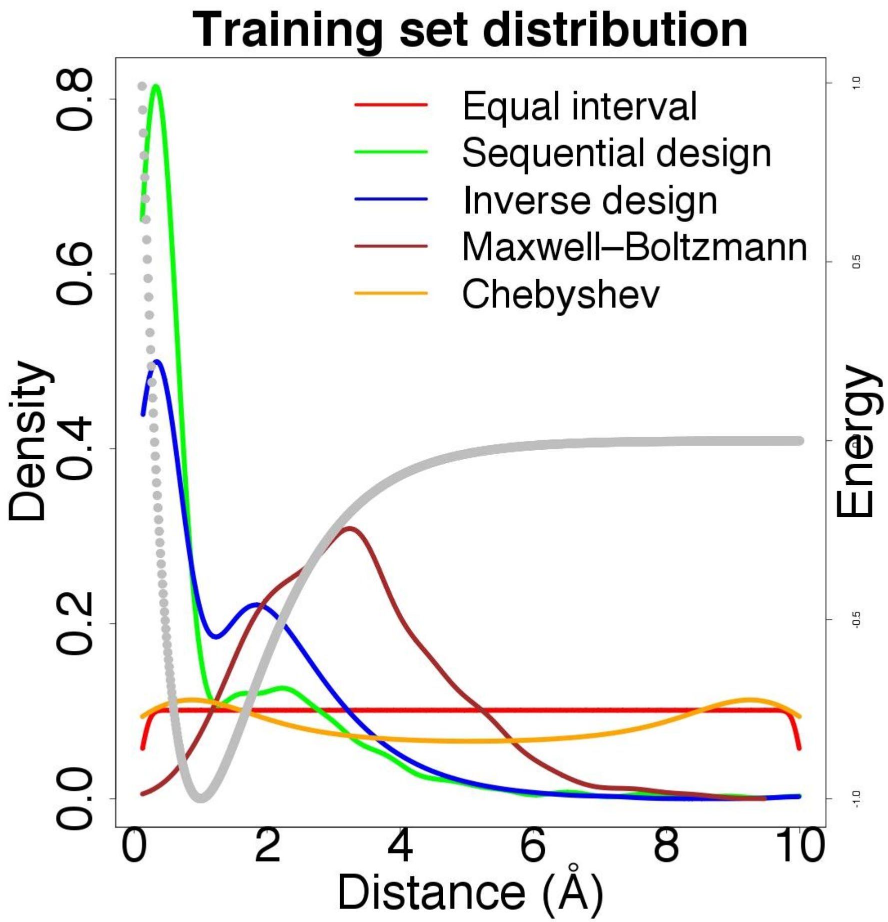
Equal interval sampling after 10 points

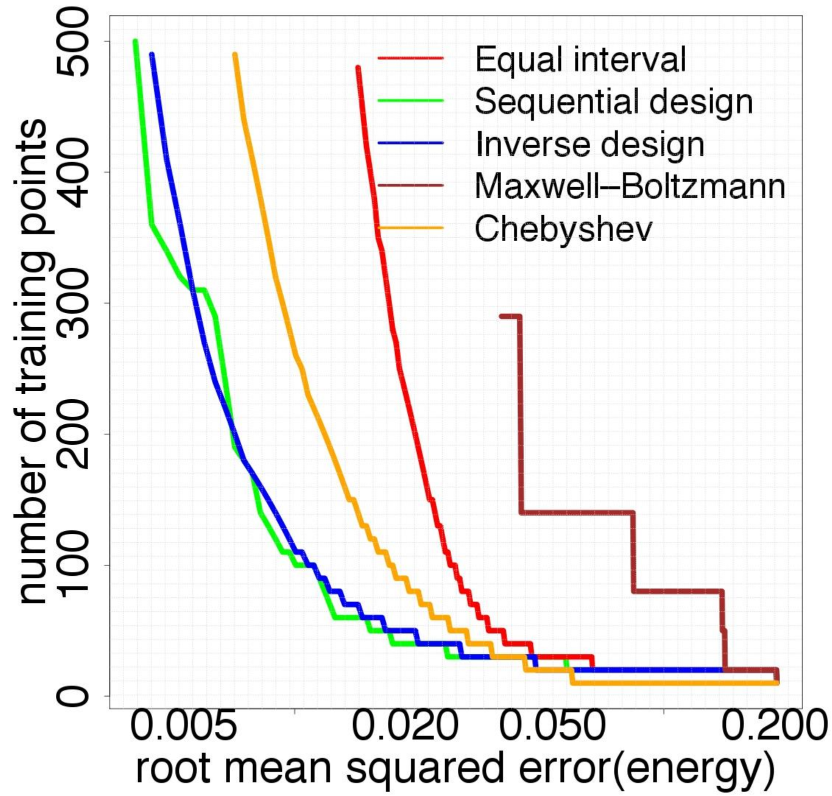


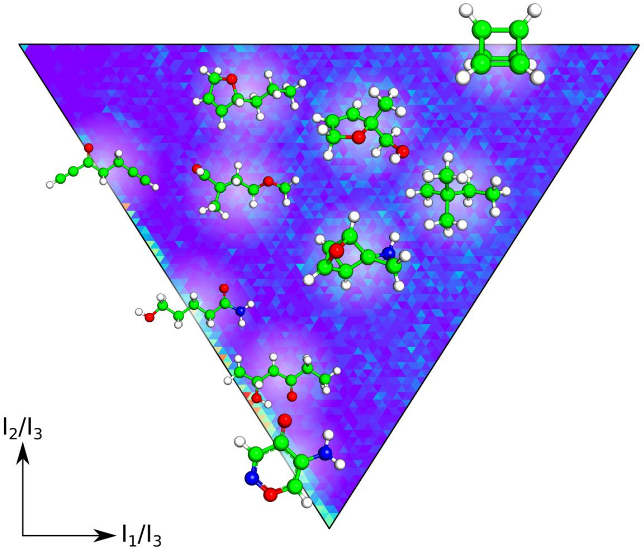

Sequential sampling after 100 points
Energy
Distance (Å)
1.0
0.5
0.0
-0.5
-1.0
0
2
4
6
8
10

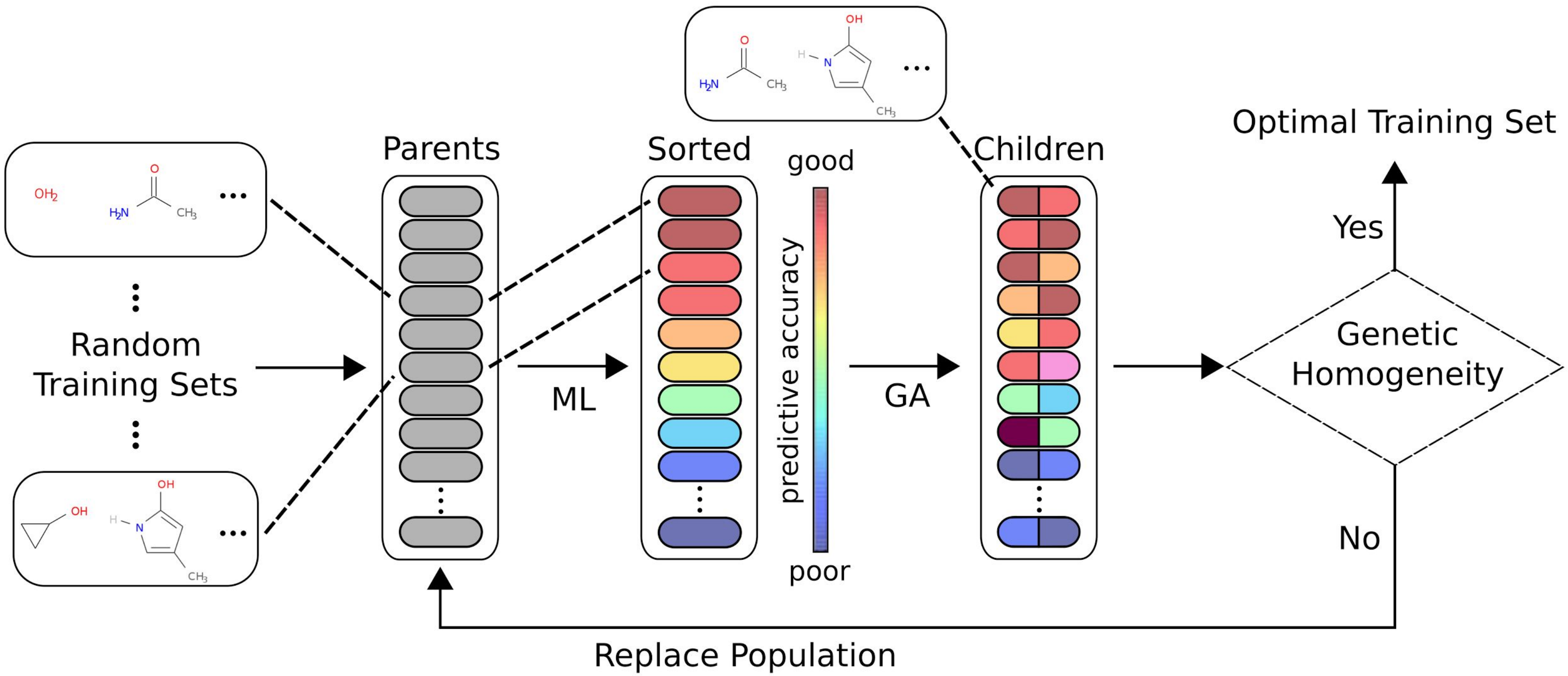


Training set distribution
Equal interval
Sequential design
Inverse design
Maxwell–Boltzmann
Chebyshev
Density
Energy
Distance (Å)
0.0
0.2
0.4
0.6
0.8
1.0
0.5
-0.5
-1.0
0
2
4
6
8
10

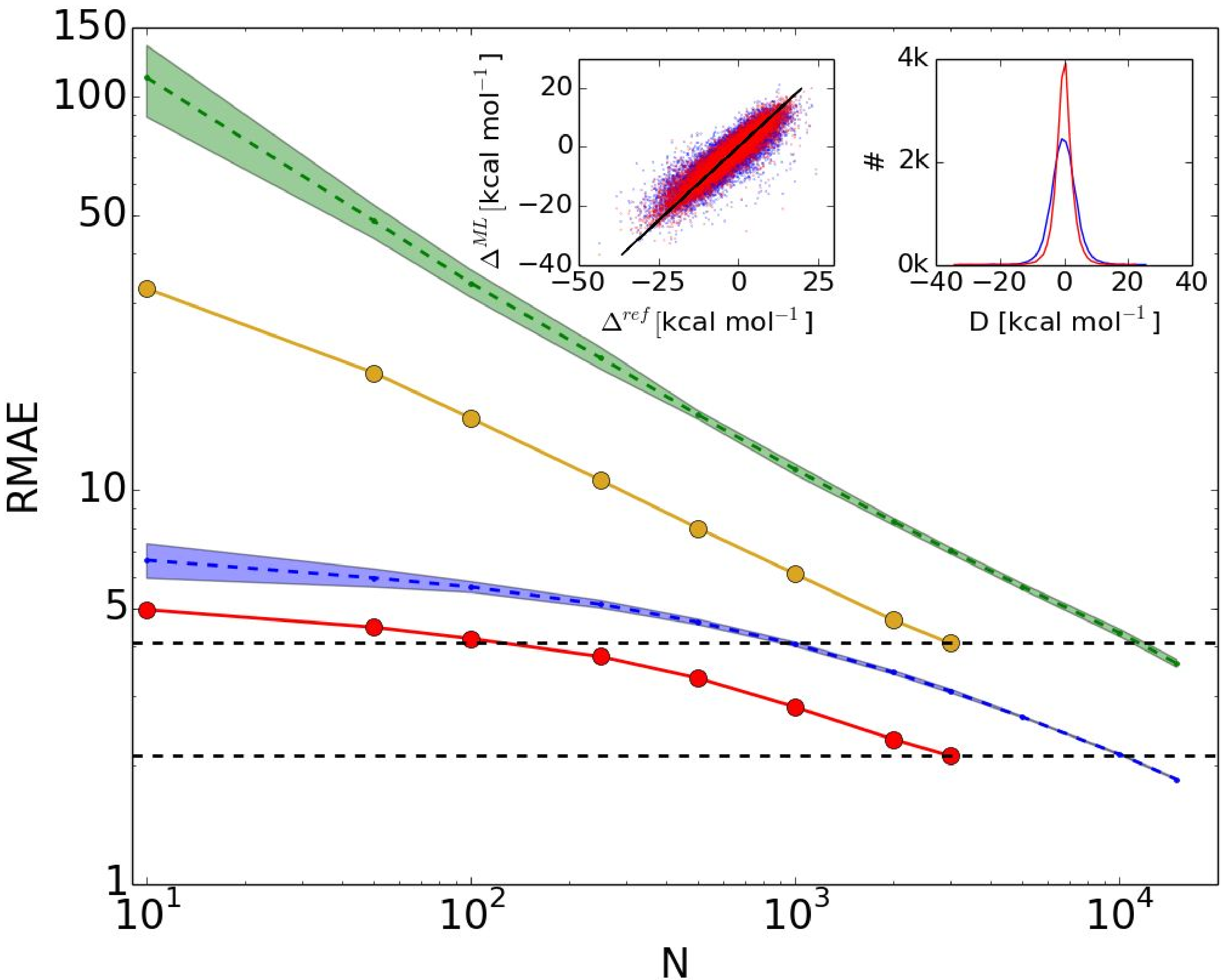

Training set distribution
Equal interval
Sequential design
Inverse design
Maxwell–Boltzmann
Chebyshev
Density
0.0
0.2
0.4
0.6
0.8
Energy
0
2
4
6
8
10
Distance (Å)


Equal interval
Sequential design
Inverse design
Maxwell–Boltzmann
Chebyshev
number of training points
0
100
200
300
400
500
0.005
0.020
0.050
0.200
root mean squared error(energy)





"Genetic optimization of training sets for improved machine learning models of molecular properties", N. J. Browning, R. Ramakrishnan, OAvL, U. Rothlisberger *J. Phys. Chem. Lett* (2017)

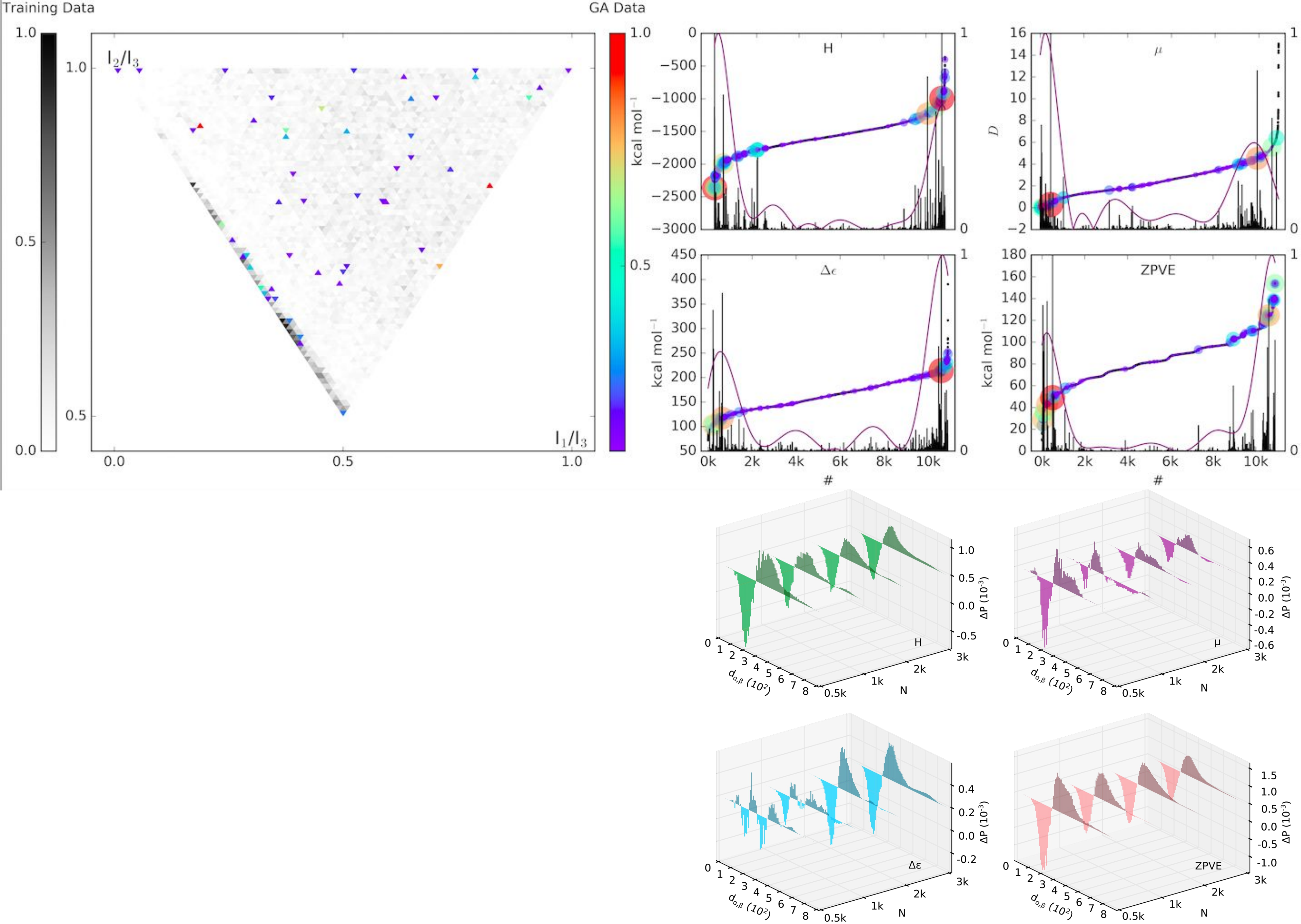


"Genetic optimization of training sets for improved machine learning models of molecular properties", N. J. Browning, R. Ramakrishnan, OAvL, U. Rothlisberger *J. Phys. Chem. Lett* (2017)

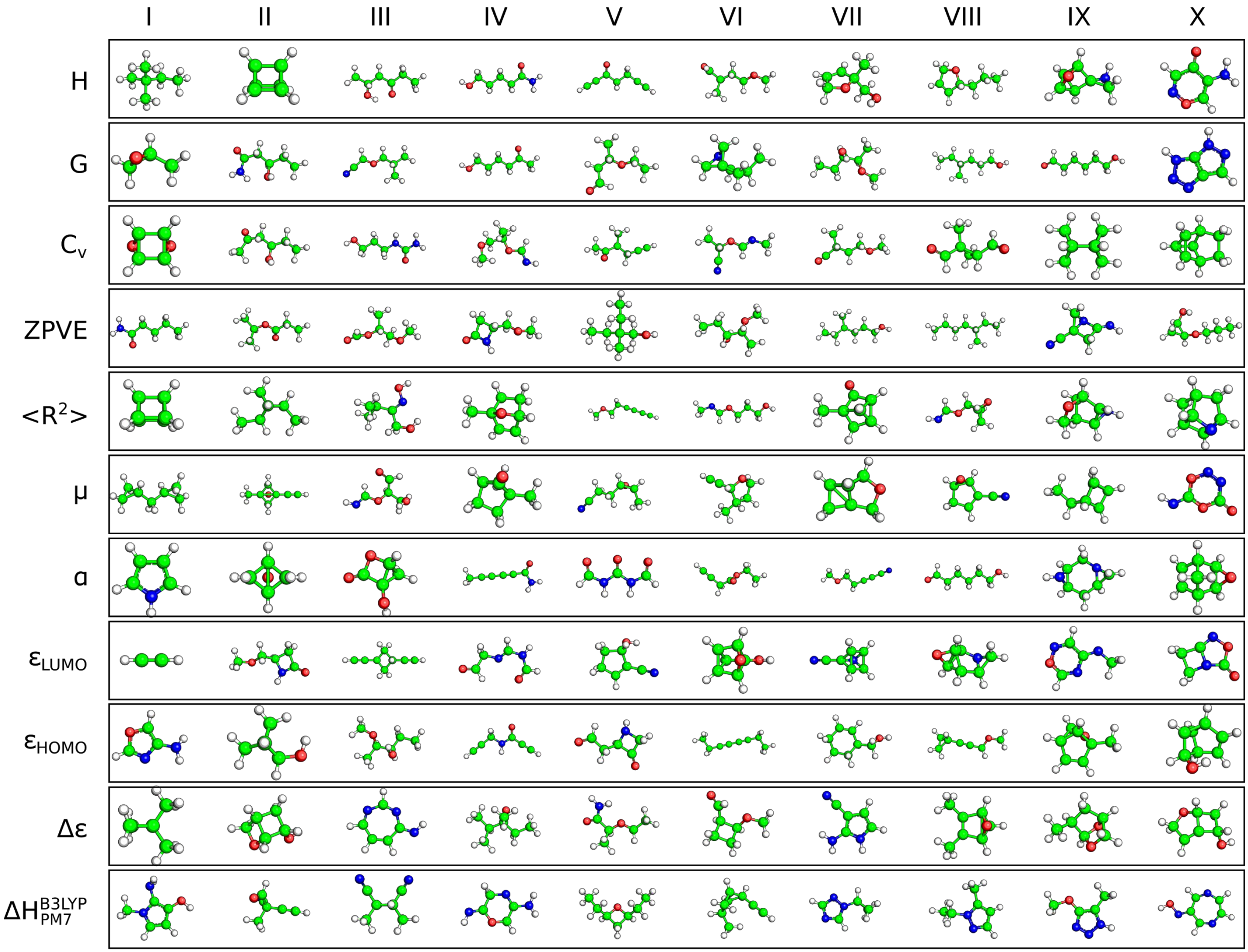
Table I. Randomized and GA-optimized out-of-sample relative mean absolute errors (RMAEs) for all properties. All target chemical accuracies are 1 kcal/mol, except for ZPVE, dipole moment and isotropic polarizability, which target accuracies of 10cm^{-1}, $0.1D$ and $0.1a_0^3$ respectively. GA-optimized RMAEs are denoted by P^{GA} while randomly generated training set MAEs are denoted as P^{ML}. Final row corresponds to out-of-sample RMAEs for enthalpy of atomization H using $\Delta_{\text{PM7}}^{\text{B3LYP}}$-learning.

P_{rand} (P_{GA})	N						
	10	50	100	500	1k	2k	3k
H	113.0 (31.6)	48.0 (18.3)	33.3 (14.3)	14.8 (7.5)	10.2 (5.8)	6.8 (4.5)	5.1 (3.9)
G	101.8 (28.8)	44.0 (17.7)	31.4 (14.1)	14.3 (7.5)	9.9 (5.6)	6.7 (4.3)	5.0 (3.9)
C_v	27.3 (14.5)	18.2 (9.4)	14.6 (7.8)	7.4 (4.0)	5.2 (2.9)	3.4 (2.3)	2.5 (2.0)
ZPVE	10.1 (2.4)	4.3 (1.1)	2.8 (0.8)	0.9 (0.4)	0.6 (0.3)	0.4 (0.2)	0.3 (0.1)
$<R^2>$	168.5 (92.2)	117.0 (44.2)	85.6 (33.2)	35.7 (19.1)	25.7 (15.5)	18.3 (12.7)	14.5 (11.6)
μ	11.3 (8.5)	10.3 (7.7)	9.9 (7.4)	8.4 (6.3)	7.5 (5.7)	6.2 (5.1)	5.2 (4.7)
α	40.8 (16.3)	23.1 (12.0)	18.5 (10.8)	11.8 (7.8)	9.6 (6.5)	7.2 (5.4)	5.8 (4.9)
ϵ_{HOMO}	13.0 (9.0)	11.2 (8.1)	10.4 (7.3)	7.7 (5.2)	6.3 (4.5)	4.9 (3.8)	4.0 (3.5)
ϵ_{LUMO}	22.3 (15.8)	18.8 (12.7)	17.0 (11.1)	11.9 (8.0)	9.7 (6.7)	7.4 (5.6)	5.9 (5.0)
gap	24.0 (17.8)	20.8 (15.0)	19.5 (13.5)	14.3 (9.8)	11.8 (8.1)	9.0 (6.8)	7.3 (6.2)
ΔH	6.6 (5.0)	6.0 (4.4)	5.7 (4.1)	4.6 (3.2)	4.1 (2.6)	3.4 (2.1)	3.1 (1.9)

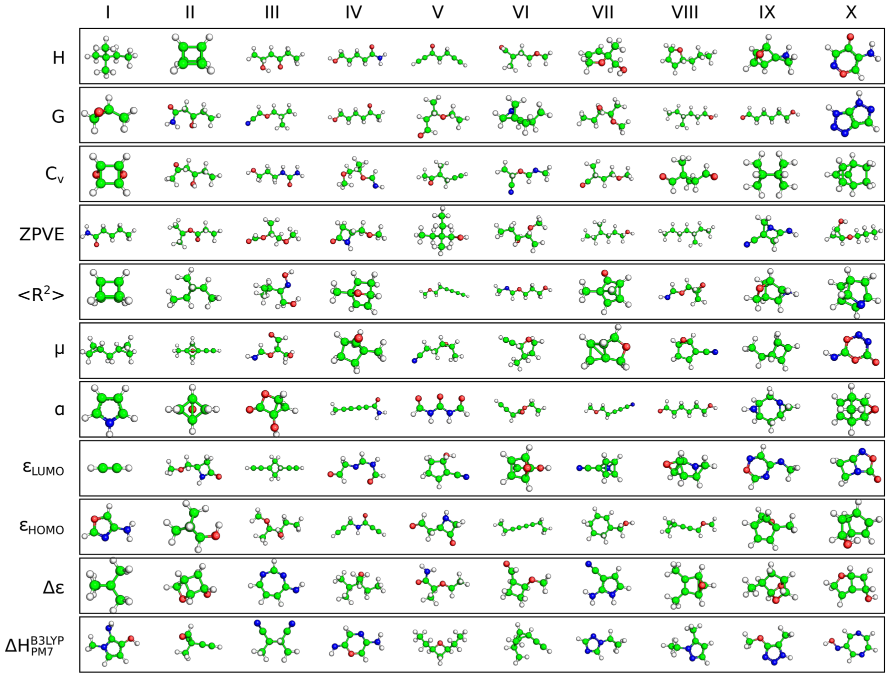
"Genetic optimization of training sets for improved machine learning models of molecular properties", N. J. Browning, R. Ramakrishnan, OAvL, U. Rothlisberger J. Phys. Chem. Lett (2017)

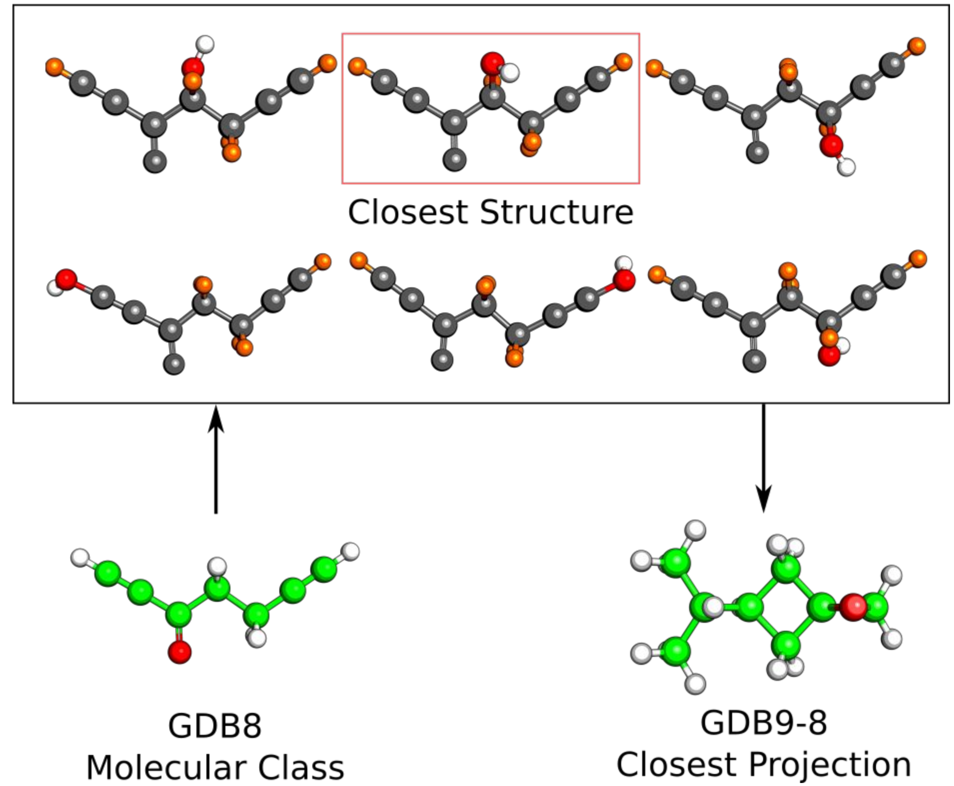


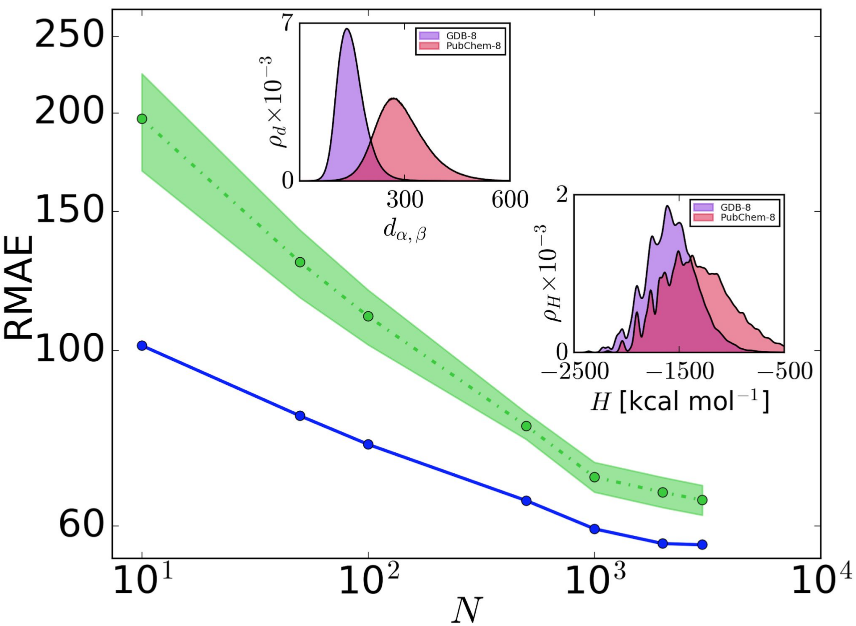
"Genetic optimization of training sets for improved machine learning models of molecular properties", N. J. Browning, R. Ramakrishnan, OAvL, U. Rothlisberger J. Phys. Chem. Lett (2017)



"Genetic optimization of training sets for improved machine learning models of molecular properties", N. J. Browning, R. Ramakrishnan, OAvL, U. Rothlisberger *J. Phys. Chem. Lett* (2017)



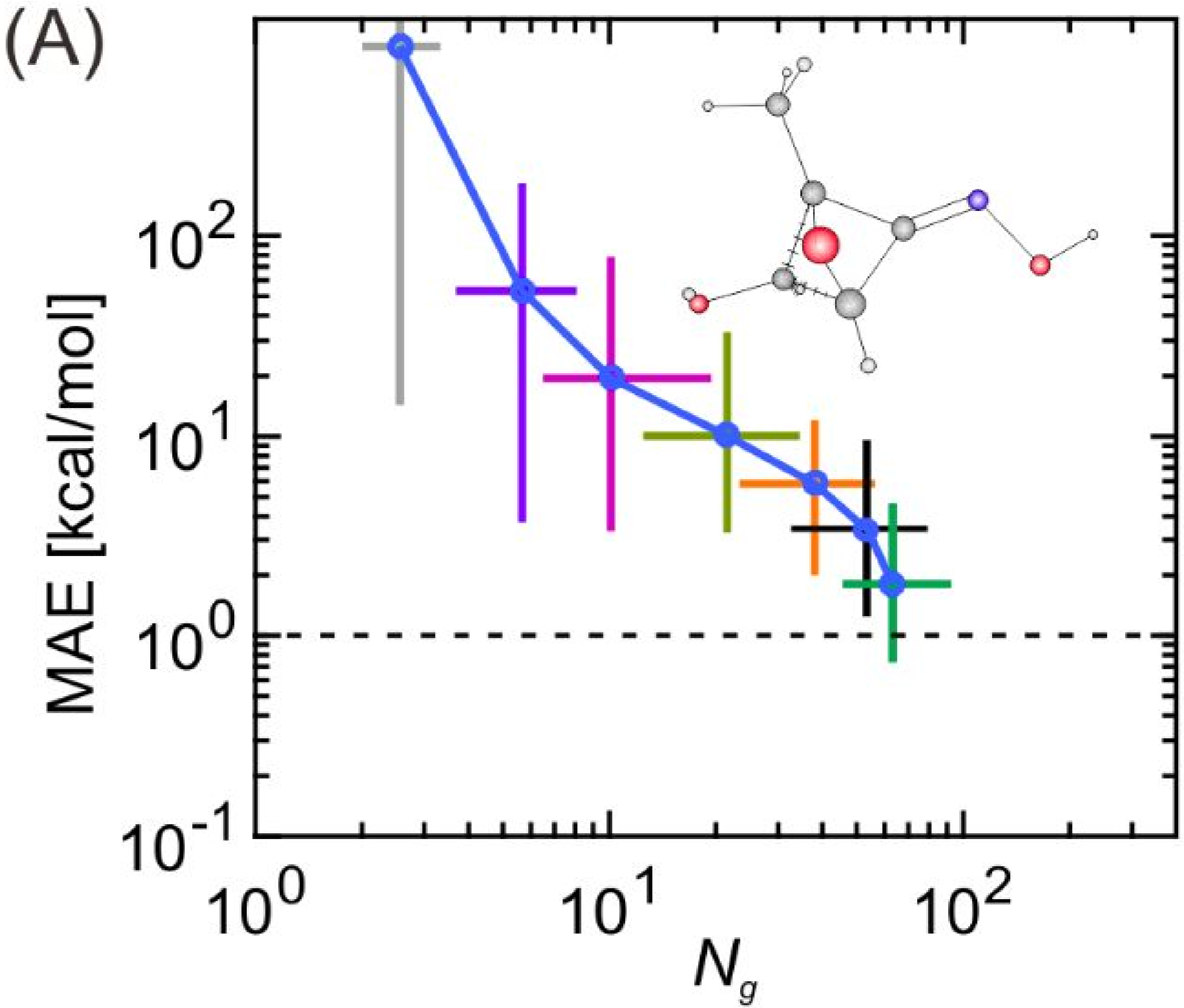




"Genetic optimization of training sets for improved machine learning models of molecular properties", N. J. Browning, R. Ramakrishnan, OAvL, U. Rothlisberger *J. Phys. Chem. Lett* (2017)

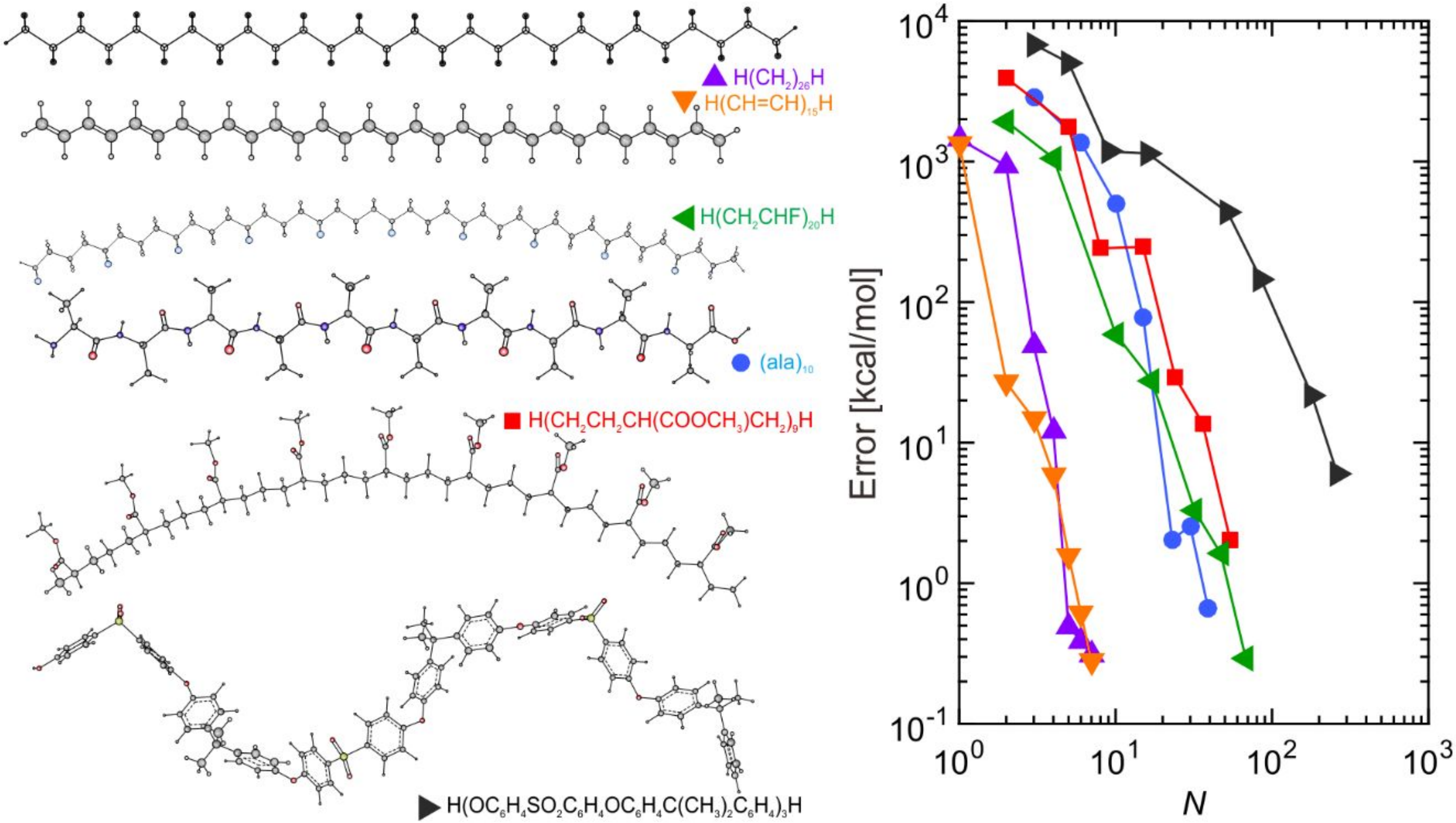
Effective

Errors on 1k QM9 predictions
LATM representation
Gaussian kernel



B. Huang, OAvL, *to be submitted* (2017)

Effective



B. Huang, OAvL, *to be submitted* (2017)

Data affects a and b

1. High dimensional function
2. Redundancy bias
3. Effective dimensionality
4. …

More ways to be wrong than right

Conclusions

1. Instantaneous QM quality predictions

2. Learning curves reveal quality of ML model

3. Representations

4. Data sets

Conclusions II

Scientific method - proper way to gain knowledge

Inductive (Data)

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

Fast (ms)
Arbitrary reference
Automatic improvement

Transferable?
Minimally condensed

Deductive (Laws)

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

Slow (depending on approx.)
Approximation dependent
Human improvement

Transferable?
Maximally condensed