Quantum Mechanics, Chemical Space, And Machine Learning

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

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 is beyond any capacity for all but the simplest systems and properties due to the combinatorial nature of chemical space, i.e. all compositional, constitutional, and conformational isomers. Consequently, efficient exploration algorithms need to exploit all implicit redundancies present in chemical space. I will discuss recently developed statistical learning approaches for interpolating quantum mechanical observables in compositional and constitutional space.

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
UNITED STATES AIR FORCE

U.S. DEPARTMENT OF ENERGY
INCITE
LEADERSHIP COMPUTING

Institute for Pure & Applied Mathematics
IPAM
University of California, Los Angeles

MARVEL
NATIONAL CENTRE OF COMPETENCE IN RESEARCH

UNI
BASEL

FNSNF
FONDS NATIONAL SUISSE
SCHWEIZERISCHER NATIONALFONDS
FONDO NAZIONALE SVIZZERO
SWISS NATIONAL SCIENCE FOUNDATION

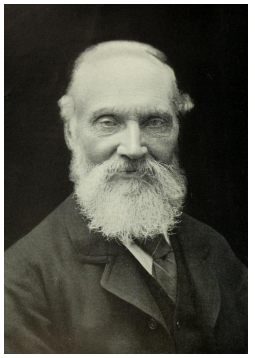
How do we do physics?

1. Guess a law
2. If it doesn't compare to experiment it's wrong

Feynman

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 understand a molecule when we can predict its properties with a degree of accuracy considered quantitative.*" 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)

From Industry 1.0 to Industry 4.0

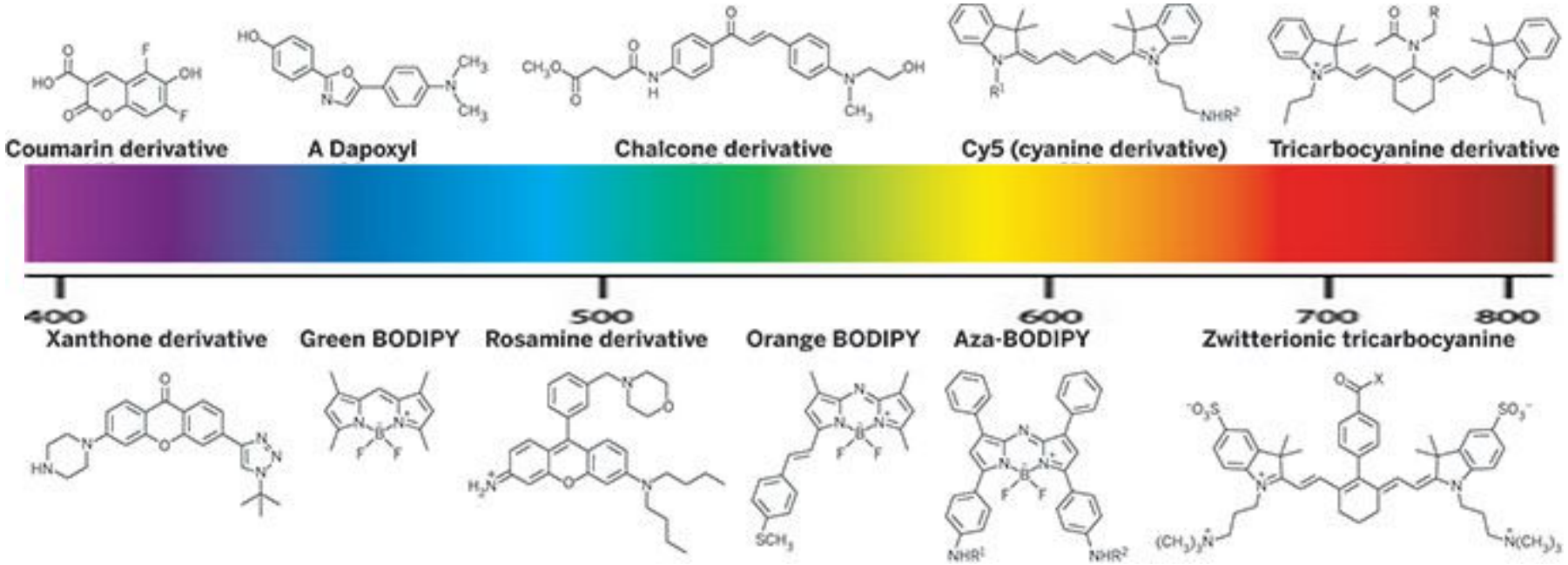
1.0 | **1784** | based on mechanical production equipment driven by water and steam power

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

Configuration + Composition → Chemical Space



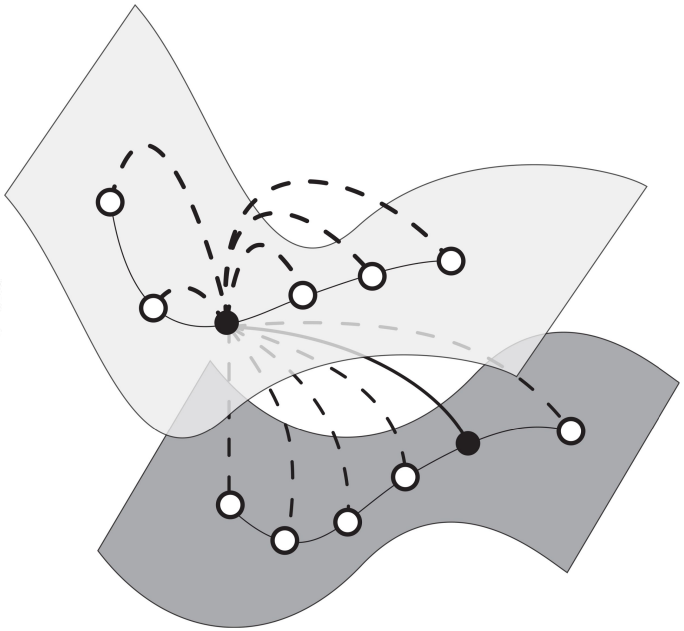
Young-Tae Chang et al *C&E News* **93 (12) 39-40** (2015)

correlations (inductive)

supervised learning

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

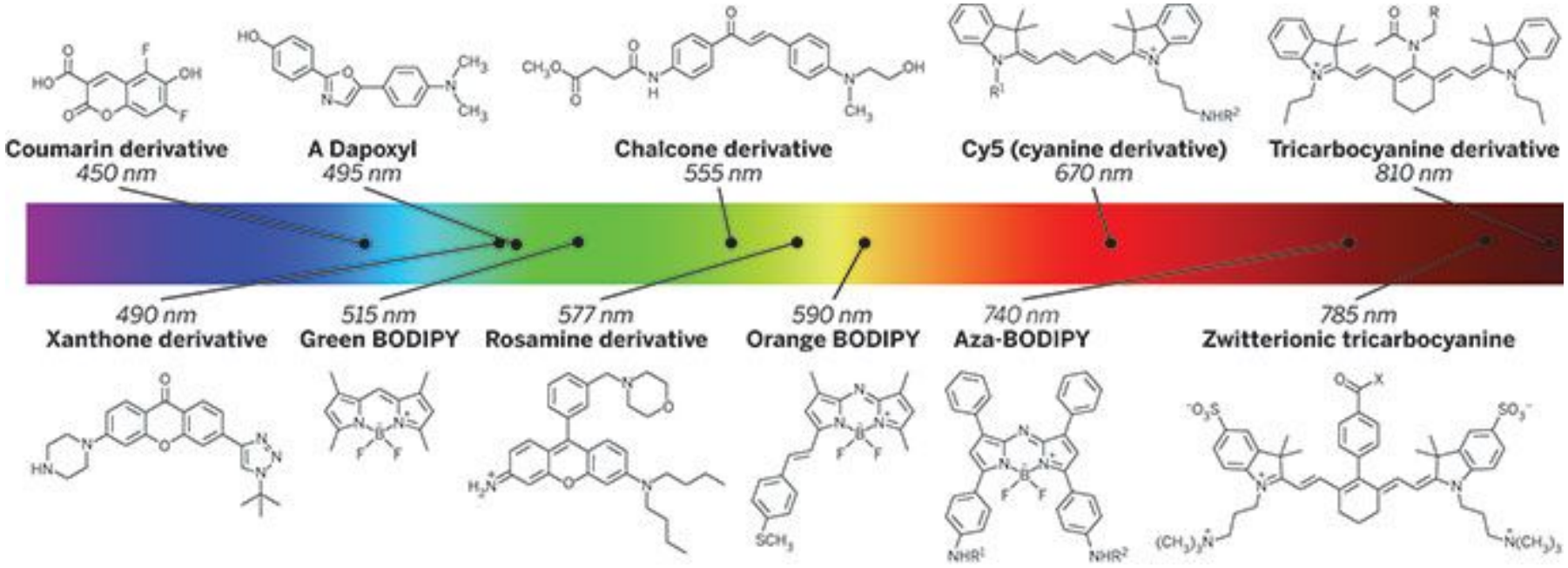
$$H(\{Z_I, \mathbf{R}_I\}) \overset{\Psi}{\longmapsto} E$$



Chang, OAvL, *CHIMIA* (2014)

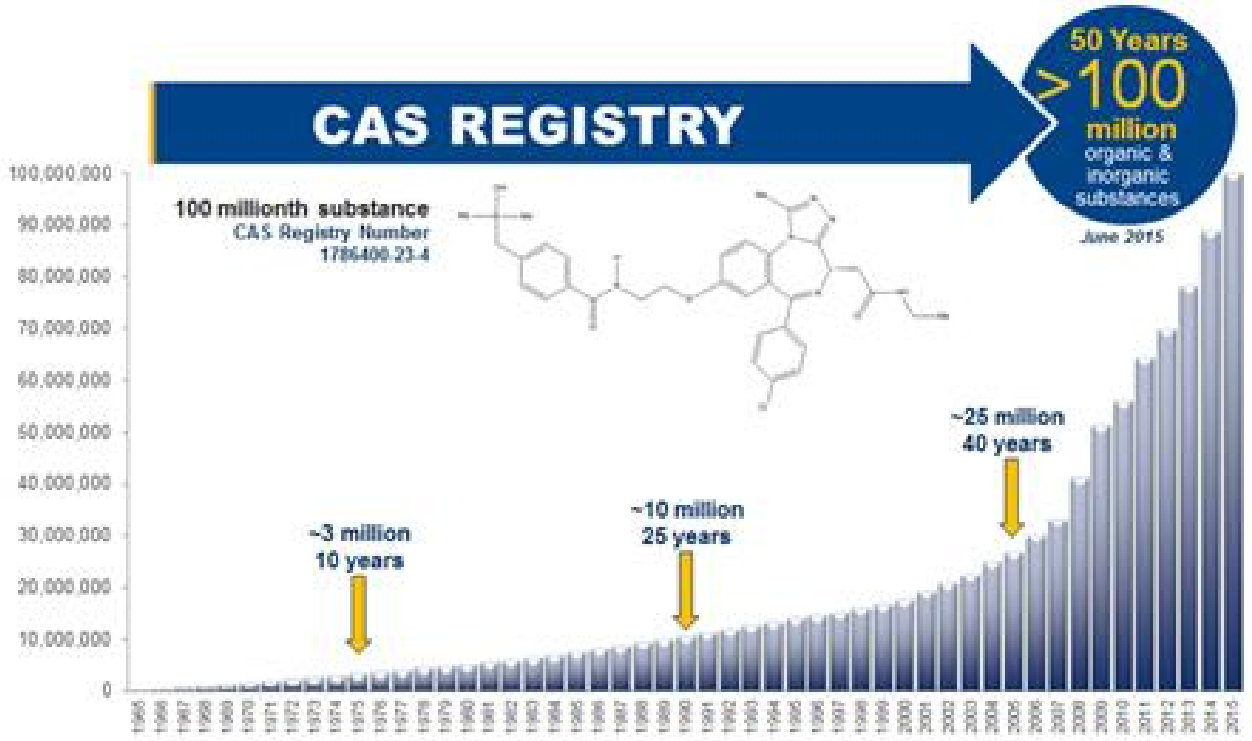
$$H(\{Z_I, \mathbf{R}_I\})\Psi(\mathbf{r}) = E\Psi(\mathbf{r})$$

Configuration + Composition → Chemical Space



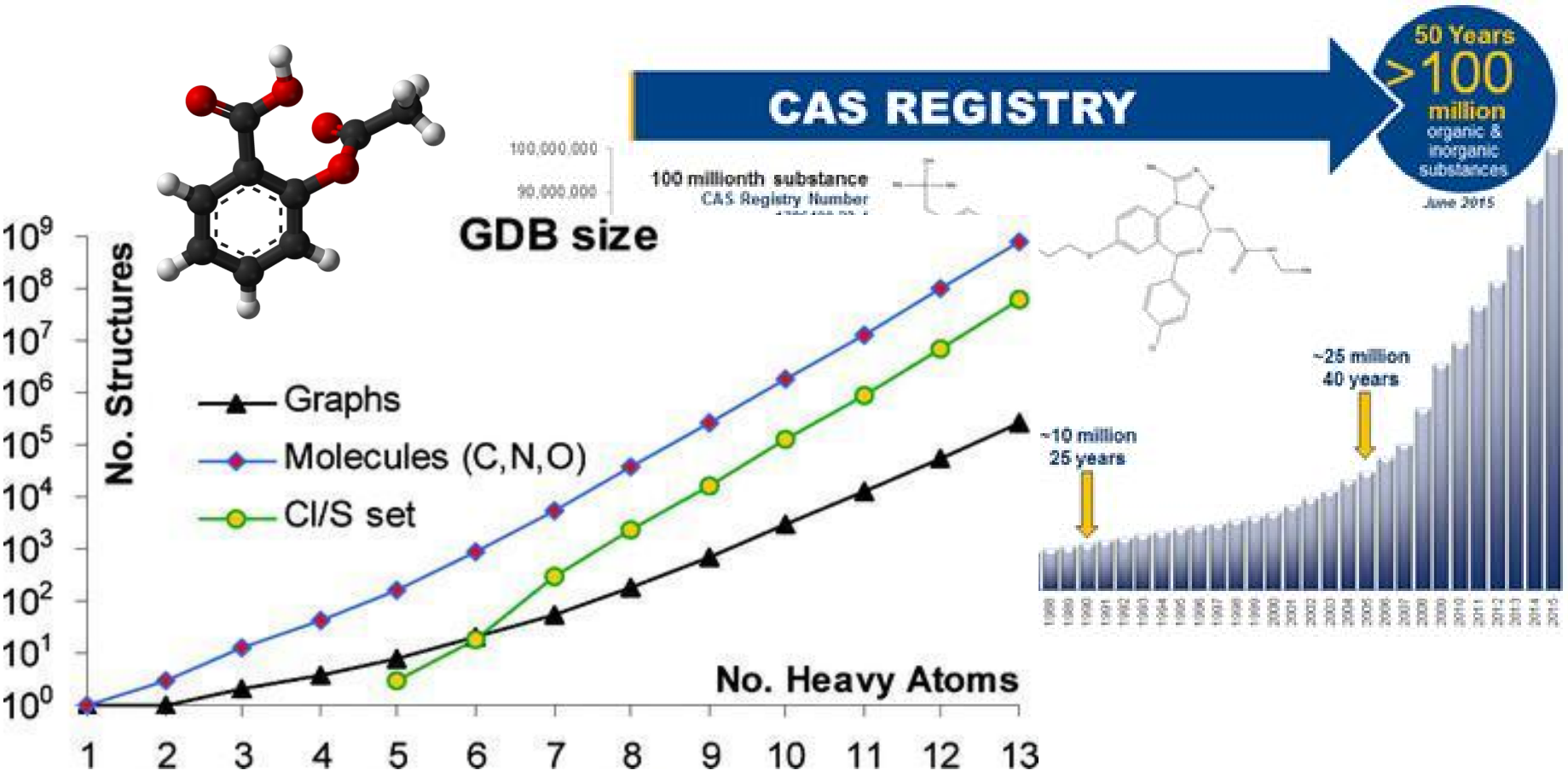
Young-Tae Chang et al *C&E News* **93 (12) 39-40** (2015)

How many are possible?



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

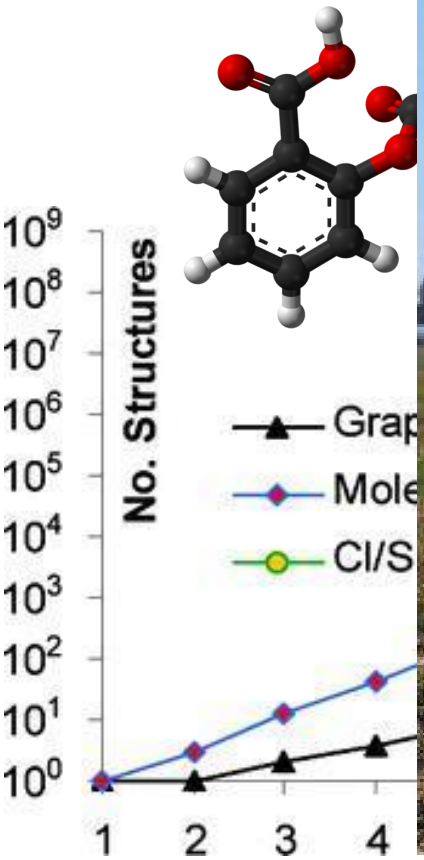
Al Bartlett, U of Colorado Boulder



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

``The great[illegible]s our inability to [illegible]n''

Al Bartlett, U c[illegible]



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. Rate and offset depend on
   a. Baseline: Δ-ML (prior expert knowledge about property)
   b. Representation: Uniqueness and energy (property independent)
   c. Domain of applicability: Data (properties and structures)

Kernel Ridge Regression

Kernel

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

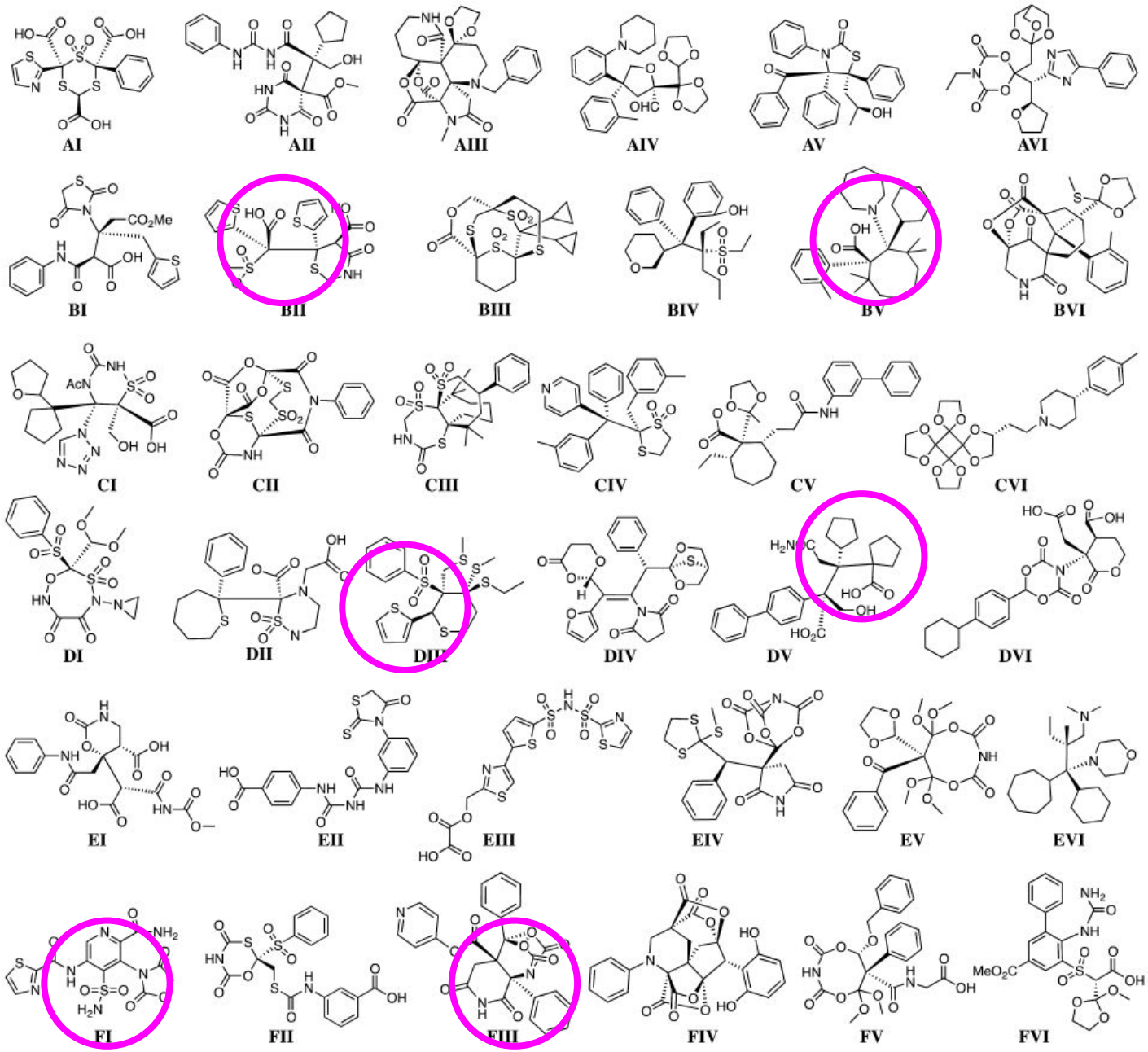
e.g. $k(\mathbf{M}, \mathbf{M}') = \exp\left(-\frac{d(\mathbf{M}, \mathbf{M}')^2}{2\sigma^2}\right)$

Regression

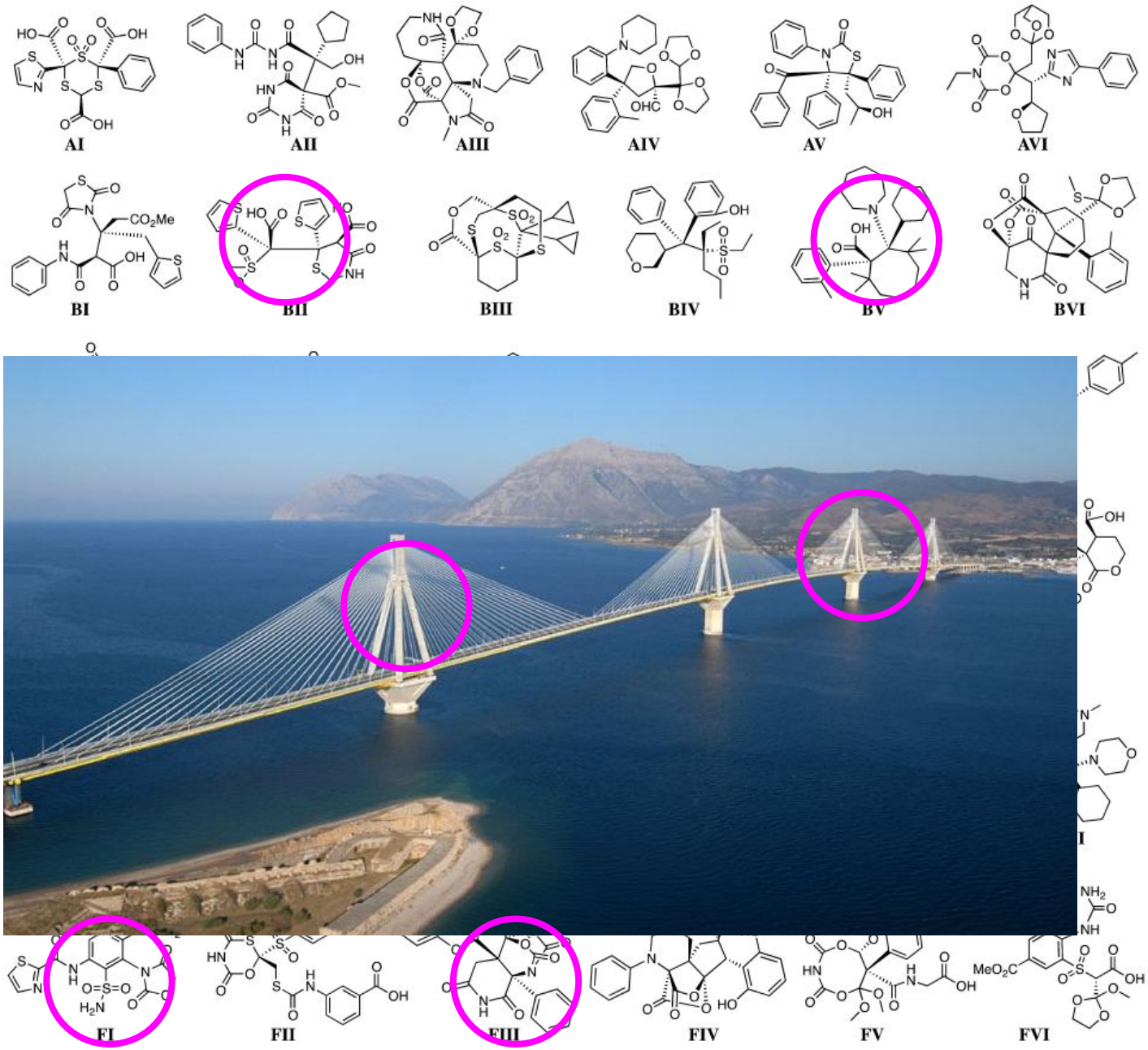
$$\min_{\alpha} \left(\sum_{i} \left(E^{est}(\mathbf{M}_i) - E_i^{ref} \right)^2 + \lambda \sum_{ij} \alpha_i \alpha_j k(\mathbf{M}_i, \mathbf{M}_j) \right)$$

Solution

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



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

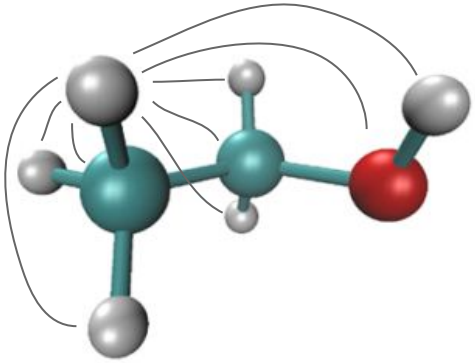


Virshup, Yang, Beratan et al *J Am Chem Soc* (2013)

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



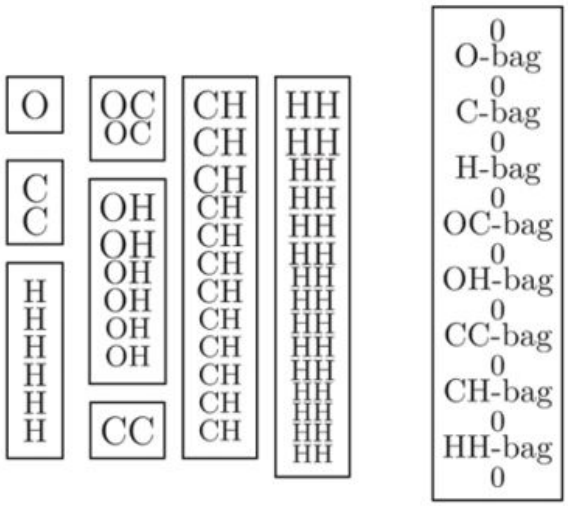
Coulomb matrix (CM)

	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)

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

Bag of Bonds (BoB)

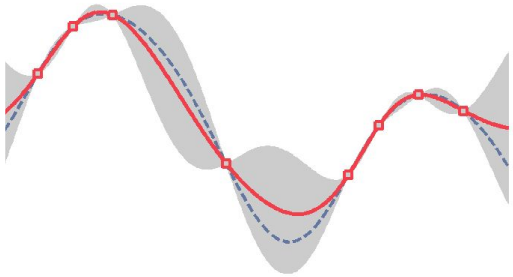


Hansen et al, *J Phys Chem Lett* (2015)

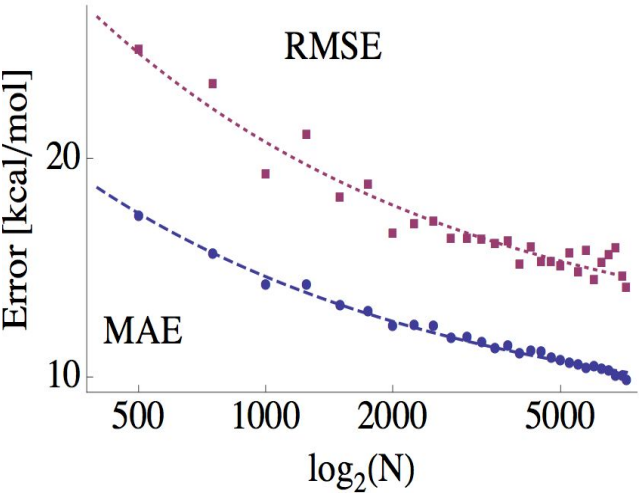
- Not unique (homometricity)
- Invariant (Tra&Rot)
- Compact
- Physical meaning
- Fast
- Simple metrics are smooth

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

2016

1. "Machine learning for many-body physics: efficient solution of dynamical mean-field theory", L.-F. Arsenault, OAvL, A. J. Millis, *submitted* , arxiv.org/abs/1506.08858
2. "Constant Size Molecular Descriptors For Use With Machine Learning", C. R. Collins, G. J. Gordon, OAvL, D. J. Yaron, submitted to *J. Chem. Inf. Model.*
3. "Machine Learning, Quantum Mechanics, and Chemical Compound Space" R. Ramakrishnan, OAvL (in press) *Reviews in Computational Chemistry*, arxiv.org/abs/1510.07512
4. **"Understanding molecular representations in machine learning: The role of uniqueness and target similarity"**, B. Huang, OAvL, accepted *J. Chem. Phys.* (2016), arxiv.org/abs/1608.06194
5. **"Machine Learning Energies of 2 M Elpasolite (ABC2D6) Crystals"**, F. Faber, A. Lindmaa, OAvL, R. Armiento, *Phys. Rev. Lett.* **117** 135502 (2016) arxiv.org/abs/1508.05315

2015

6. "Electronic Spectra from TDDFT and Machine Learning in Chemical Space", R. Ramakrishnan, M. Hartmann, E. Tapavicza, OAvL, *J. Chem. Phys.* **143** 084111 (2015) arxiv.org/abs/1504.01966
7. **"Machine Learning for Quantum Mechanical Properties of Atoms in Molecules"**, M. Rupp, R. Ramakrishnan, OAvL, *J. Phys. Chem. Lett.* **6** 3309 (2015). arxiv.org/abs/1505.00350
8. "Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Non-Locality in Chemical Space", K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, OAvL, K.-R. Mueller, A. Tkatchenko, *J. Phys. Chem. Lett.* **6** 2326 (2015).
9. "Transferable atomic multipole machine learning models for small organic molecules", T. Bereau, D. Andrienko, OAvL, *J. Chem. Theory Comput.* **11** 3225 (2015) arxiv.org/abs/1503.05453
10. "Machine Learning of Parameters for Accurate Semiempirical Quantum Chemical Calculations", P. Dral, OAvL, W. Thiel, *J. Chem. Theory Comput.* **11** 2120 (2015)
11. **"Big Data meets Quantum Chemistry Approximations: The Δ-Machine Learning Approach",** R. Ramakrishnan, P. O. Dral, M. Rupp, OAvL, *J. Chem. Theory Comput.* **11** 2087 (2015), arxiv.org/abs/1503.04987
12. "Crystal Structure Representations for Machine Learning Models of Formation Energies", F. Faber, A. Lindmaa, OAvL, R. Armiento, *Int. J. Quantum Chem.* doi:10.1002/qua.24917 (2015), arxiv.org/abs/1503.07406
13. **"Many Molecular Properties from One Kernel in Chemical Space"**, R. Ramakrishnan, OAvL, *CHIMIA* **69** 182 (2015) ; arxiv.org/abs/1502.04563
14. "Fourier series of atomic radial distribution functions: A molecular fingerprint for machine learning models of quantum chemical properties", OAvL, R. Ramakrishnan, M. Rupp, A. Knoll, *Int. J. Quantum Chem.* doi:10.1002/qua.2491 (2015) arxiv.org/abs/1307.2918

2014

15. "Machine learning for Many-Body Physics : The case of the Anderson impurity model", L.-F. Arsenault, A. Lopez-Bezanilla, OAvL, A. Millis, *Phys. Rev. B* **90** 155136 (2014) arxiv.org/abs/1408.1143
16. "Quantum chemistry structures and properties of 134 kilo molecules", R. Ramakrishnan, P. O. Dral, M. Rupp, OAvL, *Scientific Data* **1** 140022 (2014)
17. "Modeling electronic quantum transport with machine learning", A. Lopez-Bezanilla, OAvL, *Phys Rev B* **89** 235411 (2014), arxiv.org/abs/1401.8277

2013

18. "Assessment and validation of machine learning methods for predicting molecular atomization energies", K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, OAvL, A. Tkatchenko, K-R. Mueller *J Chem Theory Comput* **9** 3404 (2013)
19. "Machine Learning of Molecular Electronic Properties in Chemical Compound Space", G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K-R. Mueller, OAvL, *New J. Phys.* **15** 095003 (2013) arxiv.org/abs/1305.7074

2012

20. "Learning Invariant Representations of Molecules for Atomization Energy Prediction", G. Montavon, K. Hansen, S. Fazli, M. Rupp, F. Biegler, A. Ziehe, A. Tkatchenko, OAvL, K.-R. Mueller, *Advances in Neural Information Processing Systems* 449-457 **25** (2012)
21. "Fast and accurate modeling of molecular atomization energies with machine learning", M. Rupp, A. Tkatchenko, K.-R. Mueller, OAvL *Phys. Rev. Lett.* **108** 058301 (2012) arxiv.org/abs/1109.2618E

2011

22. "Towards quantitative structure-property relationships for charge transfer rates of polycyclic aromatic hydrocarbons" M. Misra, D. Andrienko, B. Baumeier , J.-L. Faulon, OAvL *J. Chem. Theory Comput.* **7** 2549 (2011)

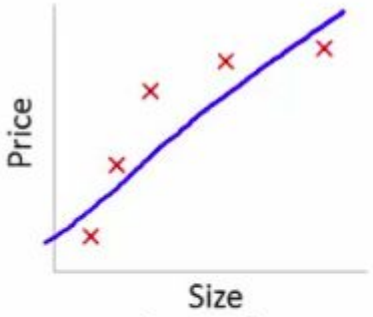
"Navigating Chemical Compound Space", IPAM, UCLA

Conclusions

1. Instantaneous QM quality predictions

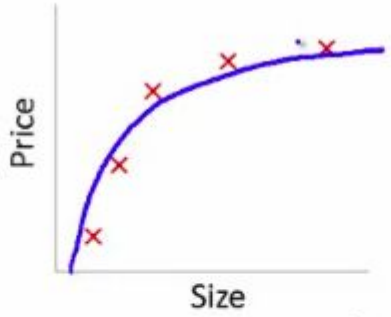
2. Learning curves reveal quality of ML model

3. Rate and offset depend on
   a. Baseline: Δ-ML (prior expert knowledge about property)
   b. Representation: Uniqueness and energy (property independent)
   c. Domain of applicability: Data (properties and structures)



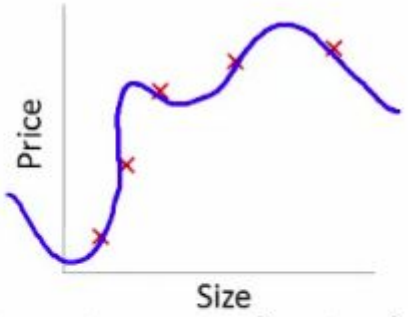
$\theta_0 + \theta_1 x$

High bias (underfit)



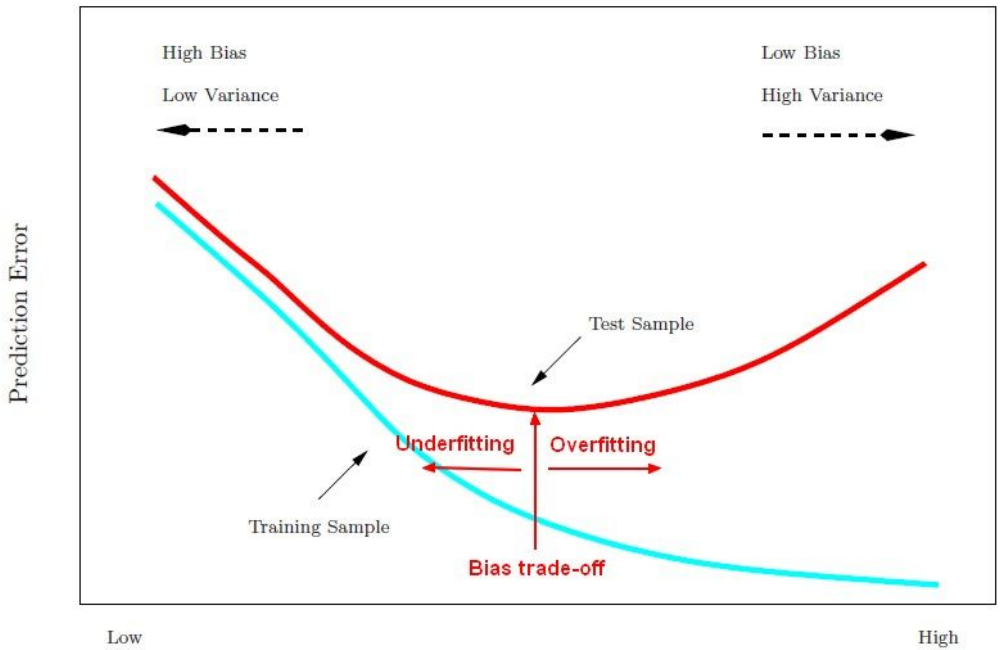
$\theta_0 + \theta_1 x + \theta_2 x^2$

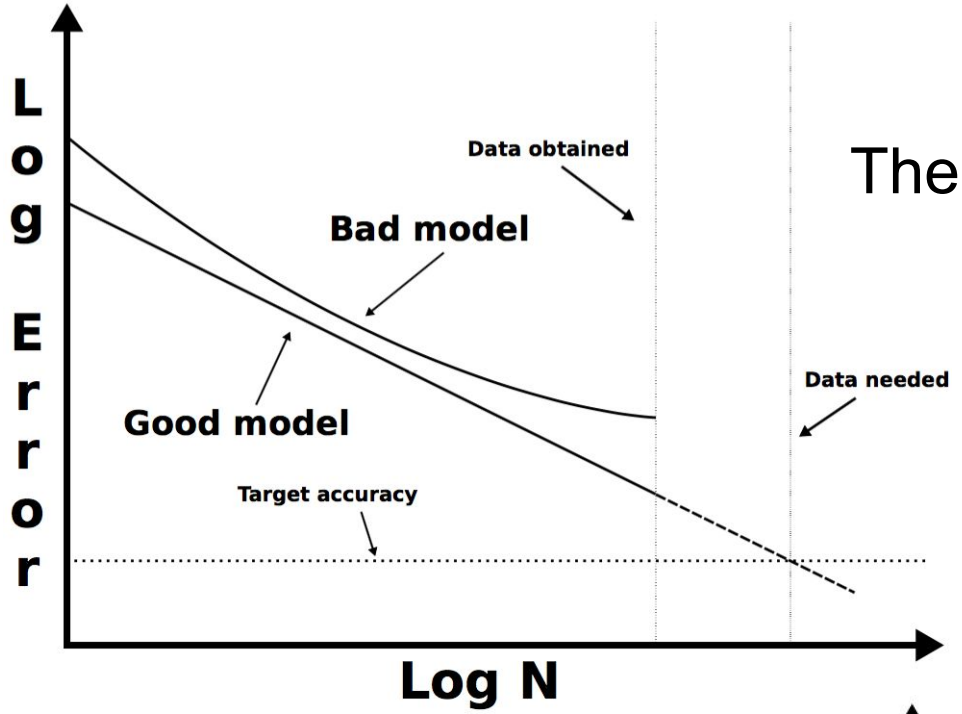
"Just right"



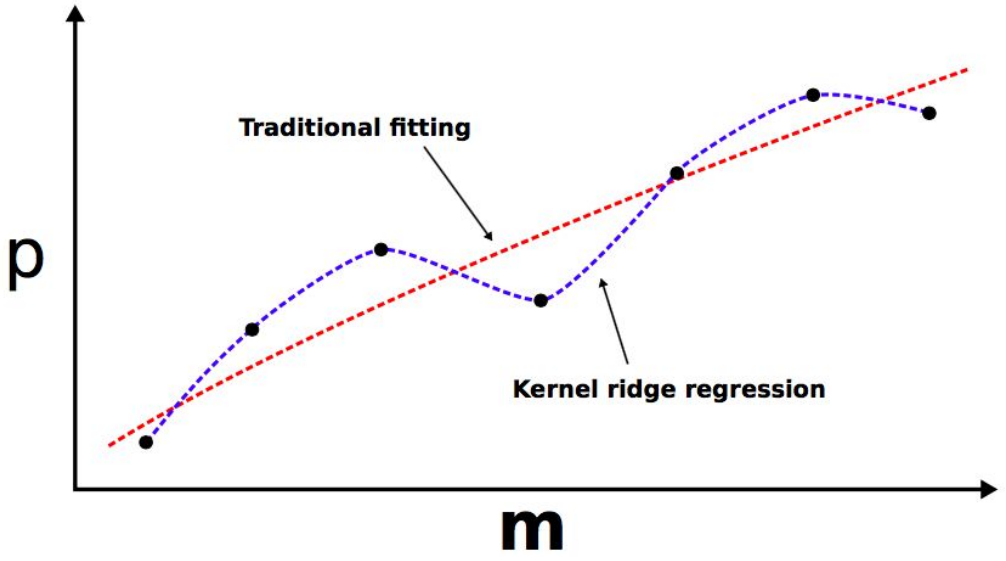
$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

High variance (overfit)





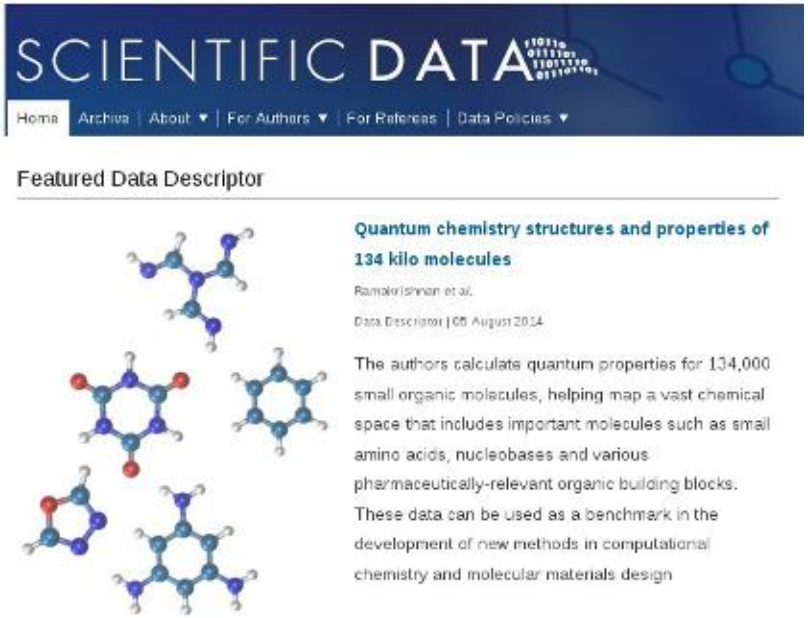
The bigger the data the better …



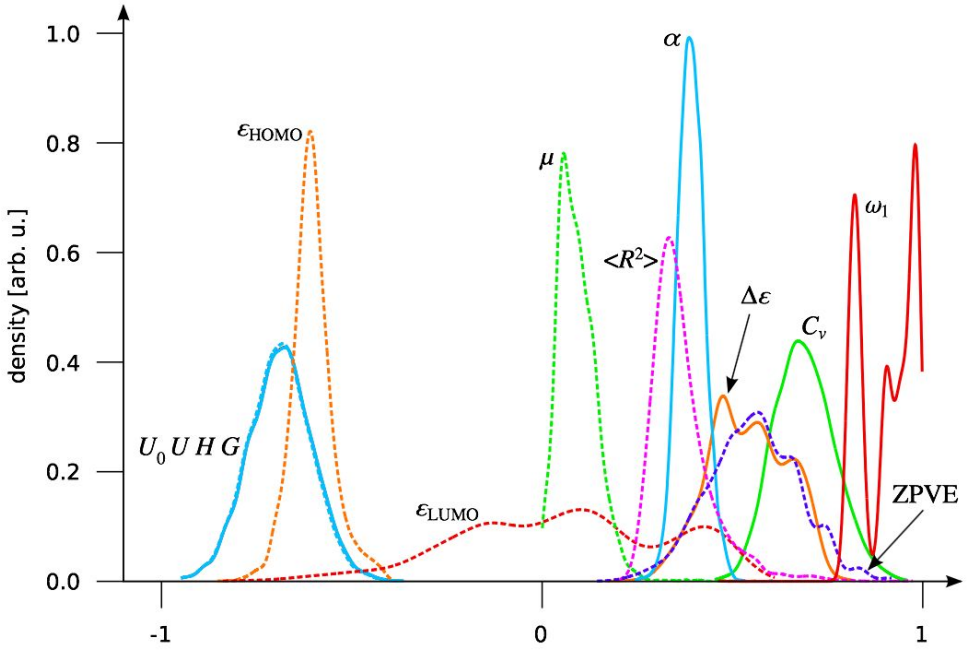
Error ~ a/N^b

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

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

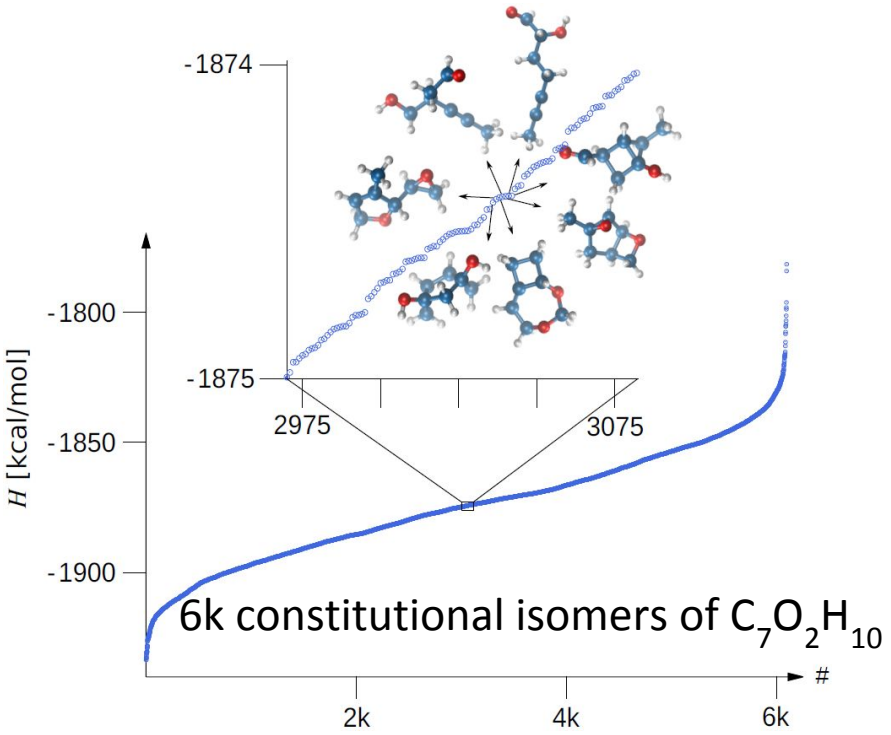


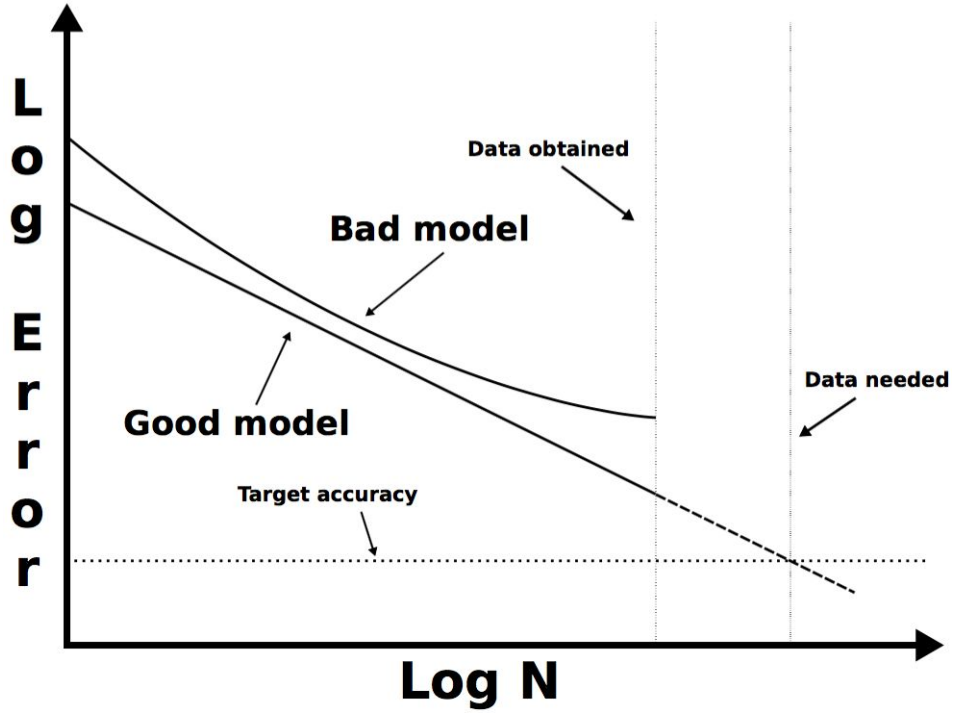
Ramakrishnan et al, *Scientific Data* (2014)

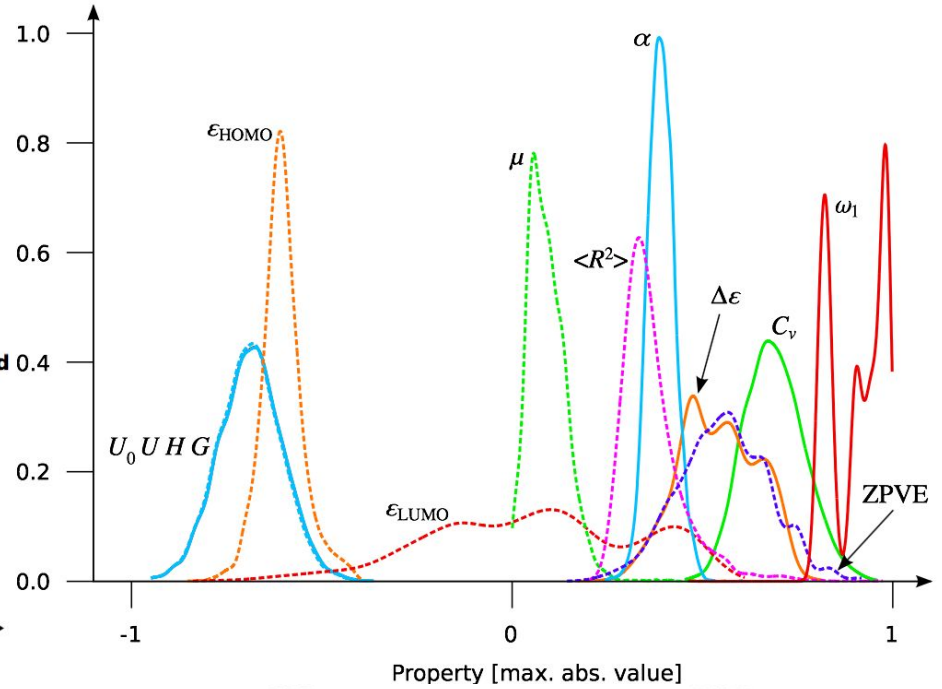


``Enumeration surpasses imagination''
J.-L. Reymond

Reymond, *Acc Chem Res* (2015)

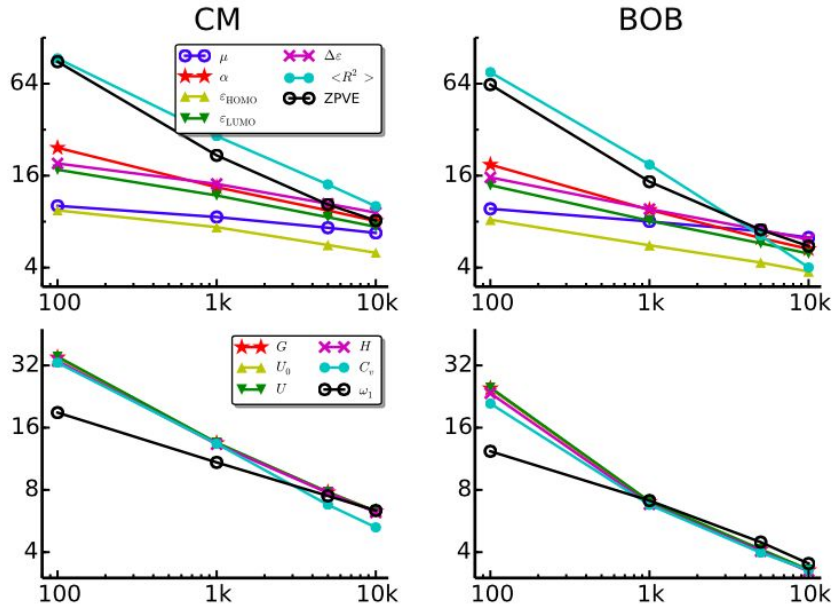






$$P^{\text{est}}(\mathbf{M}) = \sum_i \alpha_i k(\mathbf{M}, \mathbf{M}_i)$$
$$\vec{\alpha} = \mathbf{K}^{-1} \vec{P}^{\text{ref}}$$

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

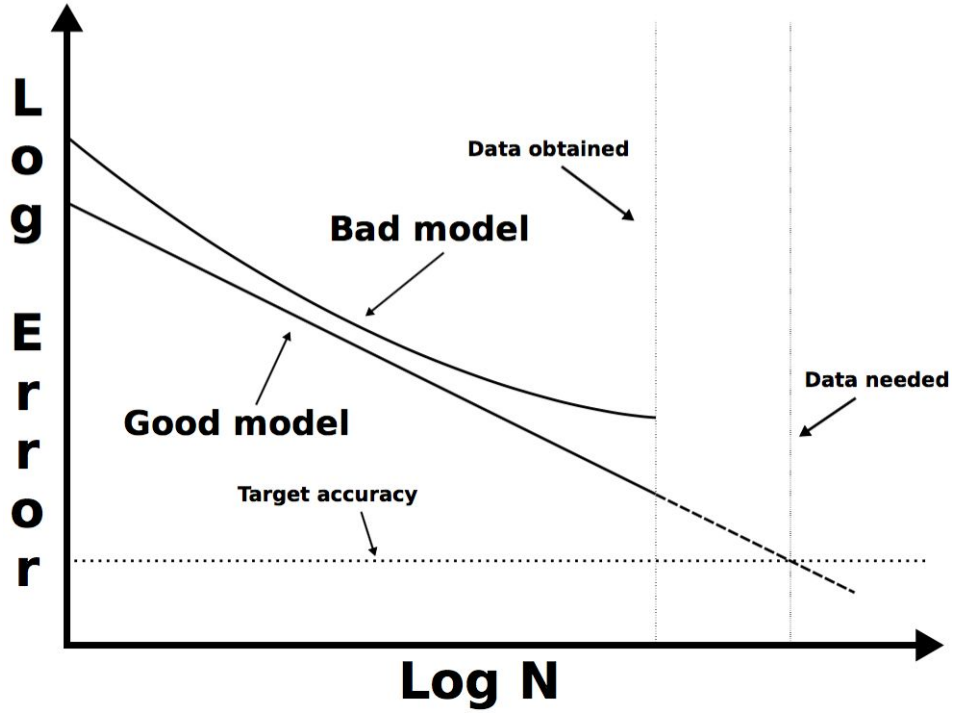


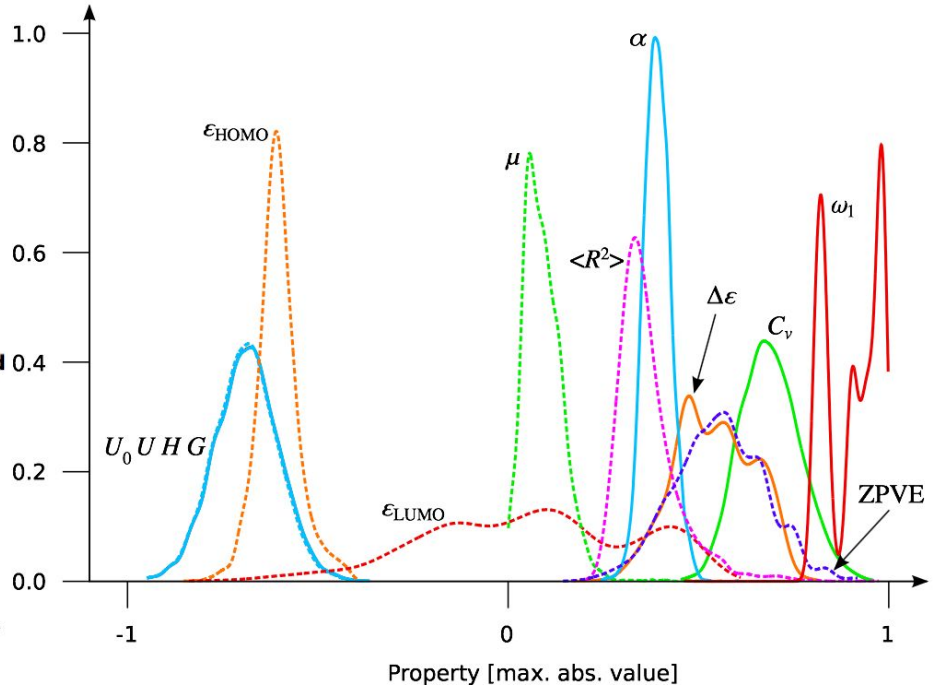
Error ~ a/N^b

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

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

Ramakrishnan, OAvL, *CHIMIA* (2015)





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

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

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

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

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

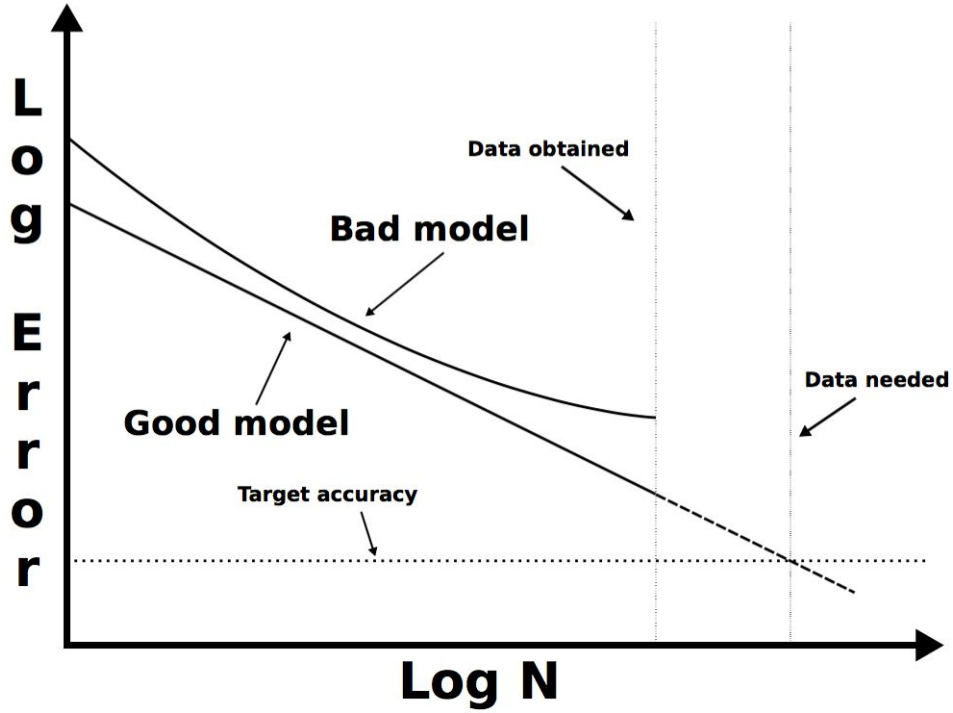
K.-R. Mueller et al, *Neural Comput* (1996)

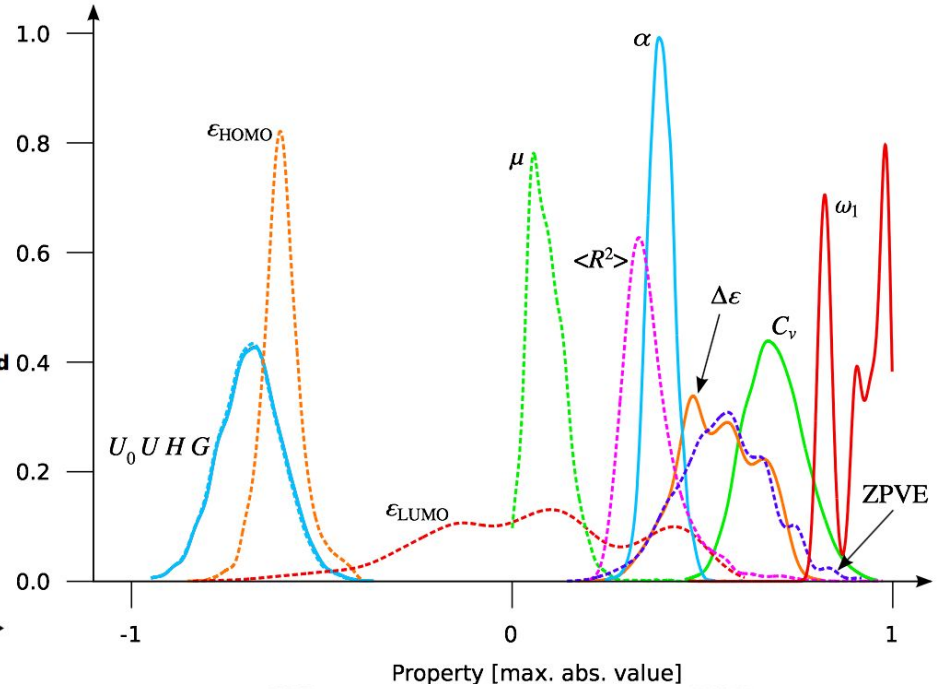
$\rightarrow \log(\text{Error}) = a - b\log(N)$

κ ~ Ψ

α ~ Ô

Ramakrishnan, OAvL, *CHIMIA* (2015)





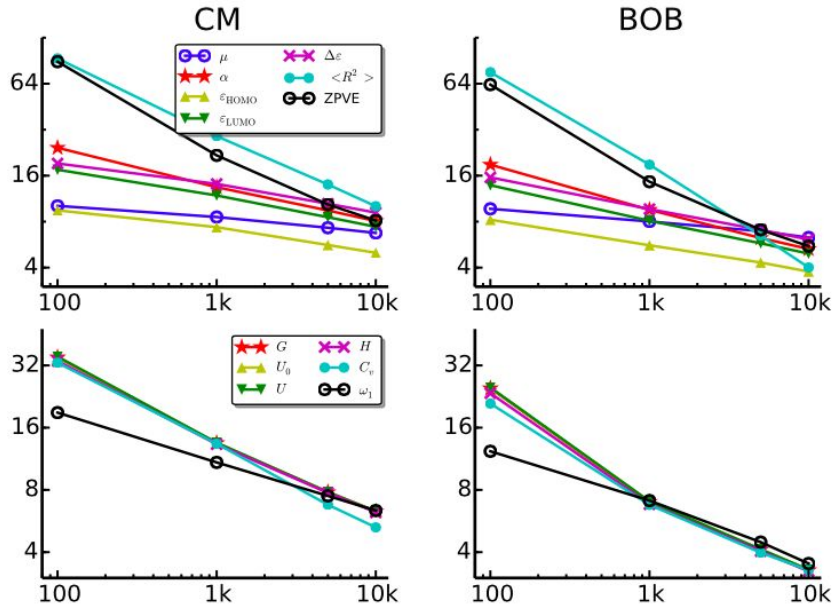
$$P^{\text{est}}(\mathbf{M}) = \sum_i \alpha_i k(\mathbf{M}, \mathbf{M}_i)$$
$$\vec{\alpha} = \mathbf{K}^{-1} \vec{P}^{\text{ref}}$$

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

Error ~ $a/(N')^b$, e.g. $N' = N/3$

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

$\rightarrow$ log(Error) = $a + 3b$ - b log(N)



Ramakrishnan, OAvL, *CHIMIA* (2015)

Conclusions

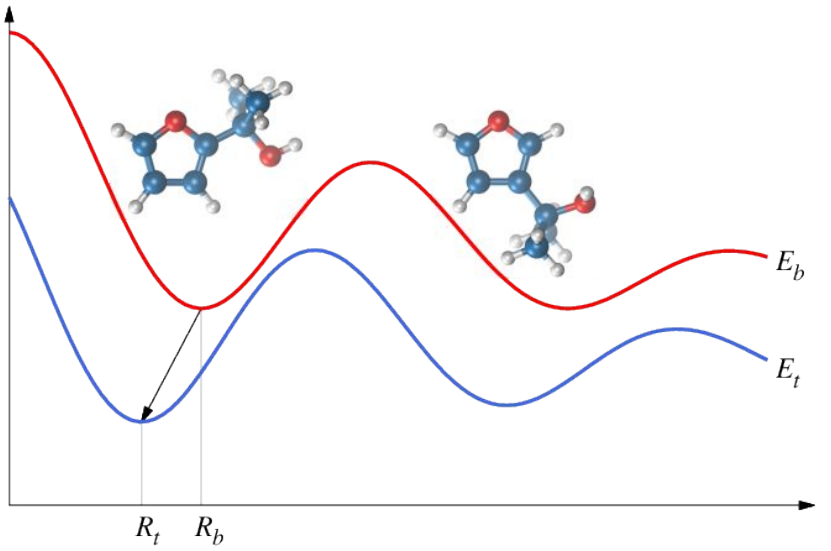
1. Instantaneous QM quality predictions

2. Learning curves reveal quality of ML model

3. Rate and offset depend on
   a. Baseline: Δ-ML (prior expert knowledge about property)
   b. Representation: Uniqueness and energy (property independent)
   c. Domain of applicability: Data (properties and structures)

Δ-ML

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



JCTC Journal of Chemical Theory and Computation Article
pubs.acs.org/JCTC

Big Data Meets Quantum Chemistry Approximations: The Δ-Machine Learning Approach

Raghunathan Ramakrishnan,[†] Pavlo O. Dral,[¶,‡] Matthias Rupp,[†] and O. Anatole von Lilienfeld*[,†,§]

[†]Institute of Physical Chemistry and National Center for Computational Design and Discovery of Novel Materials, Department of Chemistry, University of Basel, Klingelbergstraße 80, CH-4056 Basel, Switzerland

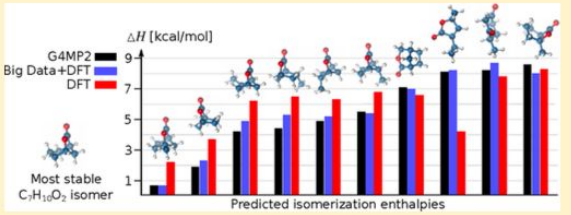
[¶]Max-Planck-Institut für Kohlenforschung, Kaiser-Wilhelm-Platz 1, 45470 Mülheim an der Ruhr, Germany

[‡]Computer-Chemie-Centrum and Interdisciplinary Center for Molecular Materials, Department Chemie und Pharmazie, Friedrich-Alexander-Universität Erlangen-Nürnberg, Nägelsbachstraße 25, 91052 Erlangen, Germany

[§]Argonne Leadership Computing Facility, Argonne National Laboratory, 9700 S. Cass Avenue, Lemont, Illinois 60439, United States

Supporting Information

ABSTRACT: Chemically accurate and comprehensive studies of the virtual space of all possible molecules are severely limited by the computational cost of quantum chemistry. We introduce a composite strategy that adds machine learning corrections to computationally inexpensive approximate legacy quantum methods. After training, highly accurate predictions of enthalpies, free energies, entropies, and electron correlation energies are possible, for significantly larger molecular sets than used for training. For thermochemical properties of up to 16k isomers of $C_7H_{10}O_2$ we present numerical evidence that chemical accuracy can be reached. We also predict electron correlation energy in post Hartree−Fock methods, at the computational cost of Hartree−Fock, and we establish a qualitative relationship between molecular entropy and electron correlation. The transferability of our approach is demonstrated, using

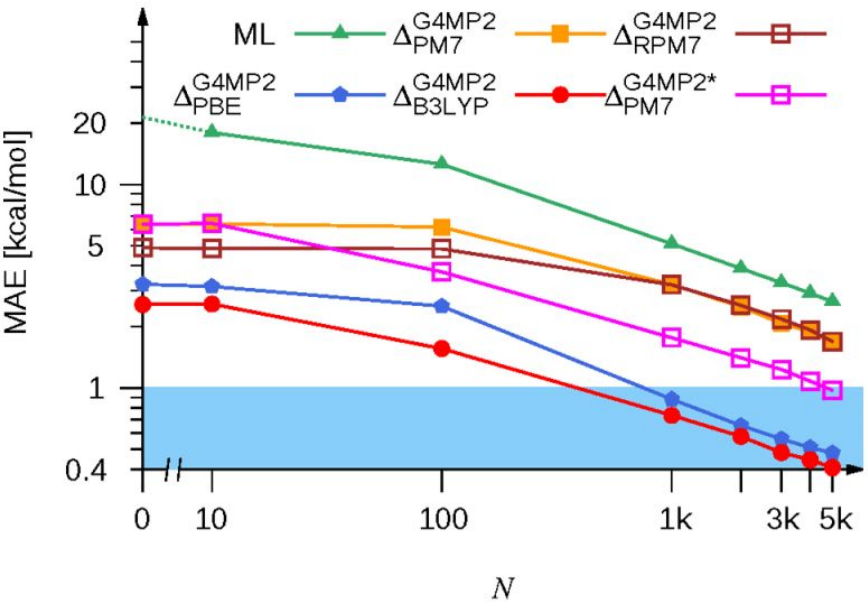


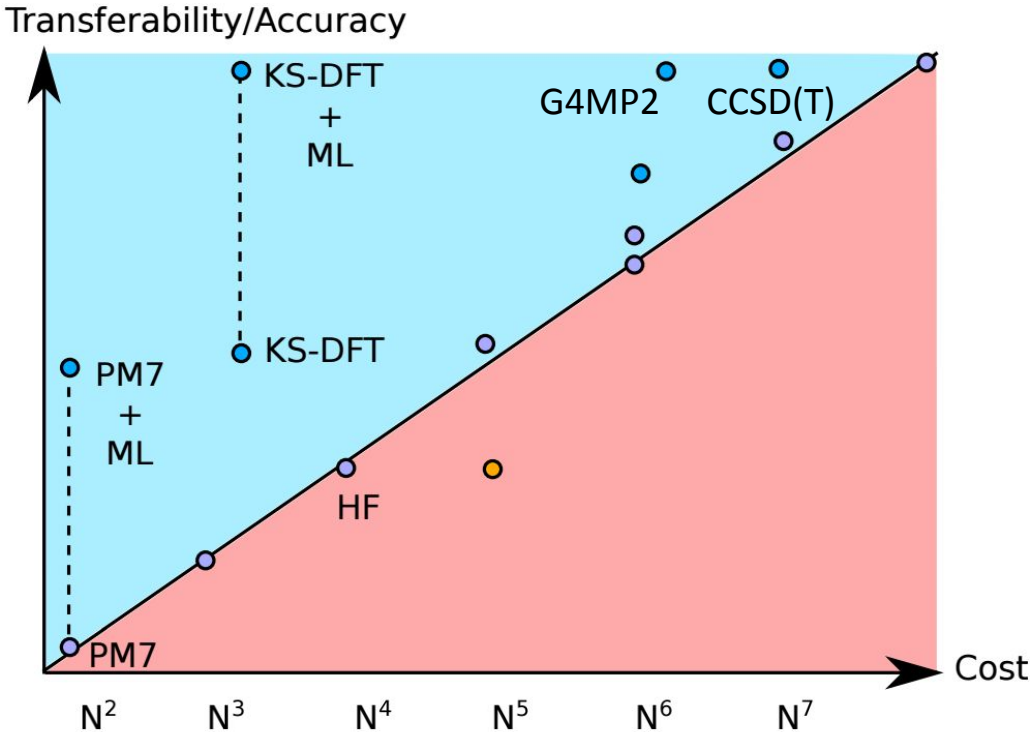
Ramakrishnan et al, *J Chem Theory Comput* (2015)

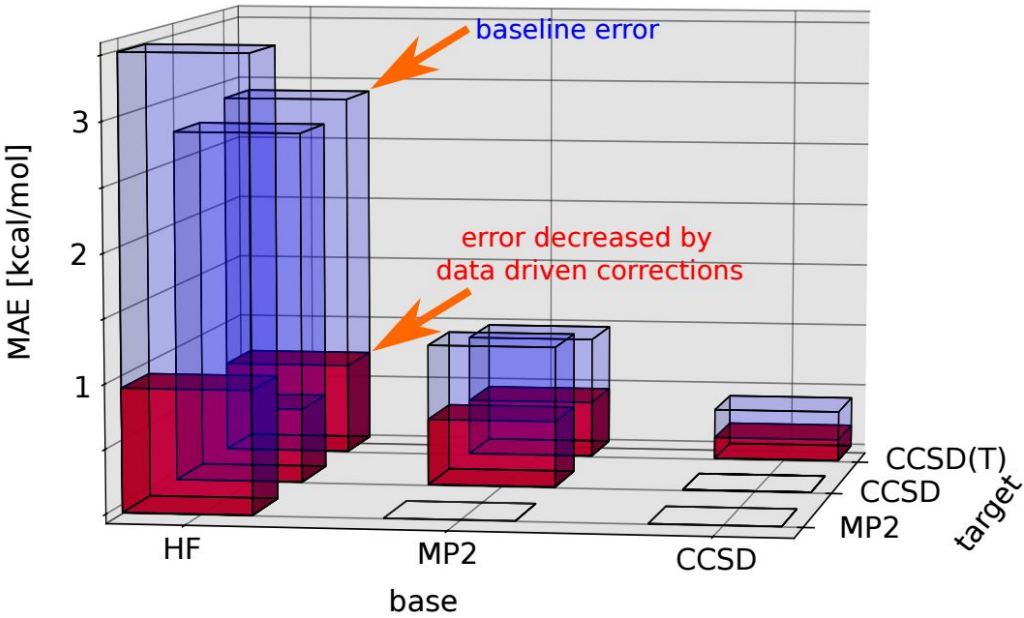
Δ-ML

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

6k constitutional isomers of $C_7O_2H_{10}$





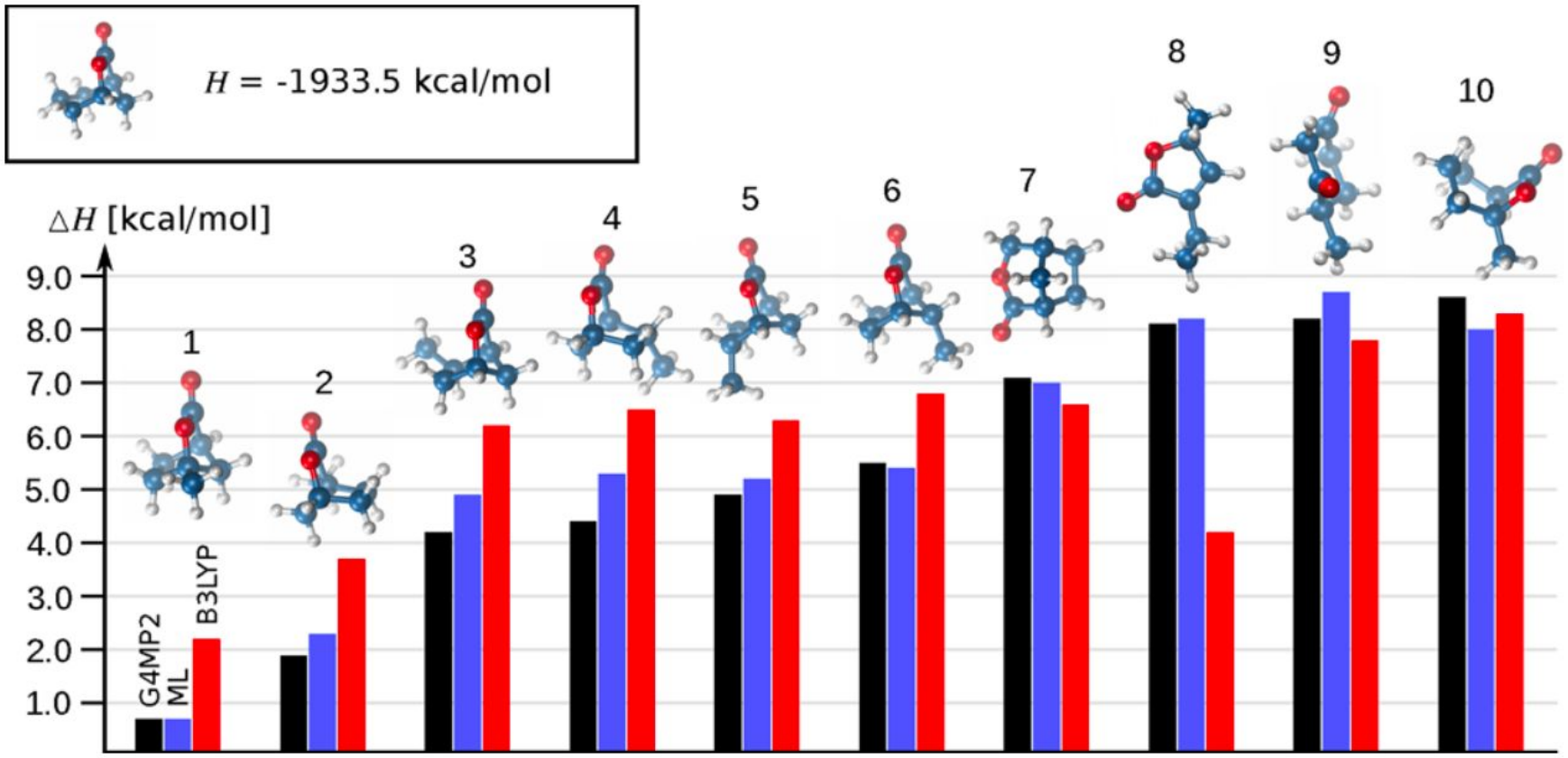


Ramakrishnan et al, *J Chem Theory Comput* (2015)

Δ-ML

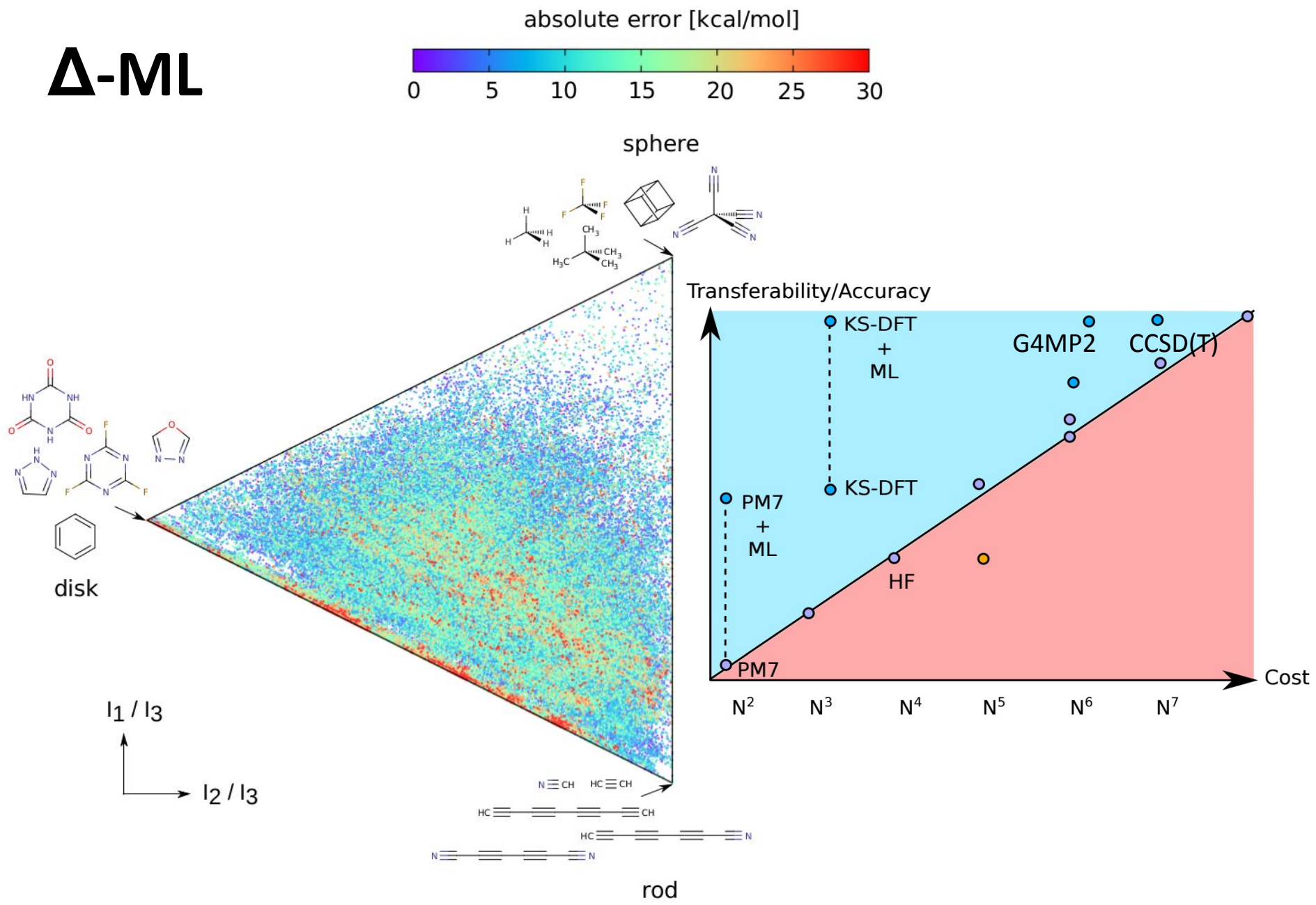
$$E_t(R_t) \approx E_b(R_b) + \Delta_b^t(R_b)$$

Ranking 10k diastereomers derived from 6k constitutional isomers of $C_7O_2H_{10}$
→ Global minimum, and its 10 closest isomers …



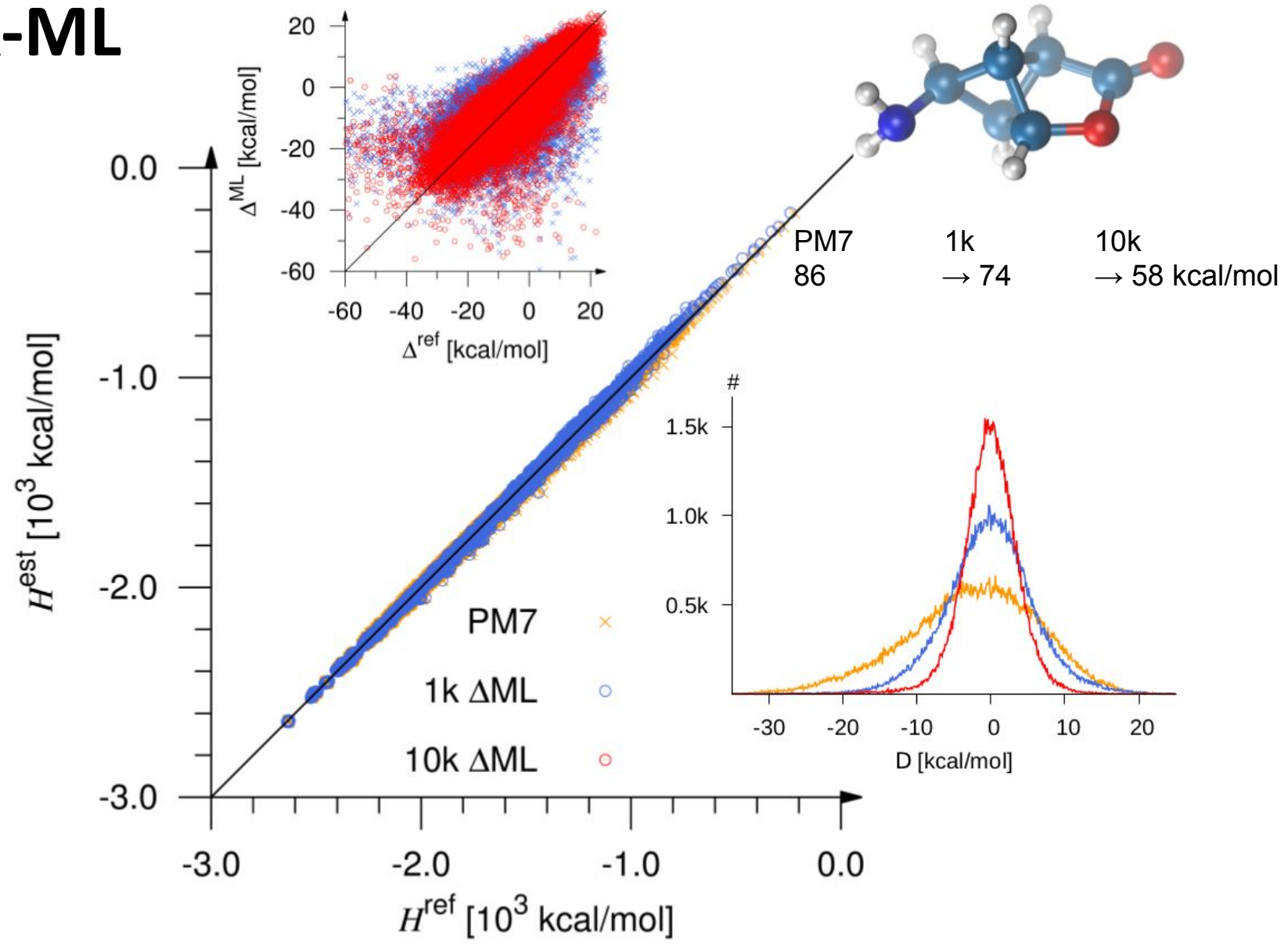
Ramakrishnan et al, *J Chem Theory Comput* (2015)

Δ-ML



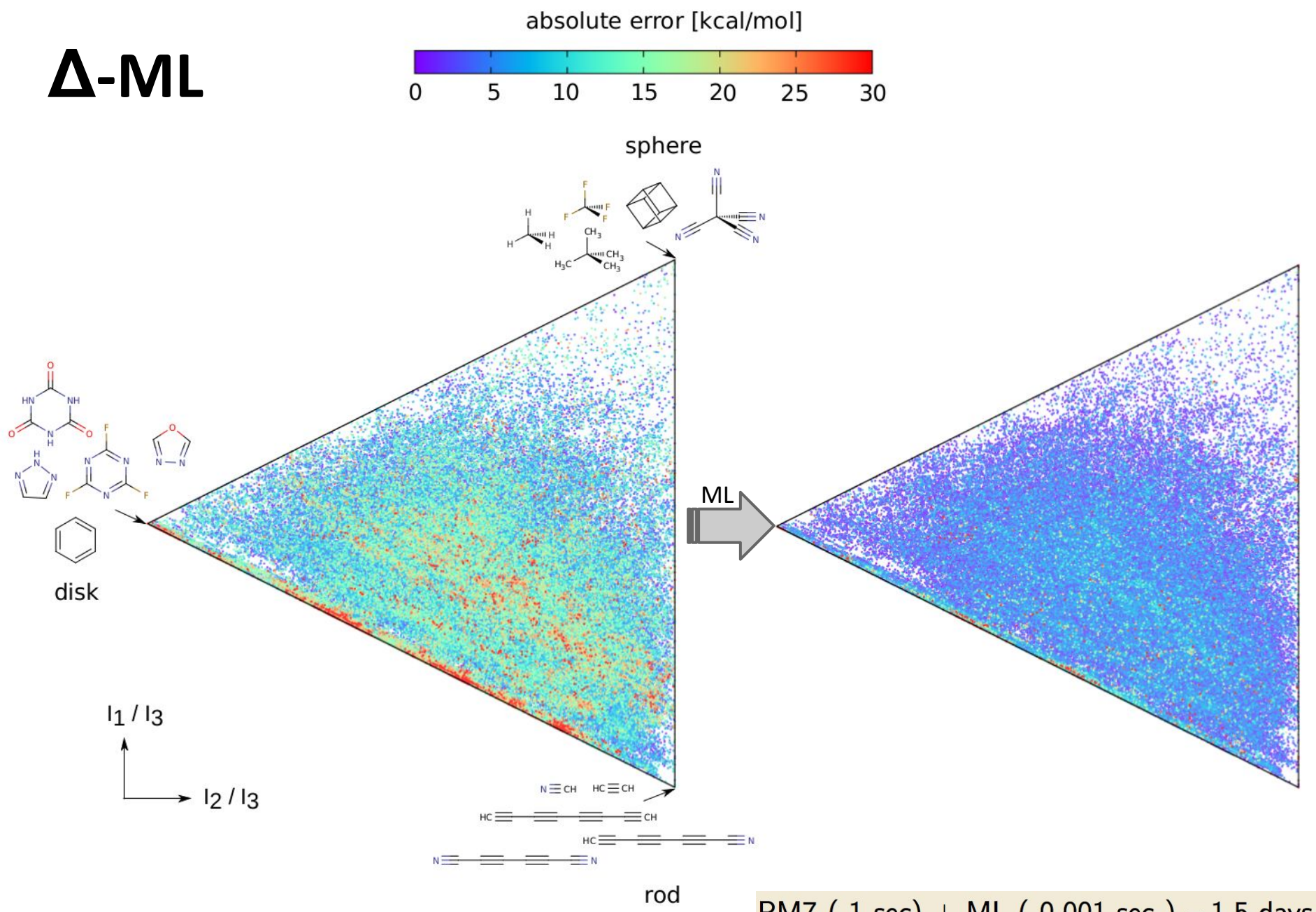
R Ramakrishnan et al *JCTC* (2015)

Δ-ML



R Ramakrishnan et al *JCTC* (2015)

Δ-ML



PM7 (1 sec) + ML (0.001 sec) – 1.5 days

DFT (30 min) – 8 yrs !

R Ramakrishnan et al *JCTC* (2015)

Conclusions

1. Instantaneous QM quality predictions

2. Learning curves reveal quality of ML model

3. Rate and offset depend on
   a. Baseline: Δ-ML (prior expert knowledge about property)
   b. Representation: Uniqueness and energy (property independent)
   c. Domain of applicability: Data (properties and structures)

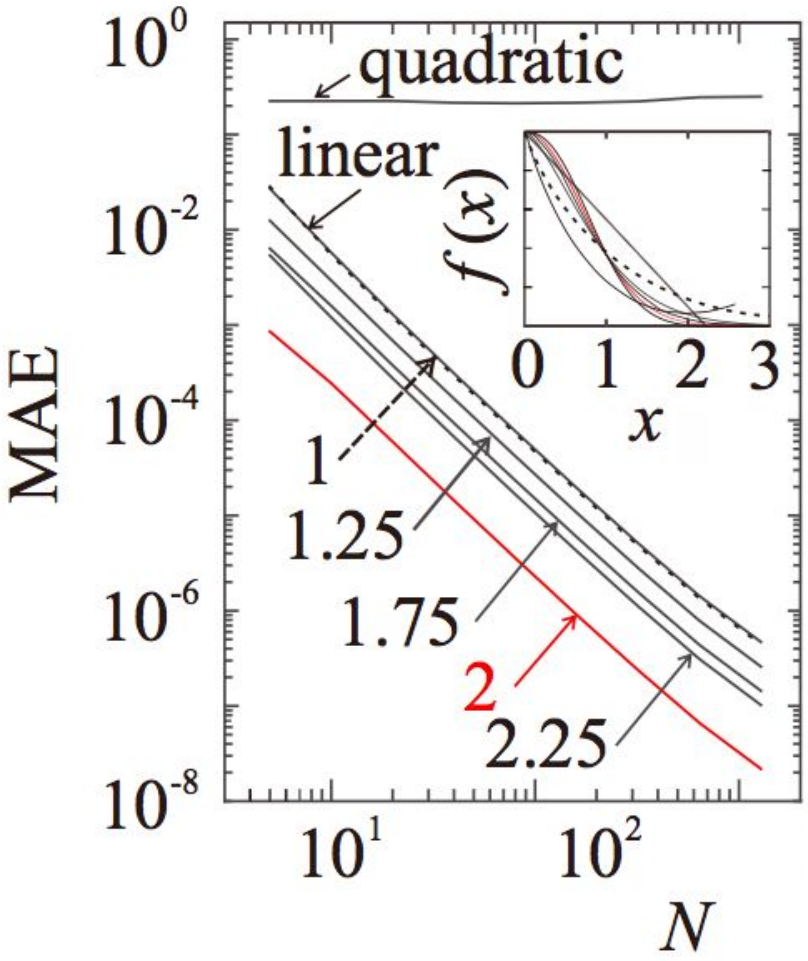
Representation

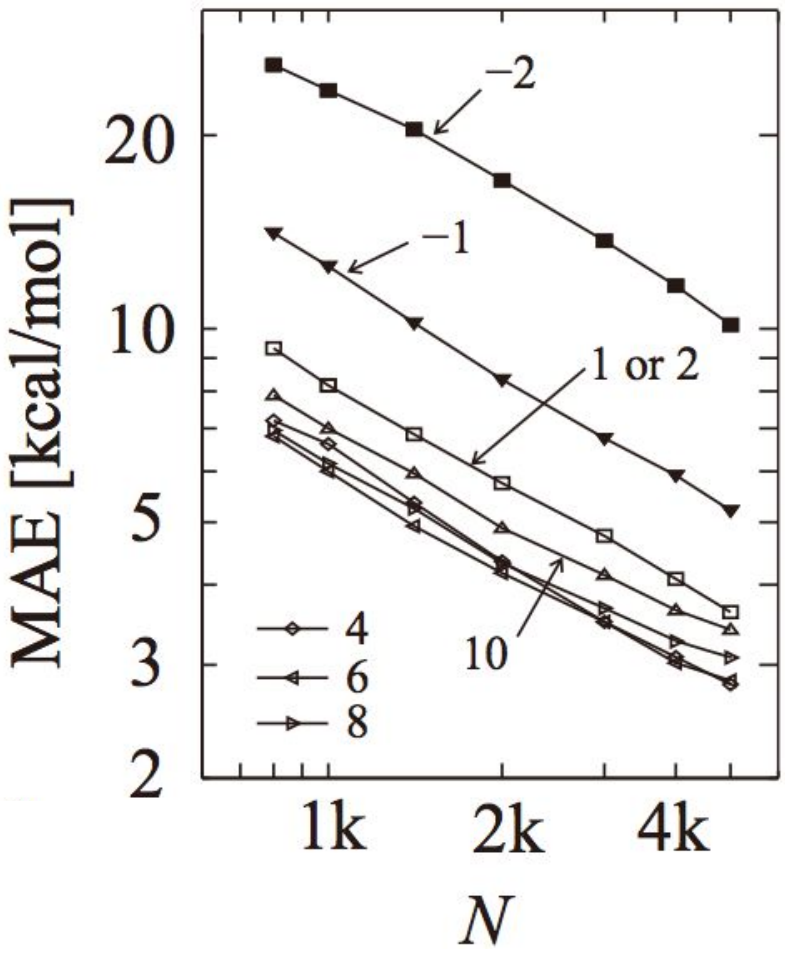
$$\log(\text{Error}) = a - b\log(N)$$

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

$$\text{CM}_{IJ}^{(n)} = \frac{Z_I Z_J}{R_{IJ}^n}$$

target similarity





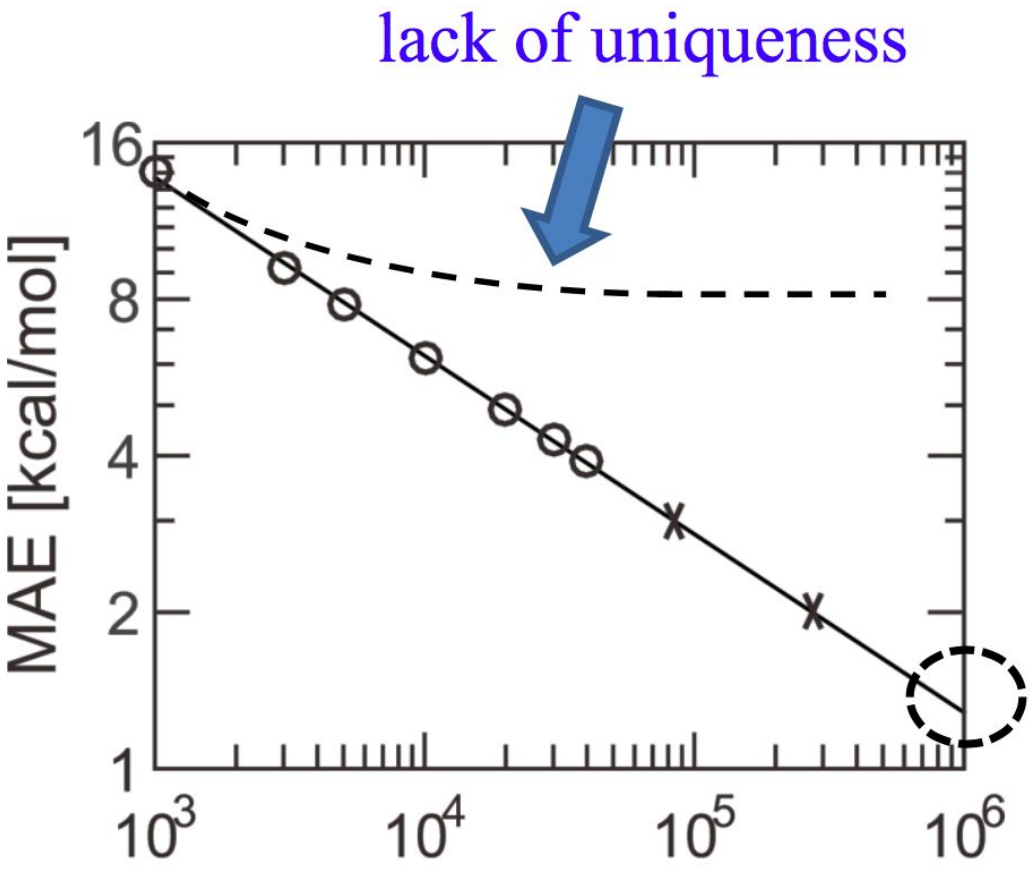
Huang, OAvL, *accepted in J Chem Phys Comm* (2016) arxiv.org/abs/1608.06194

Representation

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

target similarity

uniqueness



lack of uniqueness → absurd results → noise in training

OAvL et al, *IJQC* (2013)

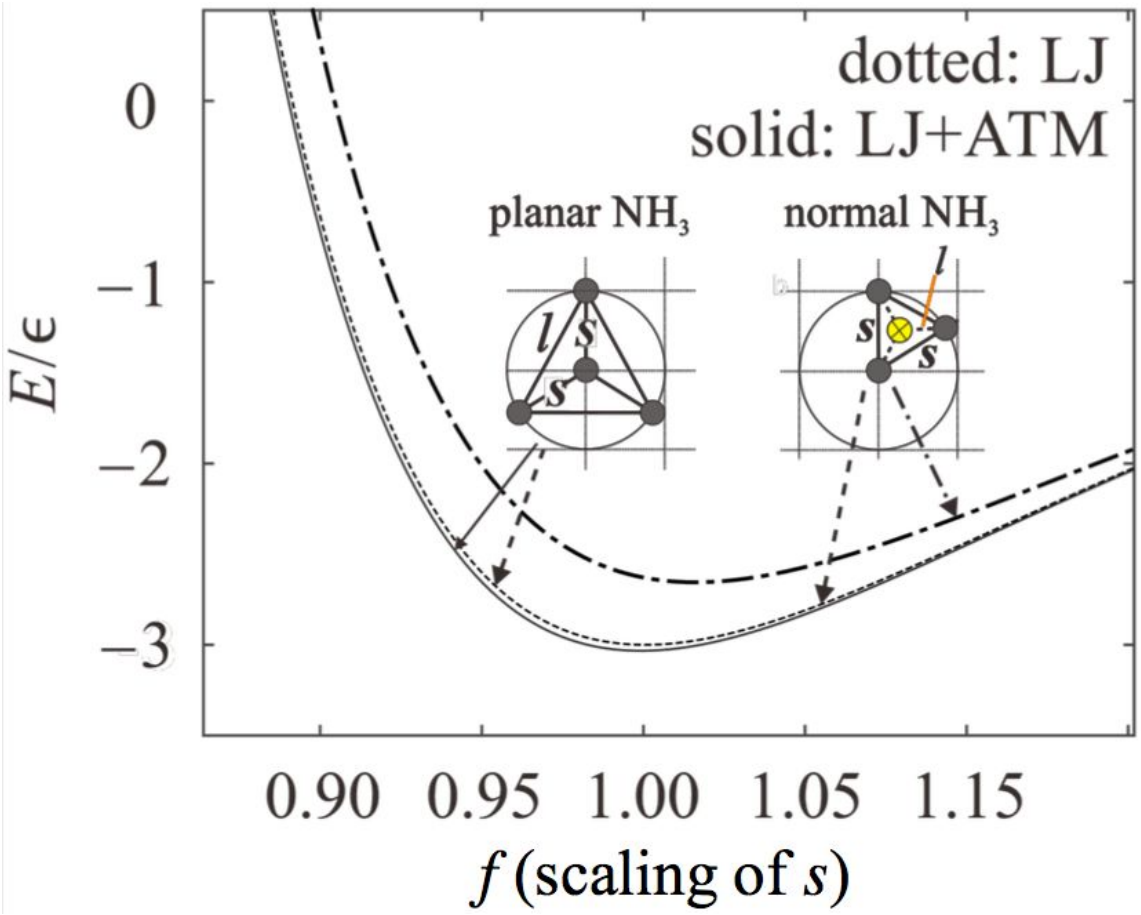
Huang, OAvL, *accepted in J Chem Phys Comm* (2016) arxiv.org/abs/1608.06194

Representation

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

target similarity

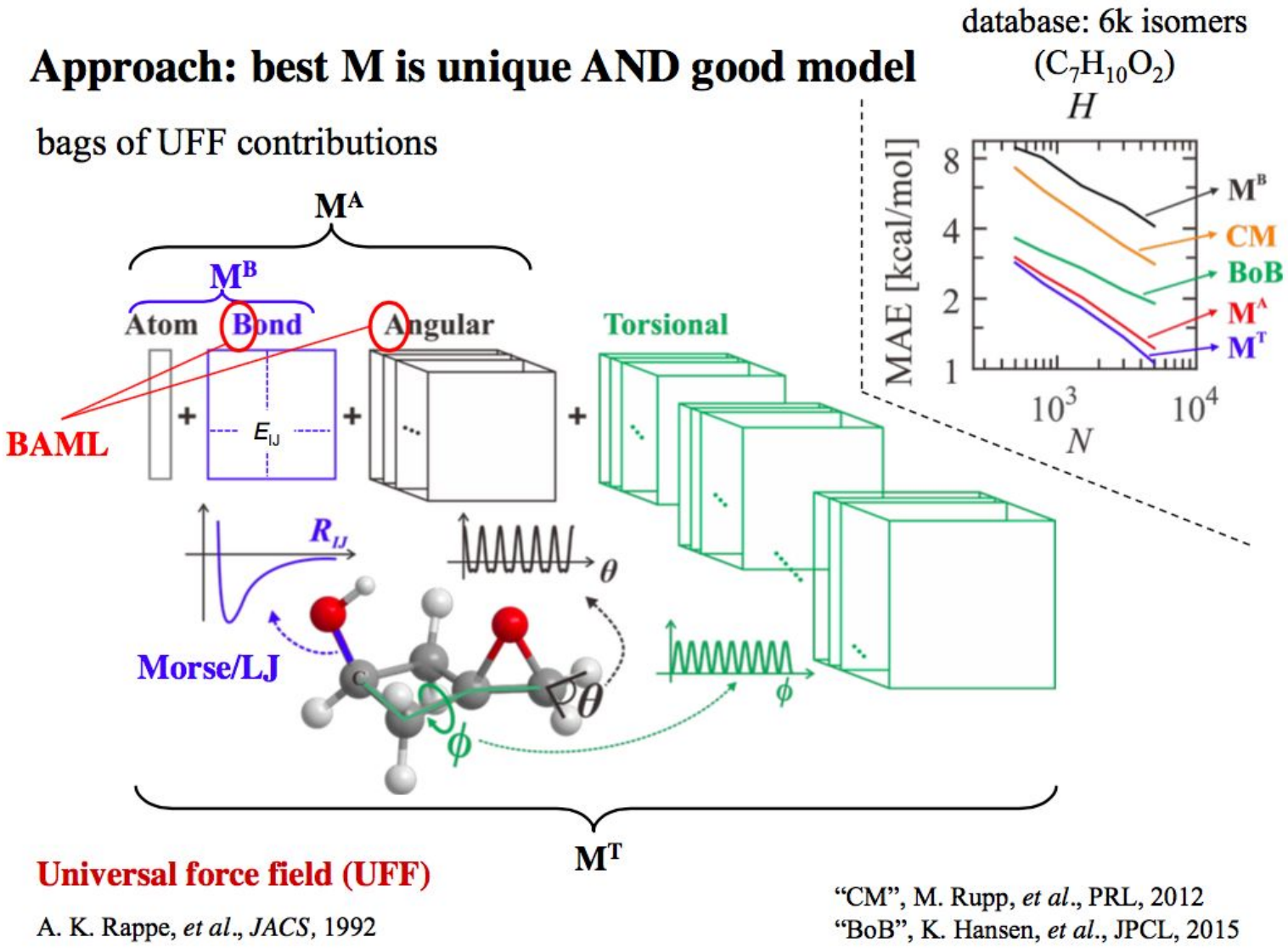
uniqueness



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

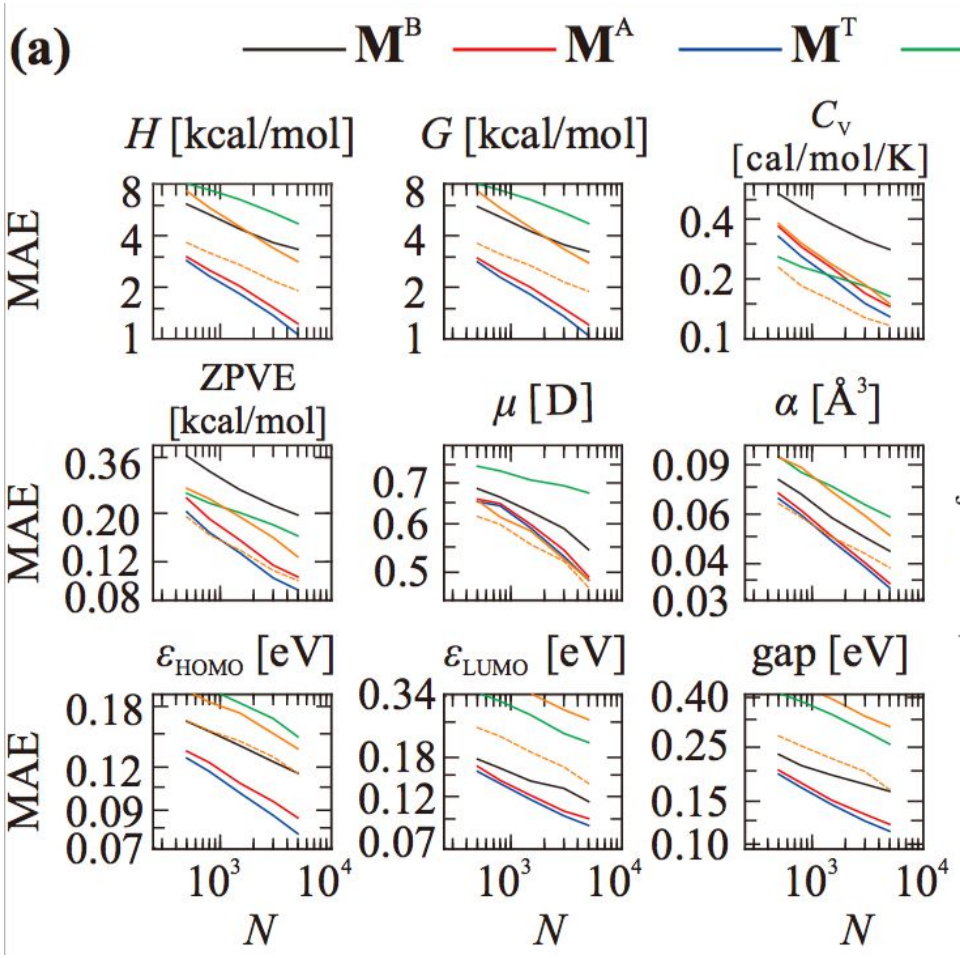
Huang, OAvL, *accepted in J Chem Phys Comm* (2016) arxiv.org/abs/1608.06194

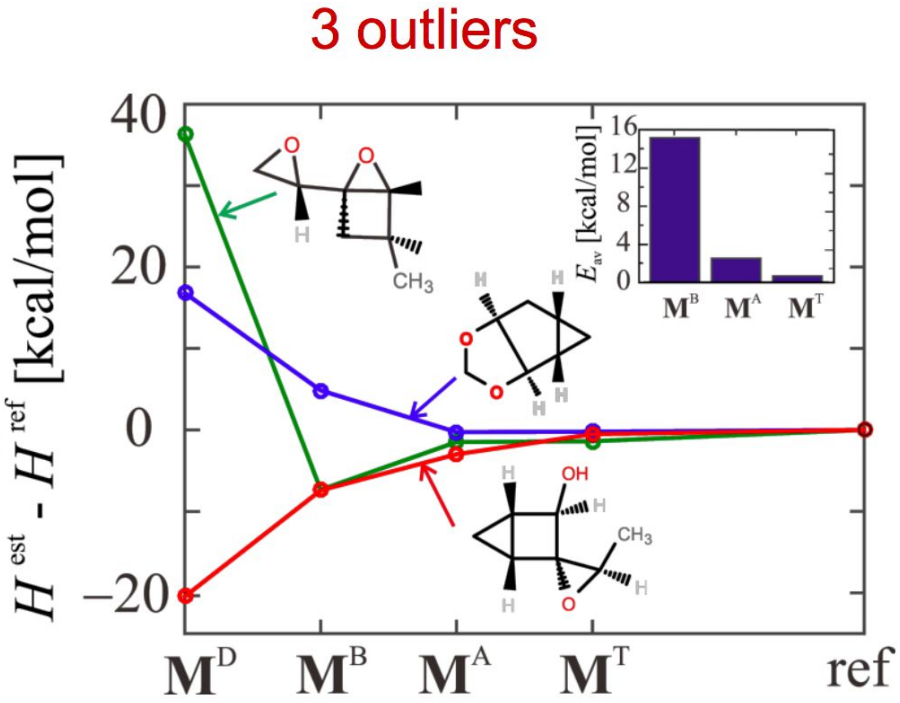
Representation



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

Representation

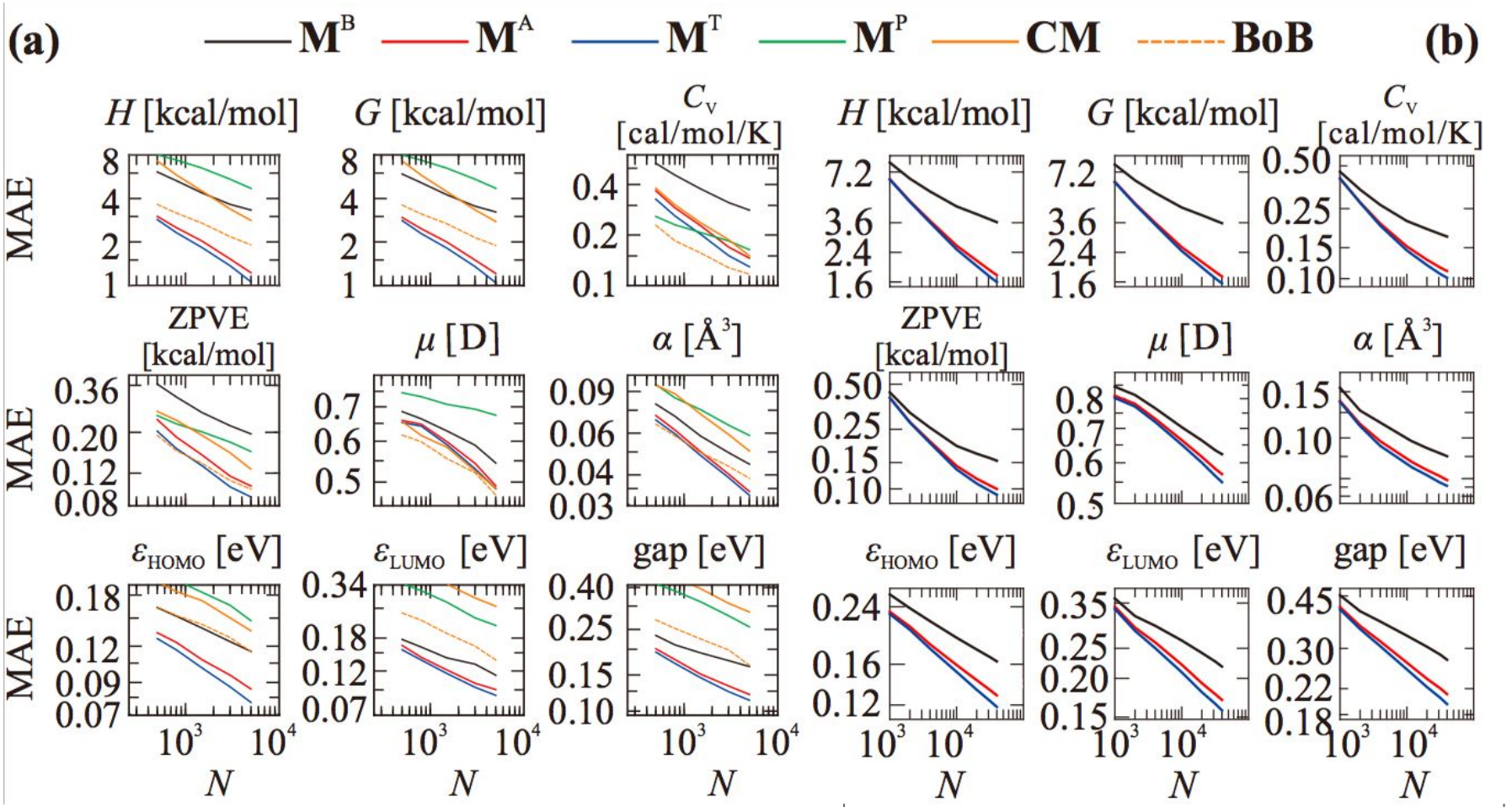




6k constitutional isomers of $C_7O_2H_{10}$

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

Representation



6k constitutional isomers of $C_7O_2H_{10}$

QM9 (134k molecules)

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

Representation

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, *accepted in J Chem Phys Comm* (2016) arxiv.org/abs/1608.06194

Conclusions

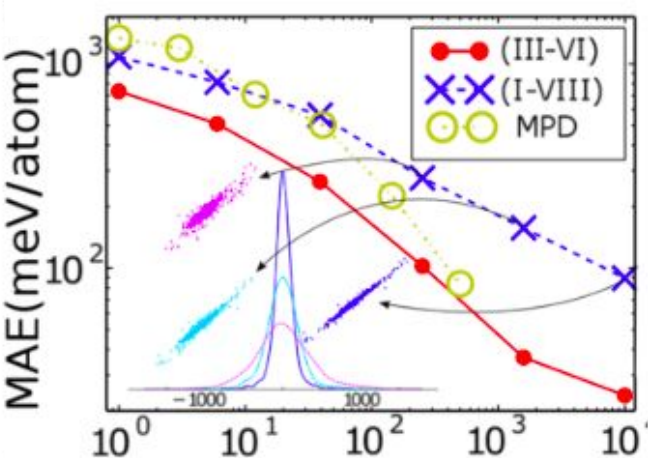
1. ML makes instantaneous QM quality predictions

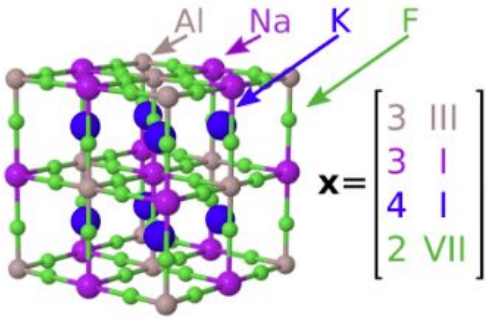
2. Learning curves reveal quality of ML model

3. Rate and offset depend on
   a. Uniqueness and target similarity (model dependent)
   b. Properties and structures (data dependent)
   c. Baseline (expert dependent)

Crystals

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. Some Elpasolites emit light when exposed to ionic radiation. This makes them interesting material candidates for scintillator devices.





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

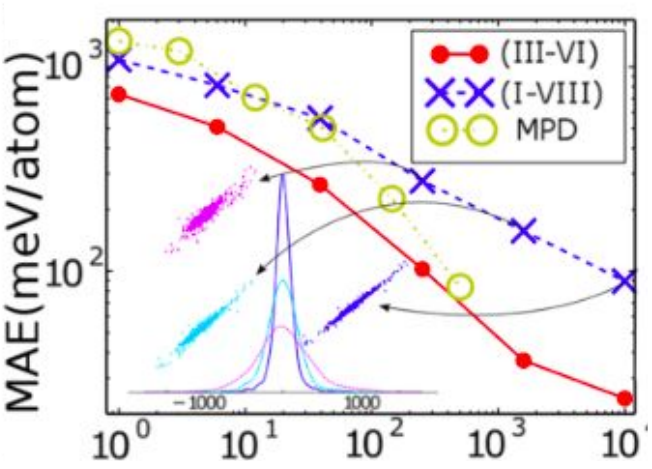
$$E^{est}(\mathbf{M}) = \sum_i^N \alpha_i k(\mathbf{M}, \mathbf{M}_i)$$

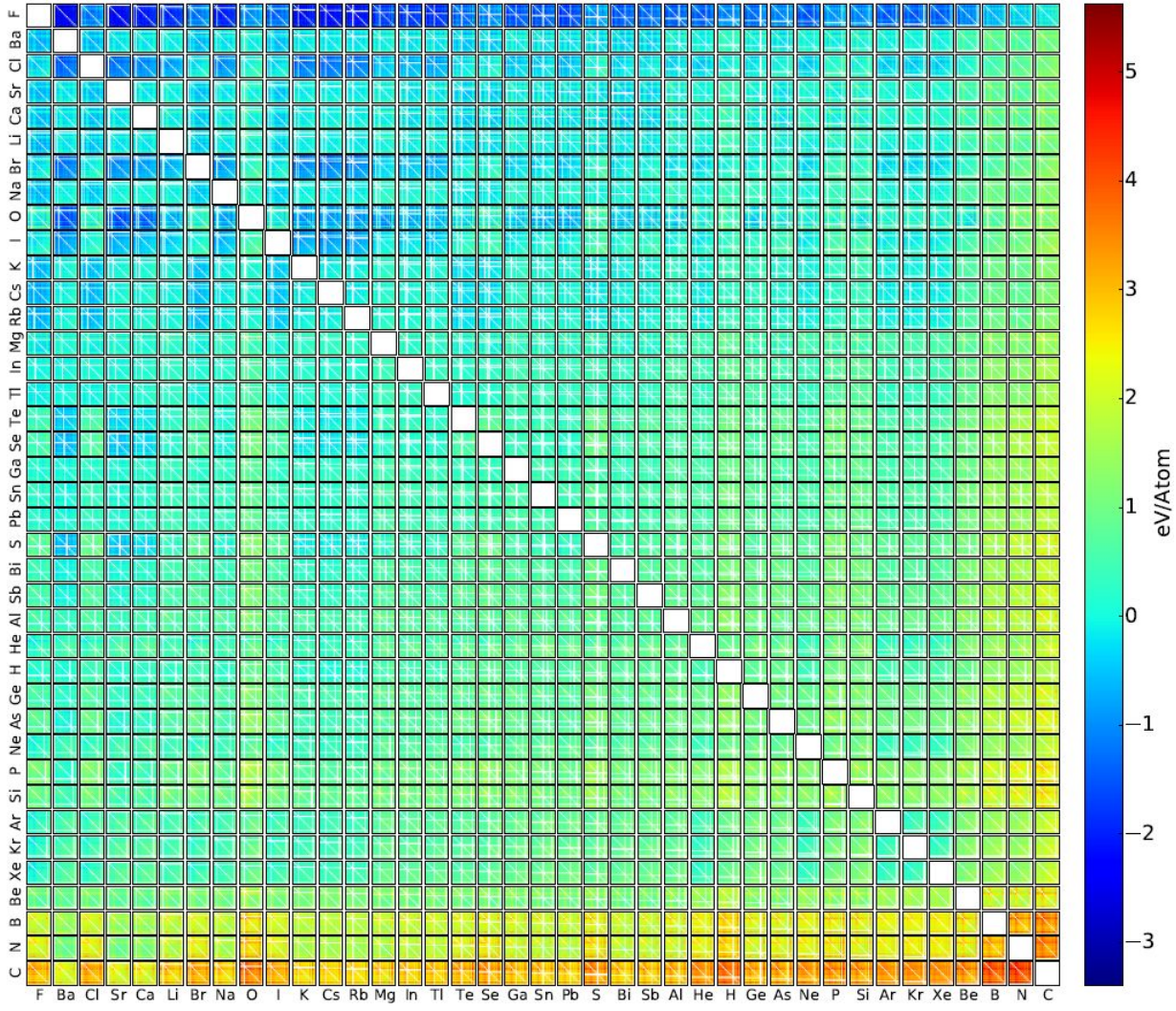
$$\alpha = (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{E}^{ref}$$

Faber et al, *Phys Rev Lett* (2016) arxiv.org/abs/1508.05315

Crystals

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. Some Elpasolites emit light when exposed to ionic radiation. This makes them interesting material candidates for scintillator devices.





Reduction in cost:
DFT: ~20 M CPU hours (optimized)
ML: ~20 CPU hours (not optimized)

Faber et al, *Phys Rev Lett* (2016) arxiv.org/abs/1508.05315

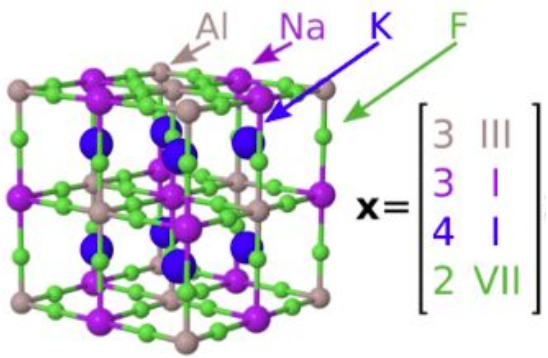
Crystals

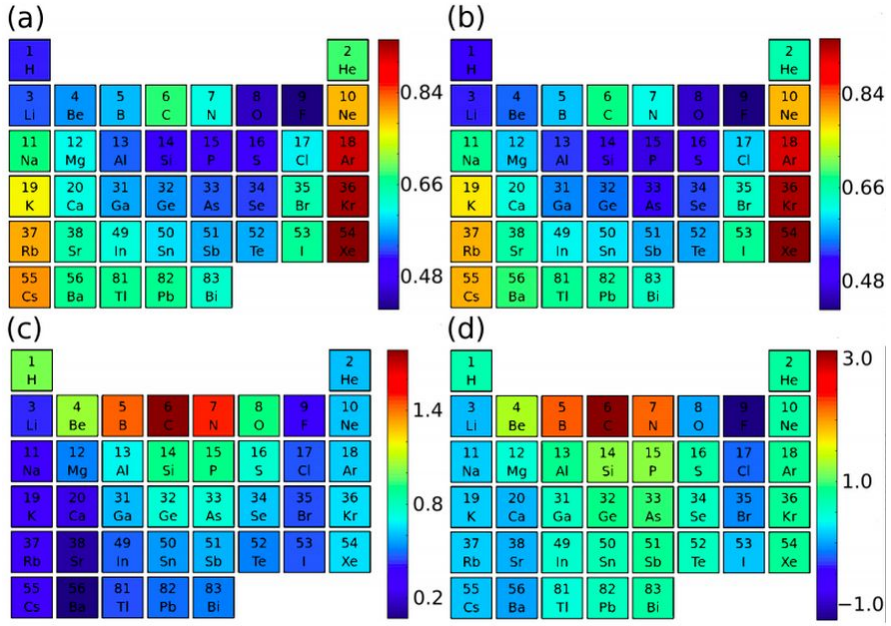
TABLE VI – *Continued from previous page*

#	Elpasolite				ΔE (eV/atom)	Type	$\Delta\epsilon$ (eV)	MP ID	Competing mixed phase in MP
	x_1	x_2	x_3	x_4					
50	Cl	N	Tl	Ba	−0.0312	C	0.00	mp-989542	$0.5\,Ba_6N_2 + 0.5\,Ba_1Cl_2 + 0.5\,Ba_2Tl_4 + 1.5\,Ba_1$
51	Tl	K	Cs	F	−0.0329	I	3.82	mp-989526	$0.6667\,Cs_3Tl_1F_6 + 0.0417\,K_{24}Tl_8F_{48}$
52	N	O	Sn	Sr	−0.0333	C	0.00	mp-989540	$Sr_2N_1 + Sr_3Sn_1O_1 + 0.5\,Sr_2Sn_2$
53	S	F	Cs	Cl	−0.0336	C	0.00	mp-989521	$0.1458\,Cl_{16} + 0.1667\,S_1F_6 + 0.1042\,S_8Cl_{16} + 2\,Cs_1Cl_1$
54	Se	Cl	Cs	F	−0.0349	C	0.03	mp-989544	$0.5\,Cs_1Cl_1 + 0.125\,Cl_4F_4 + 1.5\,Cs_1F_1 + 0.25\,Se_4F_{16}$
55	Ga	K	Cs	F	−0.0359	I	6.04	mp-989531	$0.5\,Cs_1F_1 + 0.0833\,Cs_{18}Ga_{12}F_{54} + K_1F_1$
56	Pb	Rb	Cs	F	−0.0370	C	0.00	mp-989525	$2\,Cs_1F_1 + 0.25\,Pb_4F_{12} + Rb_1F_1$
57	F	Br	Rb	Cl	−0.0382	S	0.92	mp-989573	$0.5\,Cl_4 + 0.5\,Br_2Cl_2 + 2\,Rb_1Cl_1 + 0.25\,Cl_4F_4$
58	N	Rb	Cs	F	−0.0384	S	2.96	mp-989519	$0.5\,Rb_2F_6 + 0.125\,N_8 + 2\,Cs_1F_1 + 0.25\,F_4$
59	N	Li	Na	F	−0.0396	S	2.72	mp-989504	$0.375\,F_8 + 0.5\,N_2 + Li_1F_1 + 2\,Na_1F_1$
60	Bi	Na	Rb	Cl	−0.0416	I	3.73	mp-989520	$0.25\,Bi_4Cl_{12} + 2\,Rb_1Cl_1 + Na_1Cl_1$
61	Na	Mg	Cs	F	−0.0449	C	0.00	mp-989568	$0.25\,F_4 + 0.6667\,Cs_1F_1 + 0.1667\,Cs_8Mg_6F_{20} + Na_1F_1$
62	Tl	Al	Rb	H	−0.0451	S	0.72	mp-989539	$H_1 + 0.5\,Al_2H_6 + Tl_1 + 2\,Rb_1H_1$
63	S	Br	Rb	Cl	−0.0452	C	0.00	mp-989518	$0.25\,Cl_4 + 0.5\,Br_2Cl_2 + 2\,Rb_1Cl_1 + 0.125\,S_8Cl_{16}$
64	N	F	Sn	Sr	−0.0454	C	0.00	mp-989592	$0.5\,Sr_8Sn_4 + 0.0625\,Sr_{32}N_{16}F_{16}$
65	As	Na	Rb	F	−0.0467	I	4.55	mp-989523	$2\,Rb_1F_1 + 0.25\,As_4F_{12} + Na_1F_1$
66	Pb	K	Cs	F	−0.0486	C	0.00	mp-989585	$2\,Cs_1F_1 + 0.25\,Pb_4F_{12} + K_1F_1$
67	In	Al	Cs	H	−0.0488	S	0.61	mp-989535	$0.5\,Al_2H_6 + 0.1111\,Cs_3In_9 + 1.6667\,Cs_1H_1 + 1.3333\,H_1$
68	Pb	Na	Cs	F	−0.0513	C	0.00	mp-989556	$2\,Cs_1F_1 + 0.25\,Pb_4F_{12} + Na_1F_1$
69	Ga	Na	Tl	F	−0.0514	I	4.39	mp-989561	$0.2\,Ga_2F_6 + 0.5\,Tl_4F_4 + 0.1\,Na_{10}Ga_6F_{28}$
70	In	Tl	Rb	Cl	−0.0528	S	2.40	mp-989550	$0.5\,In_2Cl_6 + Tl_1Cl_1 + 2\,Rb_1Cl_1$
71	In	Na	Rb	Cl	−0.0538	I	3.05	mp-989547	$0.3333\,In_2Cl_6 + 0.1667\,Na_6In_2Cl_{12} + 2\,Rb_1Cl_1$
72	In	Na	Cs	H	−0.0565	S	1.26	mp-989610	$3.3333\,H_1 + 0.1111\,Cs_3In_9 + 1.6667\,Cs_1H_1 + Na_1H_1$
73	Li	In	Rb	Cl	−0.0603	S	2.83	mp-989583	$0.0556\,Li_{18}In_6Cl_{36} + 2\,Rb_1Cl_1 + 0.3333\,In_2Cl_6$
74	In	Na	Tl	F	−0.0604	I	4.27	mp-989533	$Na_1F_1 + 0.5\,Tl_4F_4 + 0.5\,In_2F_6$
75	Br	F	Cs	Cl	−0.0610	S	0.95	mp-989543	$0.5\,Cl_4 + 2\,Cs_1Cl_1 + 0.5\,Br_2Cl_2 + 0.25\,Cl_4F_4$
76	In	Ga	Rb	F	−0.0616	I	3.27	mp-989566	$0.1667\,Ga_4 + 0.0833\,Rb_8In_{12}F_{44} + 0.1667\,Ga_2F_6 + 1.3333\,Rb_1F_1$
77	N	K	Cs	F	−0.0618	I	-	mp-989580	$0.25\,F_4 + 0.125\,K_8F_{24} + 0.0625\,N_{16} + 2\,Cs_1F_1$
78	Li	Na	Cs	F	−0.0634	C	0.00	mp-989559	$0.5\,F_4 + Cs_1F_1 + 0.25\,Cs_4Li_4F_8 + Na_1F_1$
79	Na	In	Rb	F	−0.0672	I	5.34	mp-989578	$0.0833\,Rb_8In_{12}F_{44} + Na_1F_1 + 1.3333\,Rb_1F_1$
80	O	N	Sn	Ca	−0.0686	C	0.00	mp-989584	$0.0625\,Ca_{24}N_{16} + Ca_3Sn_1O_1 + 0.0227\,Ca_{62}Sn_{40} + 0.0455\,Ca_2Sn_2$
81	N	F	Sn	Ca	−0.0738	S	0.14	mp-989590	$0.5\,Ca_4N_2F_2 + 0.5\,Ca_8Sn_4$
82	In	Rb	Cs	F	−0.0776	I	5.37	mp-989605	$0.0833\,Rb_8In_{12}F_{44} + 0.3333\,Rb_1F_1 + 2\,Cs_1F_1$
83	S	Br	Cs	Cl	−0.0848	C	0.00	mp-989517	$2\,Cs_1Cl_1 + 0.5\,Br_2Cl_2 + 0.25\,Cl_4 + 0.125\,S_8Cl_{16}$
84	In	K	Cs	F	−0.0875	I	5.46	mp-989639	$0.025\,K_{24}In_8F_{48} + 0.1\,K_4In_8F_{28} + 2\,Cs_1F_1$
85	Tl	Al	Cs	H	−0.0884	S	1.14	mp-989575	$2\,Cs_1H_1 + 0.5\,Al_2H_6 + Tl_1 + H_1$
86	Tl	Ga	Rb	F	−0.0945	I	4.40	mp-989565	$2\,Rb_1F_1 + 0.5\,Ga_2F_6 + 0.25\,Tl_4F_4$
87	Ga	Na	Rb	F	−0.1008	I	5.90	mp-989400	$2\,Rb_1F_1 + 0.1\,Na_{10}Ga_6F_{28} + 0.2\,Ga_2F_6$
88	Al	Na	Cs	H	−0.1019	S	2.14	mp-989642	$2\,Cs_1H_1 + 0.5\,Na_2Al_2H_8$
89	N	Na	Cs	F	−0.1064	S	2.80	mp-989527	$0.75\,F_4 + 0.5\,N_2 + Na_1F_1 + 2\,Cs_1F_1$

Faber et al, *Phys Rev Lett* (2016) arxiv.org/abs/1508.05315

Crystals



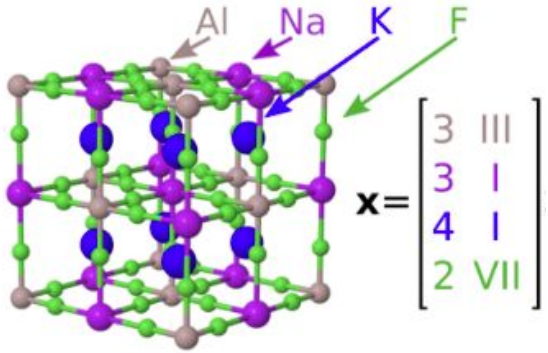


Element	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7
H					✓		✓						
He						✓							
Li							✓						
Be							✓	✓					
B	✓						✓	✓	✓				
C		✓	✓	✓	✓	✓	✓	✓	✓	✓			
N	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓		
O				✓	✓		✓	✓					
F					✓								
Ne						✓							
Na					✓		✓						
Mg							✓	✓					
Al							✓	✓	✓				
Si		✓	✓	✓	✓		✓	✓	✓	✓			
P			✓	✓	✓		✓	✓	✓	✓	✓		
S				✓	✓		✓	✓	✓	✓	✓	✓	
Cl					✓		✓		✓	✓	✓	✓	✓
Ar						✓							
K					✓		✓						
Ca					✓		✓	✓					

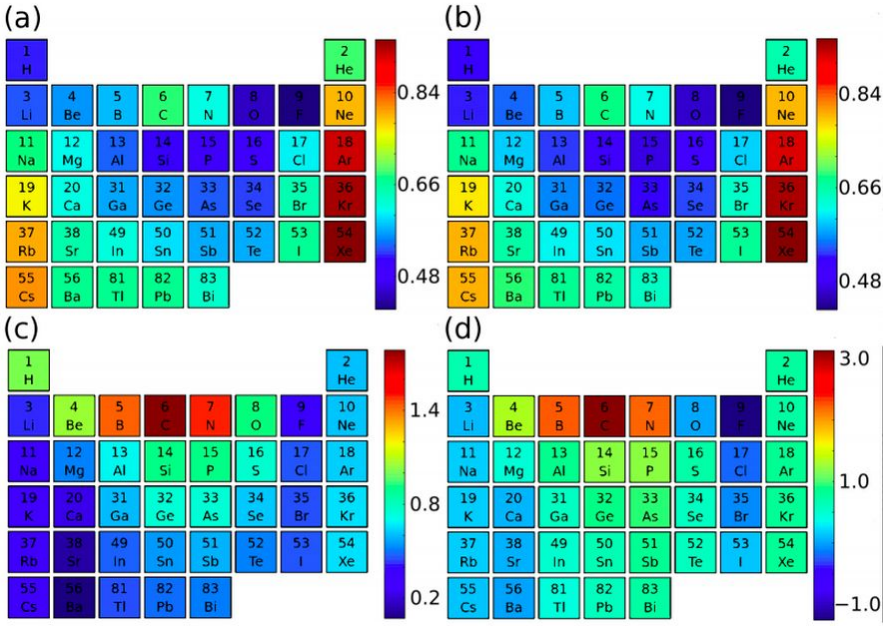
Element	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8
Ga	✓	✓		✓			✓	✓	✓					
Ge		✓	✓	✓	✓	✓	✓	✓	✓	✓				
As			✓				✓	✓	✓		✓			
Se				✓			✓	✓		✓		✓		
Br					✓		✓		✓	✓	✓		✓	
Kr						✓	✓	✓						
Rb					✓		✓							
Sr							✓	✓						
In	✓						✓	✓	✓					
Sn		✓					✓	✓	✓	✓				
Sb			✓			✓	✓		✓		✓			
Te				✓		✓	✓	✓		✓	✓	✓		
I					✓		✓				✓		✓	
Xe							✓	✓		✓		✓		✓
Cs					✓		✓							
Ba								✓						
Ti	✓						✓	✓	✓					
Pb							✓	✓	✓	✓				
Bi							✓	✓	✓	✓	✓			

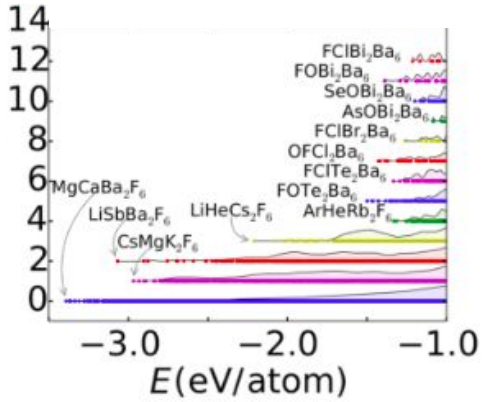
Faber et al, *Phys Rev Lett* (2016) arxiv.org/abs/1508.05315

Crystals



Formula	LPTOS	E_{ML}	E_{DFT}	q_1	q_2	q_3	q_4
$MgSbBa_2F_6$	1	-2.88	-2.70	1.66	**0.42**	1.63	-0.89
$CaTeBa_2F_6$	1	-2.90	-2.68	1.58	**0.31**	1.67	-0.87
$TeCaBa_2F_6$	1	-2.83	-2.68	**0.31**	1.59	1.67	-0.87
$LiSbBa_2F_6$	2	-3.06	-2.62	0.89	**1.06**	1.62	-0.86
$CsMgRb_2F_6$	1	-2.93	-2.61	0.98	1.67	0.92	-0.75
$BeSbBa_2F_6$	2	-2.88	-2.60	1.68	**0.35**	1.62	-0.88
$CsMgK_2F_6$	1	-2.97	-2.58	1.01	1.68	0.92	-0.75
$SrSbBa_2F_6$	2	-2.90	-2.56	1.48	**0.60**	1.59	-0.88
$SrTeBa_2F_6$	2	-2.89	-2.55	1.70	**0.40**	1.66	-0.90



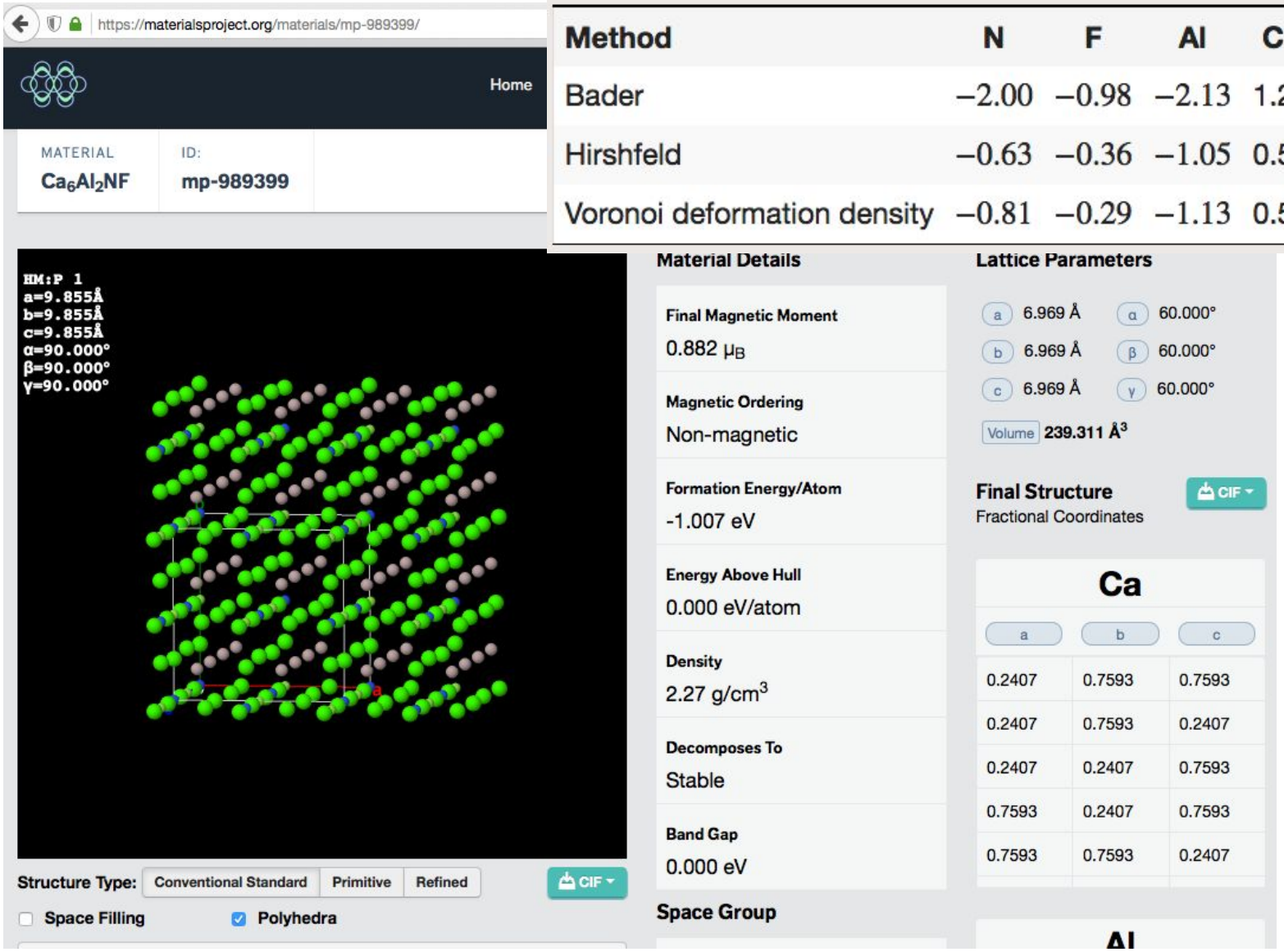


Faber et al, *Phys Rev Lett* (2016) arxiv.org/abs/1508.05315

Crystals

Calculated atomic charges in $NFAl_2Ca_6$ elpasolite using different methods (obtained using SIESTA[43]).

Method	N	F	Al	Ca
Bader	−2.00	−0.98	−2.13	1.20
Hirshfeld	−0.63	−0.36	−1.05	0.52
Voronoi deformation density	−0.81	−0.29	−1.13	0.56



Faber et al, *Phys Rev Lett* (2016) arxiv.org/abs/1508.05315

Quantum Machine: Elpasolite Crystal Energy Predictions

Input panel

Single Prediction

Group Prediction

Crystal Design

Element in first position:

Element in second position:

Element in third position:

Element in forth position:

Target formation energy:

Random steps:

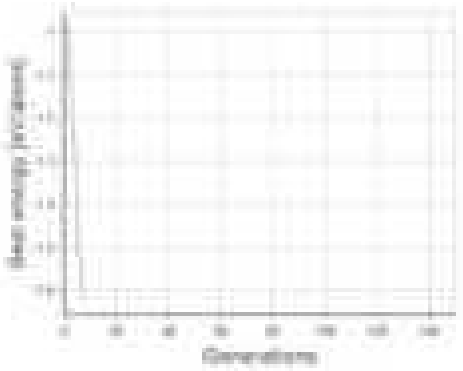
Population size:

Submit

Output panel

Optimizing crystal with
target value: -5.00
max generations: 100
population size: 20

Optimization progress



Conclusions

1. ML makes instantaneous QM quality predictions

2. Learning curves reveal quality of ML model

3. Rate and offset depend on
   a. Uniqueness and target similarity (model dependent)
   b. Properties and structures (data dependent)
   c. Baseline (expert dependent)

Conclusions II

Scientific method - how to gain knowledge

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