

Preaching on first principles views on chemical compound space:

Atom centered potentials and statistical learning

O. Anatole von Lilienfeld

*Institute of Physical Chemistry, Department of Chemistry, University of Basel, Switzerland
Argonne Leadership Computing Facility, Argonne National Laboratory, Illinois, USA*

``First principles view on chemical compound space: Gaining rigorous atomistic control of molecular properties",
O. A. von Lilienfeld, Int J Quant Chem (2013), <http://arxiv.org/abs/1209.5033>



If, in some cataclysm, all scientific knowledge were to be destroyed, and only one sentence passed on to the next generation of creatures, what statement would contain the most information in the fewest words?



If, in some cataclysm, all scientific knowledge were to be destroyed, and only one sentence passed on to the next generation of creatures, what statement would contain the most information in the fewest words?

I believe it is the atomic hypothesis (or atomic fact, or whatever you wish to call it) that all things are made of atoms — little particles that move around in perpetual motion, attracting each other when they are a little distance apart, but repelling upon being squeezed into one another. In that one sentence you will see an enormous amount of information about the world, if just a little imagination and thinking are applied.

Feynman Lectures of Physics (1964)

“QMC is not a black-box”
(M. Foulkes)

“QMC is not a black-box”
(M. Foulkes)

...

“One material every 2 s by 2016”
(T. Mueller)

“QMC is not a black-box”
(M. Foulkes)

...

“One material every 2 s by 2016”
(T. Mueller)

“QMC can inform improvements of current DFTs (or other methods)”
(M. Gillan, A. Tkatchenko, ...)

“QMC is not a black-box”
(M. Foulkes)

...

“One material every 2 s by 2016”
(T. Mueller)

“QMC can inform improvements of current DFTs (or other methods)”
(M. Gillan, A. Tkatchenko, ...)

Increase DFT’s transferability to properly account for

- spin states
- excited states
- van der Waals

“QMC is not a black-box”
(M. Foulkes)

...

“One material every 2 s by 2016”
(T. Mueller)

“QMC can inform improvements of current DFTs (or other methods)”
(M. Gillan, A. Tkatchenko, ...)

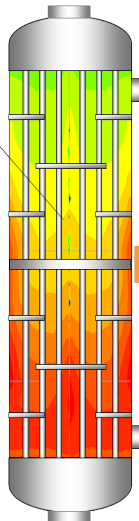
Increase DFT’s transferability to properly account for

- spin states
- excited states
- van der Waals

 Define transferability!

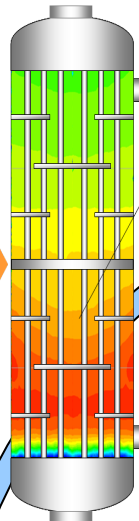
A method is called transferable if its error is invariant wrt changes in atomic configuration and composition == chemical compound space

Uneven radial
temperature
distribution



Initial

Even radial
temperature
distribution



Optimal

Smaller high
temperature
area

Argonne Blue Gene/P hardware

Computational design ...

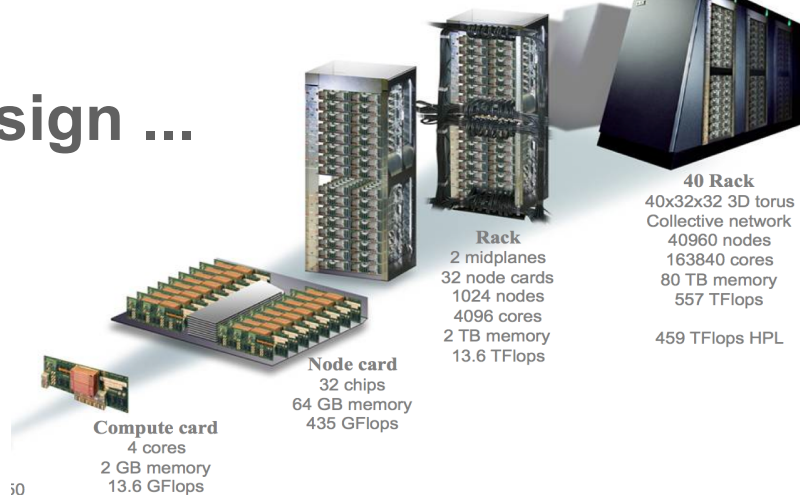
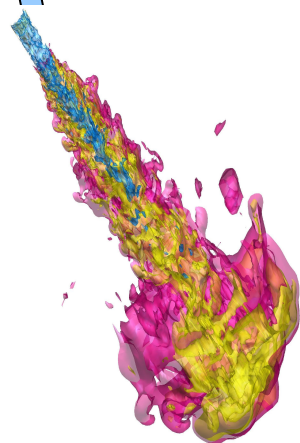
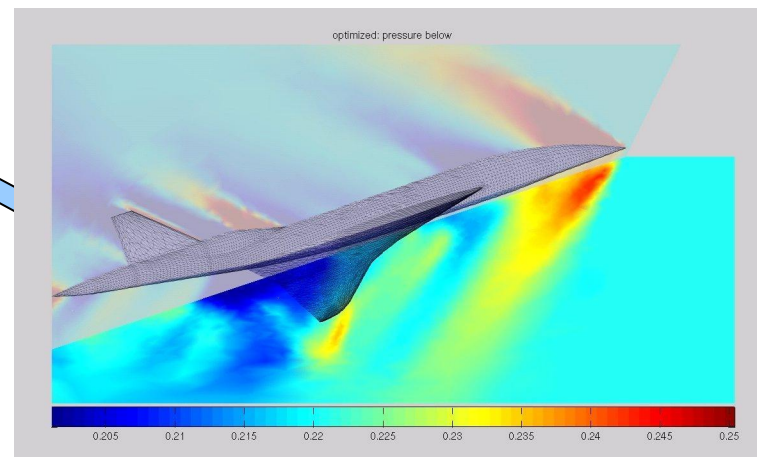
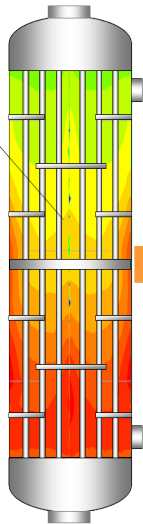


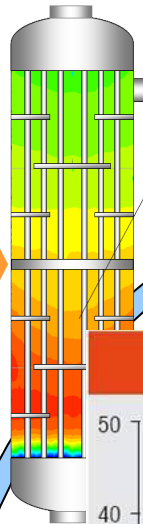
Figure 2: EcoBoost—with direct injection, fuel is injected into each cylinder or an engine in small, precise amounts.

Uneven radial temperature distribution



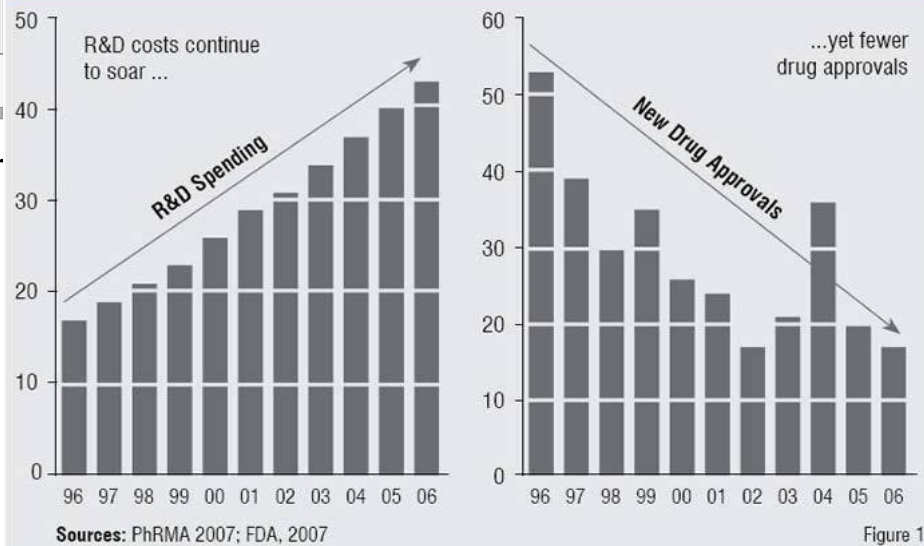
Initial

Even radial temperature distribution



Optim

R&D spending vs. FDA approvals, 1996-2006



- Infections
- Metabolic syndrome
- Aging
- Cancer
- ...

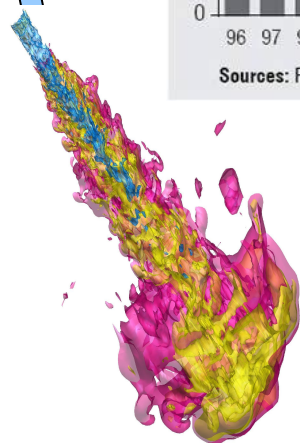
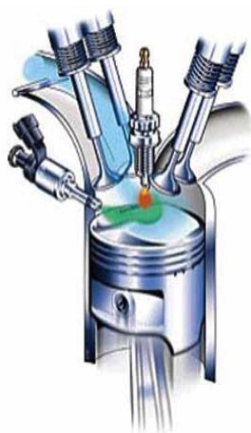
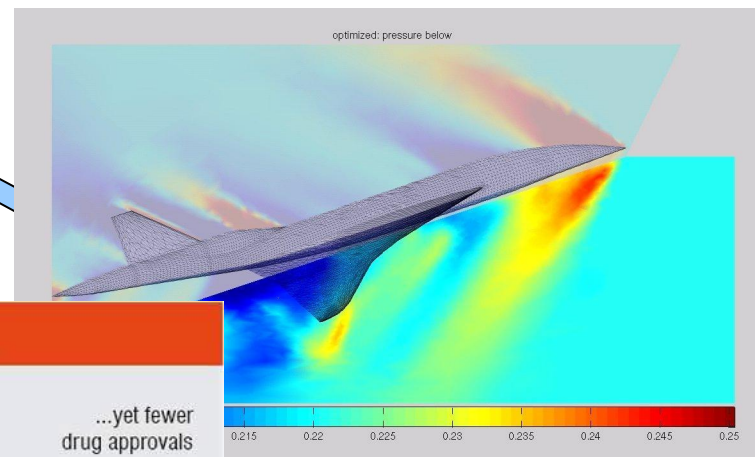


Figure 2: EcoBoost—with direct injection, fuel is injected into each cylinder of an engine in small, precise amounts.

Why is this hard?

Combinatorial catastrophe

number of small organic molecules $> 10^{60}$

Nature Insight on chemical space (2004)

Assume 1 property evaluation ~ 1 s

→ exhaustive screening $\sim 10^{52}$ yrs
(age of universe $\sim 10^{10}$ yrs)



Edisonian approach

1878



Combinatorial problem

hydrogen 1 H 1.0079																	helium 2 He 4.0026	
lithium 3 Li 6.941	beryllium 4 Be 9.0122											boron 5 B 10.811	carbon 6 C 12.011	nitrogen 7 N 14.007	oxygen 8 O 15.999	fluorine 9 F 18.998	neon 10 Ne 20.180	
sodium 11 Na 22.990	magnesium 12 Mg 24.305											aluminium 13 Al 26.982	silicon 14 Si 28.086	phosphorus 15 P 30.974	sulfur 16 S 32.065	chlorine 17 Cl 35.453	argon 18 Ar 39.948	
potassium 19 K 39.098	calcium 20 Ca 40.078	scandium 21 Sc 44.956	titanium 22 Ti 47.867	vanadium 23 V 50.942	chromium 24 Cr 51.996	manganese 25 Mn 54.938	iron 26 Fe 55.845	cobalt 27 Co 58.933	nickel 28 Ni 58.693	copper 29 Cu 63.546	zinc 30 Zn 65.39	gallium 31 Ga 69.723	germanium 32 Ge 72.61	arsenic 33 As 74.922	selenium 34 Se 78.96	bromine 35 Br 79.904	krypton 36 Kr 83.80	
rubidium 37 Rb 85.468	strontium 38 Sr 87.62	yttrium 39 Y 88.906	zirconium 40 Zr 91.224	niobium 41 Nb 92.906	molybdenum 42 Mo 95.94	technetium 43 Tc [98]	ruthenium 44 Ru 101.07	rhodium 45 Rh 102.91	palladium 46 Pd 106.42	silver 47 Ag 107.87	cadmium 48 Cd 112.41	indium 49 In 114.82	tin 50 Sn 118.71	antimony 51 Sb 121.76	tellurium 52 Te 127.60	iodine 53 I 126.90	xenon 54 Xe 131.29	
caesium 55 Cs 132.91	barium 56 Ba 137.33	57-70 ✱	lutetium 71 Lu 174.97	hafnium 72 Hf 178.49	tantalum 73 Ta 180.95	tungsten 74 W 183.84	rhenium 75 Re 186.21	osmium 76 Os 190.23	iridium 77 Ir 192.22	platinum 78 Pt 195.08	gold 79 Au 196.97	mercury 80 Hg 200.59	thallium 81 Tl 204.38	lead 82 Pb 207.2	bismuth 83 Bi 208.98	polonium 84 Po [209]	astatine 85 At [210]	radon 86 Rn [222]
francium 87 Fr [223]	radium 88 Ra [226]	89-102 ✱ ✱	lawrencium 103 Lr [262]	rutherfordium 104 Rf [261]	dubnium 105 Db [262]	seaborgium 106 Sg [266]	bohrium 107 Bh [264]	hassium 108 Hs [269]	meitnerium 109 Mt [268]	ununilium 110 Uun [271]	unununium 111 Uuu [272]	ununbium 112 Uub [277]	ununquadium 114 Uuq [289]					

* Lanthanide series

** Actinide series

lanthanum 57 La 138.91	cerium 58 Ce 140.12	praseodymium 59 Pr 140.91	neodymium 60 Nd 144.24	promethium 61 Pm [145]	samarium 62 Sm 150.36	europium 63 Eu 151.96	gadolinium 64 Gd 157.25	terbium 65 Tb 158.93	dysprosium 66 Dy 162.50	holmium 67 Ho 164.93	erbium 68 Er 167.26	thulium 69 Tm 168.93	ytterbium 70 Yb 173.04
actinium 89 Ac [227]	thorium 90 Th 232.04	protactinium 91 Pa 231.04	uranium 92 U 238.03	neptunium 93 Np [237]	plutonium 94 Pu [244]	americium 95 Am [243]	curium 96 Cm [247]	berkelium 97 Bk [247]	californium 98 Cf [251]	einsteinium 99 Es [252]	fermium 100 Fm [257]	mendelevium 101 Md [258]	nobelium 102 No [259]

B₈
ScH₁₉
O₅
H₄₀
ZnNe
He₂₀
Ca₂
Zr

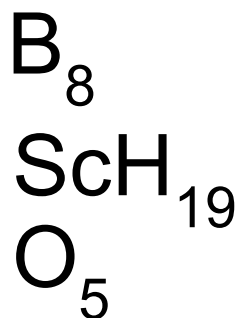
<div>hydrogen</div> <div>1</div> <div>H</div> <div>1.0079</div>																		<div>helium</div> <div>2</div> <div>He</div> <div>4.0026</div>					
<div>lithium</div> <div>3</div> <div>Li</div> <div>6.941</div>	<div>beryllium</div> <div>4</div> <div>Be</div> <div>9.0122</div>																	<div>boron</div> <div>5</div> <div>B</div> <div>10.811</div>	<div>carbon</div> <div>6</div> <div>C</div> <div>12.011</div>	<div>nitrogen</div> <div>7</div> <div>N</div> <div>14.007</div>	<div>oxygen</div> <div>8</div> <div>O</div> <div>15.999</div>	<div>fluorine</div> <div>9</div> <div>F</div> <div>18.998</div>	<div>neon</div> <div>10</div> <div>Ne</div> <div>20.180</div>
<div>sodium</div> <div>11</div> <div>Na</div> <div>22.990</div>	<div>magnesium</div> <div>12</div> <div>Mg</div> <div>24.305</div>																	<div>aluminium</div> <div>13</div> <div>Al</div> <div>26.982</div>	<div>silicon</div> <div>14</div> <div>Si</div> <div>28.086</div>	<div>phosphorus</div> <div>15</div> <div>P</div> <div>30.974</div>	<div>sulfur</div> <div>16</div> <div>S</div> <div>32.065</div>	<div>chlorine</div> <div>17</div> <div>Cl</div> <div>35.453</div>	<div>argon</div> <div>18</div> <div>Ar</div> <div>39.948</div>
<div>potassium</div> <div>19</div> <div>K</div> <div>39.098</div>	<div>calcium</div> <div>20</div> <div>Ca</div> <div>40.078</div>	<div>scandium</div> <div>21</div> <div>Sc</div> <div>44.956</div>	<div>titanium</div> <div>22</div> <div>Ti</div> <div>47.867</div>	<div>vanadium</div> <div>23</div> <div>V</div> <div>50.942</div>	<div>chromium</div> <div>24</div> <div>Cr</div> <div>51.996</div>	<div>manganese</div> <div>25</div> <div>Mn</div> <div>54.938</div>	<div>iron</div> <div>26</div> <div>Fe</div> <div>55.845</div>	<div>cobalt</div> <div>27</div> <div>Co</div> <div>58.933</div>	<div>nickel</div> <div>28</div> <div>Ni</div> <div>58.693</div>	<div>copper</div> <div>29</div> <div>Cu</div> <div>63.546</div>	<div>zinc</div> <div>30</div> <div>Zn</div> <div>65.39</div>	<div>gallium</div> <div>31</div> <div>Ga</div> <div>69.723</div>	<div>germanium</div> <div>32</div> <div>Ge</div> <div>72.61</div>	<div>arsenic</div> <div>33</div> <div>As</div> <div>74.922</div>	<div>selenium</div> <div>34</div> <div>Se</div> <div>78.96</div>	<div>bromine</div> <div>35</div> <div>Br</div> <div>79.904</div>	<div>krypton</div> <div>36</div> <div>Kr</div> <div>83.80</div>						
<div>rubidium</div> <div>37</div> <div>Rb</div> <div>85.468</div>	<div>strontium</div> <div>38</div> <div>Sr</div> <div>87.62</div>	<div>yttrium</div> <div>39</div> <div>Y</div> <div>88.906</div>	<div>zirconium</div> <div>40</div> <div>Zr</div> <div>91.224</div>	<div>niobium</div> <div>41</div> <div>Nb</div> <div>92.906</div>	<div>molybdenum</div> <div>42</div> <div>Mo</div> <div>95.94</div>	<div>technetium</div> <div>43</div> <div>Tc</div> <div>[98]</div>	<div>ruthenium</div> <div>44</div> <div>Ru</div> <div>101.07</div>	<div>rhodium</div> <div>45</div> <div>Rh</div> <div>102.91</div>	<div>palladium</div> <div>46</div> <div>Pd</div> <div>106.42</div>	<div>silver</div> <div>47</div> <div>Ag</div> <div>107.87</div>	<div>cadmium</div> <div>48</div> <div>Cd</div> <div>112.41</div>	<div>indium</div> <div>49</div> <div>In</div> <div>114.82</div>	<div>tin</div> <div>50</div> <div>Sn</div> <div>118.71</div>	<div>antimony</div> <div>51</div> <div>Sb</div> <div>121.76</div>	<div>tellurium</div> <div>52</div> <div>Te</div> <div>127.60</div>	<div>iodine</div> <div>53</div> <div>I</div> <div>126.90</div>	<div>xenon</div> <div>54</div> <div>Xe</div> <div>131.29</div>						
<div>caesium</div> <div>55</div> <div>Cs</div> <div>132.91</div>	<div>barium</div> <div>56</div> <div>Ba</div> <div>137.33</div>	<div>57-70</div> <div>★</div>	<div>lutetium</div> <div>71</div> <div>Lu</div> <div>174.97</div>	<div>hafnium</div> <div>72</div> <div>Hf</div> <div>178.49</div>	<div>tantalum</div> <div>73</div> <div>Ta</div> <div>180.95</div>	<div>tungsten</div> <div>74</div> <div>W</div> <div>183.84</div>	<div>rhenium</div> <div>75</div> <div>Re</div> <div>186.21</div>	<div>osmium</div> <div>76</div> <div>Os</div> <div>190.23</div>	<div>iridium</div> <div>77</div> <div>Ir</div> <div>192.22</div>	<div>platinum</div> <div>78</div> <div>Pt</div> <div>195.08</div>	<div>gold</div> <div>79</div> <div>Au</div> <div>196.97</div>	<div>mercury</div> <div>80</div> <div>Hg</div> <div>200.59</div>	<div>thallium</div> <div>81</div> <div>Tl</div> <div>204.38</div>	<div>lead</div> <div>82</div> <div>Pb</div> <div>207.2</div>	<div>bismuth</div> <div>83</div> <div>Bi</div> <div>208.98</div>	<div>polonium</div> <div>84</div> <div>Po</div> <div>[209]</div>	<div>astatine</div> <div>85</div> <div>At</div> <div>[210]</div>	<div>radon</div> <div>86</div> <div>Rn</div> <div>[222]</div>					
<div>francium</div> <div>87</div> <div>Fr</div> <div>[223]</div>	<div>radium</div> <div>88</div> <div>Ra</div> <div>[226]</div>	<div>89-102</div> <div>★ ★</div>	<div>lawrencium</div> <div>103</div> <div>Lr</div> <div>[262]</div>	<div>rutherfordium</div> <div>104</div> <div>Rf</div> <div>[261]</div>	<div>dubnium</div> <div>105</div> <div>Db</div> <div>[262]</div>	<div>seaborgium</div> <div>106</div> <div>Sg</div> <div>[266]</div>	<div>bohrium</div> <div>107</div> <div>Bh</div> <div>[264]</div>	<div>hassium</div> <div>108</div> <div>Hs</div> <div>[269]</div>	<div>meitnerium</div> <div>109</div> <div>Mt</div> <div>[268]</div>	<div>unnilium</div> <div>110</div> <div>Uun</div> <div>[271]</div>	<div>unununium</div> <div>111</div> <div>Uuu</div> <div>[272]</div>	<div>ununbium</div> <div>112</div> <div>Uub</div> <div>[277]</div>	<div>ununquadium</div> <div>114</div> <div>Uuq</div> <div>[289]</div>										

* * Actinide series

lanthanum 57	cerium 58	praseodymium 59	neodymium 60	promethium 61	samarium 62	europium 63	gadolinium 64	terbium 65	dysprosium 66	holmium 67	erbium 68	thulium 69	ytterbium 70
La	Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb
138.91	140.12	140.91	144.24	[145]	150.36	151.96	157.25	158.93	162.50	164.93	167.26	168.93	173.04
actinium 89	thorium 90	protactinium 91	uranium 92	neptunium 93	plutonium 94	americium 95	curium 96	berkelium 97	californium 98	einsteinium 99	fermium 100	mendelevium 101	nobelium 102
Ac	Th	Pa	U	Np	Pu	Am	Cm	Bk	Cf	Es	Fm	Md	No
[227]	232.04	231.04	238.03	[237]	[244]	[243]	[247]	[247]	[251]	[252]	[257]	[258]	[259]

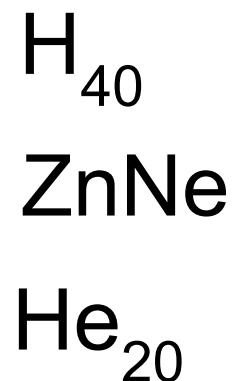
Combinatorial problem

How many stoichiometries have $N_p = 40$ protons?



Discrete number theory

- Integer partition of N_p
- Number of ways to write N_p as sum of positive integers



- Young-Ferrers diagrams

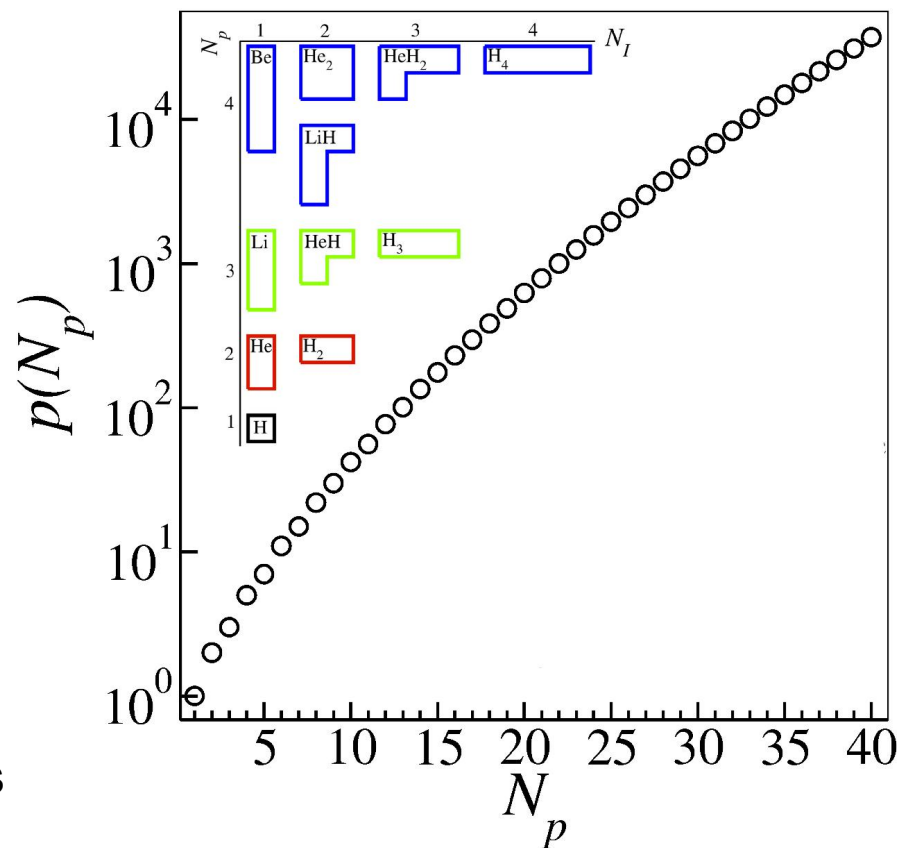
Combinatorial problem

How many stoichiometries have $N_p = 40$ protons?

B_8
 ScH_{19}
 O_5
 H_{40}
 $ZnNe$
 He_{20}
 Ca_2
 Zr

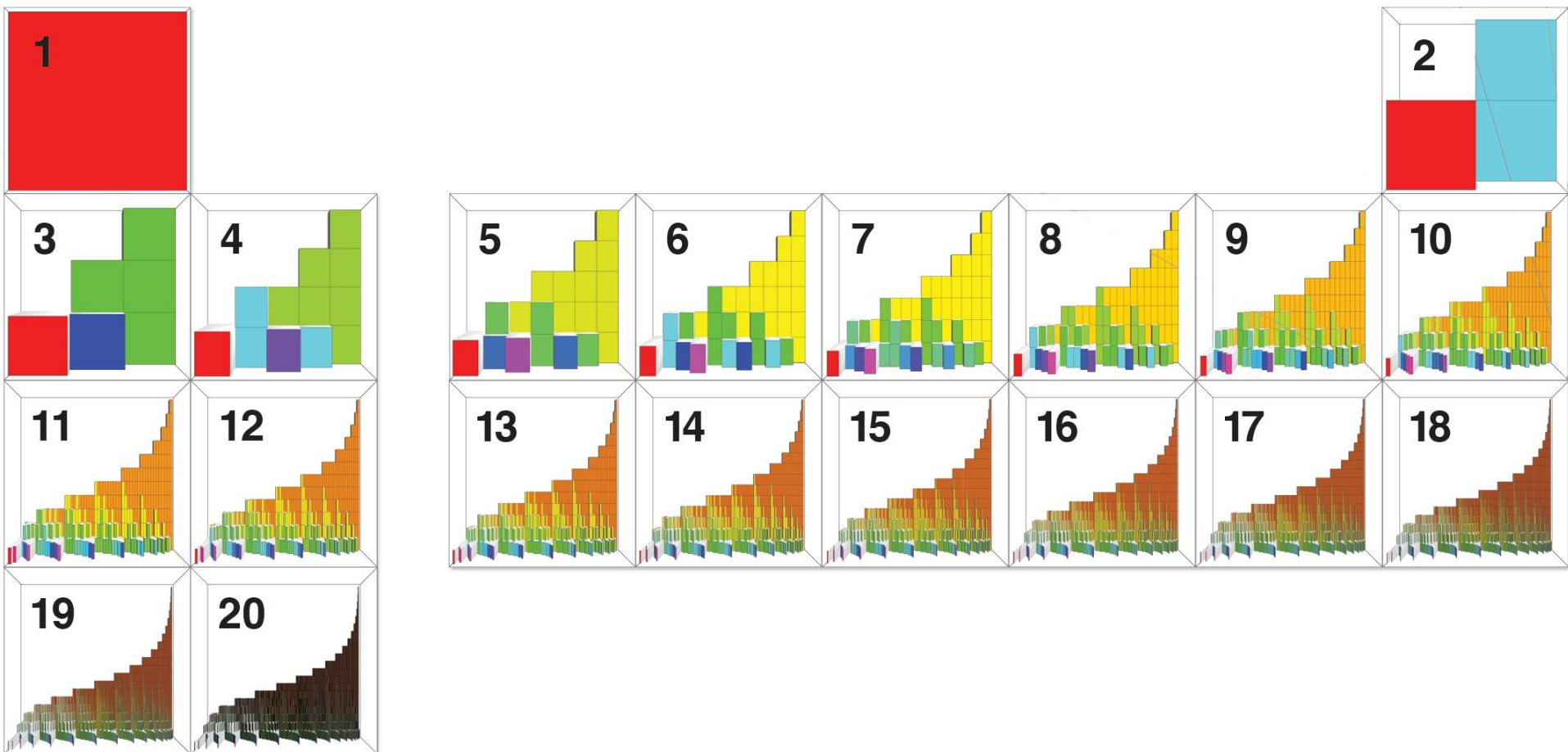
Discrete number theory

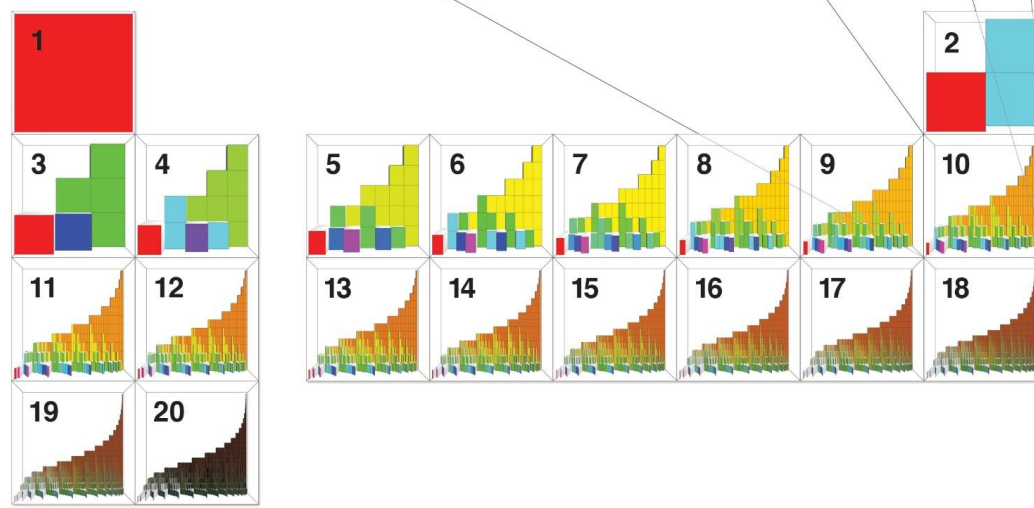
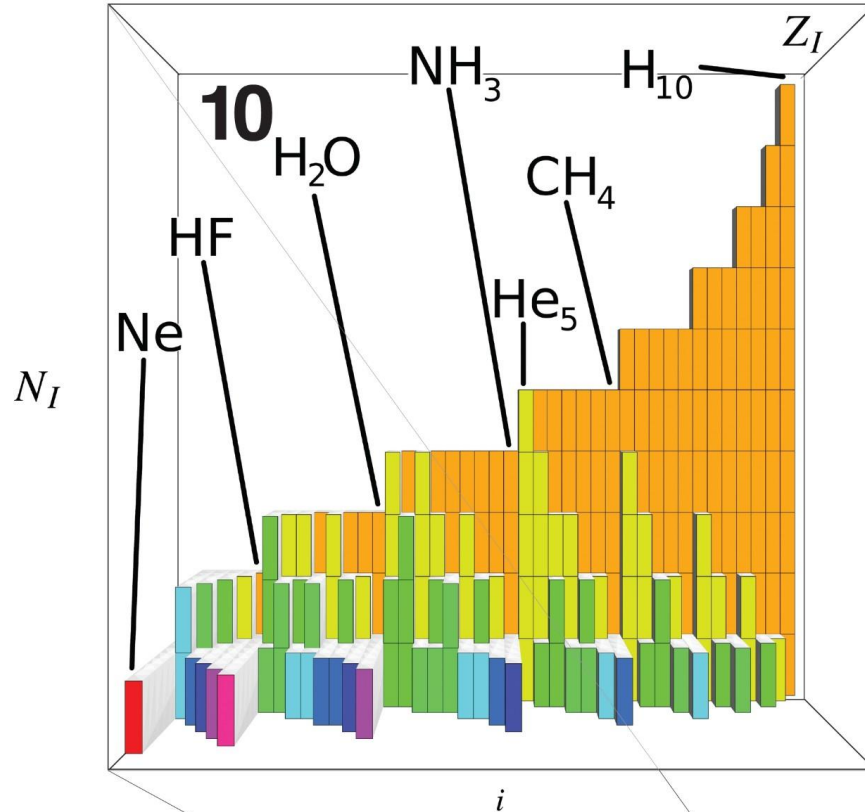
- Integer partition of N_p
- Number of ways to write N_p as sum of positive integers
- Young-Ferrers diagrams



→ 40 protons yield > 37k stoichiometries

Combinatorial problem - availability heuristic?



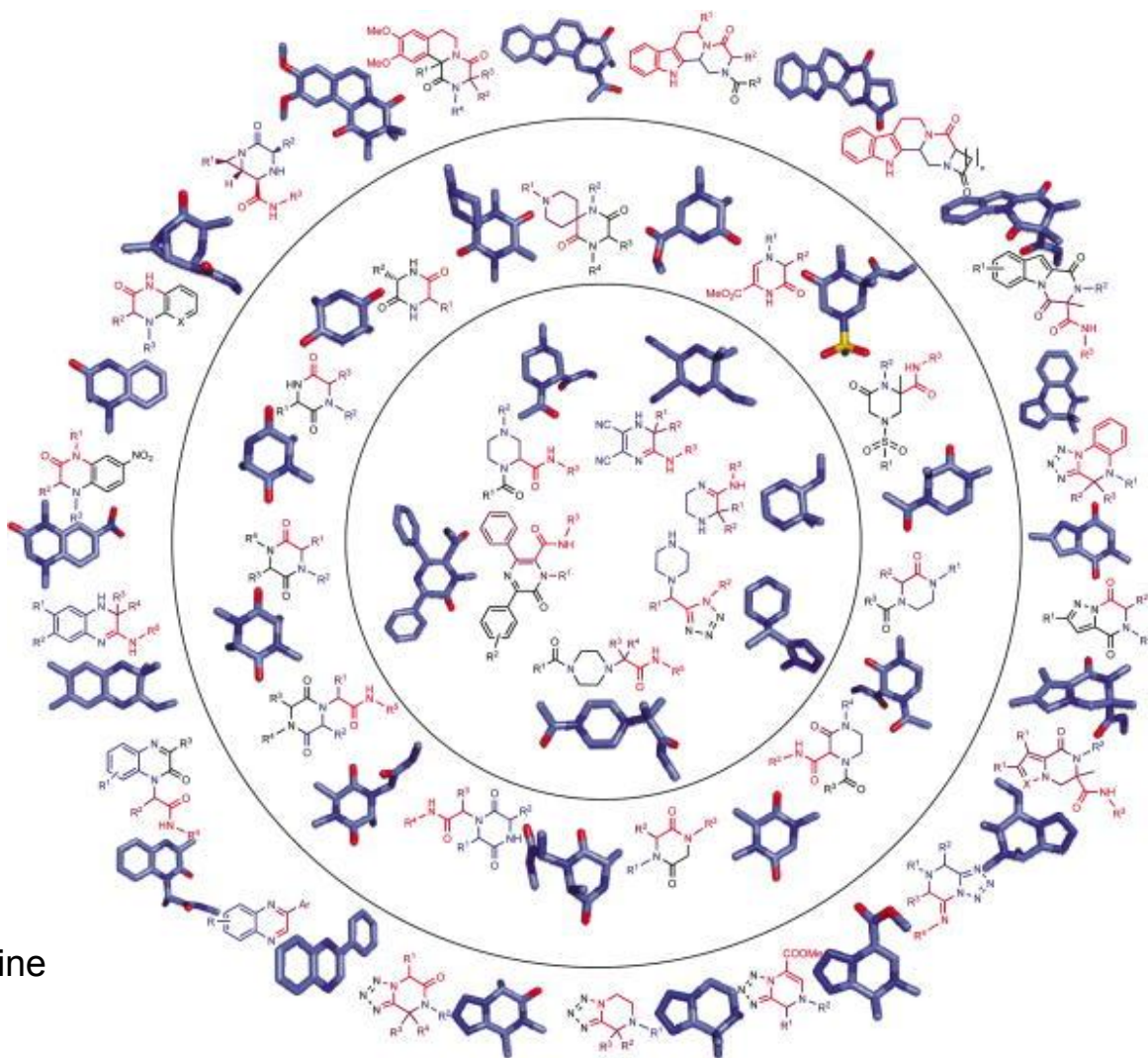


Why is this hard?

Combinatorial catastrophe

number of small organic molecules $> 10^{60}$

Nature Insight on chemical space (2004)



Why is this hard?

Combinatorial catastrophe

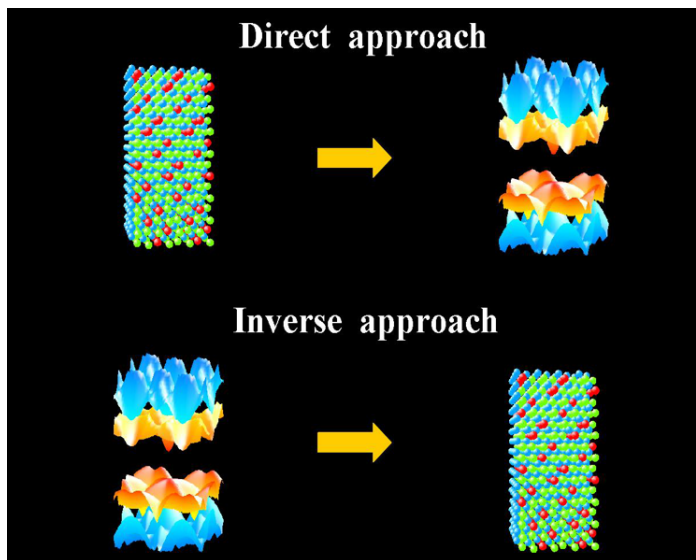
number of small organic molecules $> 10^{60}$

Nature Insight on chemical space (2004)

Assume 1 property evaluation ~ 1 s

→ exhaustive screening $\sim 10^{52}$ yrs
(age of universe $\sim 10^{10}$ yrs)

New



Franceschetti and Zunger, *Nature* (1999)



Why is this hard?

Combinatorial catastrophe

number of small organic molecules $> 10^{60}$

Nature Insight on chemical space (2004)

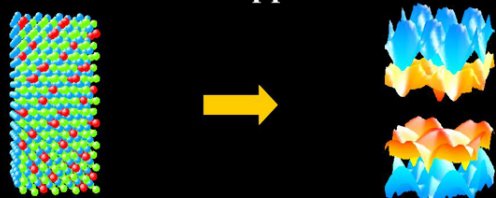
Assume 1 property evaluation ~ 1 s

→ exhaustive screening $\sim 10^{52}$ yrs
(age of universe $\sim 10^{10}$ yrs)

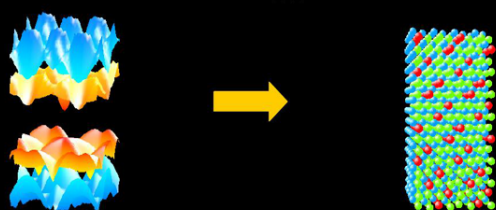
New



Direct approach



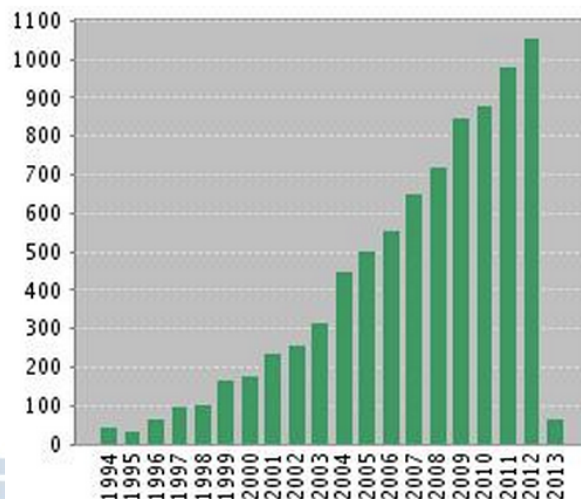
Inverse approach



Franceschetti and Zunger, *Nature* (1999)

DFT & Surface Adsorption

Published Items in Each Year



Why is this hard?

Combinatorial catastrophe

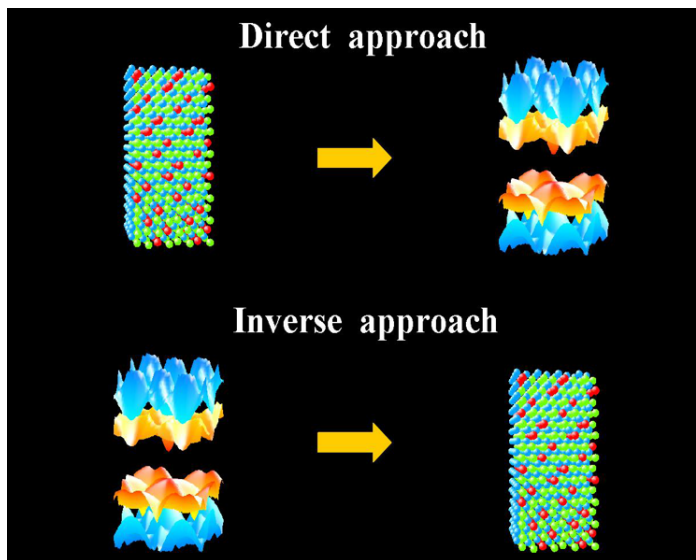
number of small organic molecules $> 10^{60}$

Nature Insight on chemical space (2004)

Assume 1 property evaluation ~ 1 s

→ exhaustive screening $\sim 10^{52}$ yrs
(age of universe $\sim 10^{10}$ yrs)

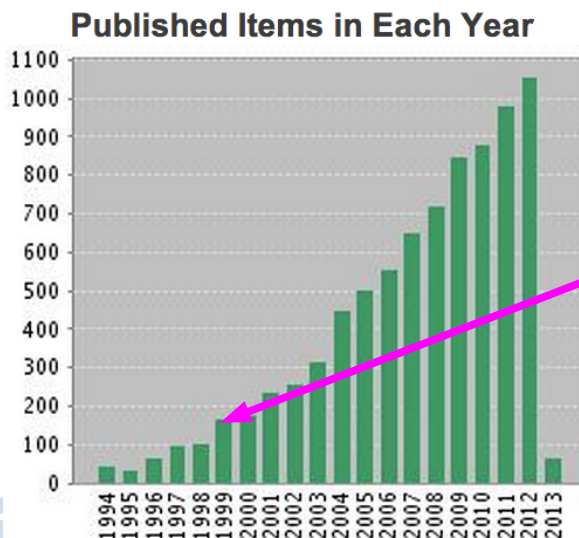
New



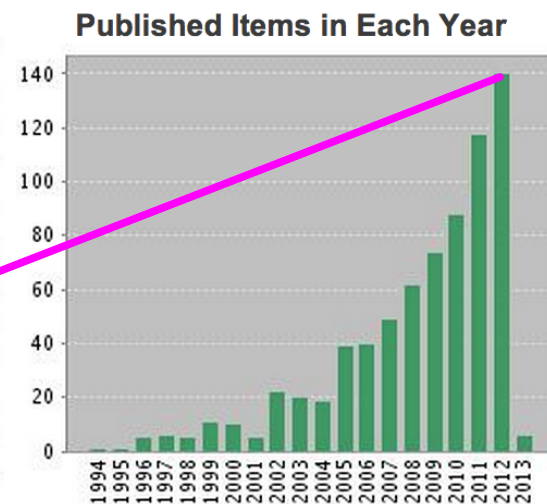
Franceschetti and Zunger, *Nature* (1999)



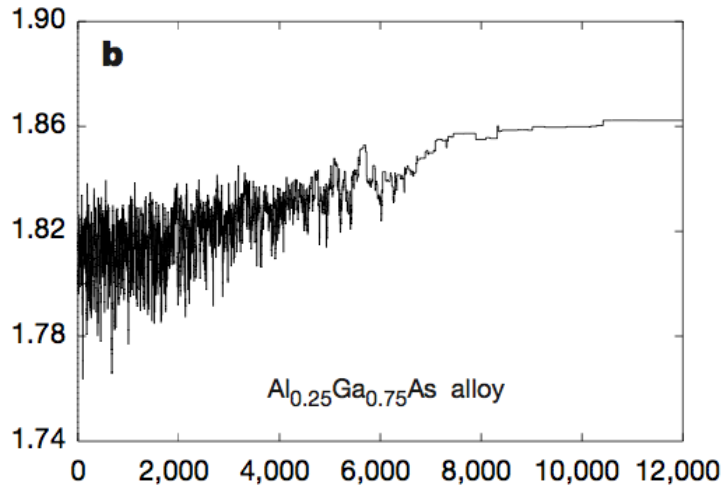
DFT & Surface Adsorption



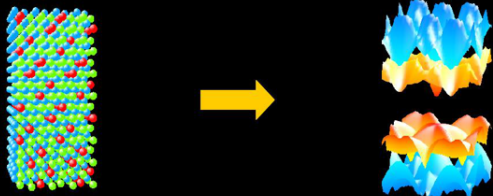
DFT & Surface Design



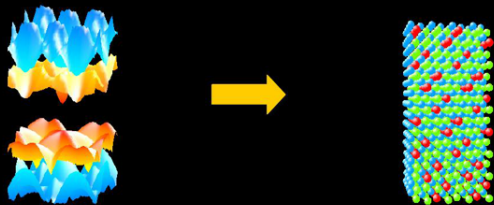
Right compound for right reason!



Direct approach



Inverse approach

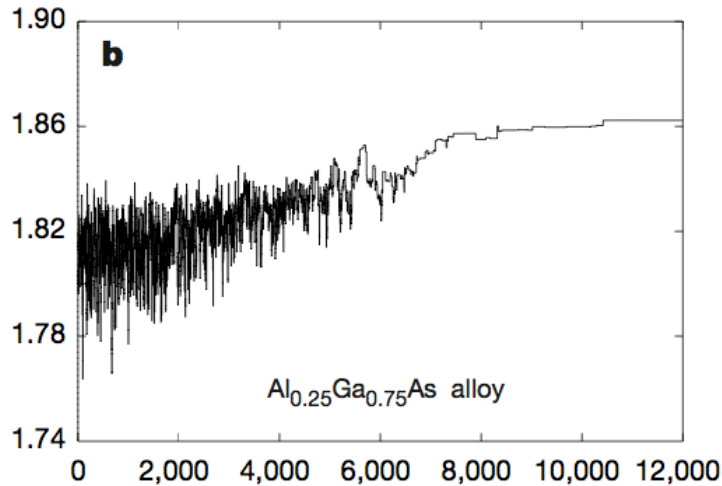


$$\min_{\{Z_I, \mathbf{R}_I\}} \sum_i \omega_i \left(P_i(\{Z_I, \mathbf{R}_I\}) - P_i^{\text{ref}} \right)^2$$

Franceschetti and Zunger, *Nature* (1999)



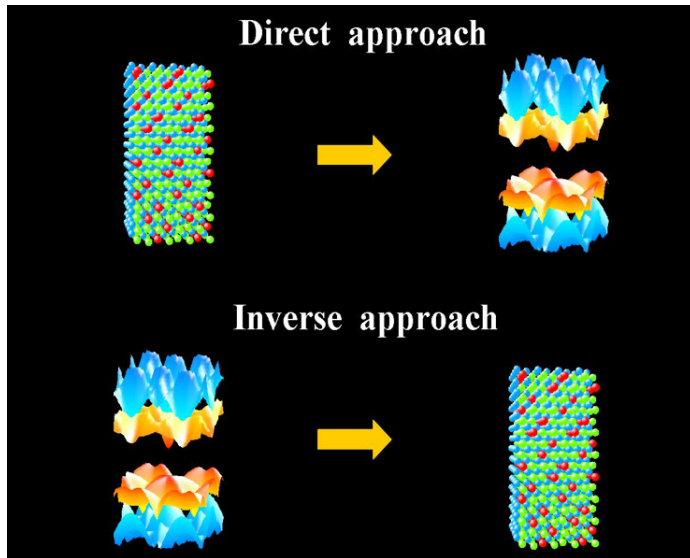
Right compound for right reason!



- No analytic solution
 - Ill-defined
 - high dimensional
 - expensive
- Iterative minimization

$$\min_{\{Z_I, \mathbf{R}_I\}} \sum_i \omega_i \left(P_i(\{Z_I, \mathbf{R}_I\}) - P_i^{\text{ref}} \right)^2$$

- 1) Transferable: First principles (QM, UFF)
- 2) Smart: Variational (dP/dX), Genetic, ...
- 3) Fast: Correlational (Machine Learning)
- 4) Muscle: Supercomputing & Data



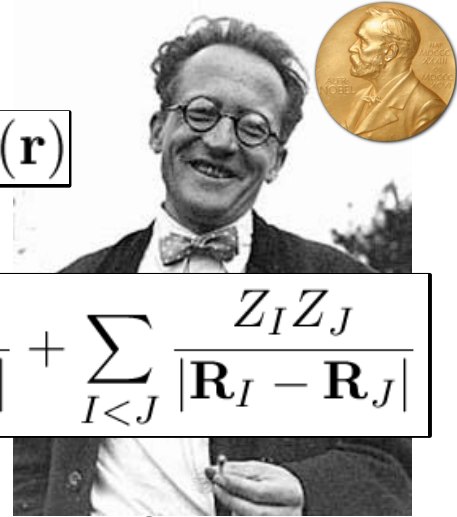
Franceschetti and Zunger, *Nature* (1999)



First Principles

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

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



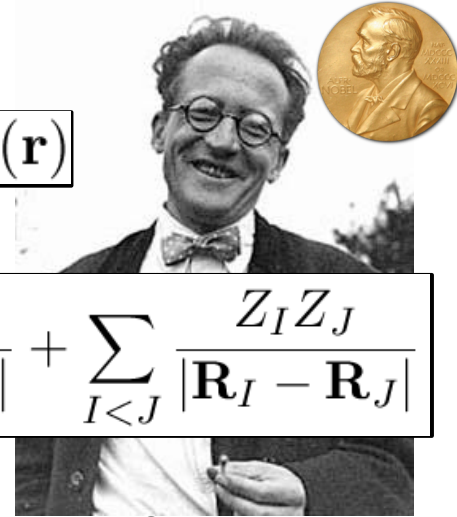
Schrödinger



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

variational (deductive)

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

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

variational (deductive)

Alchemy

1. Free energies
2. Gradients to optimize

Feynman

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

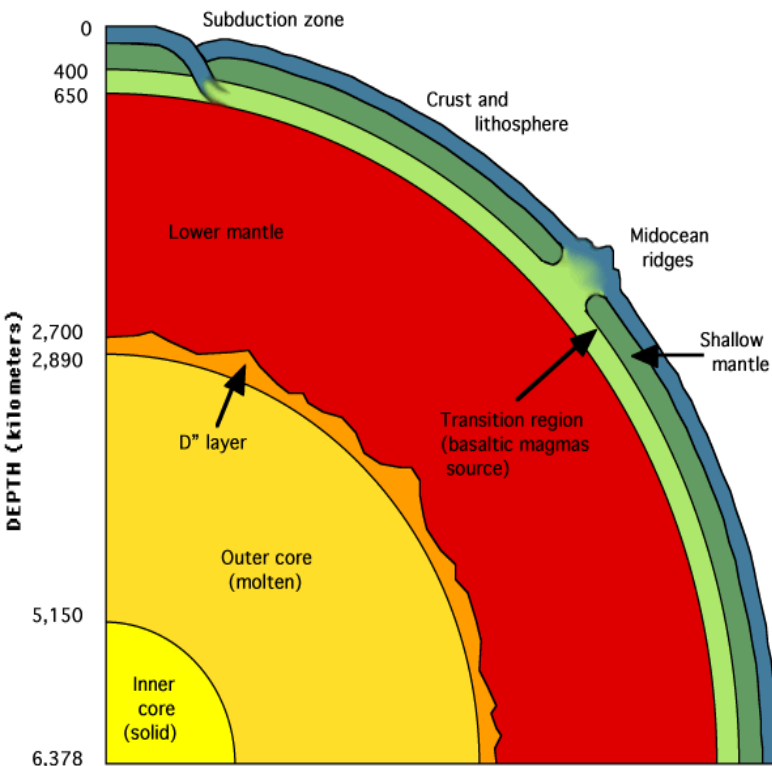
$$\frac{\partial E[H]}{\partial Z_I} = \left\langle \Psi \left| \frac{\partial H}{\partial Z_I} \right| \Psi \right\rangle$$



Constraints on the composition of the Earth's core from *ab initio* calculations

D. Alfè*, M. J. Gillan† & G. D. Price*

* Research School of Geological and Geophysical Sciences, Birkbeck College and University College London, Gower Street, London WC1E 6BT, UK
† Physics and Astronomy Department, University College London, Gower Street, London WC1E 6BT, UK



Knowledge of the composition of the Earth's core¹⁻³ is important for understanding its melting point and therefore the temperature at the inner-core boundary and the temperature profile of the core and mantle. In addition, the partitioning of light elements between solid and liquid, as the outer core freezes at the inner-core boundary, is believed to drive compositional convection⁴, which in turn generates the Earth's magnetic field. It is generally

Nature (2000)

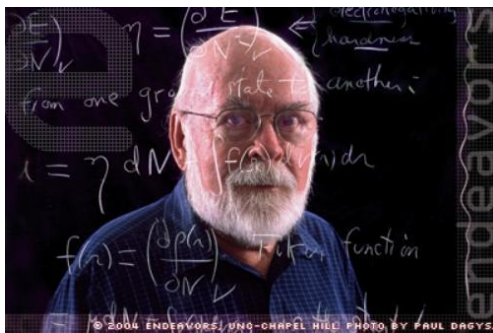
Variational: Gradients

Fractional N_e

$$\frac{\partial E[H]}{\partial N_e} = \mu_e = \epsilon$$



Mermin



Parr

Fukui function: Response of frontier orbitals to molecular changes

Conceptual DFT (Parr, Yang et al)



Fukui

Variational: Gradients

Fractional N_e

$$\frac{\partial E[H]}{\partial N_e} = \mu_e = \epsilon$$

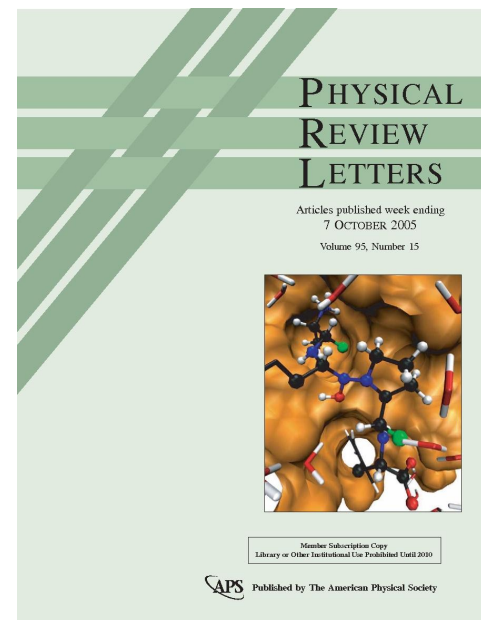
Fractional Z_I

$$\frac{\partial E[H]}{\partial Z_I} = \left\langle \Psi \left| \frac{\partial H}{\partial Z_I} \right| \Psi \right\rangle = \int d\mathbf{r} \frac{n(\mathbf{r})}{|\mathbf{r} - \mathbf{R}_I|} - \sum_J \frac{Z_J}{|\mathbf{R}_J - \mathbf{R}_I|}$$

Weigend and Ahlrichs *J Chem Phys* (2004)

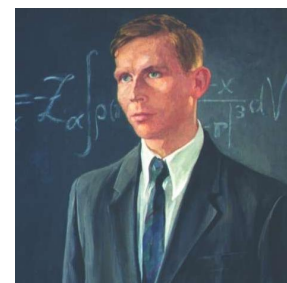
Yang & Beratan et al *JACS* (2006)

OAvL: *Phys Rev Lett* (2005), *J Chem Phys* (2006, 2009), *J Chem Theory Comput* (2007)



OAvL et al, *Phys Rev Lett* (2005)

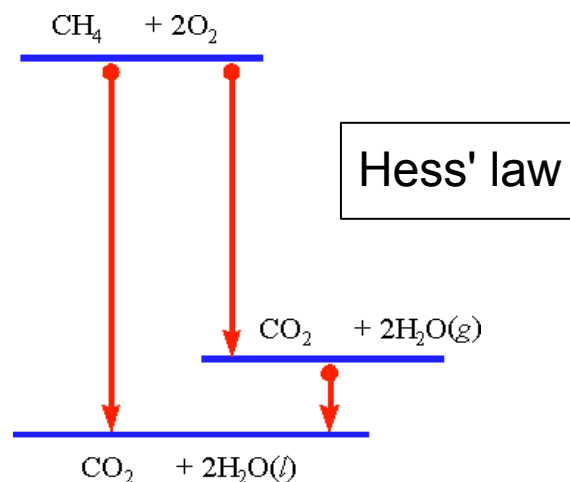
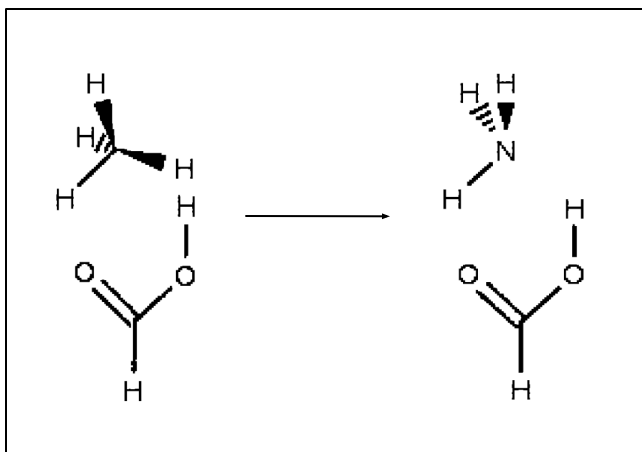
Hellmann



Feynman



Variational: Fractional nuclei



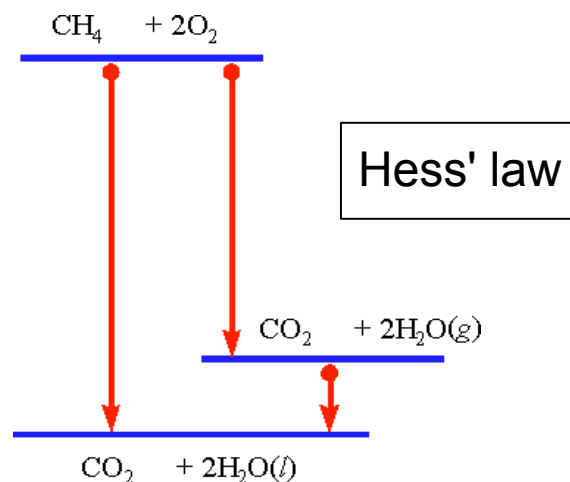
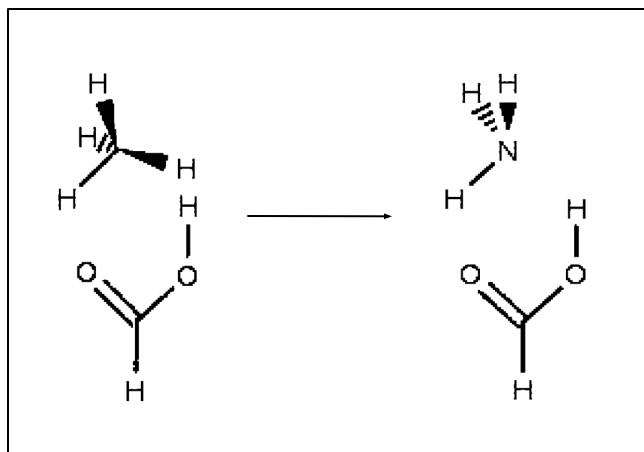
Tuckerman (NYU)

Weigend and Ahlrichs *J Chem Phys* (2004)

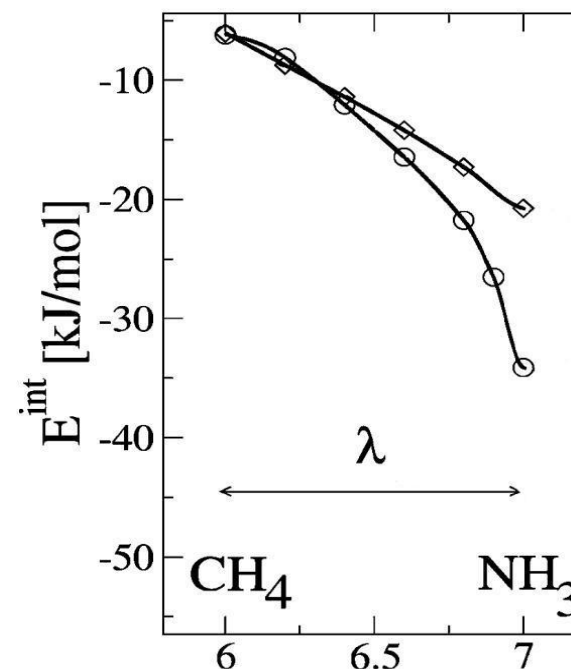
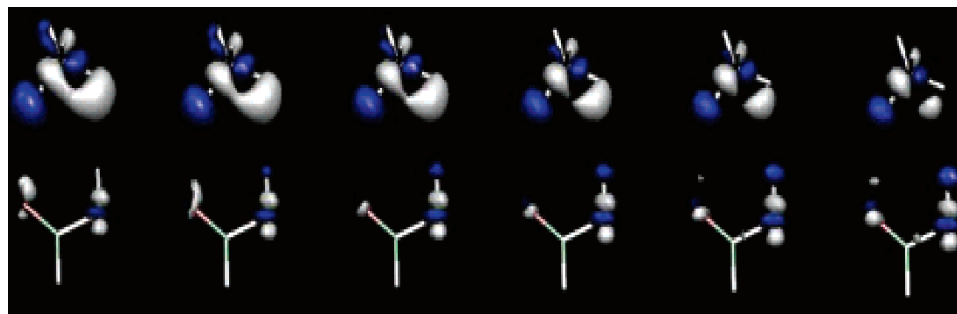
Yang & Beratan et al *JACS* (2006)

OAvL: *Phys Rev Lett* (2005), *J Chem Phys* (2006, 2009), *J Chem Theory Comput* (2007)

Variational: Fractional nuclei



Tuckerman (NYU)

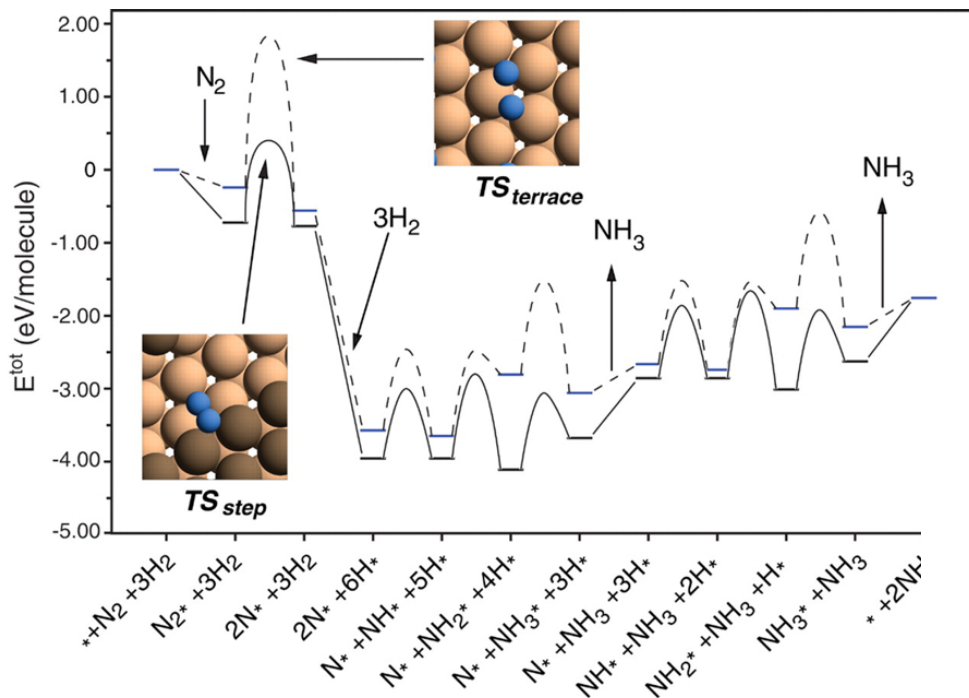
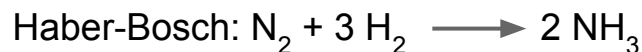


Weigend and Ahlrichs *J Chem Phys* (2004)

Yang & Beratan et al *JACS* (2006)

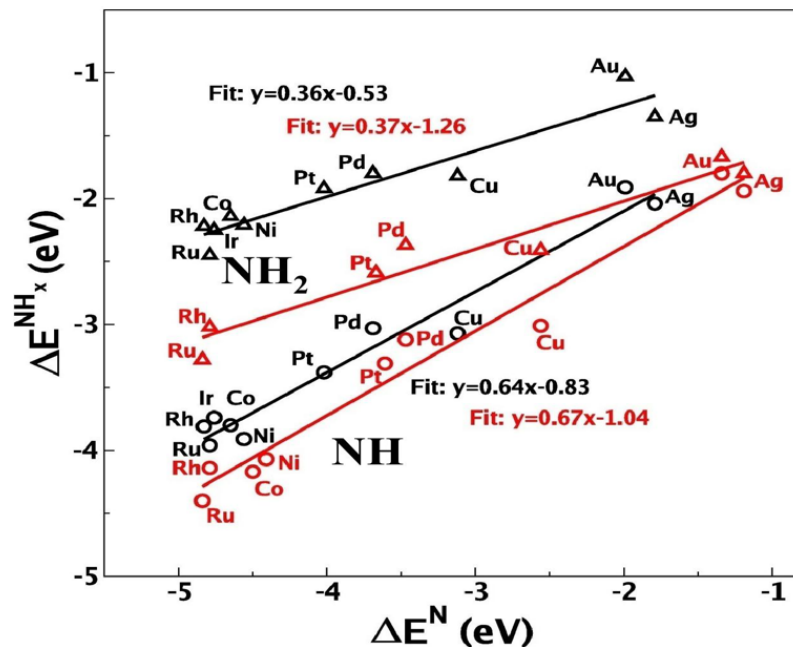
OAvL: *Phys Rev Lett* (2005), *J Chem Phys* (2006, 2009), *J Chem Theory Comput* (2007)

Adsorption

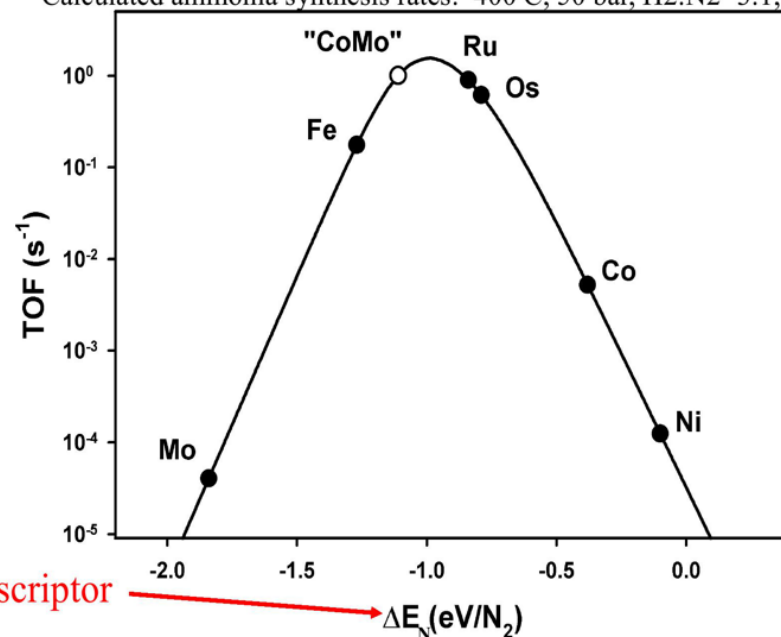


Sabatier's principle

Nørskov et al. *Nature Chemistry* (2009)
 Nørskov et al. *Science* (2005)
 Nørskov et al. *Phys Rev Lett* (2005)
 Nørskov et al. *J Am Chem Soc.* (2001)



Calculated ammonia synthesis rates: 400 C, 50 bar, $\text{H}_2:\text{N}_2=3:1$, 5% NH_3

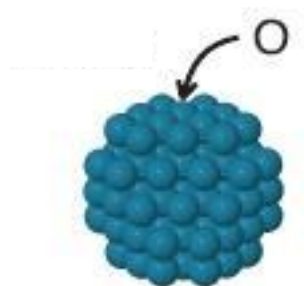


Descriptor ΔE_{N} (eV/ N_2)

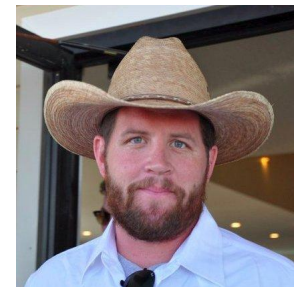
Adsorption

Volcano for oxygen reduction reaction: Oxygen binding

$$E^{\text{bind}} = E(\text{Pd}_{79}) - E(\text{Pd}_{79}\text{-O}) - 0.5 E(\text{O}_2)$$



45	46	47
Rh	Pd	Ag



Henkelman (UT) Sheppard (LANL)

How to dope?

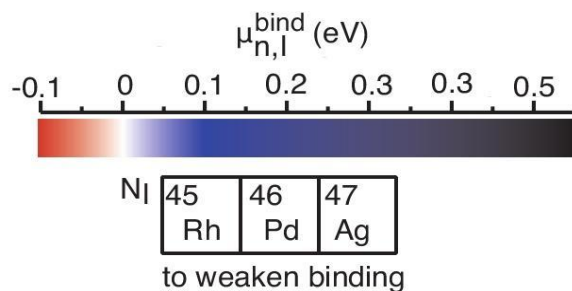
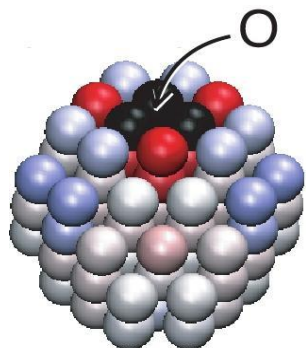
D Sheppard, G Henkelman, OAvL, *J Chem Phys* (2010)

Adsorption

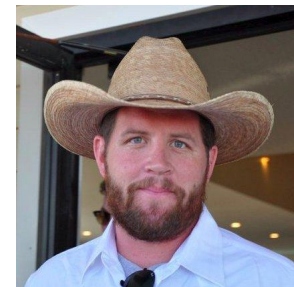
Volcano for oxygen reduction reaction: Oxygen binding

$$E^{\text{bind}} = E(\text{Pd}_{79}) - E(\text{Pd}_{79}\text{-O}) - 0.5 E(\text{O}_2)$$

$$\mu_{n,I} = \partial E^{\text{bind}} / \partial Z_I$$



Henkelman (UT)



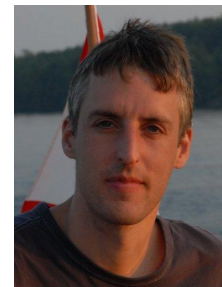
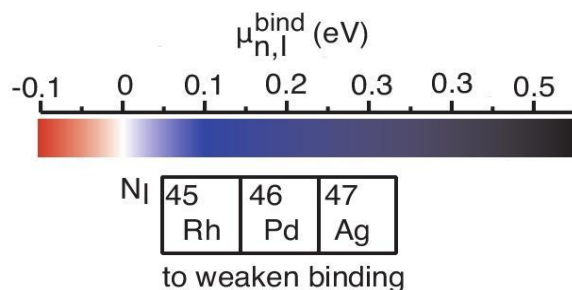
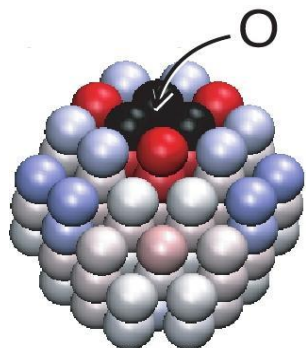
Sheppard (LANL)

Adsorption

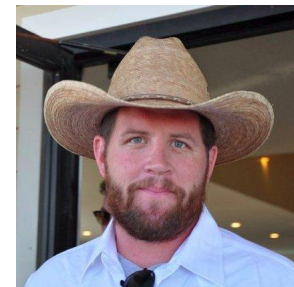
Volcano for oxygen reduction reaction: Oxygen binding

$$E^{\text{bind}} = E(\text{Pd}_{79}) - E(\text{Pd}_{79}\text{-O}) - 0.5 E(\text{O}_2)$$

$$\mu_{n,I} = \partial E^{\text{bind}} / \partial Z_I$$



Henkelman (UT)



Sheppard (LANL)

1st order expansion for 10 doped mutants

$$\partial_{\lambda} E^{\text{bind}} = \sum_I \mu_{n,I}^{\text{bind}} \partial_{\lambda} Z_I(\lambda)$$

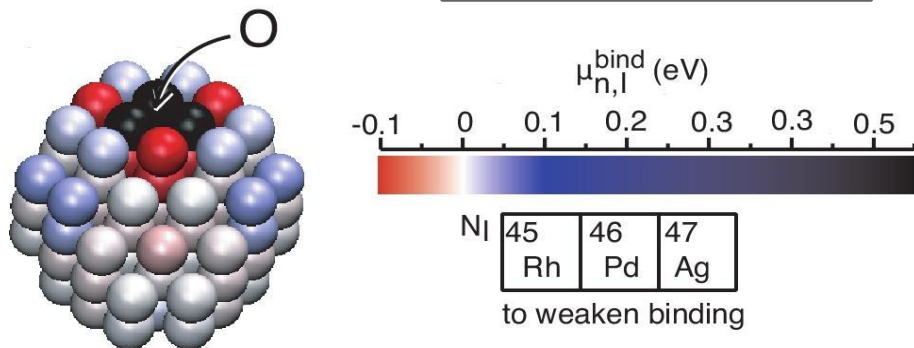
D Sheppard, G Henkelman, OAvL, *J Chem Phys* (2010)

Adsorption

Volcano for oxygen reduction reaction: Oxygen binding

$$E^{\text{bind}} = E(\text{Pd}_{79}) - E(\text{Pd}_{79}\text{-O}) - 0.5 E(\text{O}_2)$$

$$\mu_{n,I} = \partial E^{\text{bind}} / \partial Z_I$$

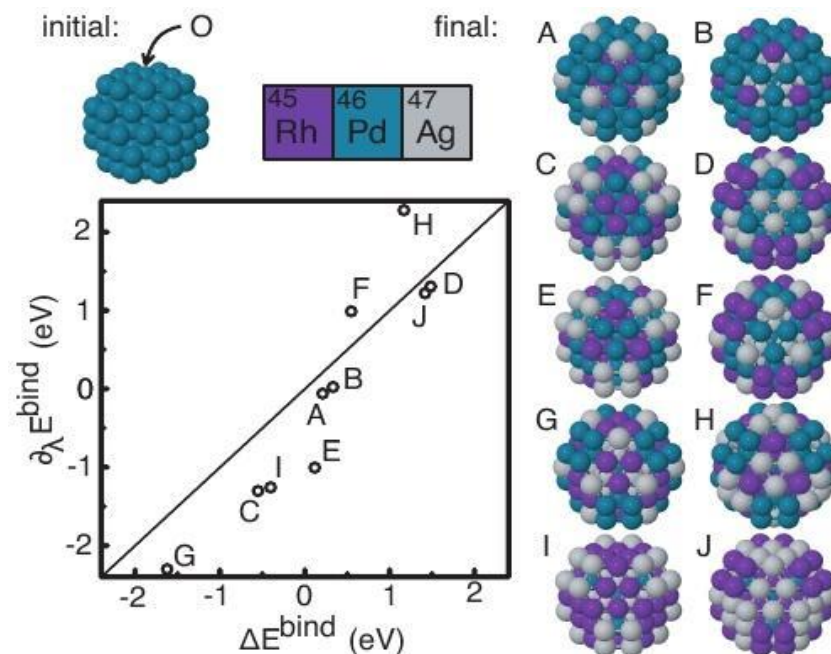


1st order expansion for 10 doped mutants

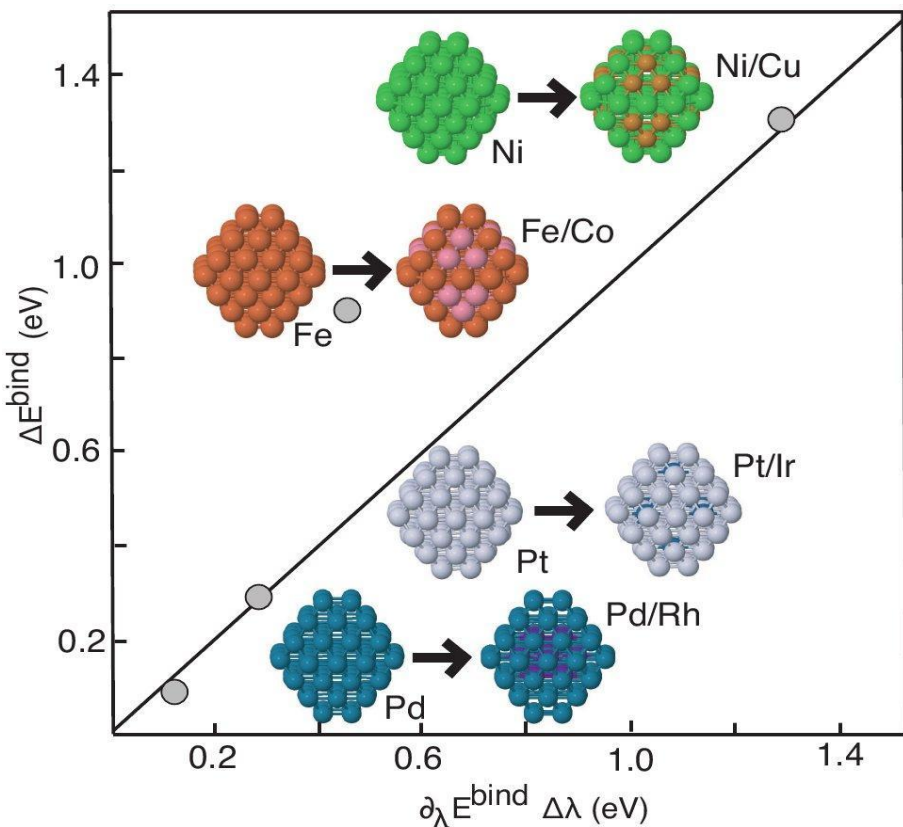
$$\partial_{\lambda} E^{\text{bind}} = \sum_I \mu_{n,I}^{\text{bind}} \partial_{\lambda} Z_I(\lambda)$$



Henkelman (UT) Sheppard (LANL)



Adsorption



Henkelman (UT)



Sheppard (LANL)

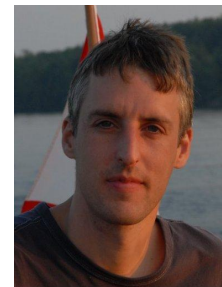
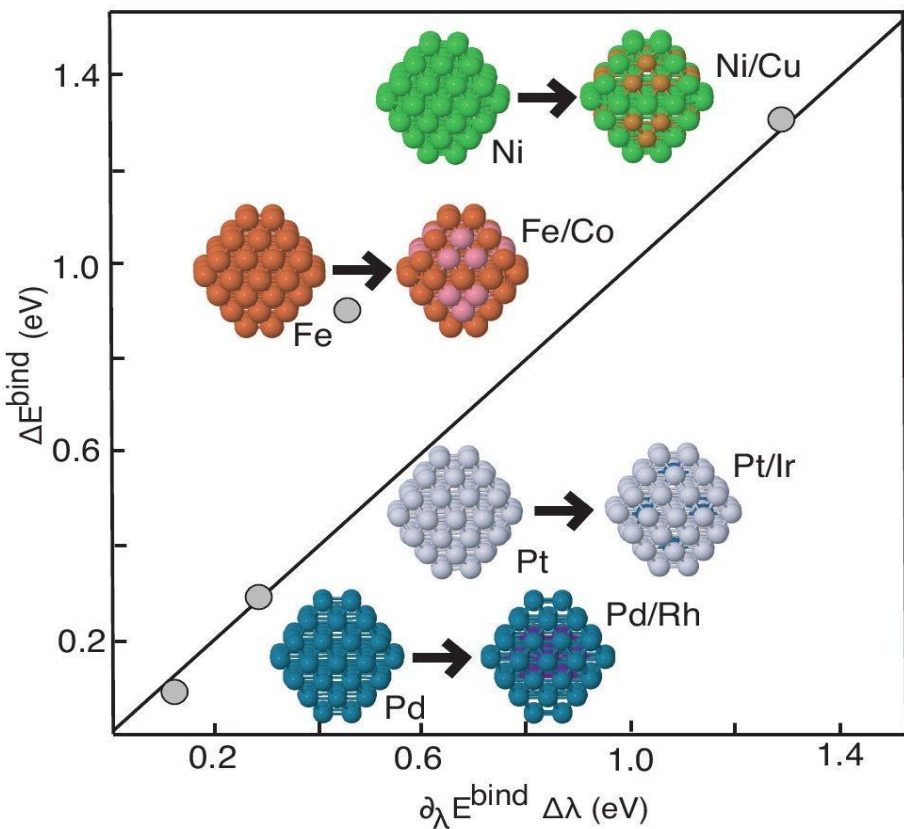
							5 B
							13 Al
4	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga
4	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In
4	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl
5	107	108	109	110			

Target oxygen binding value: 1.65 eV

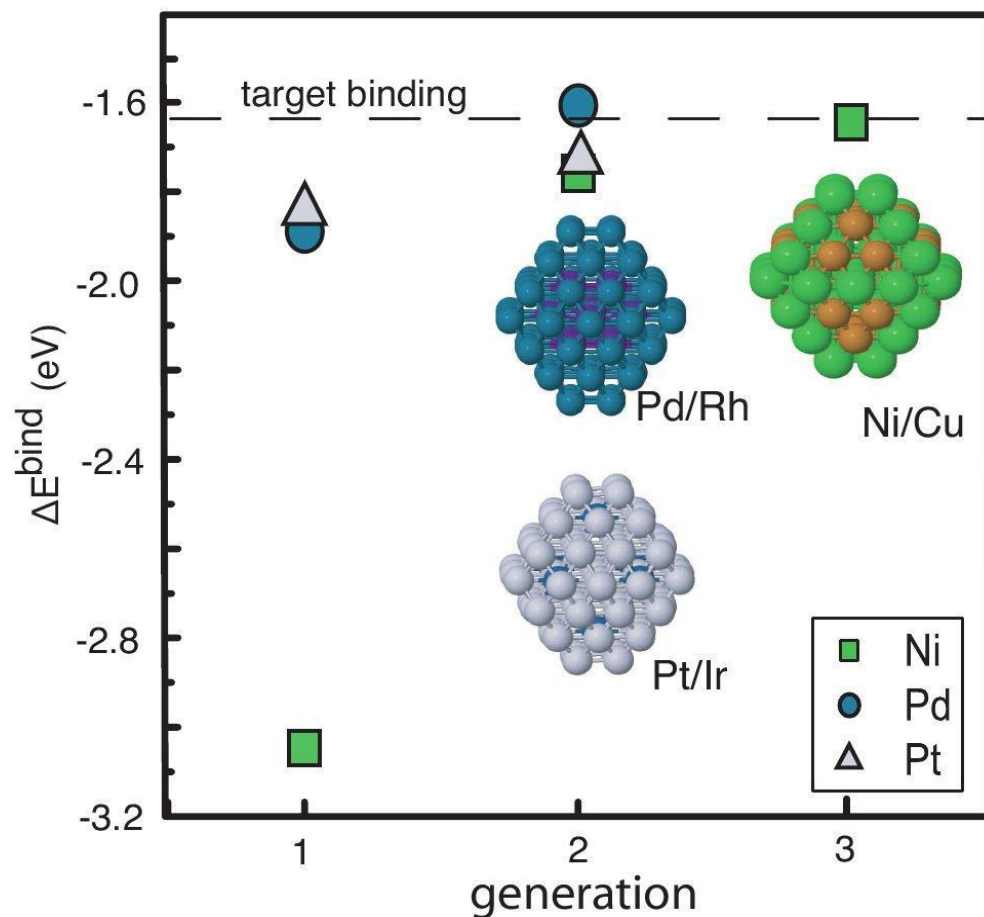
$$\min_{\{Z_I, \mathbf{R}_I\}} \left(P(\{Z_I, \mathbf{R}_I\}) - P^{\text{ref}} \right)^2$$

Dan Sheppard, PhD thesis, UT Austin 2010

Adsorption



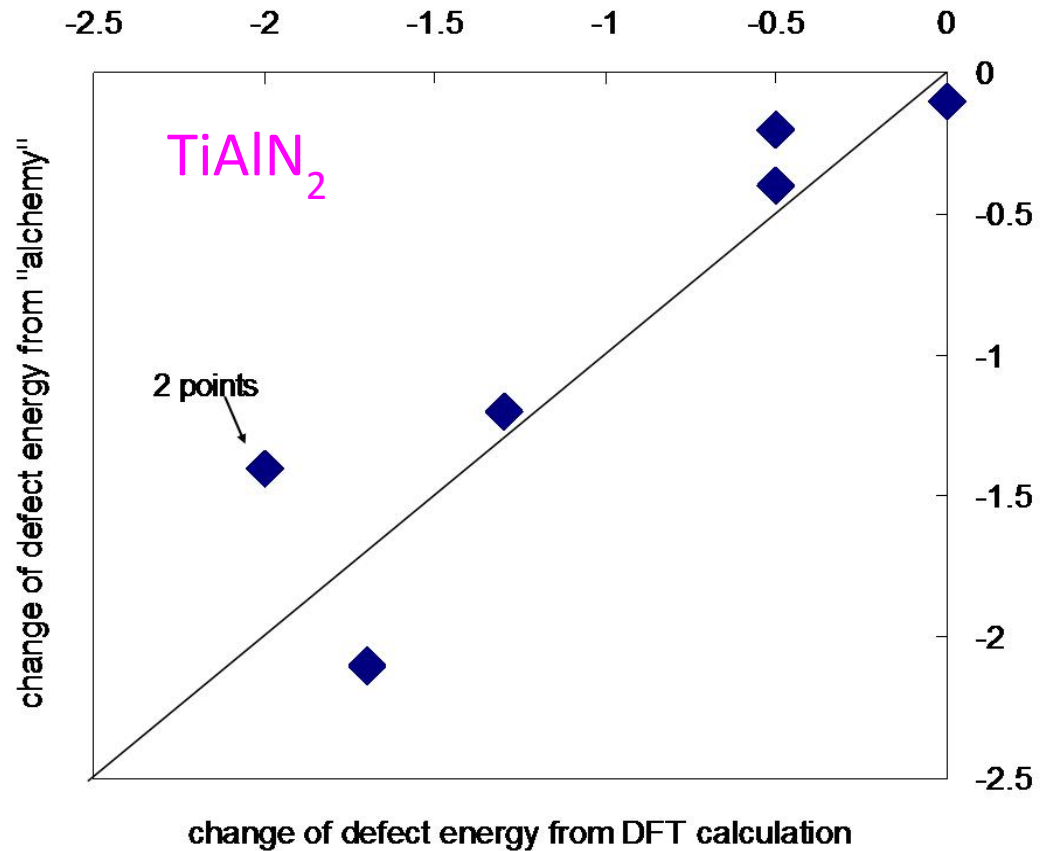
Henkelman (UT) Sheppard (LANL)



Dan Sheppard, PhD thesis, UT Austin 2010

Defects

Al-vacancy energies [eV]



Predicted changes for various N to O mutations (out of 32)

Preliminary results from Moritz to Baben (group of Prof. Schneider, RWTH Aachen)

Is Z really a good variable?

hydrogen 1 H 1.0079																	helium 2 He 4.0026	
lithium 3 Li 6.941	beryllium 4 Be 9.0122											boron 5 B 10.811	carbon 6 C 12.011	nitrogen 7 N 14.007	oxygen 8 O 15.999	fluorine 9 F 18.998	neon 10 Ne 20.180	
sodium 11 Na 22.990	magnesium 12 Mg 24.305											aluminium 13 Al 26.982	silicon 14 Si 28.086	phosphorus 15 P 30.974	sulfur 16 S 32.065	chlorine 17 Cl 35.453	argon 18 Ar 39.948	
potassium 19 K 39.098	calcium 20 Ca 40.078	scandium 21 Sc 44.956	titanium 22 Ti 47.867	vanadium 23 V 50.942	chromium 24 Cr 51.996	manganese 25 Mn 54.938	iron 26 Fe 55.845	cobalt 27 Co 58.933	nickel 28 Ni 58.693	copper 29 Cu 63.546	zinc 30 Zn 65.39	gallium 31 Ga 69.723	germanium 32 Ge 72.61	arsenic 33 As 74.922	selenium 34 Se 78.96	bromine 35 Br 79.904	krypton 36 Kr 83.80	
rubidium 37 Rb 85.468	strontium 38 Sr 87.62	yttrium 39 Y 88.906	zirconium 40 Zr 91.224	niobium 41 Nb 92.906	molybdenum 42 Mo 95.94	technetium 43 Tc [98]	ruthenium 44 Ru 101.07	rhodium 45 Rh 102.91	palladium 46 Pd 106.42	silver 47 Ag 107.87	cadmium 48 Cd 112.41	indium 49 In 114.82	tin 50 Sn 118.71	antimony 51 Sb 121.76	tellurium 52 Te 127.60	iodine 53 I 126.90	xenon 54 Xe 131.29	
caesium 55 Cs 132.91	barium 56 Ba 137.33	57-70 ★	lutetium 71 Lu 174.97	hafnium 72 Hf 178.49	tantalum 73 Ta 180.95	tungsten 74 W 183.84	rhenium 75 Re 186.21	osmium 76 Os 190.23	iridium 77 Ir 192.22	platinum 78 Pt 195.08	gold 79 Au 196.97	mercury 80 Hg 200.59	thallium 81 Tl 204.38	lead 82 Pb 207.2	bismuth 83 Bi 208.98	polonium 84 Po [209]	astatine 85 At [210]	radon 86 Rn [222]
francium 87 Fr [223]	radium 88 Ra [226]	89-102 ★ ★	lawrencium 103 Lr [262]	rutherfordium 104 Rf [261]	dubnium 105 Db [262]	seaborgium 106 Sg [266]	bohrium 107 Bh [264]	hassium 108 Hs [269]	meitnerium 109 Mt [268]	ununnium 110 Uun [271]	ununium 111 Uuu [272]	unubium 112 Uub [277]	ununquadium 114 Uuq [289]					

* Lanthanide series

lanthanum 57 La 138.91	cerium 58 Ce 140.12	praseodymium 59 Pr 140.91	neodymium 60 Nd 144.24	promethium 61 Pm [145]	samarium 62 Sm 150.36	europium 63 Eu 151.96	gadolinium 64 Gd 157.25	terbium 65 Tb 158.93	dysprosium 66 Dy 162.50	holmium 67 Ho 164.93	erbium 68 Er 167.26	thulium 69 Tm 168.93	ytterbium 70 Yb 173.04
actinium 89 Ac [227]	thorium 90 Th 232.04	protactinium 91 Pa 231.04	uranium 92 U 238.03	neptunium 93 Np [237]	plutonium 94 Pu [244]	americium 95 Am [243]	curium 96 Cm [247]	berkelium 97 Bk [247]	californium 98 Cf [251]	einsteinium 99 Es [252]	fermium 100 Fm [257]	mendelevium 101 Md [258]	nobelium 102 No [259]

** Actinide series

Is Z really a good variable?

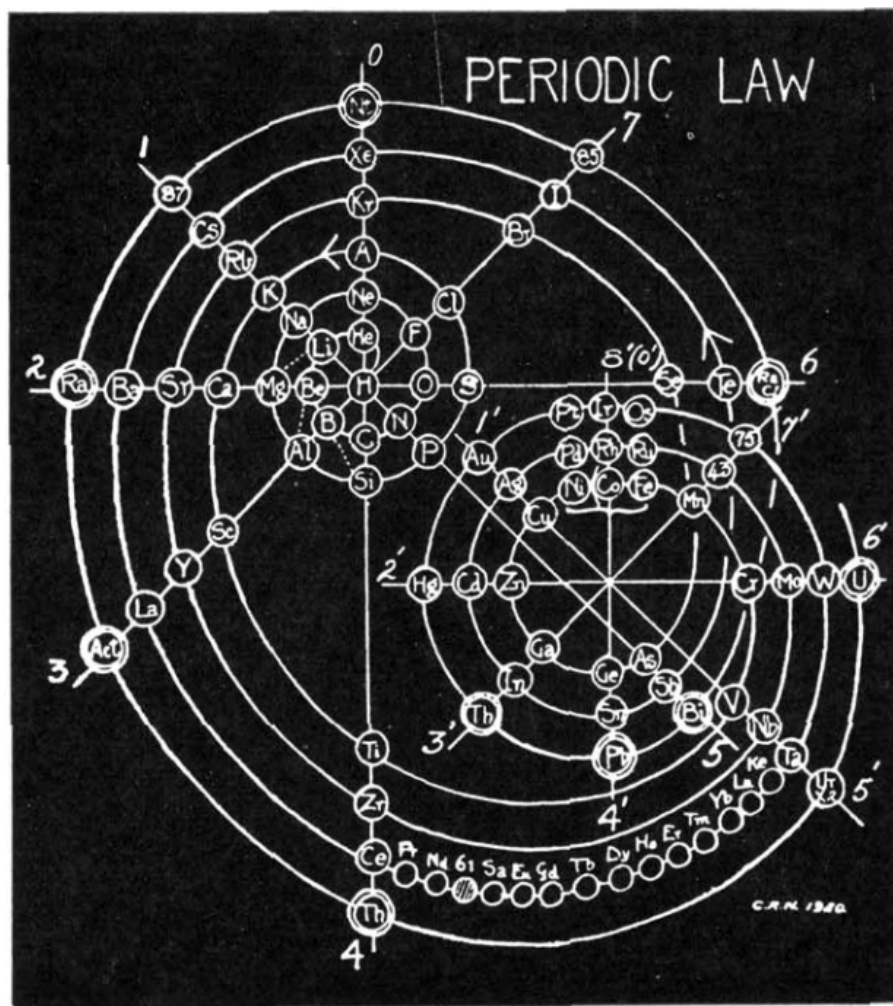


FIGURE 16.—NODDER'S PERIODIC TABLE

Quam & Quam, J Chem Educ (1934)

Is Z really a good variable?

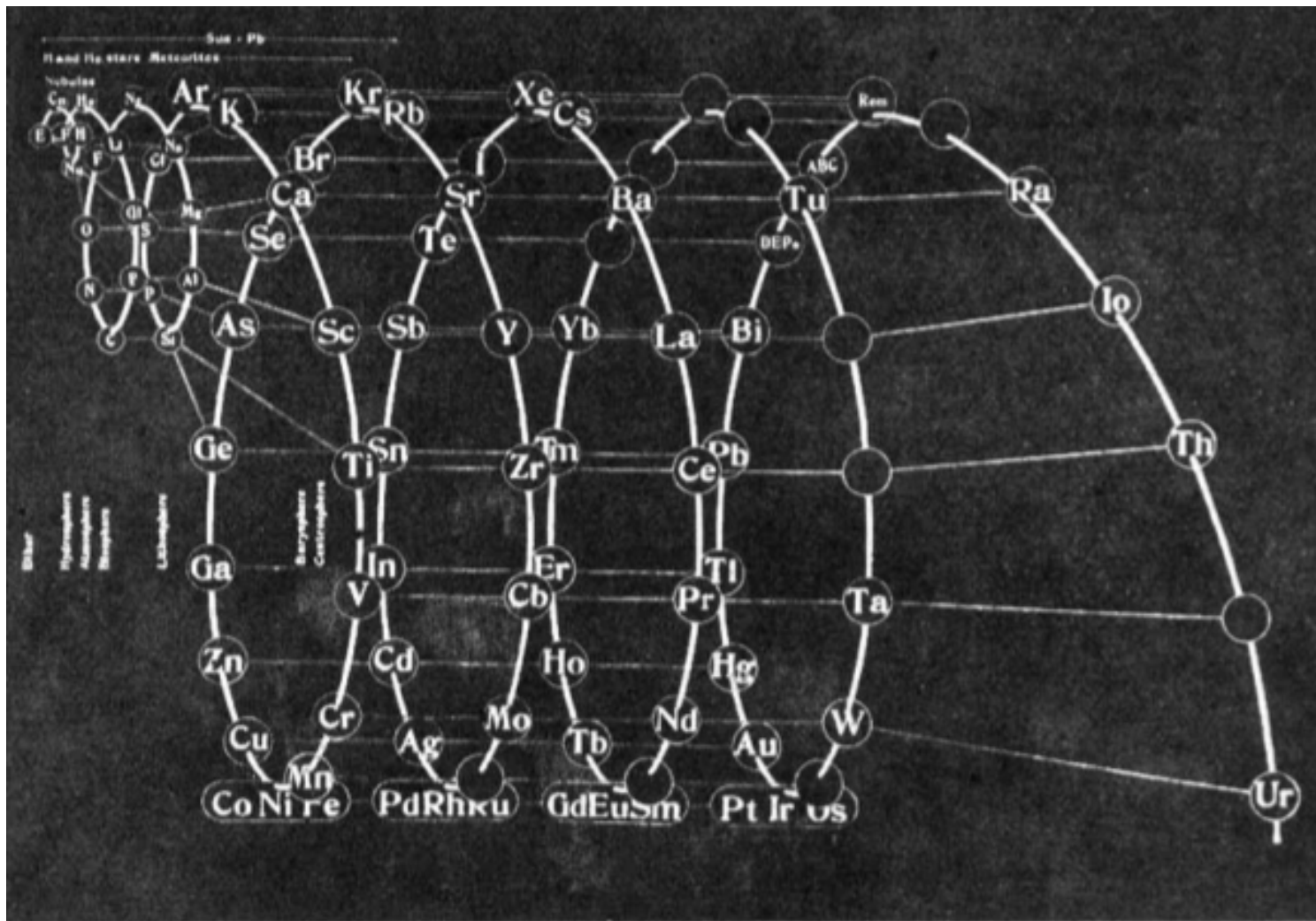


FIGURE 21.—EMERSON'S HELIX

Quam & Quam, J Chem Educ (1934)

Is Z really a good variable?

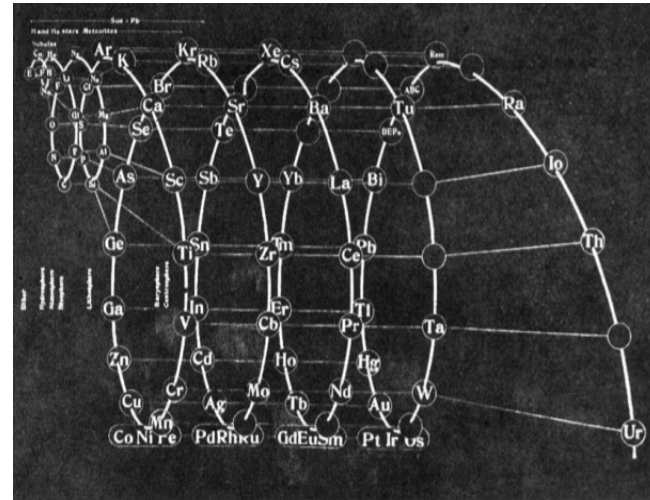
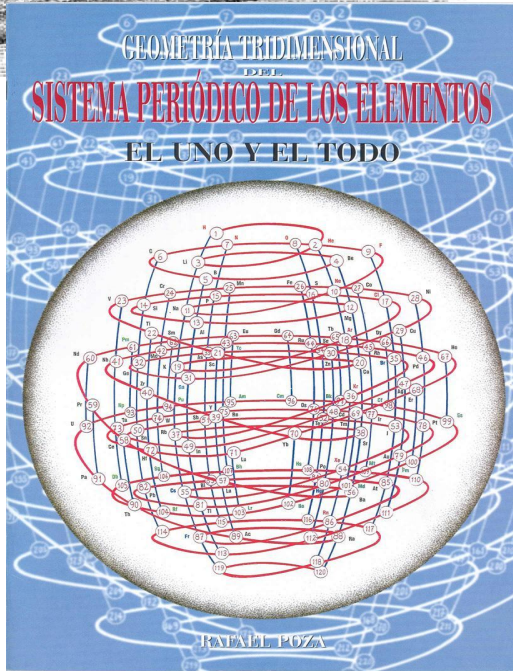
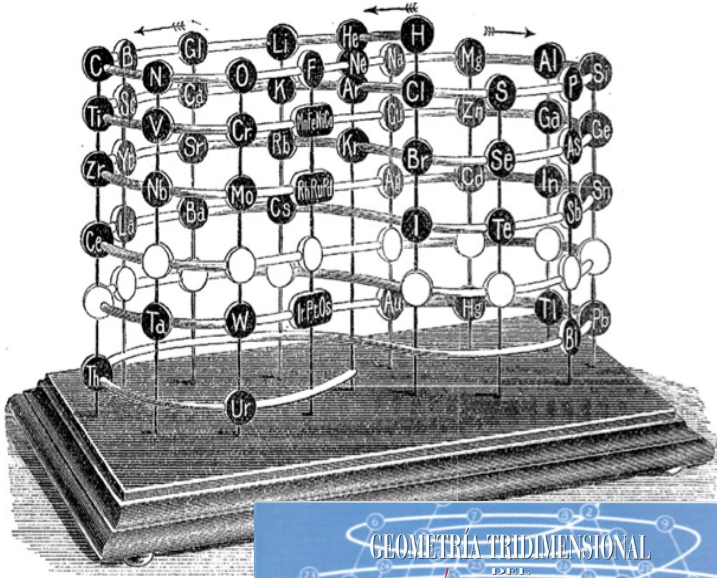
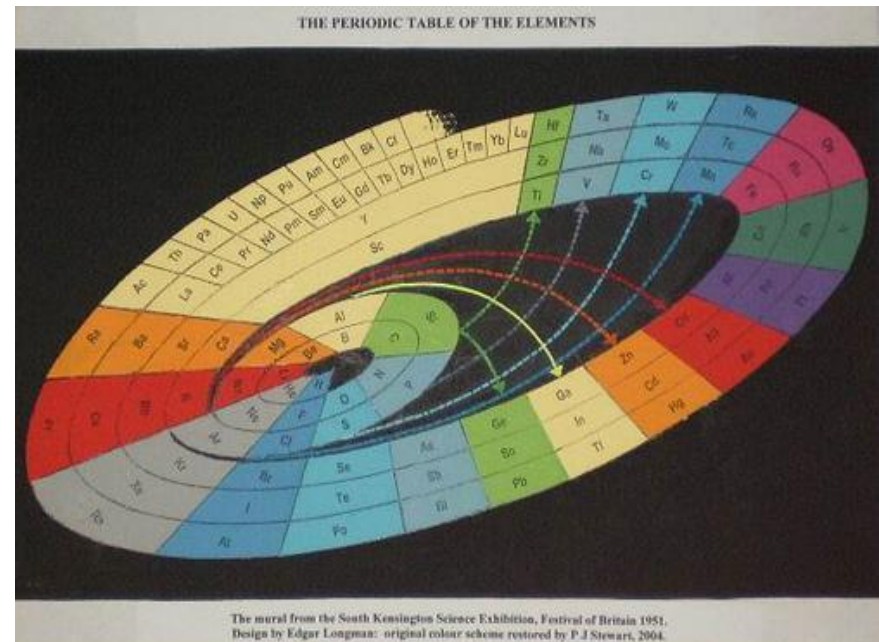
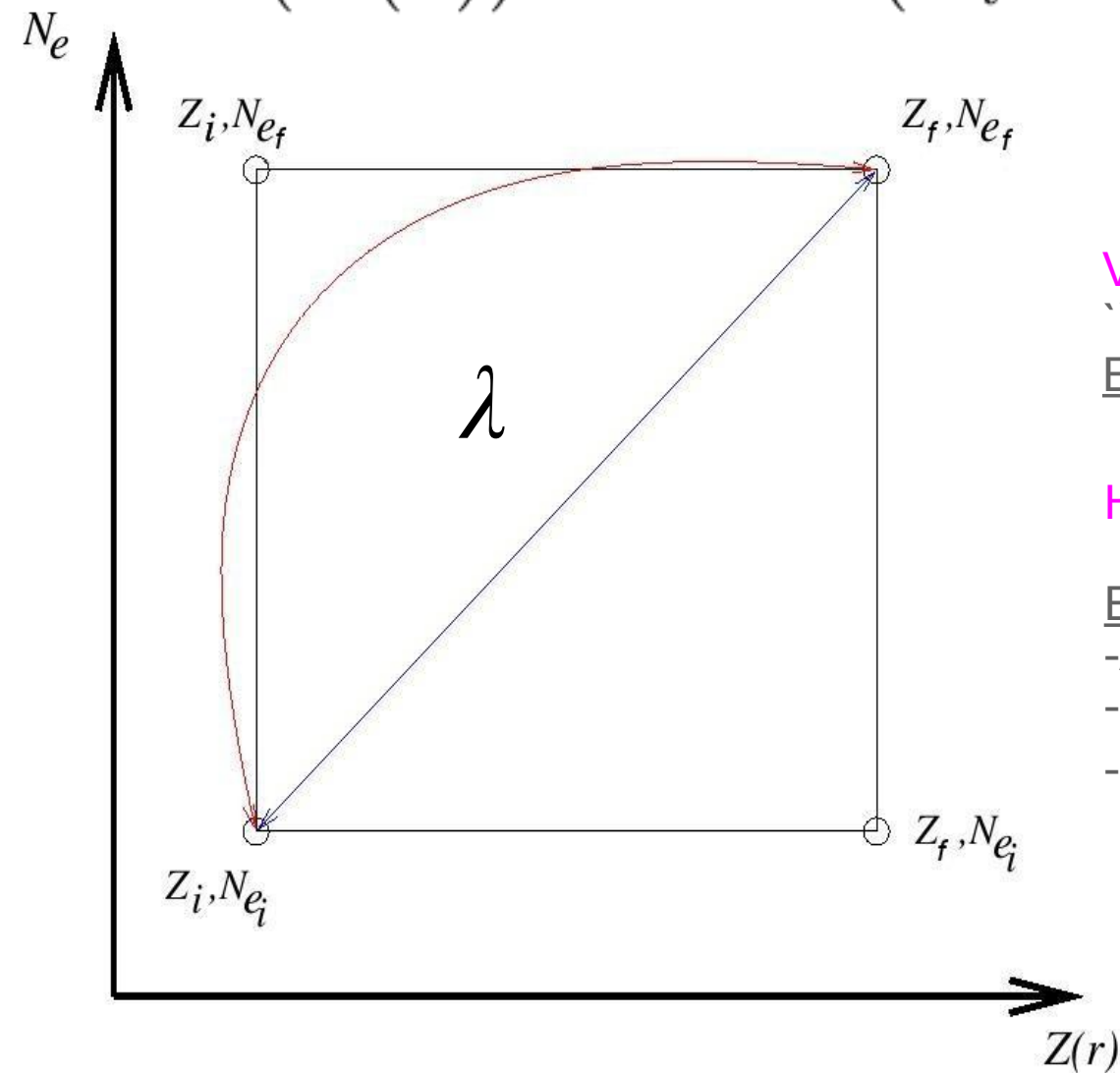


FIGURE 21.—EMERSON'S HELIX



Generalization

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



Vertical changes:

``redox''

Example: $\text{Li} \rightarrow \text{Li}^+ + \text{e}^-$

Horizontal changes:

iso-electronic & ``alchemical''

Example:

- All constitutional isomers
- hydrazine (N_2H_4) \rightarrow CH_3OH
- same number of valence electrons

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

variational (deductive)

Feynman



Alchemy

1. Free energies
2. Gradients to optimize

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

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



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

variational (deductive)

Alchemy

1. Free energies
2. Gradients to optimize

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

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

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

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

Feynman

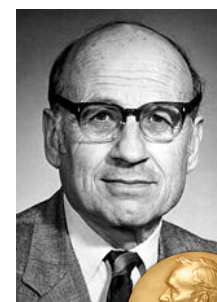


Generalization

$$\begin{aligned} E(H(\lambda)) &= E(H_i + \lambda(H_f - H_i)) \\ \frac{\partial E[H]}{\partial \lambda} &= \left\langle \Psi \left| \frac{\partial H(\lambda)}{\partial \lambda} \right| \Psi \right\rangle \\ &= \int d\mathbf{r} \, n_\lambda(\mathbf{r}) \times [v_j^{ext}(\mathbf{r}) - v_i^{ext}(\mathbf{r})] \end{aligned}$$



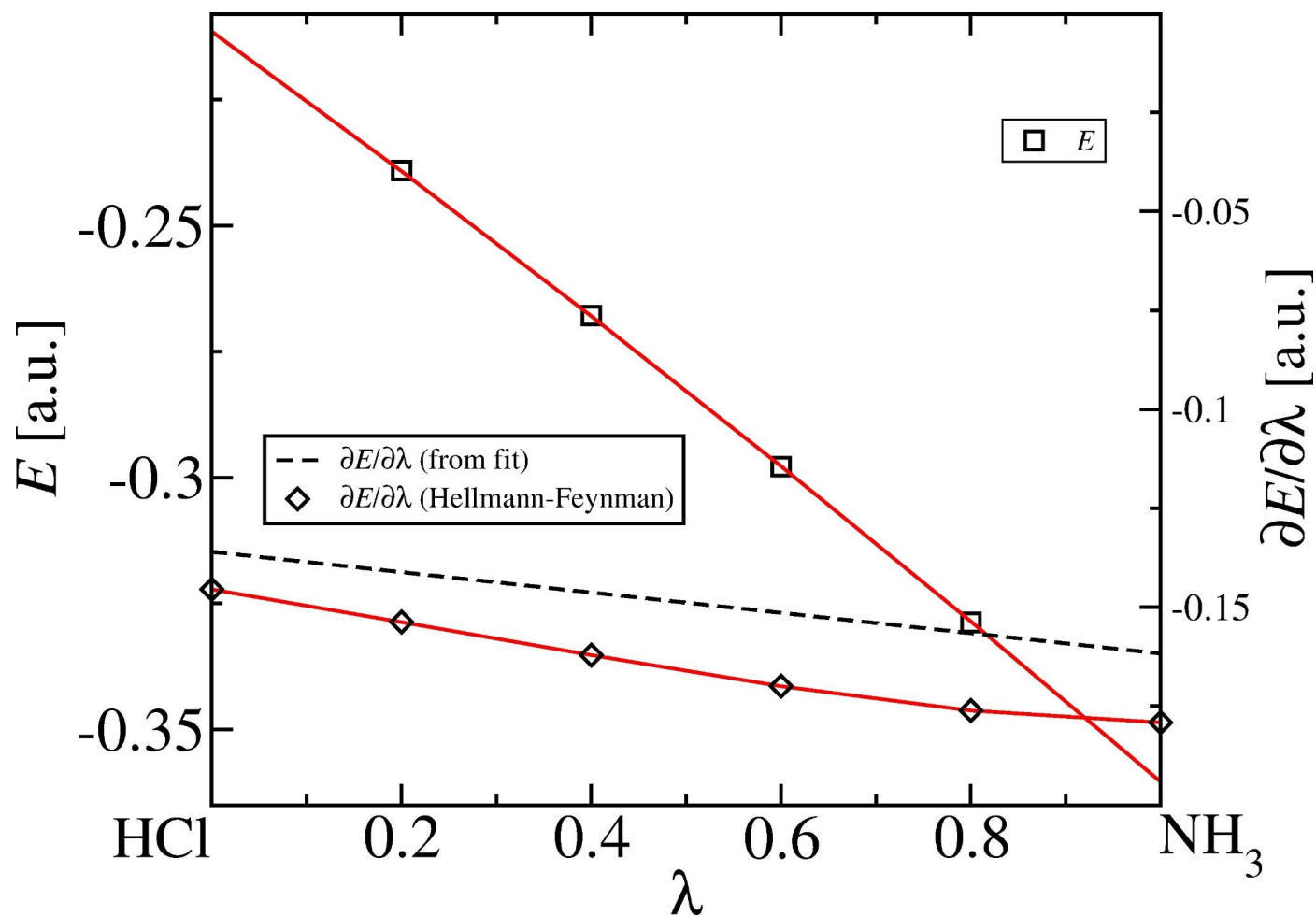
Hohenberg



Kohn



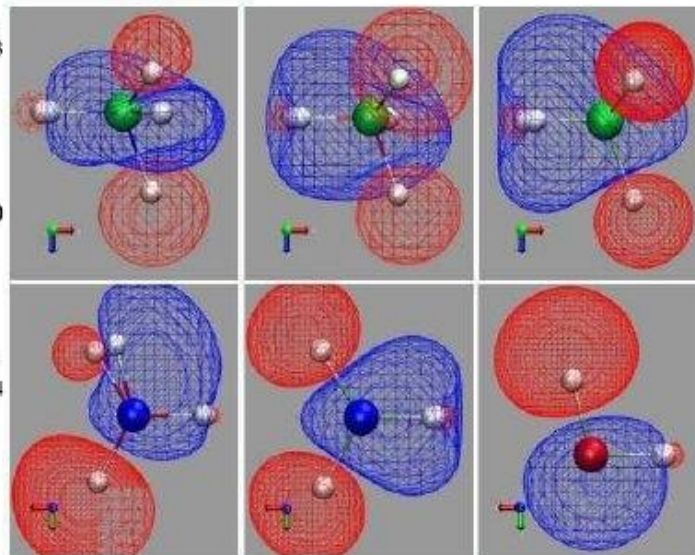
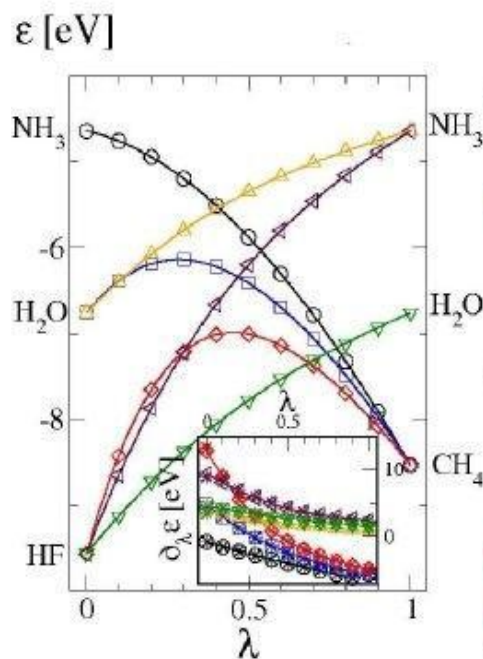
Generalization



Example?

$$\epsilon(\lambda) = \frac{1}{\delta} (E(N_e, \lambda) - E(N_e - \delta, \lambda))$$

$$\partial_\lambda E(\lambda) = \int d\mathbf{r} n_\lambda(\mathbf{r}) \times [v_j^{ext}(\mathbf{r}) - v_i^{ext}(\mathbf{r})]$$



$$\partial_\lambda \epsilon(\lambda) = \frac{1}{\delta} \left(\int d\mathbf{r} [n_\lambda(\mathbf{r}) - n_\lambda^{+\delta}(\mathbf{r})] \times [v_j^{ext}(\mathbf{r}) - v_i^{ext}(\mathbf{r})] \right)$$

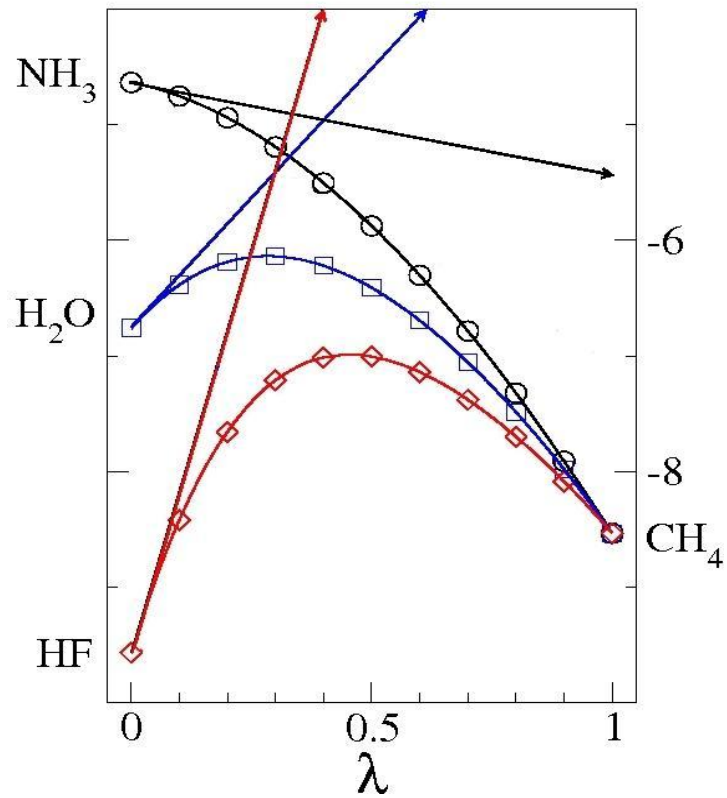
vs.

$$d\epsilon(\lambda)/d\lambda = \frac{1}{\delta} (\epsilon(\lambda + \delta) - \epsilon(\lambda))$$

OAvL JCP
(2009)

But what about prediction?

ϵ [eV]

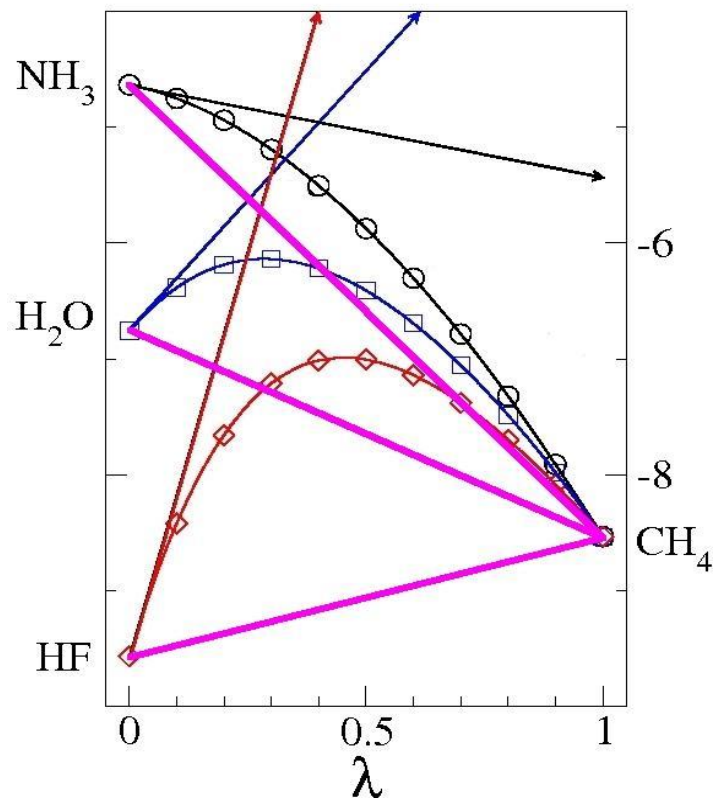


$$\epsilon_{\lambda=1} \approx \epsilon_{\lambda=0} + \left. \frac{\partial \epsilon}{\partial \lambda} \right|_{\lambda=0} \Delta\lambda + H.O.T.$$

$$\Delta\lambda = 1$$

Prediction?

ϵ [eV]



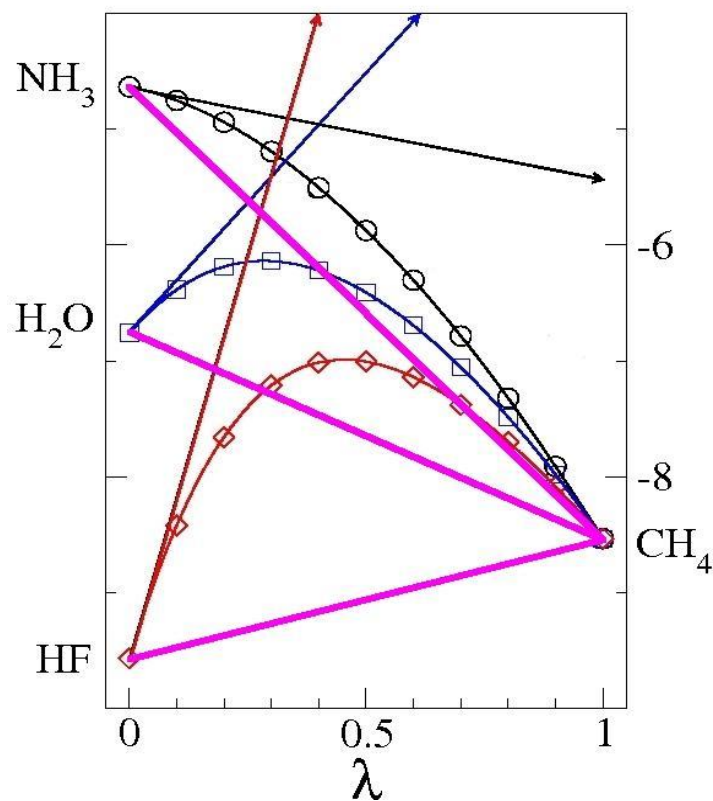
$$\epsilon_{\lambda=1} \approx \epsilon_{\lambda=0} + \left. \frac{\partial \epsilon}{\partial \lambda} \right|_{\lambda=0} \Delta\lambda + H.O.T.$$

$$\Delta\lambda = 1$$

OAvL JCP
(2009)

Prediction?

ϵ [eV]



$$\epsilon_{\lambda=1} \approx \epsilon_{\lambda=0} + \left. \frac{\partial \epsilon}{\partial \lambda} \right|_{\lambda=0} \Delta \lambda + H.O.T.$$

$$\Delta \lambda = 1$$

In analogy to:
Smith and van Gunsteren *JCP* (1994)

$$E^{lin} = E_i + \lambda \times (E_f - E_i)$$

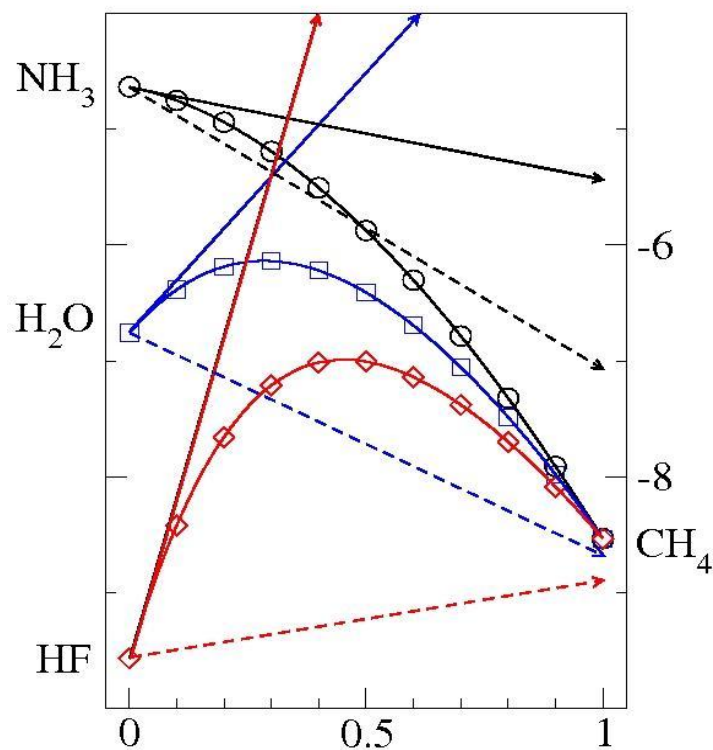
$$= \langle H_i + f_{if}(\lambda) \times (H_f - H_i) \rangle_{\lambda}$$

$$f_{if}(\lambda) = \begin{cases} 0 & \text{if } \lambda = 0 \\ 1 & \text{if } \lambda = 1 \end{cases}$$

$$f_{if}(\lambda) = a_{if}(\lambda^2 - \lambda) + \lambda$$

Prediction?

ϵ [eV]



For reference pairs:
 $\text{CH}_3\text{NH}_2 \rightarrow \text{CH}_3\text{CH}_3$
 $\text{CH}_3\text{OH} \rightarrow \text{CH}_3\text{CH}_3$
 $\text{CH}_3\text{F} \rightarrow \text{CH}_3\text{CH}_3$

OAvL *JCP*
 (2009)

$$\epsilon_{\lambda=1} \approx \epsilon_{\lambda=0} + \left. \frac{\partial \epsilon}{\partial \lambda} \right|_{\lambda=0} \Delta \lambda + H.O.T.$$

$$\Delta \lambda = 1$$

In analogy to:
 Smith and van Gunsteren *JCP* (1994)

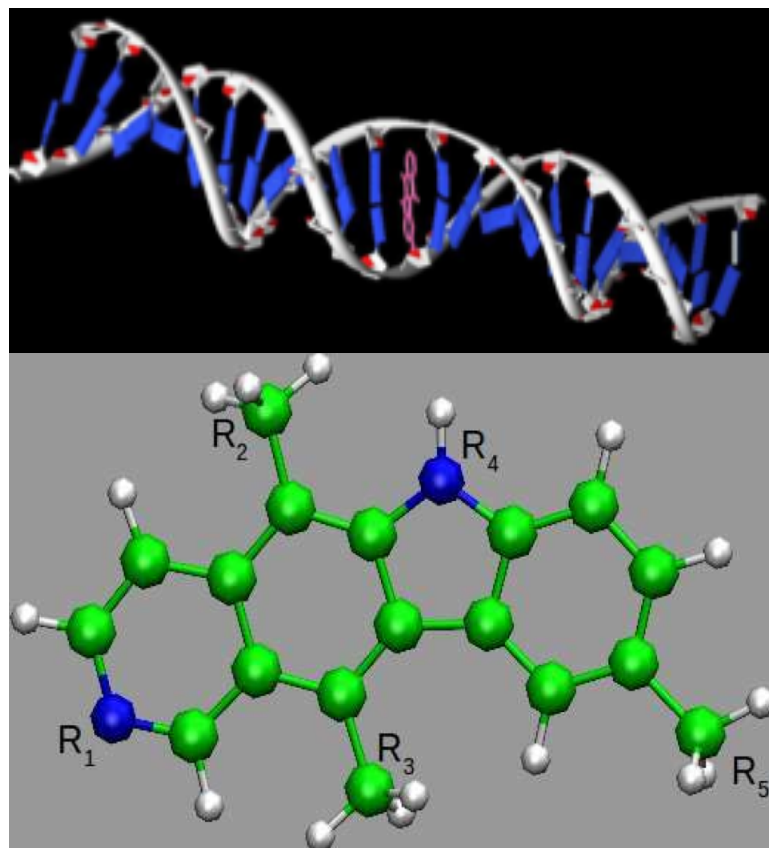
$$E^{lin} = E_i + \lambda \times (E_f - E_i)$$

$$= \langle H_i + f_{if}(\lambda) \times (H_f - H_i) \rangle_\lambda$$

$$f_{if}(\lambda) = \begin{cases} 0 & \text{if } \lambda = 0 \\ 1 & \text{if } \lambda = 1 \end{cases}$$

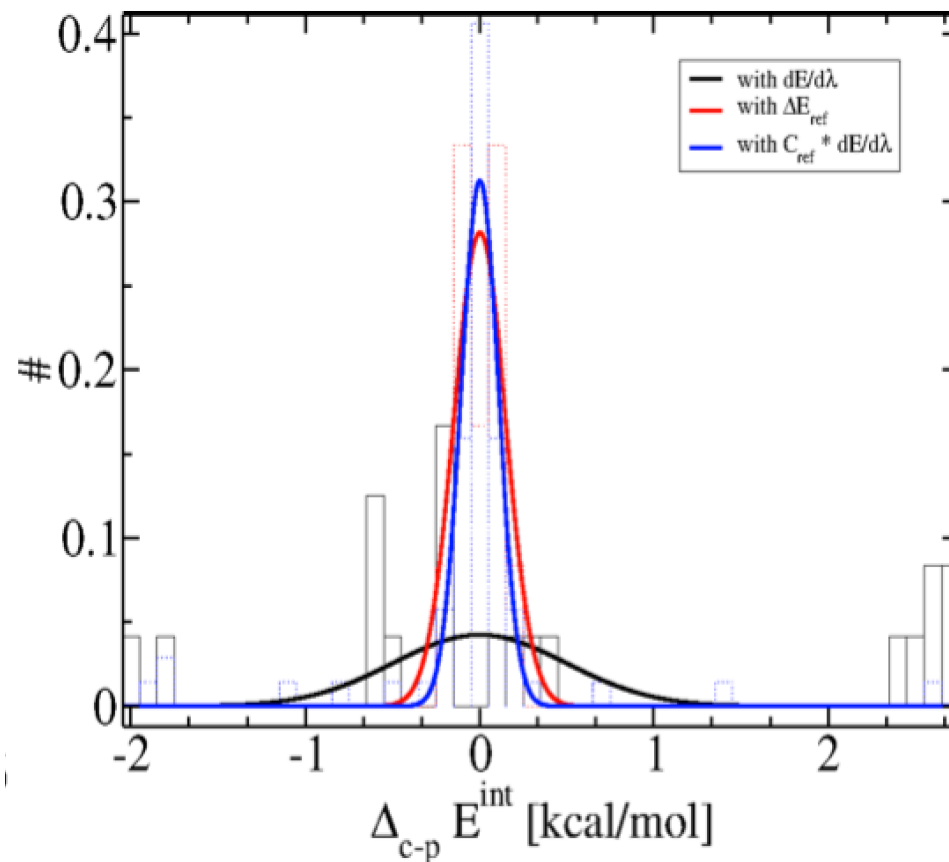
$$f_{if}(\lambda) = a_{if}(\lambda^2 - \lambda) + \lambda$$

Drug design?



Ellipticine, intercalated
between 2 Watson-Crick
base-pairs w backbone,
using vdW+DFT
(GGA+DCACP)
J Phys Chem B (2007)

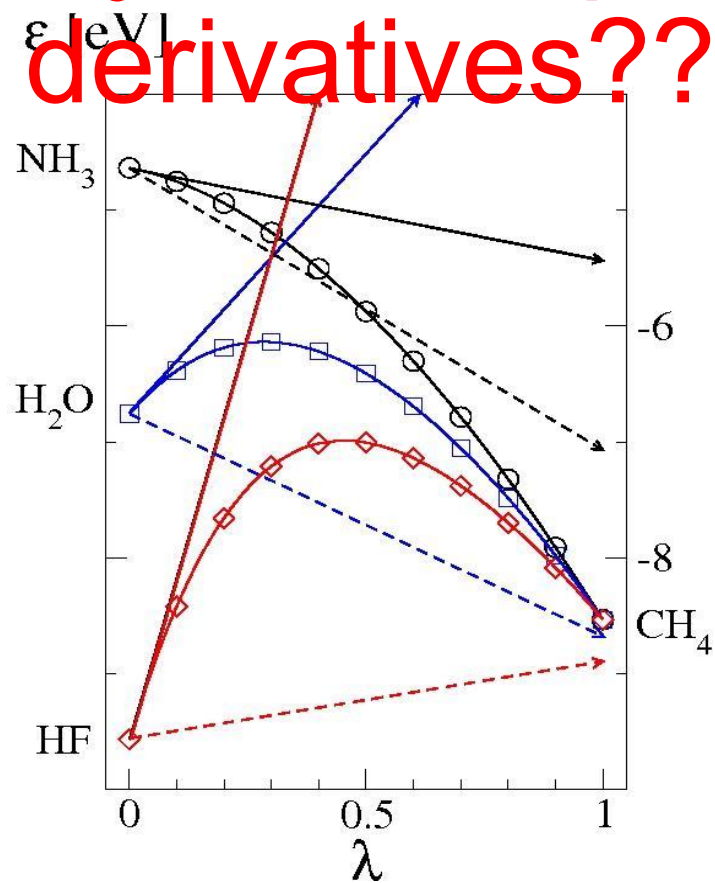
OAvL, Int J Quant Chem (2013)



site vs. group	1	2	3	4	5	6
R ₁	CH	N	SiH	P	-	-
R ₂	CH ₃	NH ₂	OH ^{left}	OH ^{right}	F	Cl
R ₃	CH ₃	NH ₂	OH ^{left}	OH ^{right}	F	Cl
R ₄	CH ₂	NH	O	SiH ₂	PH	S
R ₅	CH ₃	NH ₂	OH ^{left}	OH ^{right}	F	Cl

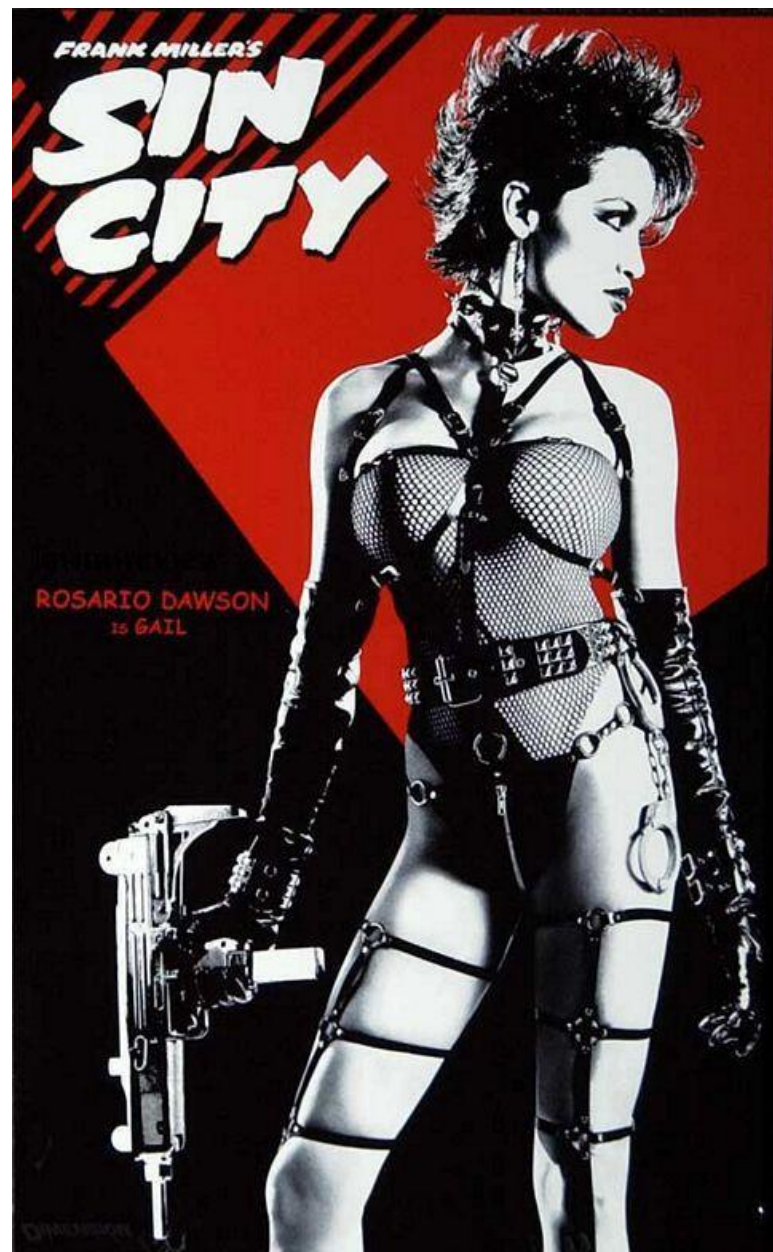
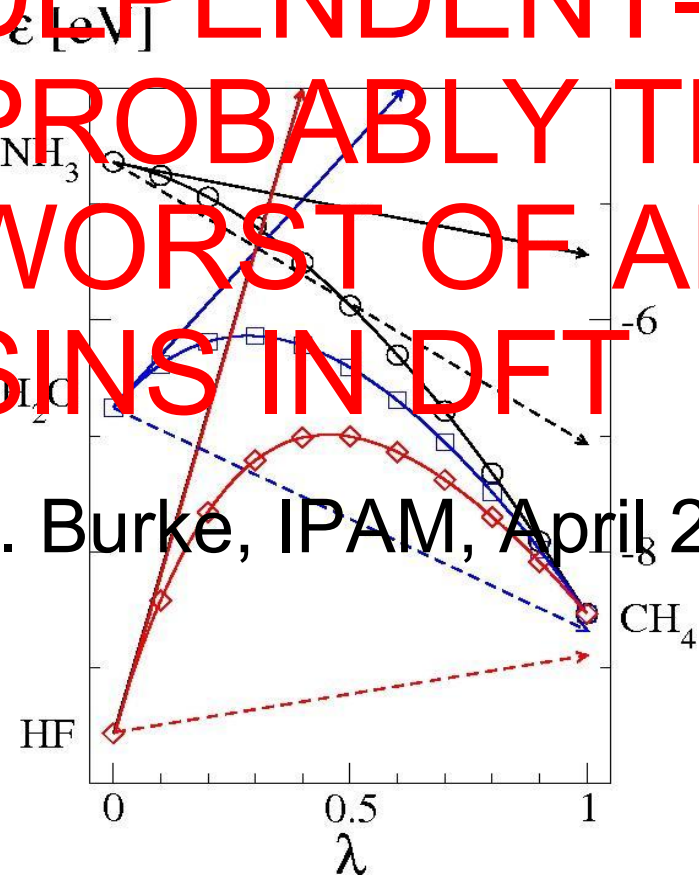
Prediction?

System dependent derivatives???



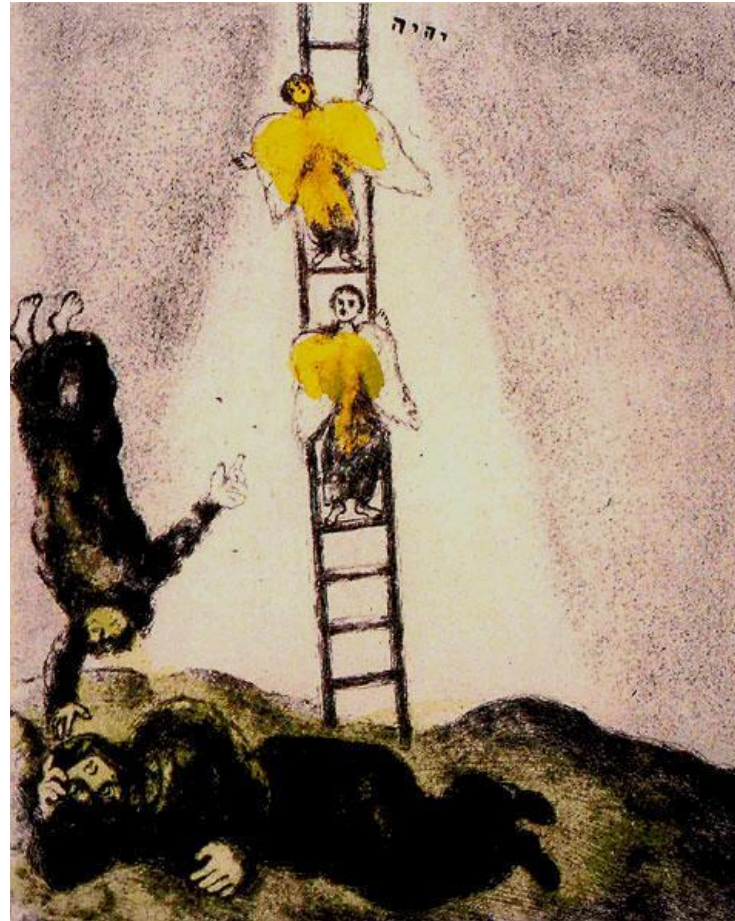
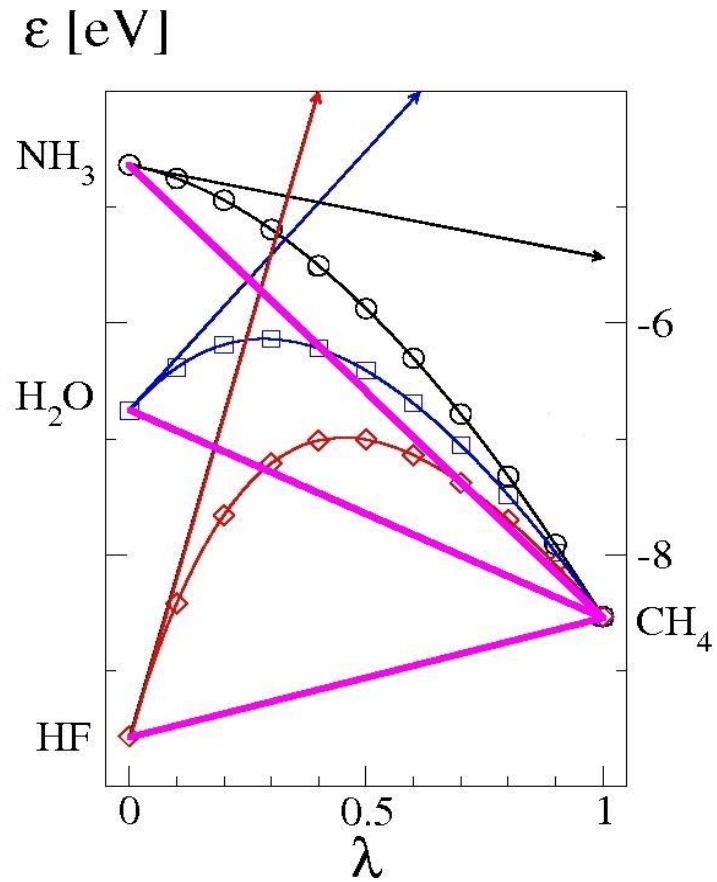
SYSTEM DEPENDENT--- PROBABLY THE WORST OF ALL SINS IN DFT

K. Burke, IPAM, April 2011

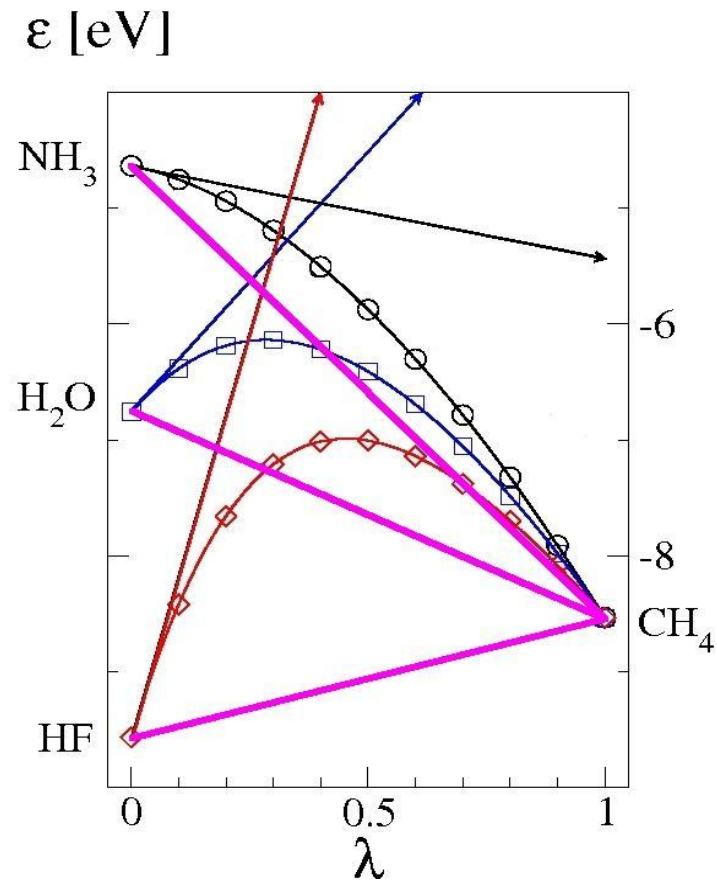


Help!

WANTED!
Jacob's Ladder for CCS



Help! → Swarm Intelligence



Erdős problems

Throughout his career, Erdős would offer US\$ prizes for solutions to unresolved problems.

<http://wikipedia.org>



Win a prize!!!

An ounce of Gold in the form of 100 shares in iShares Trust (IAU) --- currently worth a total of ~US\$1.7k

for the first person who presents a solution to this problem:

Find---or show non-existence of---a system independent (i.e. valid for all of CCS as defined above) interpolating function f for which two differing (iso-)electronic Hamiltonians transform such that

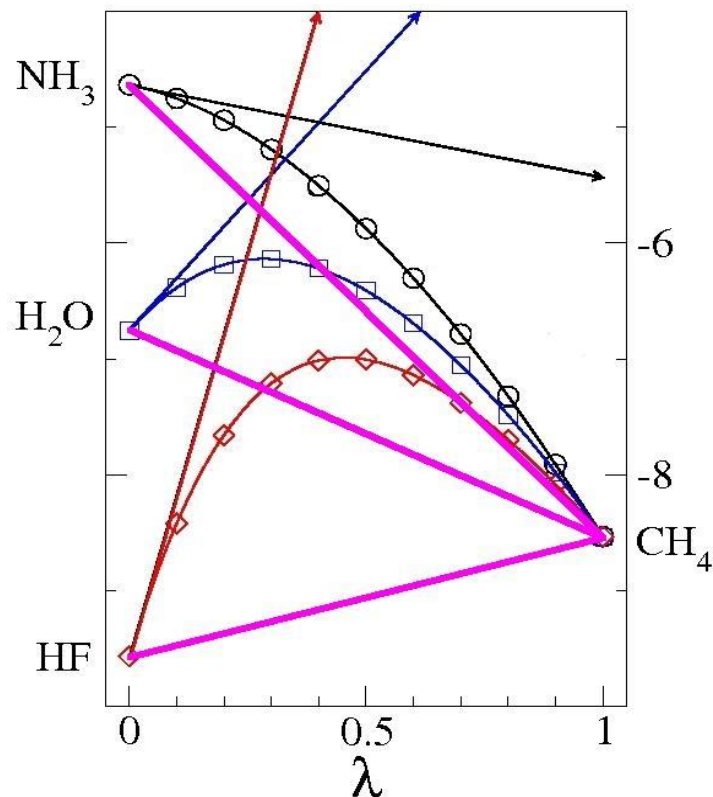
$$\begin{aligned} \left. \frac{\partial E(\lambda)}{\partial \lambda} \right|_{\lambda=0} &= \left\langle \frac{\partial H(f_{if}(\lambda))}{\partial \lambda} \right\rangle_{\lambda=0} \\ &= E_f - E_i \end{aligned}$$

where

$$0 \leq \lambda \leq 1$$

$$\begin{aligned} E(\lambda=0) &= \langle H(f(\lambda=0)) \rangle = \langle H_i \rangle = E_i \\ E(\lambda=1) &= \langle H(f(\lambda=1)) \rangle = \langle H_f \rangle = E_f \end{aligned}$$

ϵ [eV]



See

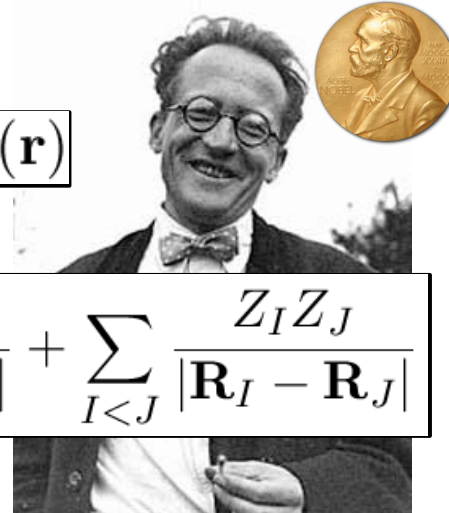
www.alcf.anl.gov/~anatole

For more info

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

variational (deductive)

Alchemy

1. Free energies
2. Gradients to optimize

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

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

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

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

Feynman



First Principles

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



Schrödinger

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

variational (deductive)

Feynman

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

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

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

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



correlational (inductive)

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

supervised
learning

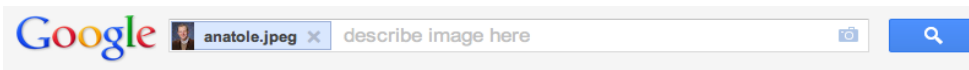
$$\{Z_I, \mathbf{R}_I\} \xrightarrow{\text{ML}} E$$



Vapnik



Correlational: Machine Learning



Search

About 4 results (0.35 seconds)

Everything

Images

Maps

Videos

News

Shopping

More



Image size:
160 × 213

No other sizes of this image found.

Pages that include matching images



[Argonne National Laboratories: Leadership Computing Facility ...](http://www.mcs.anl.gov/~anatole/)

www.mcs.anl.gov/~anatole/

Contact Dr. O. A. von Lilienfeld Assistant Computational Scientist at Leadership Computing Facility and Fellow at Computation Institute (UoC) Argonne National ...

160 × 213

Search by image

Visually similar

More sizes

Any time

Past hour

Past 24 hours

Past week

Past month

Past year

Custom range...



[Fellows | Computation Institute](http://www.ci.uchicago.edu/people/fellows.php)

www.ci.uchicago.edu/people/fellows.php

40+ items – ci. jobs |; contact us |; computing resources |; help desk ...

Igor Aronson Adjunct Professor

Gyorgy

Babnigg

Asst. Bioinformatics Spec/Biochemist Biosciences

Division

285 × 380



[People Directory | Argonne Leadership Computing Facility](http://www.alcf.anl.gov/staff-directory)

www.alcf.anl.gov/staff-directory

60+ items – The Argonne Leadership Computing Facility (ALCF) is a DOE ...

Yury Alekseev Assistant Computational Scientist, ALCF

Catalyst

630- 252 ...

Bill Allcock ALCF Director of Operations AIG

630-252-7573

500 × 545



[O. Anatole von Lilienfeld | Argonne Leadership Computing Facility](https://www.alcf.anl.gov/staff-directory/o-anatole-von-lilienfeld)

<https://www.alcf.anl.gov/staff-directory/o-anatole-von-lilienfeld>

The Argonne Leadership Computing Facility (ALCF) is a DOE leadership computing facility. The ALCF provides the computational science community with a ...

500 × 545

Visually similar images - Report images



Correlational: Machine Learning

Google

Search

Everything

Images

Maps

Videos

News

Shopping

More

Search by image

Visually similar

More sizes

Any time

Past hour

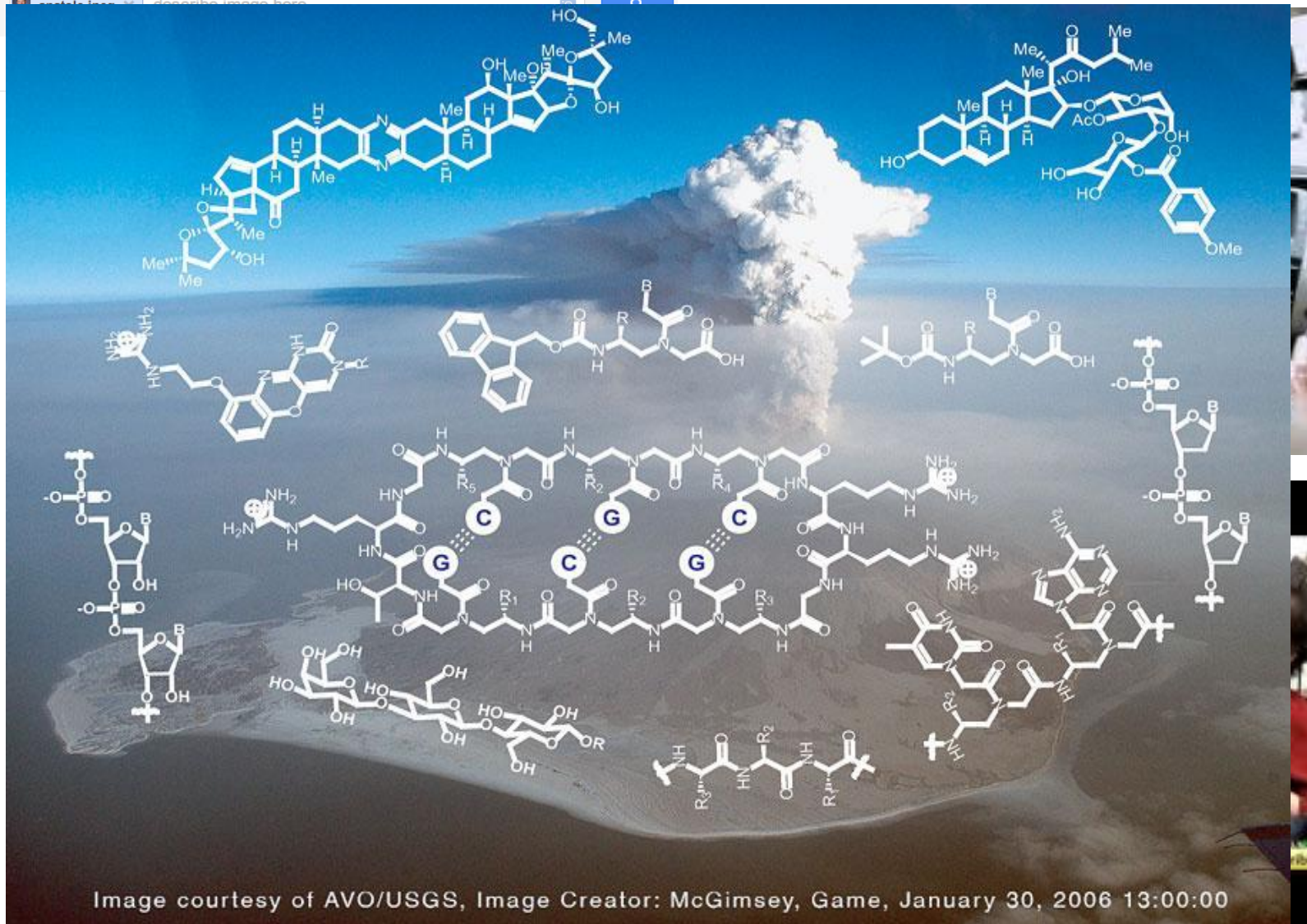
Past 24 hours

Past week

Past month

Past year

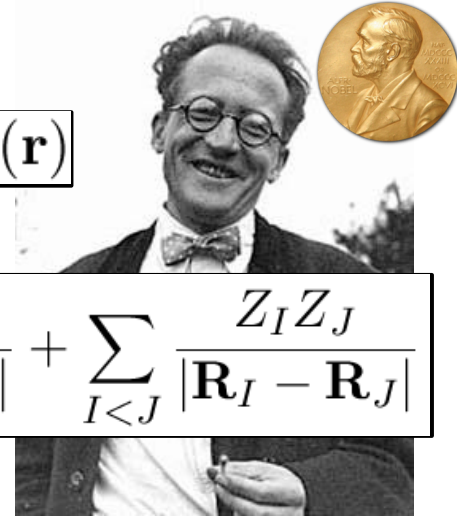
Custom range...



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

correlational (inductive)

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

supervised
learning

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

Infer solution by comparison
to previous examples

- Regression method?
- Function?
- Variables?
- Metric?
- Data?

Vapnik



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

Non-linear function

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



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

Non-linear function

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

Desirable descriptors are

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

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



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

Non-linear function

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

Desirable descriptors are

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

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

Coulomb-matrix

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

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



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

Non-linear function

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

Desirable descriptors are

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

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

Coulomb-matrix

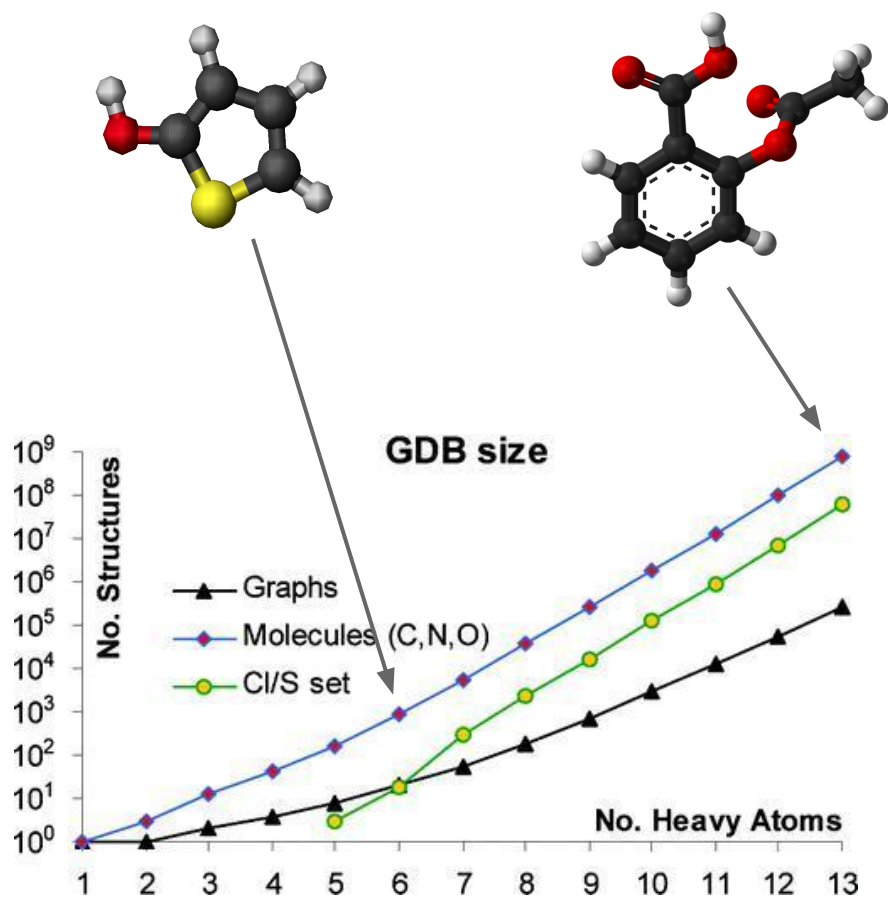
- unique
- translation
- rotation
- symmetry
- sort/diagonalize
- fill up w zeros

Euclidean distance

$$d(\mathbf{M}, \mathbf{M}_i) = \sqrt{\sum_{IJ} |M_{IJ} - M_{IJ}^{(i)}|^2}$$

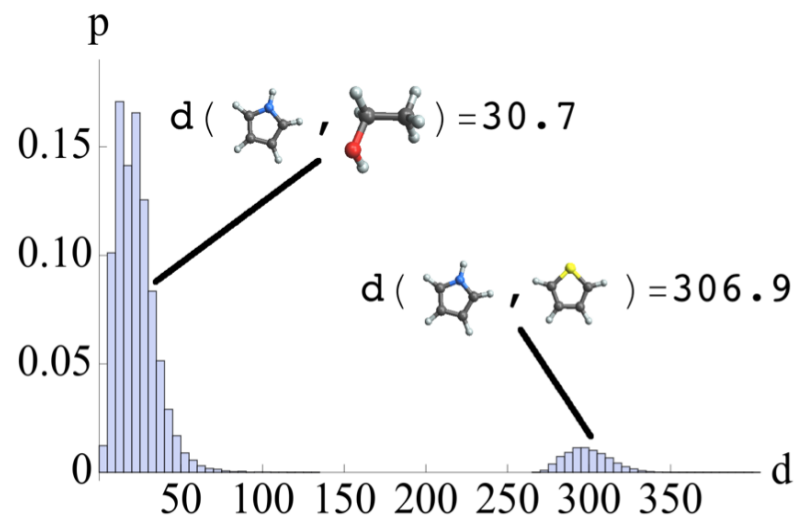


GDB: All organic molecules up to 13 atoms

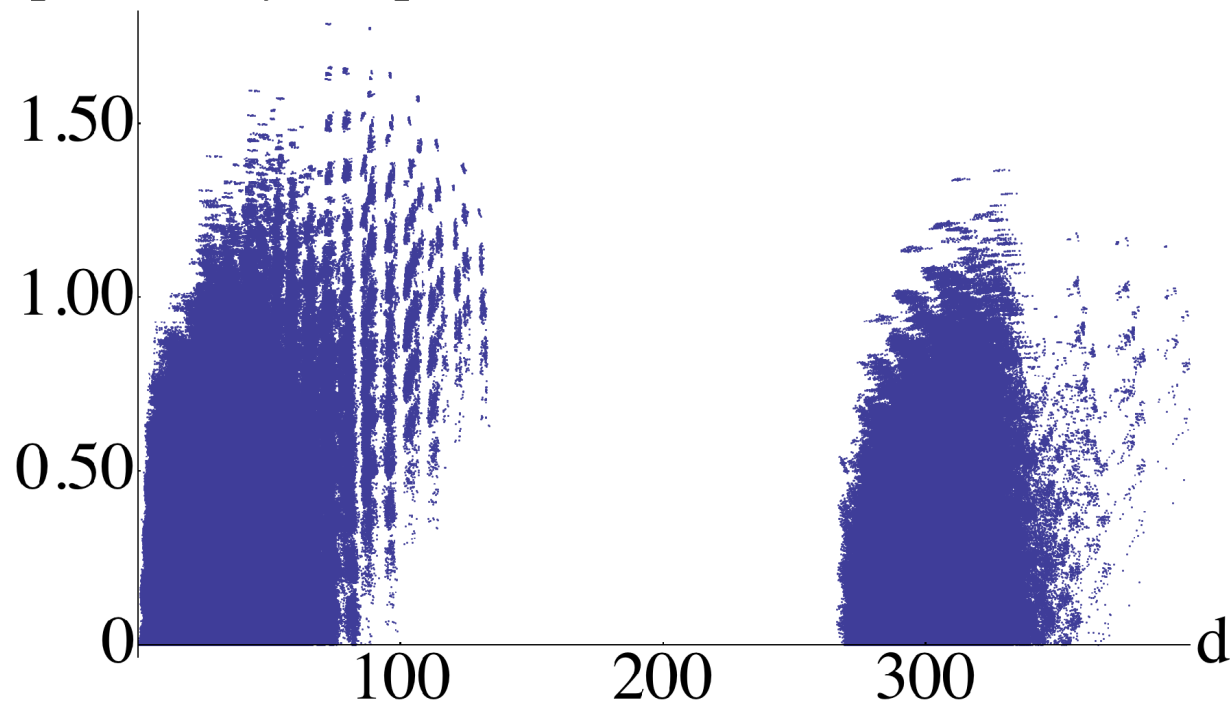


Fink, Bruggesser, Reymond *ACIE* (2005),
Blum, Reymond *JACS* (2009)

1. 7k compositional & constitutional isomers
2. Initial coordinates from universal force field [Goddard et al *JACS* (1992)]
3. Relax geometry with DFT
4. Calculate atomization energies



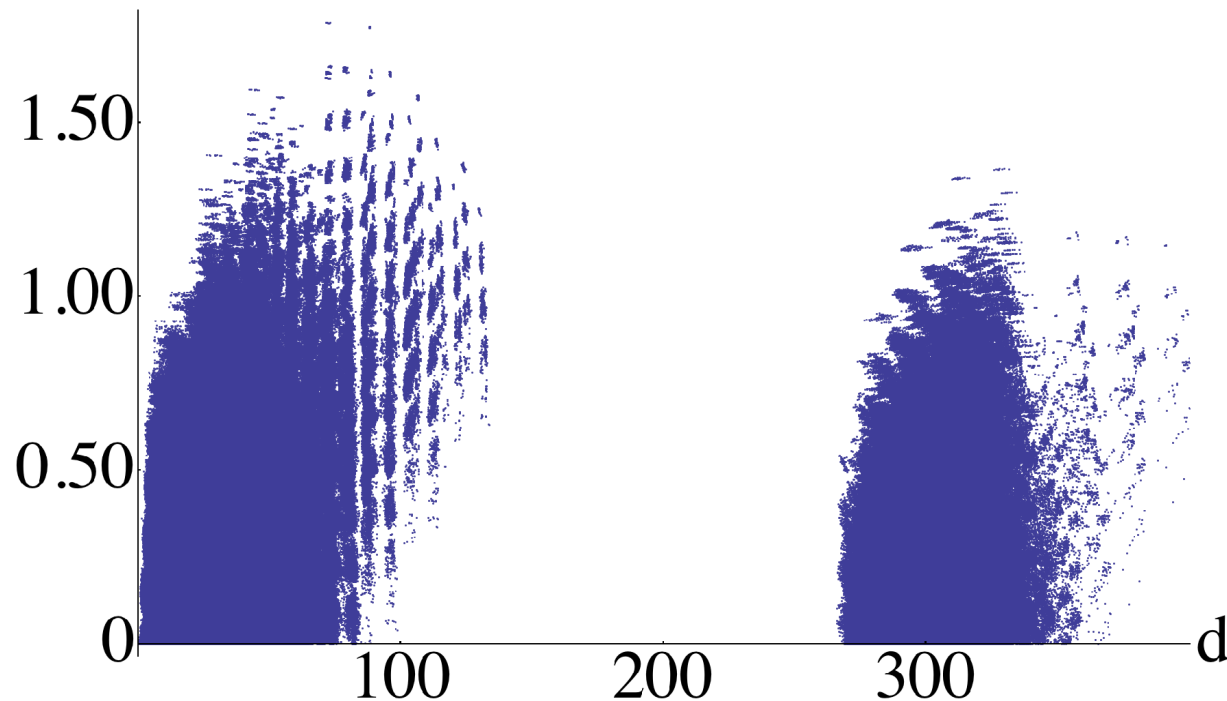
$\Delta E^{\text{ref}} [10^3 \text{ kcal/mol}]$



$$\min_{\alpha} \sum_i (E^{\text{est}}(\mathbf{M}_i) - E_i^{\text{ref}})^2 + \lambda \sum_i \alpha_i^2$$



$\Delta E^{\text{ref}} [10^3 \text{ kcal/mol}]$



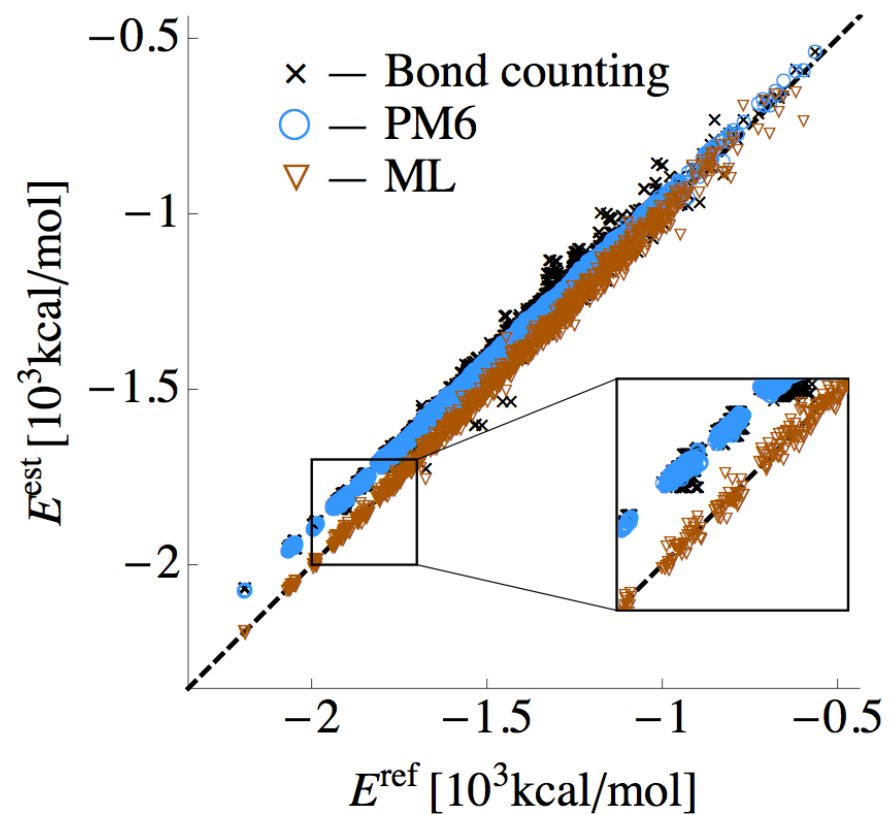
$$\min_{\alpha} \sum_i (E^{\text{est}}(\mathbf{M}_i) - E_i^{\text{ref}})^2 + \lambda \sum_i \alpha_i^2$$

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

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

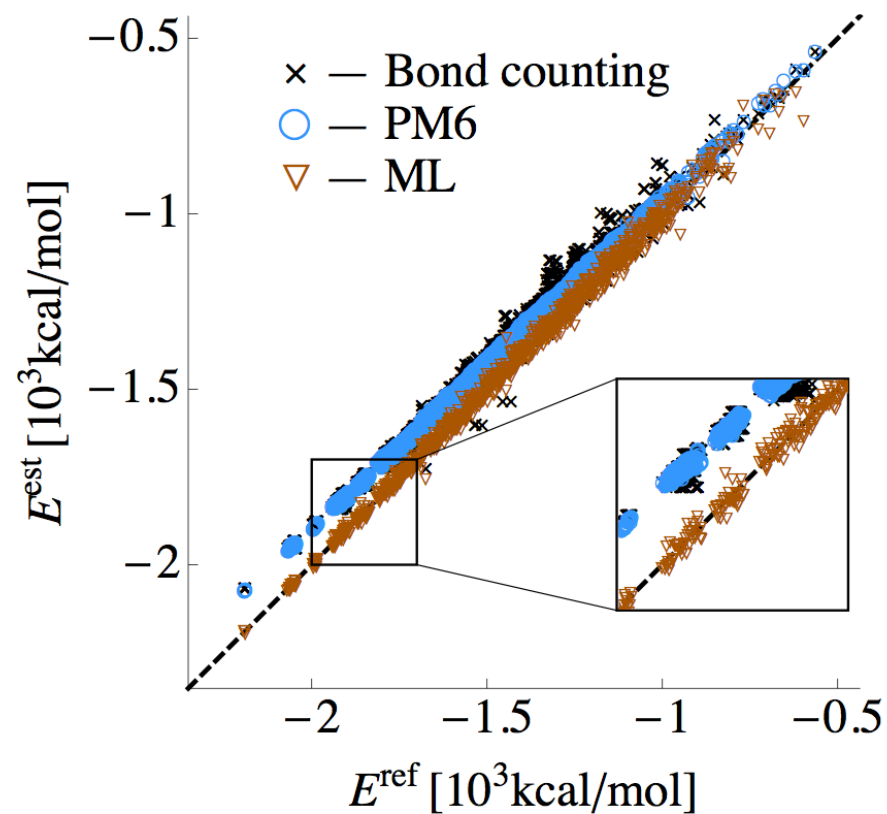
$$E^{\text{est}}(\mathbf{M}) = \sum_i \alpha_i e^{-\frac{d(\mathbf{M}, \mathbf{M}_i)^2}{2\sigma^2}}$$



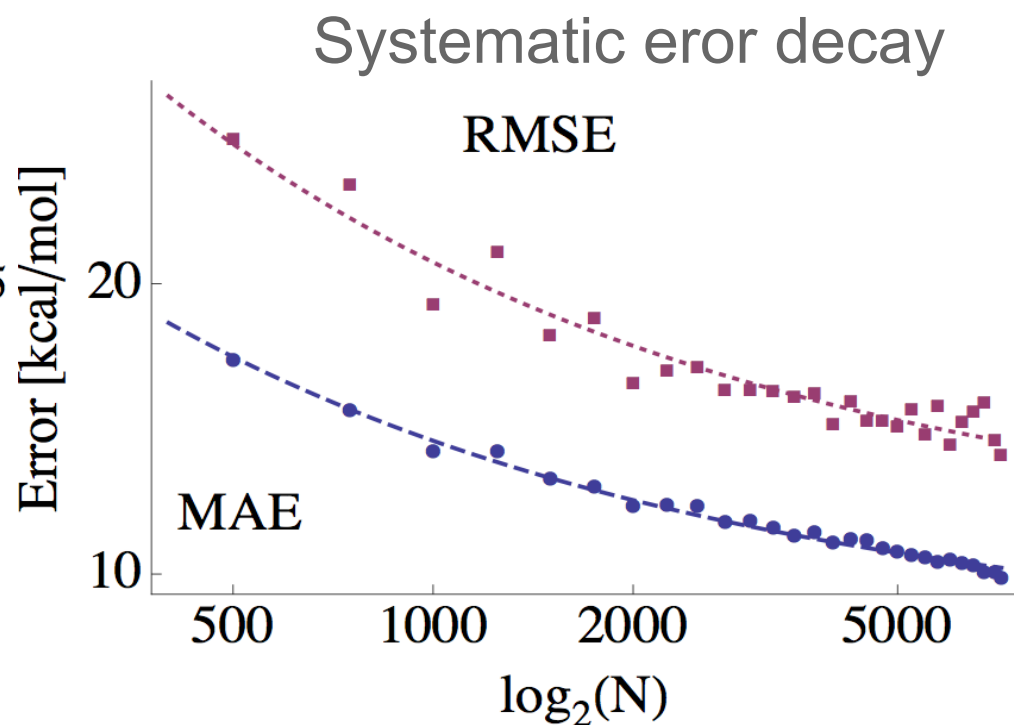


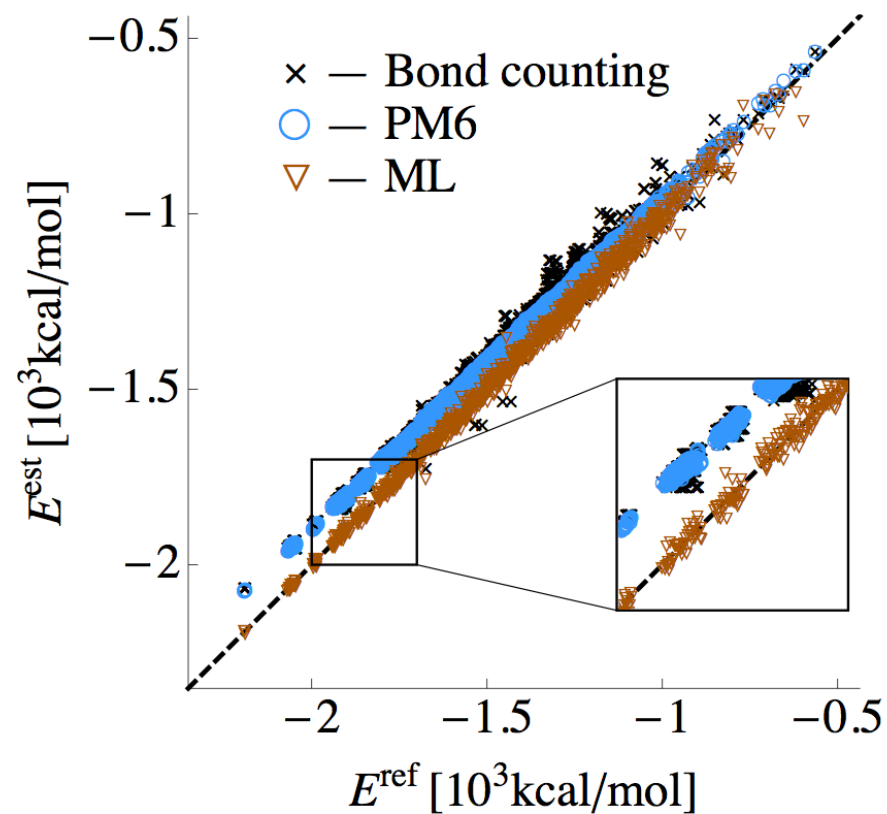
Training for $N = 1000$
molecules
MAE $\sim 15 \text{ kcal/mol}$





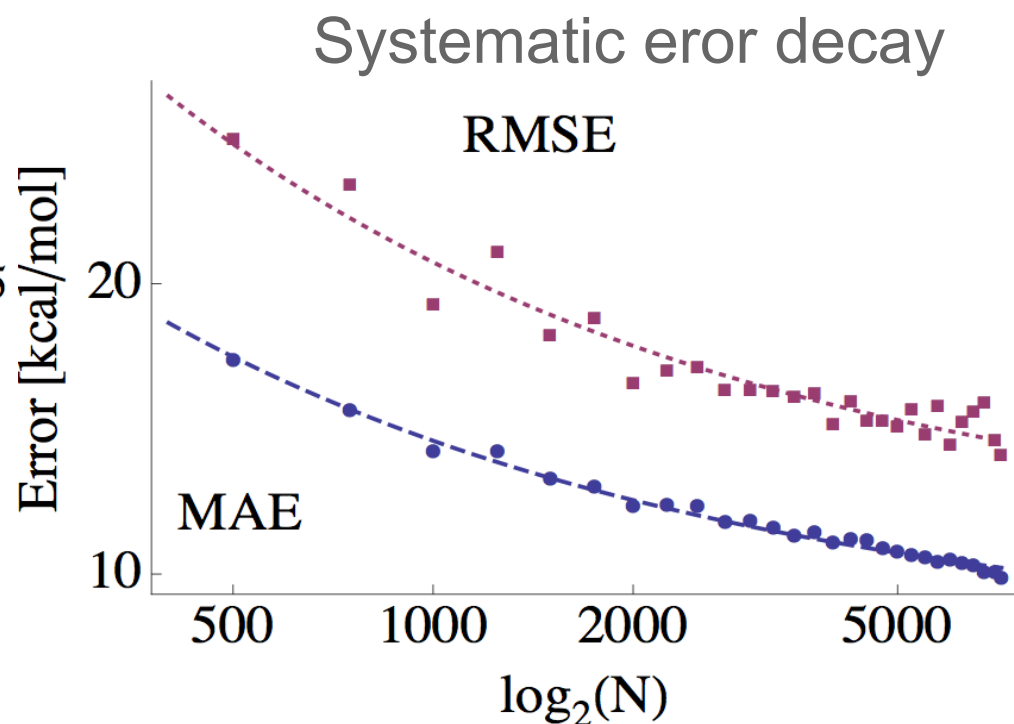
Training for $N = 1000$
molecules
MAE ~ 15 kcal/mol



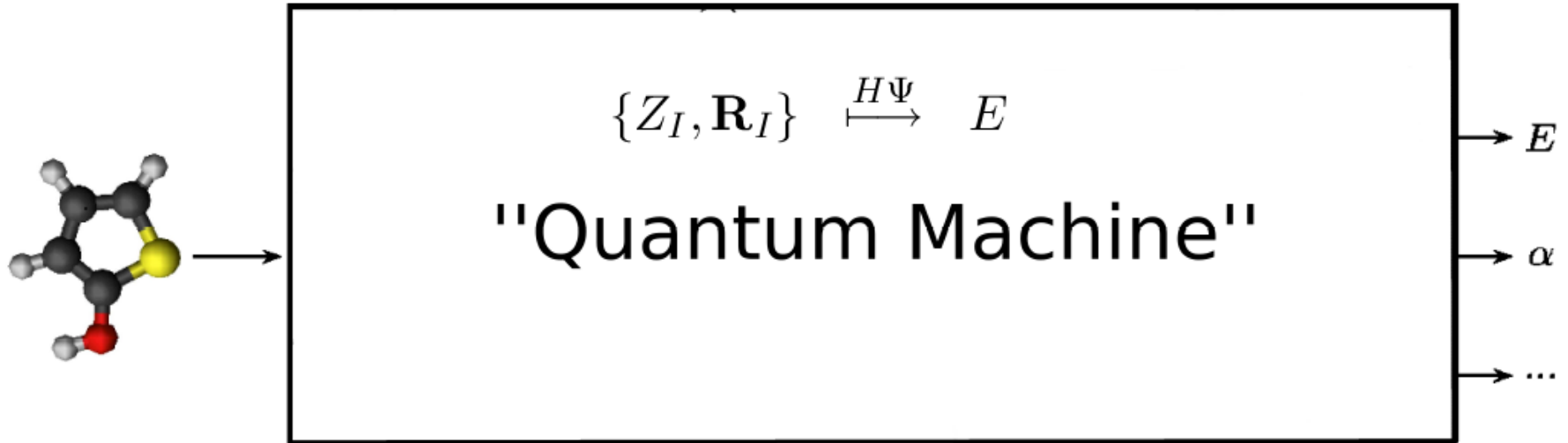


Training for N = 1000
molecules
MAE ~ 15 kcal/mol

PBE0: ~ 1000 seconds
ML: \sim milli seconds



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



Tkatchenko (FHI)



Rupp (ETHZ)



Müller (TU Berlin)

M. Rupp, A. Tkatchenko, K.-R. Müller, OAvL, *Phys Rev Lett* (2012)



Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning

Matthias Rupp,^{1,2} Alexandre Tkatchenko,^{3,2} Klaus-Robert Müller,^{1,2} and O. Anatole von Lilienfeld^{4,2,*}

¹*Machine Learning Group, Technical University of Berlin, Franklinstr 28/29, 10587 Berlin, Germany*

²*Institute of Pure and Applied Mathematics, University of California Los Angeles, Los Angeles, California 90095, USA*

³*Fritz-Haber-Institut der Max-Planck-Gesellschaft, 14195 Berlin, Germany*

⁴*Argonne Leadership Computing Facility, Argonne National Laboratory, Argonne, Illinois 60439, USA*

(Received 15 June 2011; published 31 January 2012)

We introduce a machine learning model to predict atomization energies of a diverse set of organic molecules, based on nuclear charges and atomic positions only. The problem of solving the molecular Schrödinger equation is mapped onto a nonlinear statistical regression problem of reduced complexity. Regression models are trained on and compared to atomization energies computed with hybrid density-functional theory. Cross validation over more than seven thousand organic molecules yields a mean absolute error of ~ 10 kcal/mol. Applicability is demonstrated for the prediction of molecular atomization potential energy curves.



Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning

Matthias Rupp,^{1,2} Alexandre Tkatchenko,^{3,2} Klaus-Robert Müller,^{1,2} and O. Anatole von Lilienfeld^{4,2,*}

¹Machine Learning Group, Technical University of Berlin, Franklinstr 28/29, 10587 Berlin, Germany

Comment on “Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning”

In a recent Letter [1], the authors construct a machine learning (ML) model of molecular atomization energies, which they compare to bond counting (BC) and the PM6 semiempirical method [2]. However, their ML model was trained and tested on density functional theory (DFT)

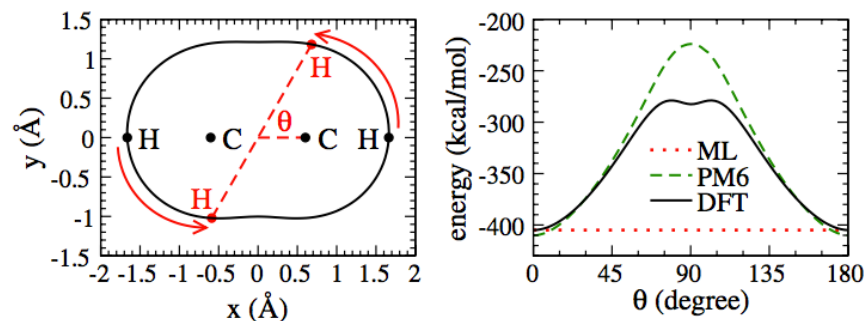


FIG. 2 (color online). A continuous deformation of acetylene. (left) Hydrogen atoms follow the closed curve with the line connecting them fixed to the origin. Carbon atoms remain near their equilibrium positions. (right) Atomization energy as a function of the H-origin-C angle.

Jonathan E. Moussa*

Sandia National Laboratories

Albuquerque, New Mexico 87185, USA

the full from
distributions

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

N = 4

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

Coulomb-matrix

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

???



Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning

Matthias Rupp,^{1,2} Alexandre Tkatchenko,^{3,2} Klaus-Robert Müller,^{1,2} and O. Anatole von Lilienfeld^{4,2,*}

¹Machine Learning Group, Technical University of Berlin, Franklinstr 28/29, 10587 Berlin, Germany

PHYSICAL REVIEW LETTERS

PRL 109, 059801 (2012)

week ending
3 AUGUST 2012

Comment on “Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning”

In a recent Letter [1], the authors construct a machine learning (ML) model of molecular atomization energies, which they compare to bond counting (BC) and the PM6 semiempirical method [2]. However, their ML model was trained and tested on density functional theory (DFT)

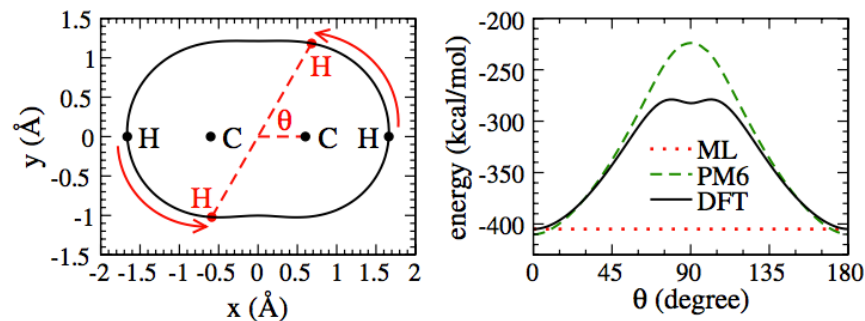


FIG. 2 (color online). A continuous deformation of acetylene. (left) Hydrogen atoms follow the closed curve with the line connecting them fixed to the origin. Carbon atoms remain near their equilibrium positions. (right) Atomization energy as a function of the H-origin-C angle.

Jonathan E. Moussa*

Sandia National Laboratories

Albuquerque, New Mexico 87185, USA

the full from
distributions

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

N = 4

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

Coulomb-matrix

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



Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning

Matthias Rupp,^{1,2} Alexandre Tkatchenko,^{3,2} Klaus-Robert Müller,^{1,2} and O. Anatole von Lilienfeld^{4,2,*}

¹Machine Learning Group, Technical University of Berlin, Franklinstr 28/29, 10587 Berlin, Germany

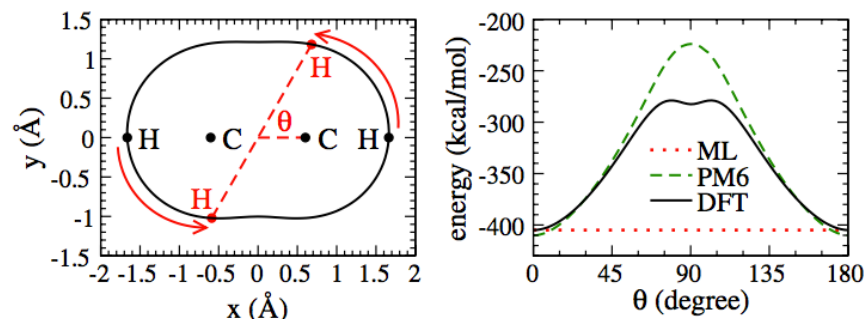
PRL 109, 059801 (2012)

PHYSICAL REVIEW LETTERS

week ending
3 AUGUST 2012

Comment on “Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning”

In a recent Letter [1], the authors construct a machine learning (ML) model of molecular atomization energies, which they compare to bond counting (BC) and the PM6 semiempirical method [2]. However, their ML model was trained and tested on d



PRL 109, 059802 (2012)

PHYSICAL REVIEW LETTERS

week ending
3 AUGUST 2012

Jonathan E. Mous
Sandia National
Albuquerque, Ne
distributions

$$M_{IJ} =$$

$$N = 4$$

$$\rightarrow 3 \cdot N - 6$$

Rupp et al. Reply: In his Comment [1], J.E. Moussa (JEM) raises concerns regarding the accuracy of our recently published Machine Learning (ML) model [2]. Our performance estimates, based on cross-validated Kernel Ridge Regression, amount to less than 10 kcal/mol mean absolute error (MAE) with respect to DFT-PBE0 [3,4] predictions of atomization energies, using a training set of more than 7000 small organic molecules from the GDB-13 data set [5]. As such, the ML model achieves an accuracy similar to generalized gradient DFT, and significantly exceeds that of Hartree-Fock or local density approximated DFT [6].

In our Letter we presented numerical evidence that ML models can be built using (i) sufficient examples and (ii) a molecular representation based on Cartesian coordinates and elemental composition *without* explicitly accounting for the electronic degrees of freedom. Therefore, performance of our

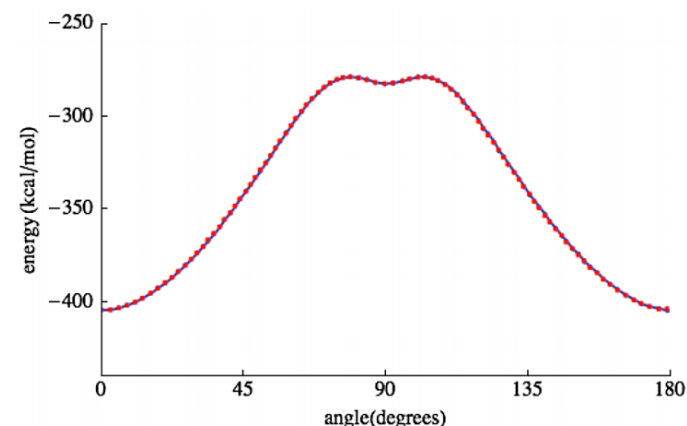


FIG. 1 (color online). Blue line: PBE0. Red dots: ML model using Frobenius norm of, and trained on, Coulomb matrices of geometries corresponding to JEM’s example.

Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning

Matthias Rupp,^{1,2} Alexandre Tkatchenko,^{3,2} Klaus-Robert Müller,^{1,2} and O. Anatole von Lilienfeld^{4,2,*}

¹Machine Learning Group, Technical University of Berlin, Franklinstr 28/29, 10587 Berlin, Germany

PRL 109, 059801 (2012)

PHYSICAL REVIEW LETTERS

week ending
3 AUGUST 2012

Comment on “Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning”

In a recent Letter [1], the authors construct a machine learning (ML) model of molecular atomization energies, which they compare to bond counting (BC) and the PM6 semiempirical method [2]. However, their ML model was trained and tested on density functional theory (DFT)

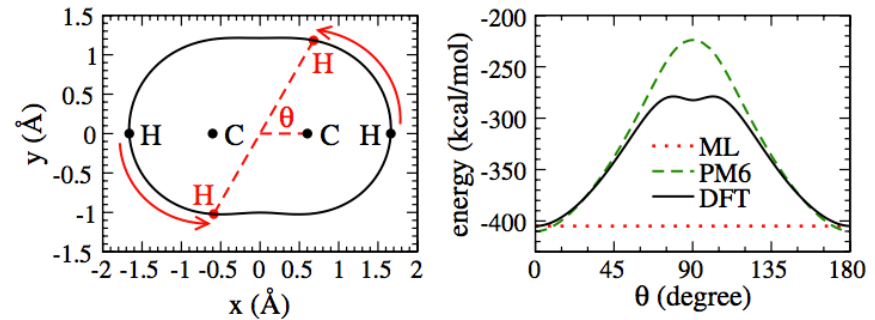


FIG. 2 (color online). A continuous deformation of acetylene. (left) Hydrogen atoms follow the closed curve with the line connecting them fixed to the origin. Carbon atoms remain near their equilibrium positions. (right) Atomization energy as a function of the H-origin-C angle.

Jonathan E. Moussa*

Sandia National Laboratories

Albuquerque, New Mexico 87185, USA

the full range
distributions

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

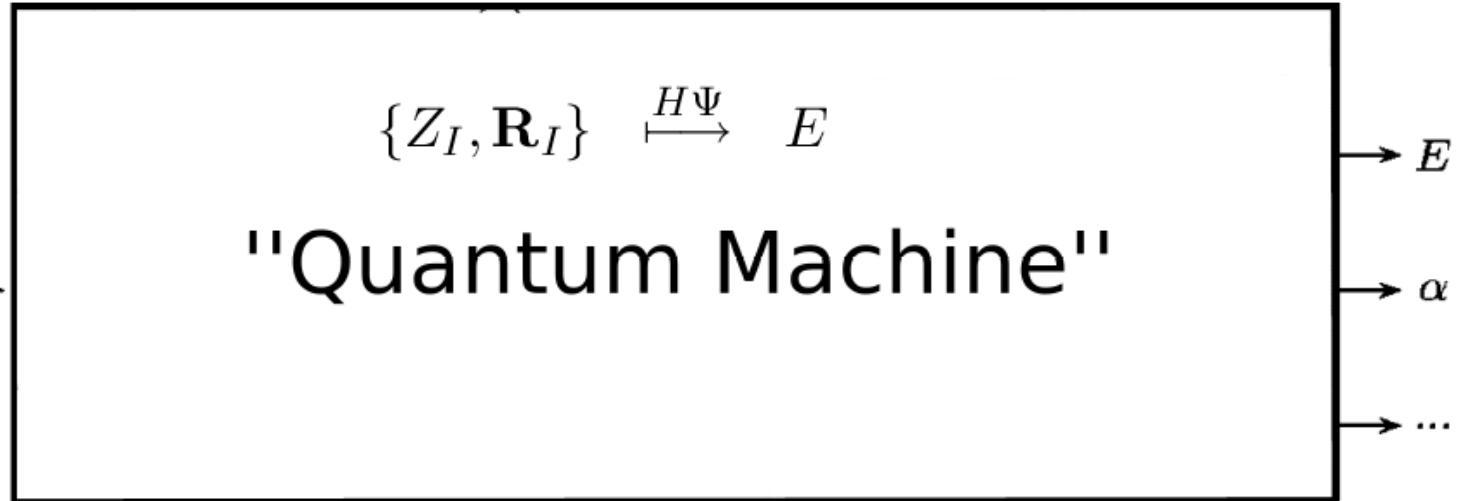
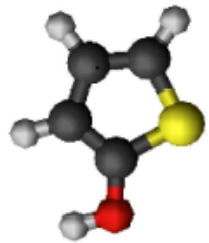
N = 4

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

Coulomb-matrix

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





Tkatchenko (FHI)



Hansen (FHI)



Rupp (ETHZ)



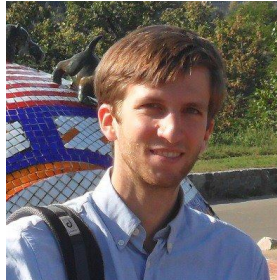
Vazquez (ANL)



Gobre (FHI)



Montavon (TU Berlin)



Müller (TU Berlin)



M. Rupp, A. Tkatchenko, K.-R. Müller, OAvL, *Phys Rev Lett* (2012);

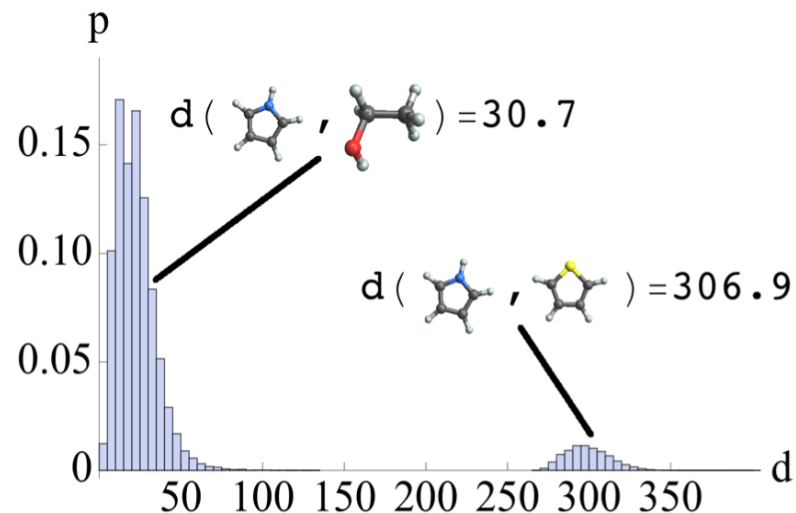
G. Montavon, M. Rupp, V. Gobre, A. Vazquez, K. Hansen, A. Tkatchenko, K.-R. Müller, OAvL, *NJP* accepted (2013); Montavon et al *NIPS proceedings* (2013)

GDB: All organic molecules up to 13 atoms

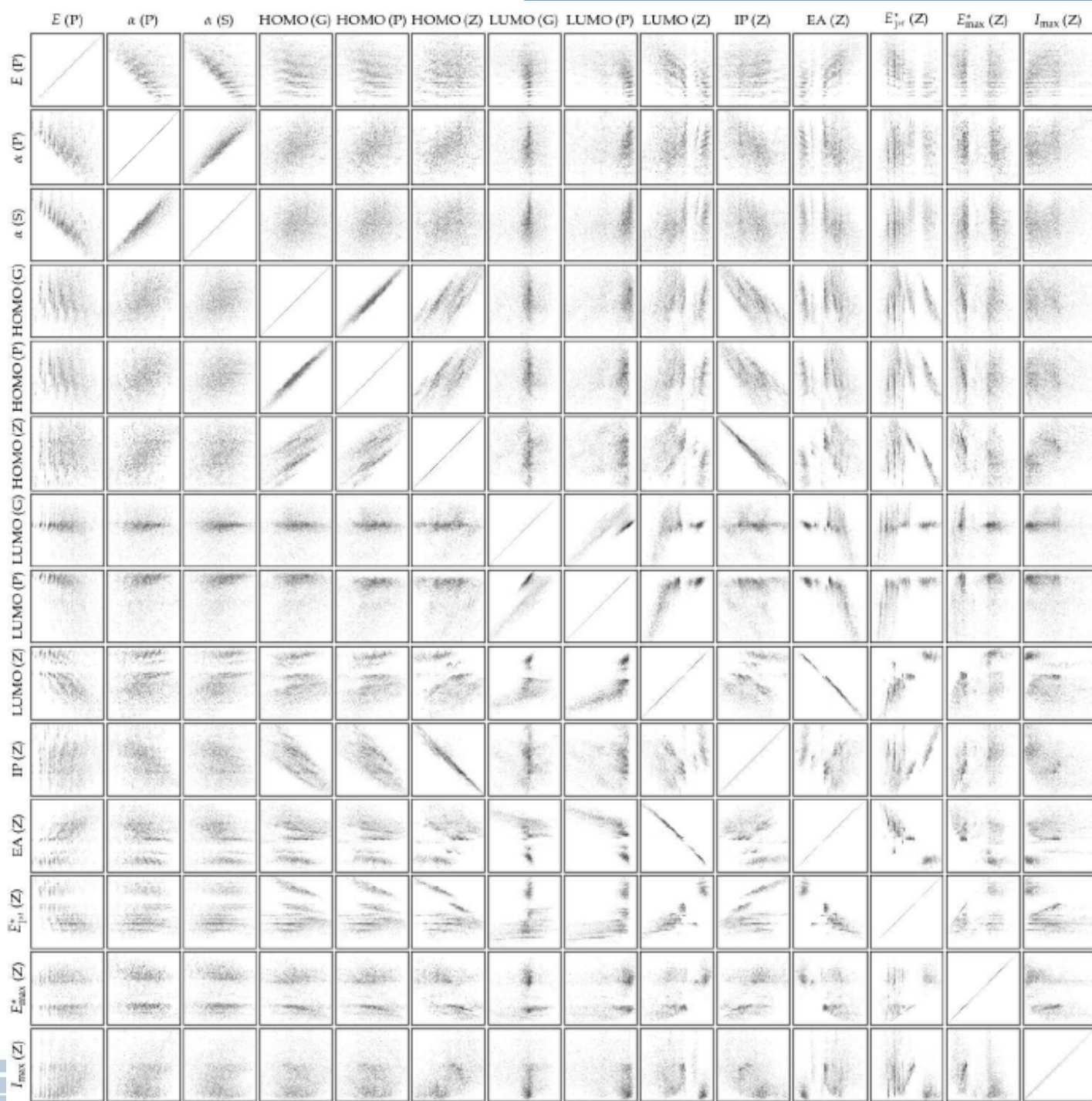
Calculate 14 properties:

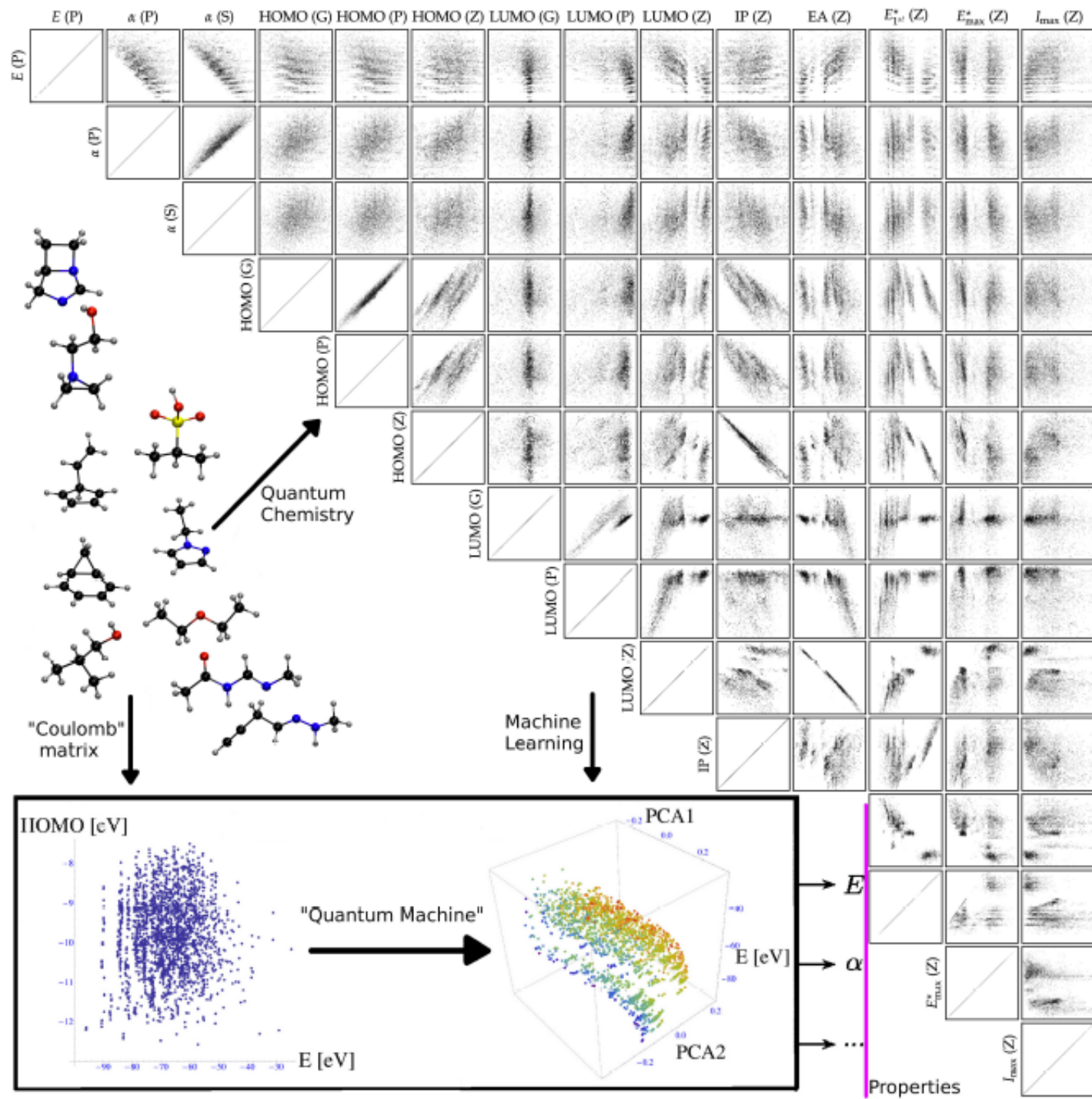
- atomization energy (PBE0)
- 2 x polarizability (PBE0/SCS)
- 6 x HOMO/LUMO (GW/PBE0/ZINDO)
- 2 x IP/EA (ZINDO)
- 3 x Excitations (ZINDO)

- 7k compositional & constitutional isomers
- Initial coordinates from universal force field [Goddard et al JACS (1992)]
- Relaxed geometry with DFT

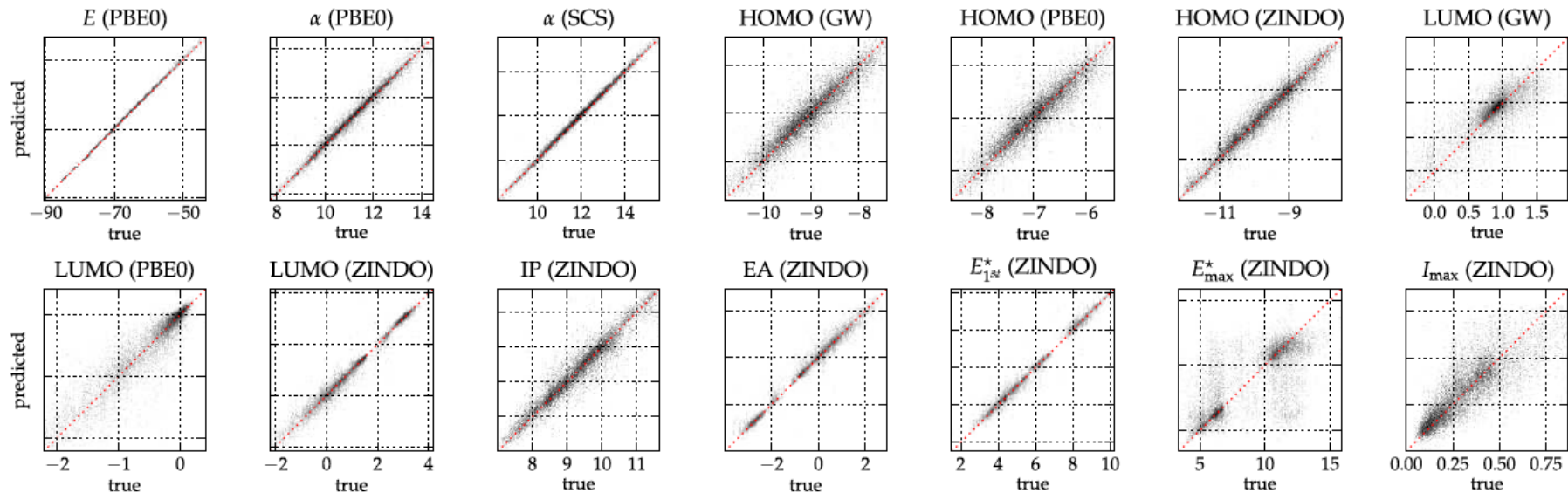


The data

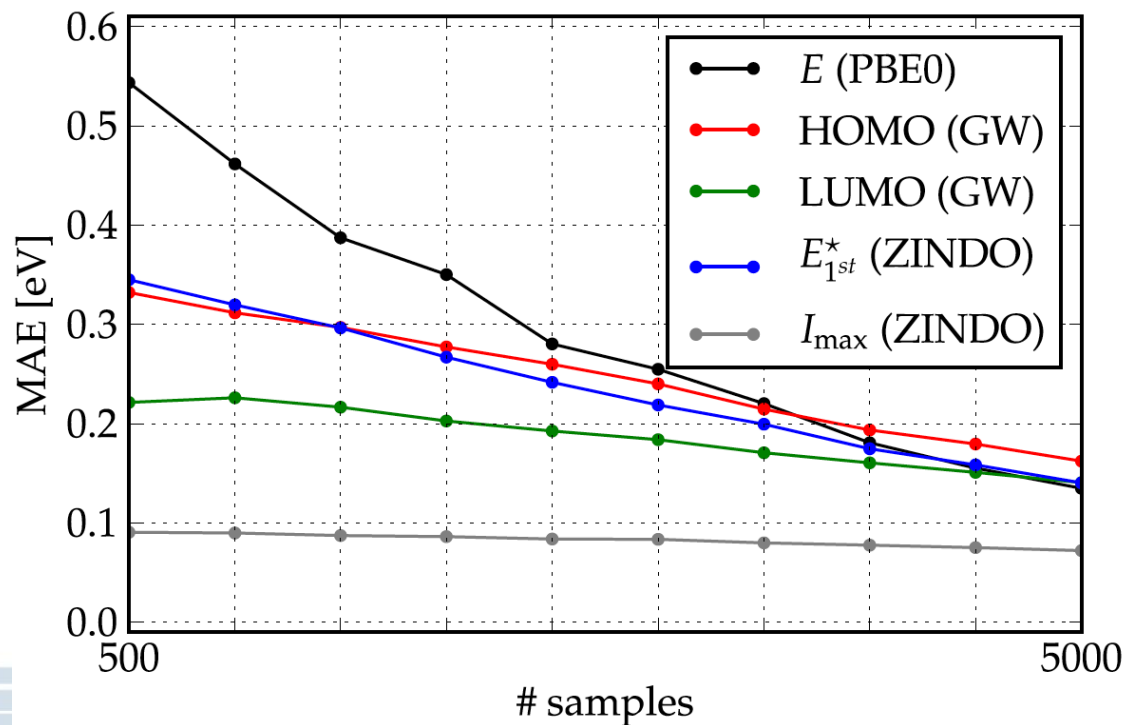
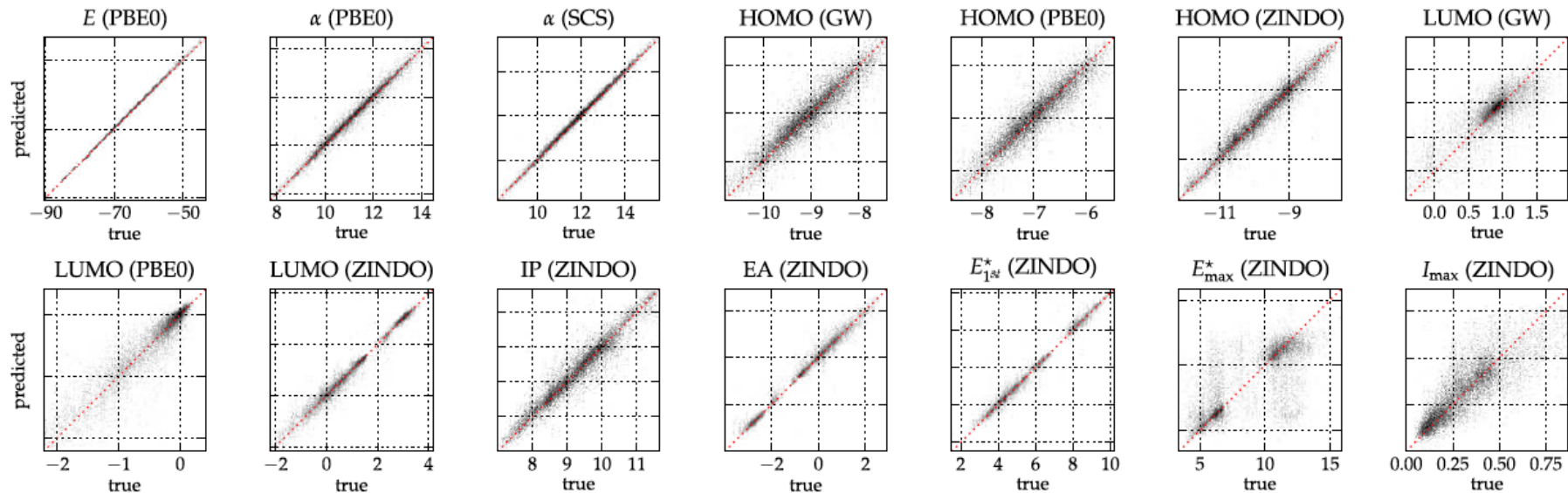




<http://www.quantum-machine.org/>



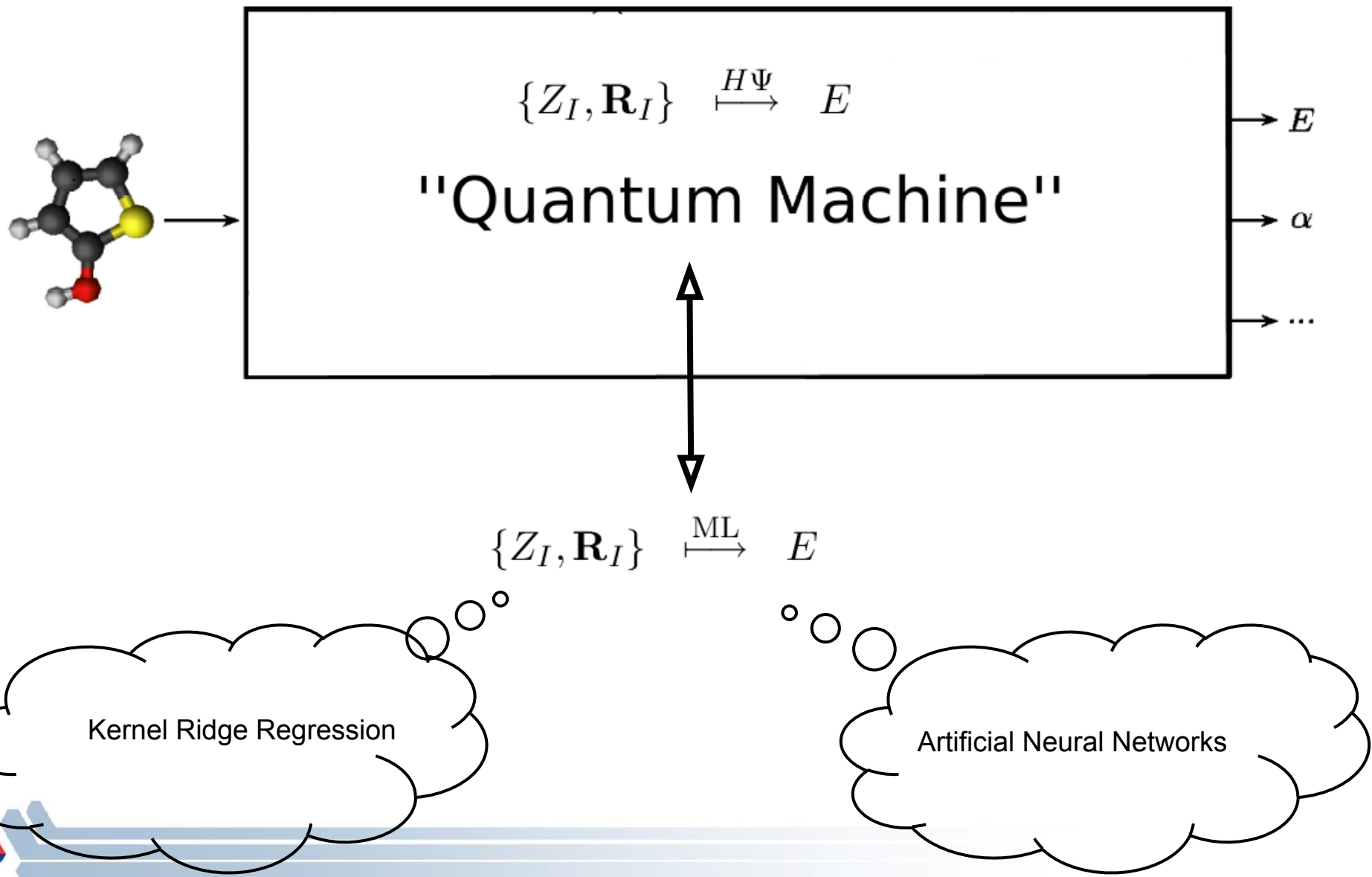
<http://www.quantum-machine.org/>



Property [eV, Å ³]	Mean	MAE	Reference MAE
E (PBE0)	-67.79	0.16	0.15 ^a , 0.23 ^b , 0.09 - 0.22 ^c
α (PBE0)	11.11	0.11	0.05-0.27 ^d , 0.04-0.14 ^e
α (SCS)	11.87	0.07	0.05-0.27 ^f , 0.04-0.14 ^g
HOMO (GW)	-9.09	0.16	-
HOMO (PBE0)	-7.01	0.15	2.08 ^h
HOMO (ZINDO)	-9.81	0.16	0.79 ^h
LUMO (GW)	0.78	0.14	-
LUMO (PBE0)	-0.52	0.12	1.30 ^h
LUMO (ZINDO)	1.05	0.11	0.93 ^h
IP (ZINDO)	9.27	0.18	0.20 ⁱ , 0.15 ^j
EA (ZINDO)	0.55	0.12	0.16 ^k , 0.11 ^l
E_{1st}^* (ZINDO)	5.58	0.13	0.18 ^m , 0.21 ⁿ
E_{max}^* (ZINDO)	8.82	1.07	-
I_{max} (ZINDO)	0.33	0.07	-

G. Montavon, M. Rupp, V. Gobre, A. Vazquez, K. Hansen, A. Tkatchenko, K.-R. Müller, OAvL, *NJP* accepted (2013)







Sumpter, Noid: Potential energy surfaces for macromolecules. A neural network technique *Chem Phys Lett* (1992)

Lorenz, Gross, Scheffler (FHI): Representing high-dimensional potential-energy surfaces for reactions at surfaces by neural networks *Chem Phys Lett* (2004)



Manzhos, T. Carrington (Montreal): Using neural networks to represent potential surfaces as sums of products, *J Chem Phys* (2006)



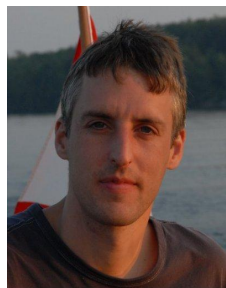
Parrinello, Behler (Bochum): Generalized neural-network representation of high-dimensional potential energy surfaces, *Phys Rev Lett* (2010)



Csanyi (Cambridge): Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons, *Phys Rev Lett* (2010)

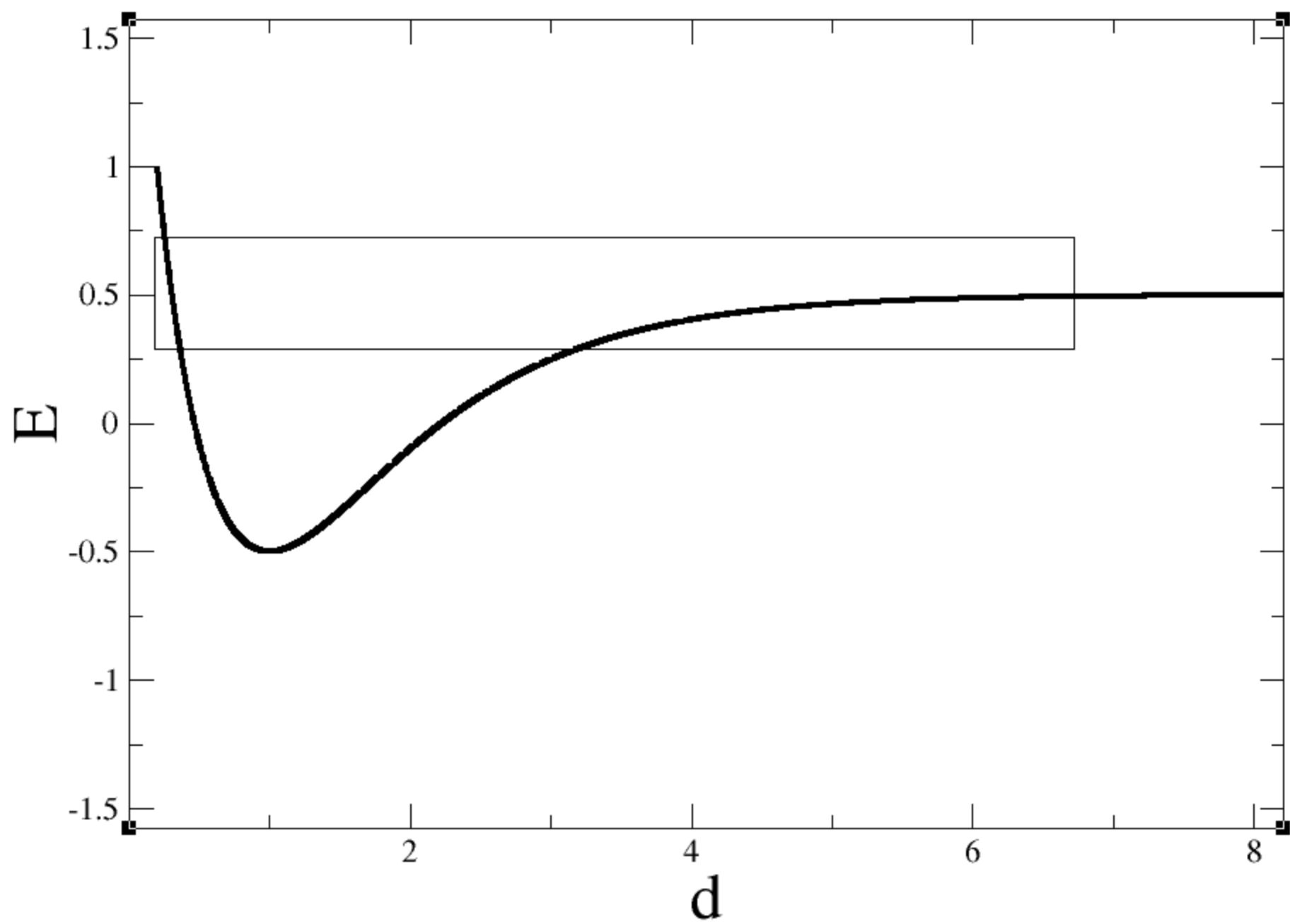


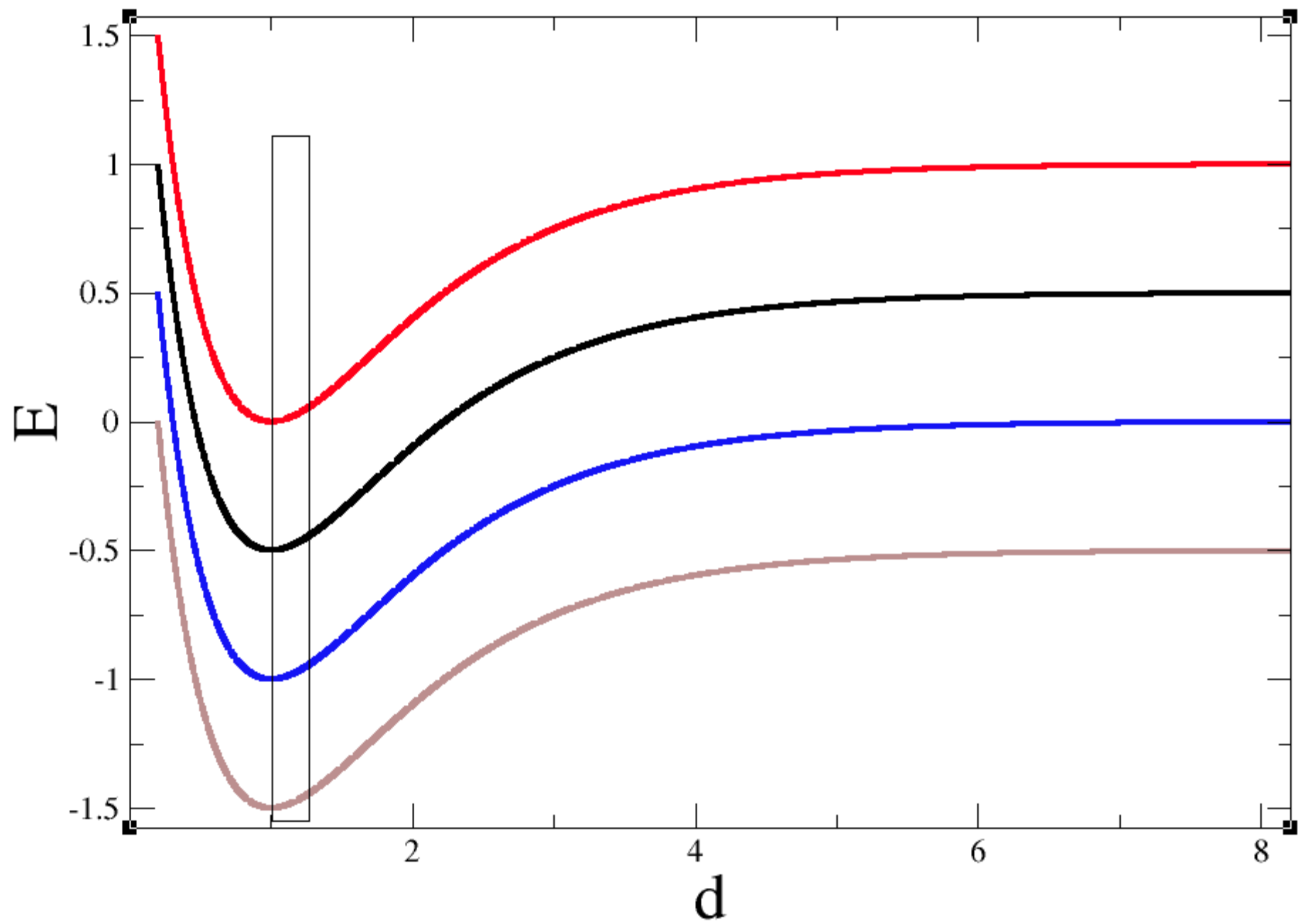
Henkelman (UT Austin): Optimizing transition states via kernel based machine learning, *J Chem Phys* (2012)



Burke (UC Irvine): Finding density functionals with machine learning, *Phys Rev Lett* (2012)



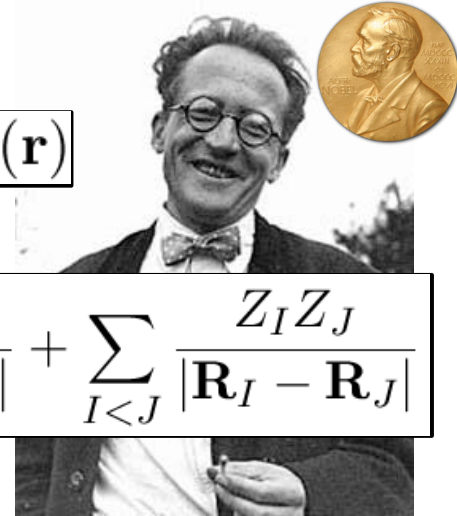




First Principles

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

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



Schrödinger

correlational (inductive)

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

supervised
learning

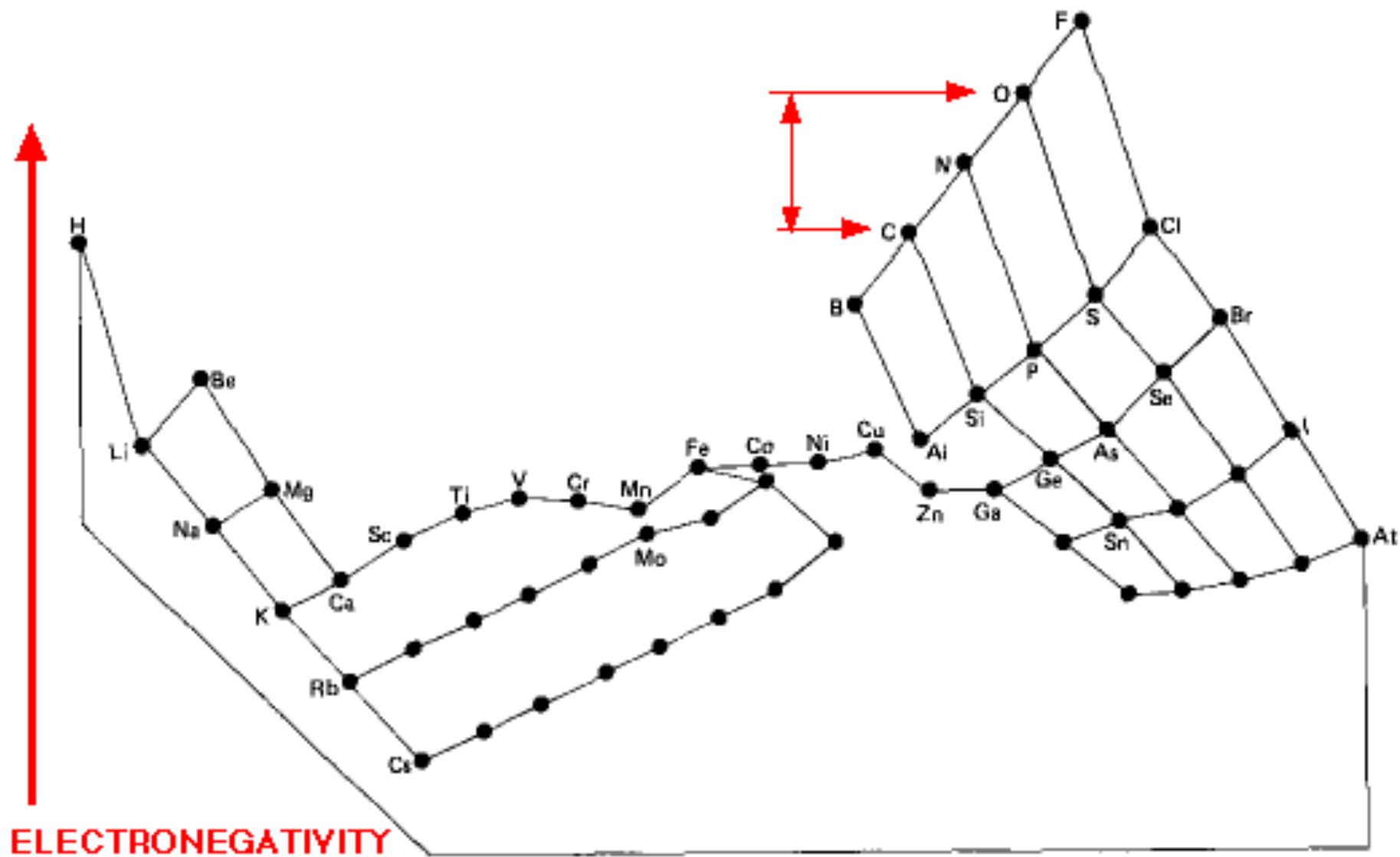
$$\{Z_I, \mathbf{R}_I\} \xrightarrow{\text{ML}} E$$

Infer solution by comparison
to previous examples

- Regression method?
- Function?
- **Variables?**
- Metric?
- Data?

Vapnik





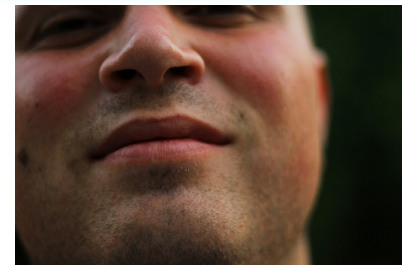
Property	Zxyz	CM	Eig(CM)
Unique	✓	✓	¬
First principles	✓	✓	✓
Transl. invariant	¬	✓	✓
Rotat. invariant	¬	✓	✓
Permutat. invariant	¬	¬	✓
Symmetry	¬	✓	✓
Size extensive	✓	✓	✓
Complete/global	✓	✓	¬
Dimensionality	$4N$	$(N^2 + N)/2$	N
Analytical	✓	✓	✓
Differentiable	N.A.	✓	✓
Uniform length	¬	¬	¬
Variable ranges	✓	✓	✓



Descriptor

$$P(\mathbf{r}) = \sum_I Z_I e^{-a|\mathbf{r} - \mathbf{R}_I|^2}$$

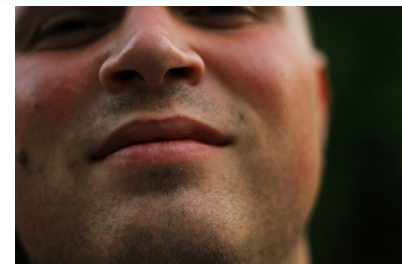
$$\mathcal{F}(P) = \frac{1}{(2a)^{3/2}} e^{\frac{\omega^2}{4a}} \sum_I Z_I e^{i\omega^T \mathbf{R}_I}$$



Aaron Knoll
(TACC)



Descriptor



Aaron Knoll
(TACC)

$$P(\mathbf{r}) = \sum_I Z_I e^{-a|\mathbf{r}-\mathbf{R}_I|^2}$$

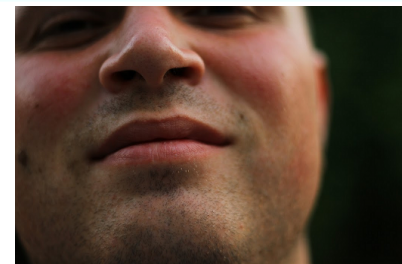
$$\mathcal{F}(P) = \frac{1}{(2a)^{3/2}} e^{\frac{\omega^2}{4a}} \sum_I Z_I e^{i\omega^T \mathbf{R}_I}$$

$$\mathcal{F} \mathcal{F}^* = \frac{1}{(2a)^3} e^{\frac{\omega^2}{2a}} \sum_J \sum_I Z_I Z_J \cos[\omega^T (\mathbf{R}_I - \mathbf{R}_J)]$$

$$M_{IJ} = Z_I Z_J \cos[\omega^T (\mathbf{R}_I - \mathbf{R}_J)]$$



Descriptor



Aaron Knoll
(TACC)

$$P(\mathbf{r}) = \sum_I Z_I e^{-a|\mathbf{r}-\mathbf{R}_I|^2}$$

$$\mathcal{F}(P) = \frac{1}{(2a)^{3/2}} e^{\frac{\omega^2}{4a}} \sum_I Z_I e^{i\omega^T \mathbf{R}_I}$$

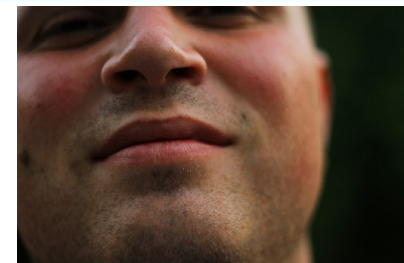
$$\mathcal{F} \mathcal{F}^* = \frac{1}{(2a)^3} e^{\frac{\omega^2}{2a}} \sum_J \sum_I Z_I Z_J \cos[\omega^T (\mathbf{R}_I - \mathbf{R}_J)]$$

$$M_{IJ} = Z_I Z_J \cos[\omega^T (\mathbf{R}_I - \mathbf{R}_J)]$$

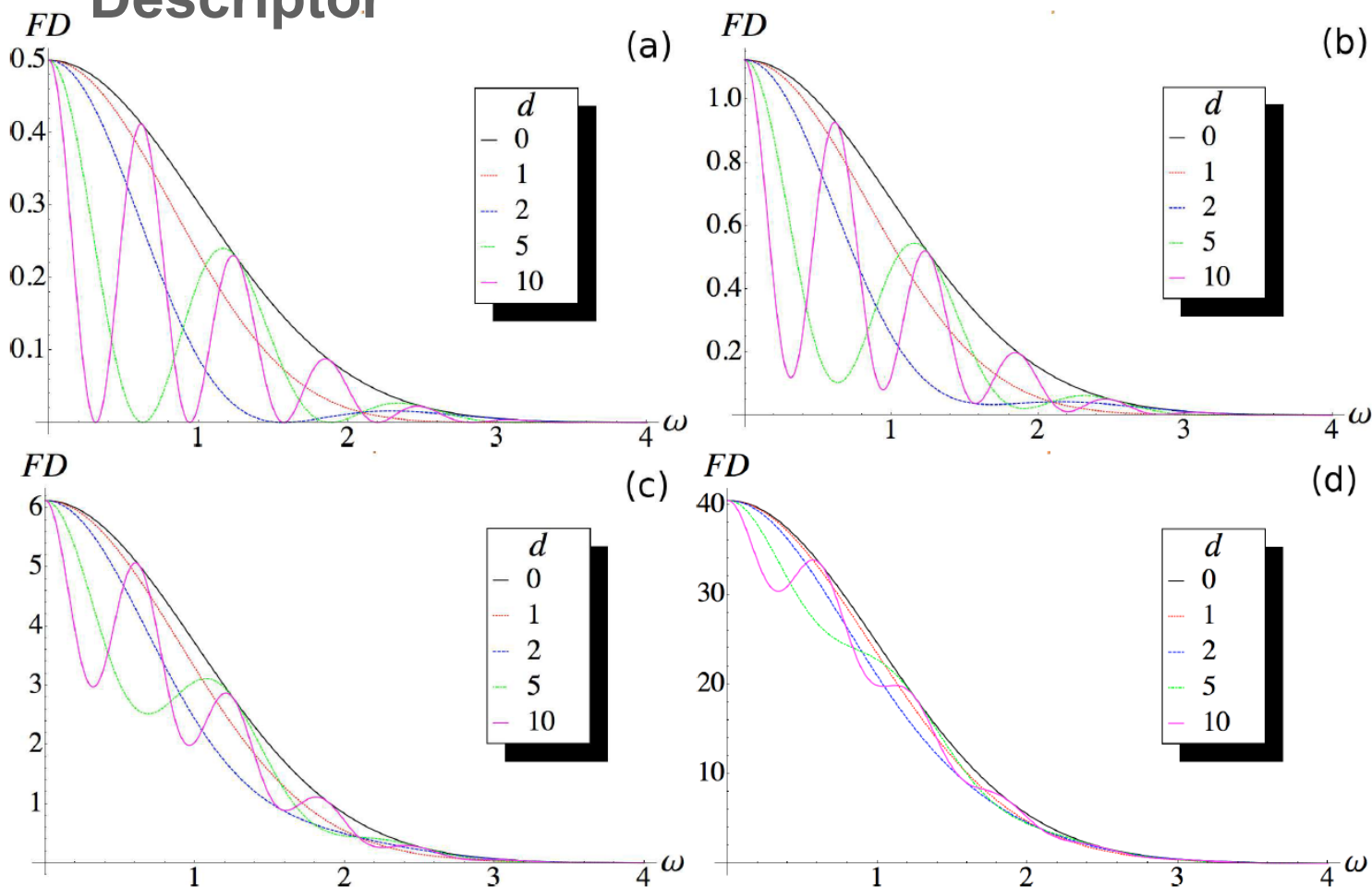
$$FD = \frac{1}{(2a)^3} e^{\frac{\omega^2}{2a}} \sum_J \sum_I Z_I Z_J \cos[\omega \times d_{IJ}],$$



Descriptor



Aaron Knoll
(TACC)



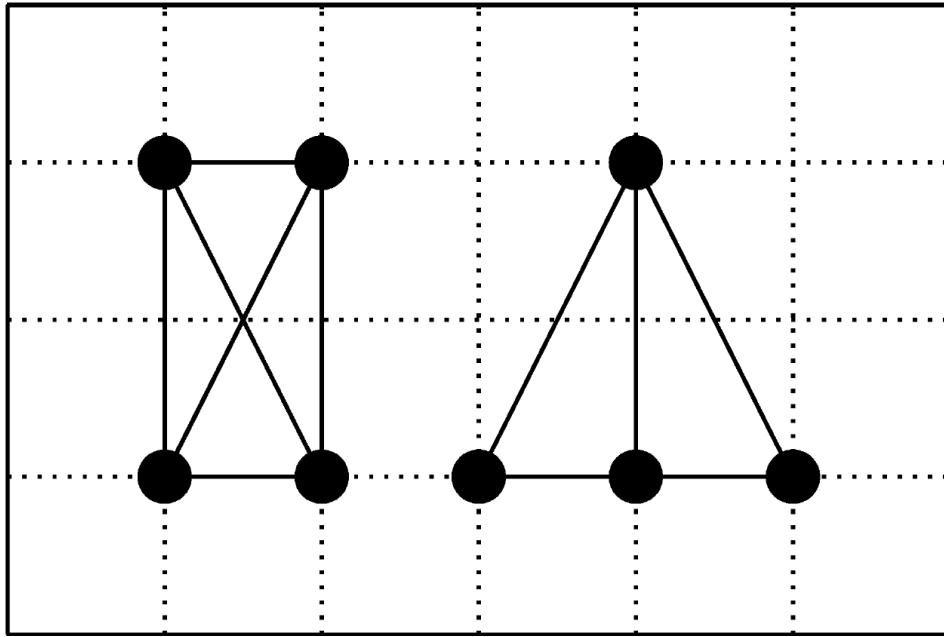
Fourier descriptor FD of four diatomics, H_2 (a), HHe (b), HC (c), and HCl (d), for five interatomic distances d



$$FD = \frac{1}{(2a)^3} e^{\frac{\omega^2}{2a}} \sum_J \sum_I Z_I Z_J \cos[\omega \times d_{IJ}],$$



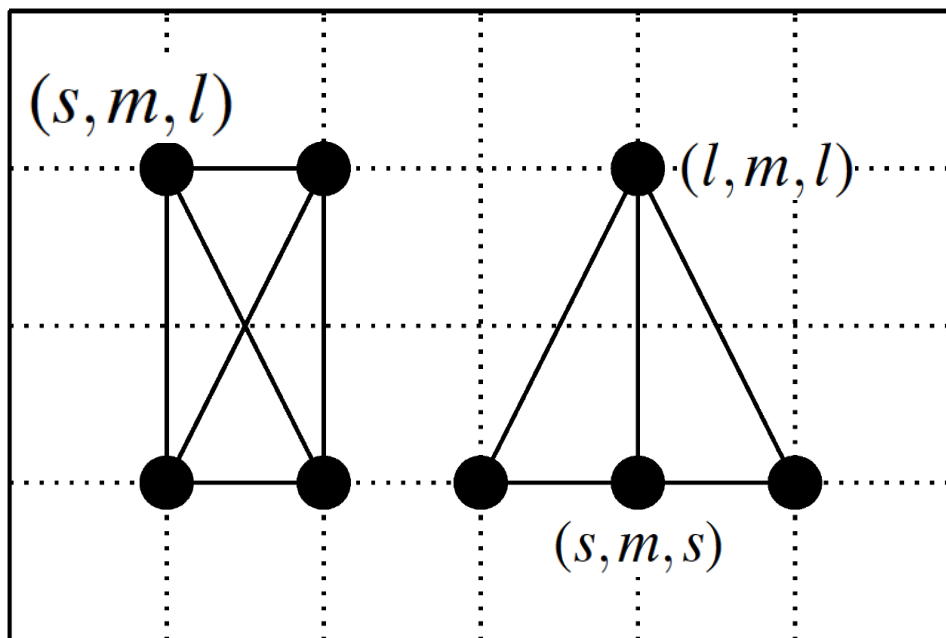
$$FD = \frac{1}{(2a)^3} e^{\frac{\omega^2}{2a}} \sum_J \sum_I Z_I Z_J \cos[\omega \times d_{IJ}],$$



Homometric molecules?



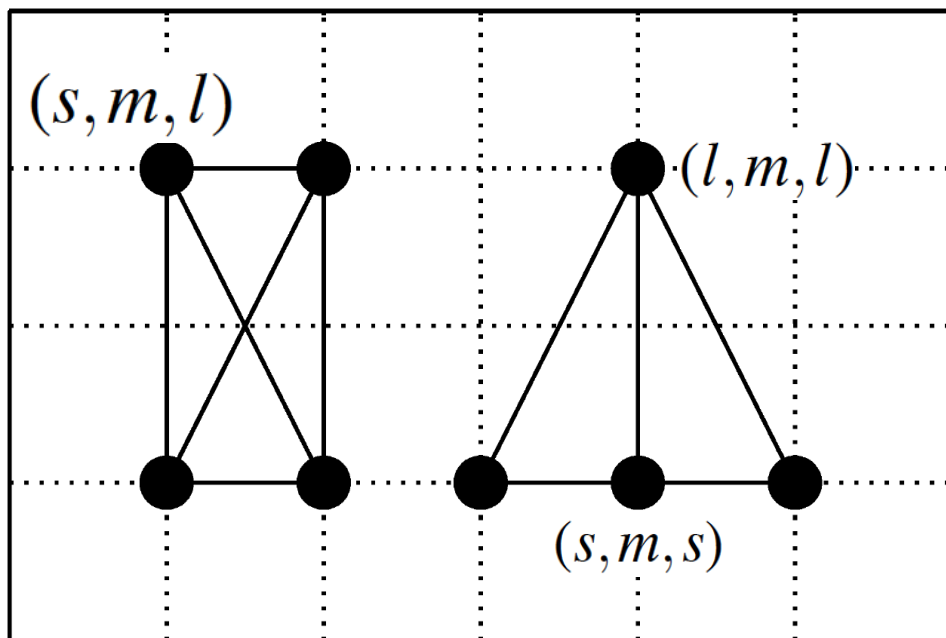
$$FD = \frac{1}{(2a)^3} e^{\frac{\omega^2}{2a}} \sum_J \sum_I Z_I Z_J \cos[\omega \times d_{IJ}],$$



Homometric molecules?

$$\sum_J Z_J e^{-b(d-d_{IJ})^2}$$

$$FD = \frac{1}{(2a)^3} e^{\frac{\omega^2}{2a}} \sum_J \sum_I Z_I Z_J \cos[\omega \times d_{IJ}],$$

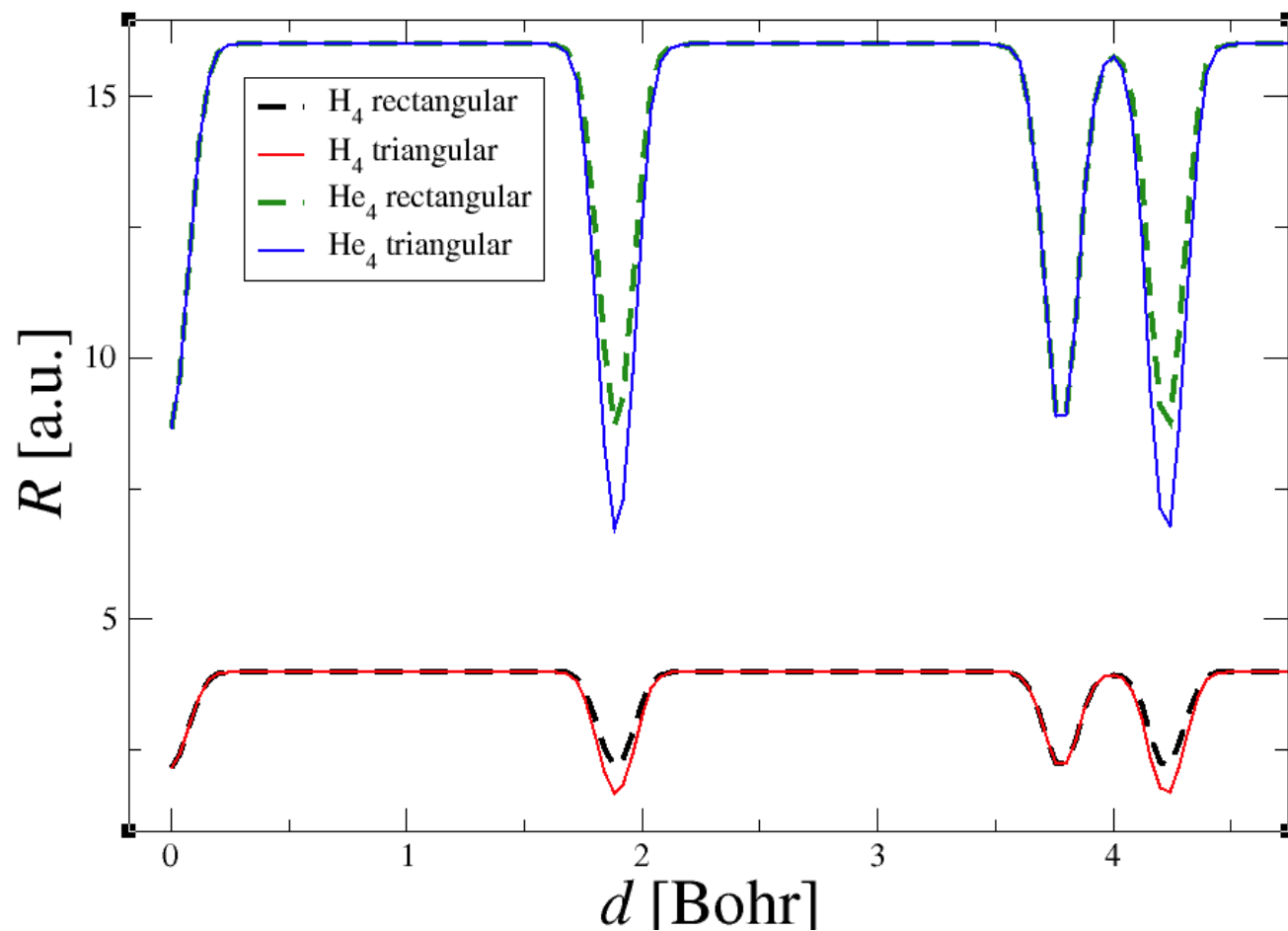


Homometric molecules?

$$\sum_J Z_J e^{-b(d-d_{IJ})^2}$$

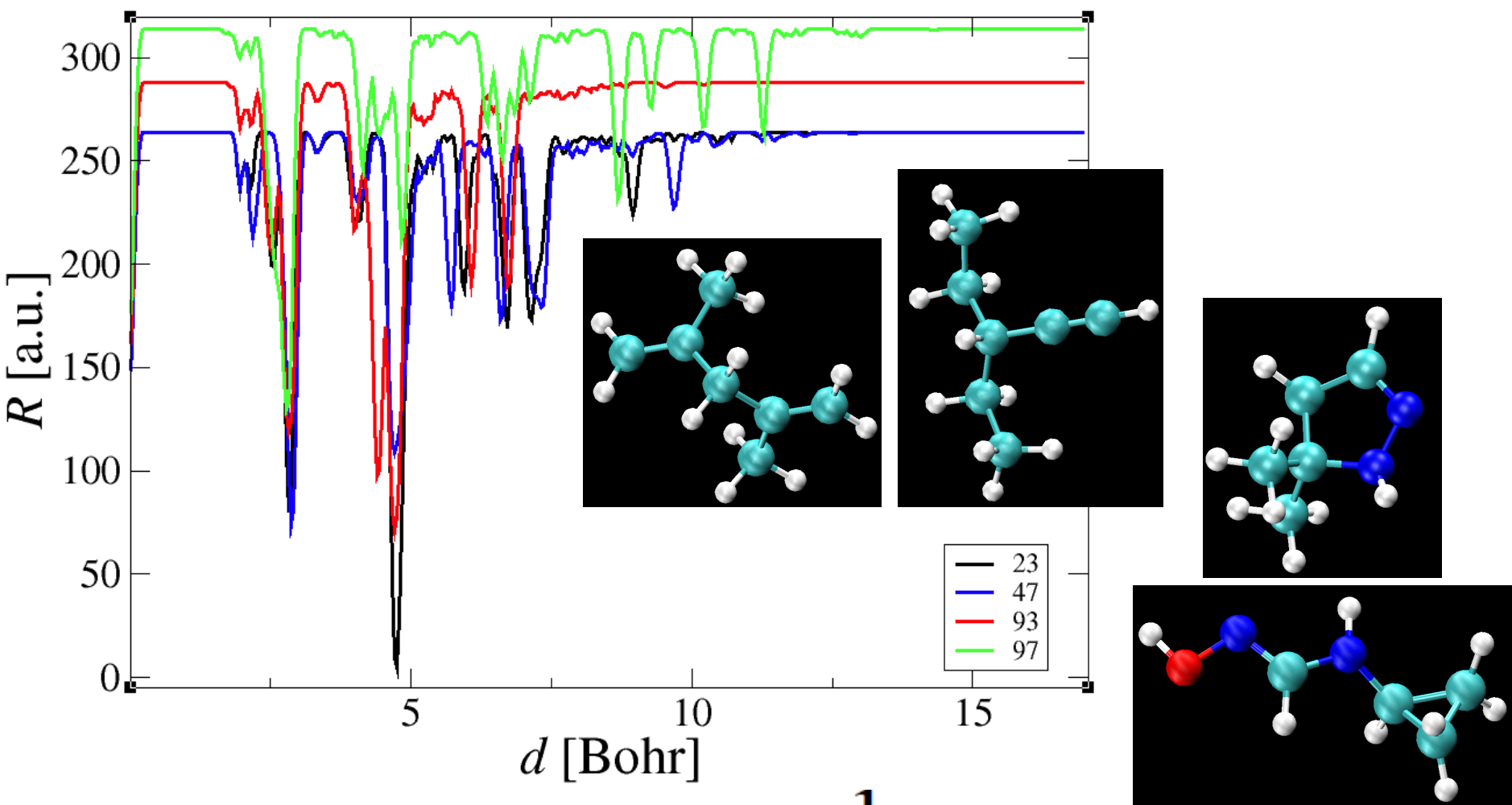
$$R(d) = \sum_I Z_I^2 \cos\left[\frac{1}{Z_I} \sum_J Z_J e^{-b(d-d_{IJ})^2}\right]$$





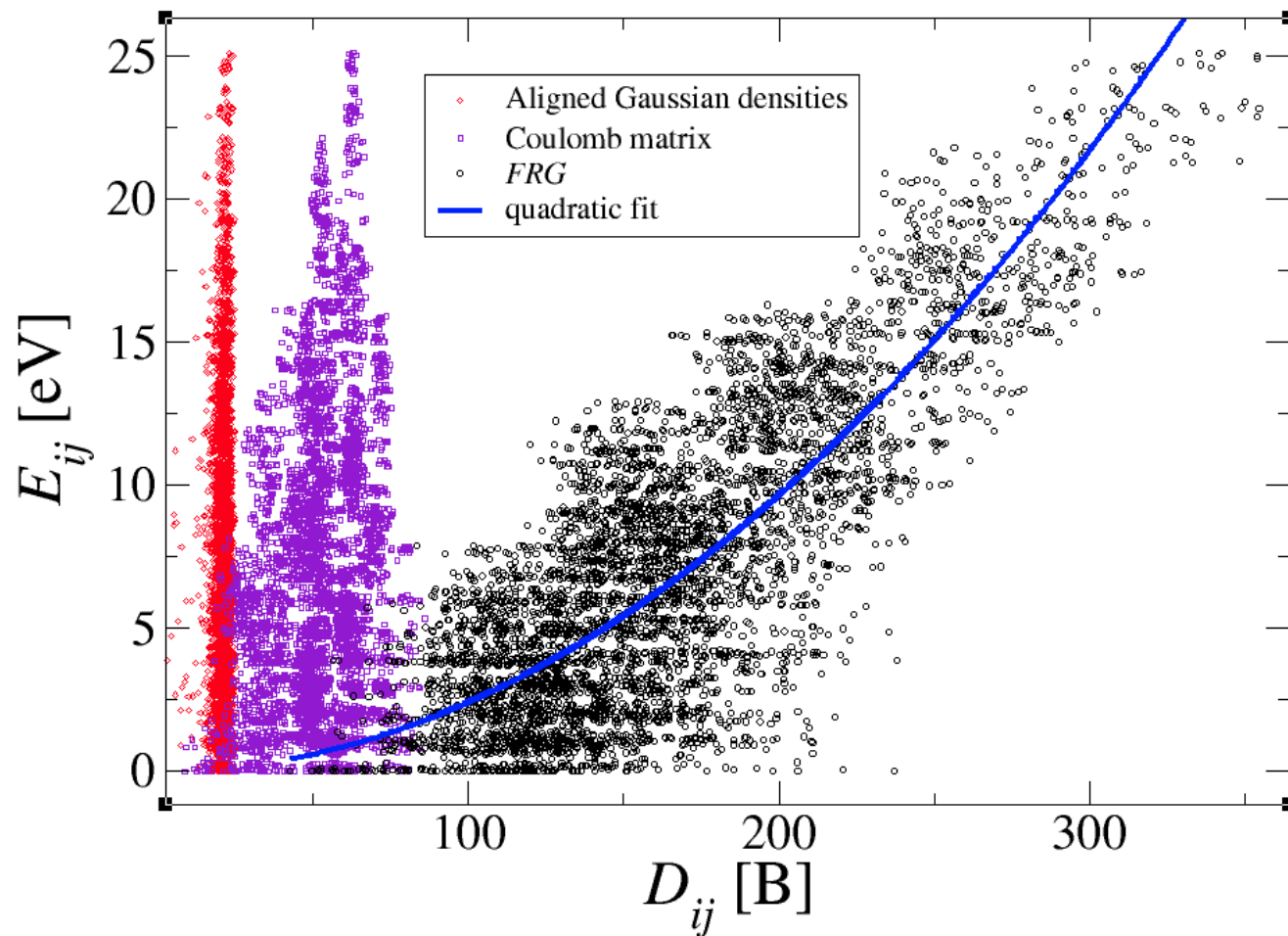
$$R(d) = \sum_I Z_I^2 \cos\left[\frac{1}{Z_I} \sum_J Z_J e^{-b(d-d_{IJ})^2}\right]$$

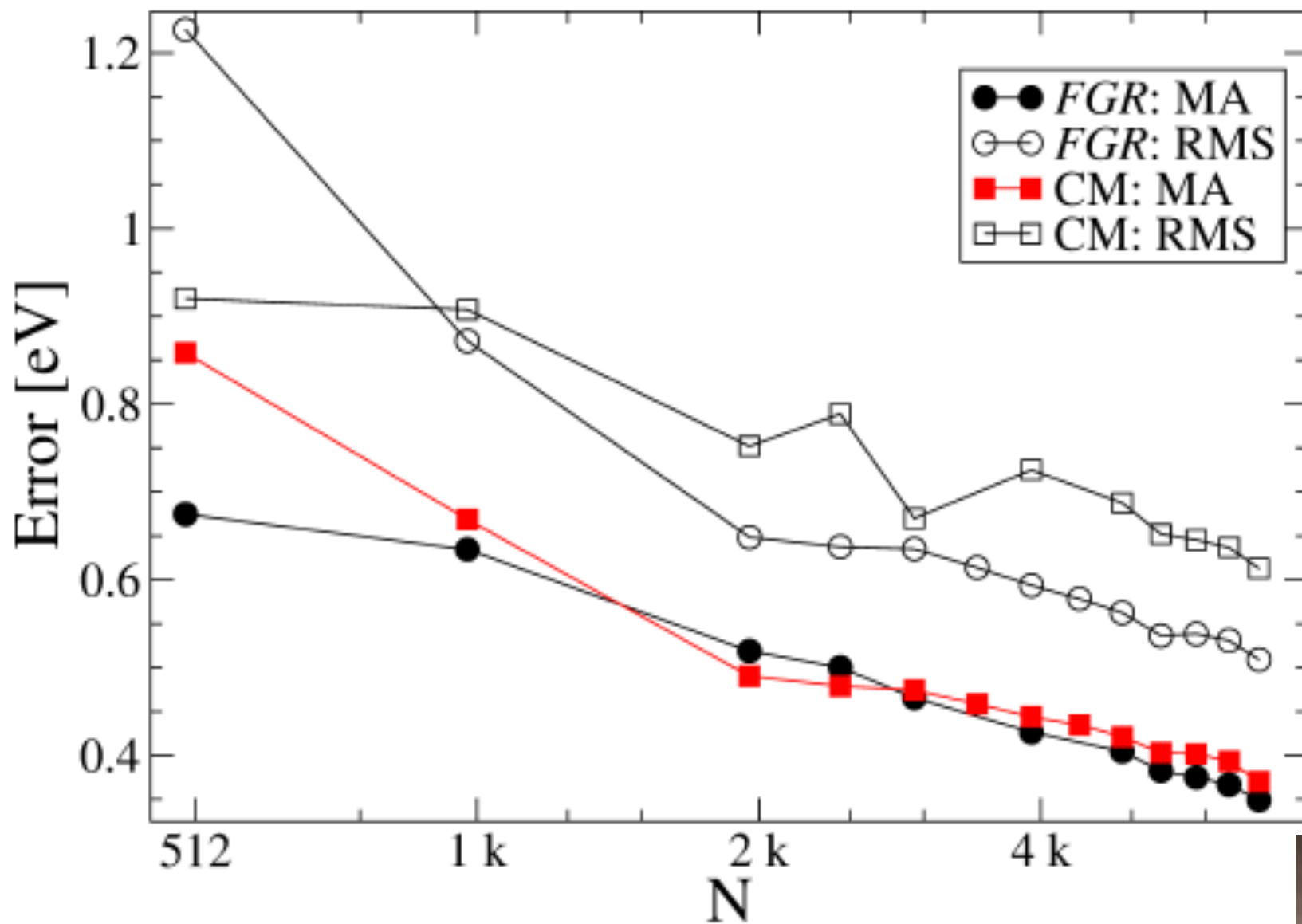




$$R(d) = \sum_I Z_I^2 \cos\left[\frac{1}{Z_I} \sum_J Z_J e^{-b(d-d_{IJ})^2}\right]$$

$$D(M_i, M_j) = \sqrt{\int_{d=0}^{d \geq d_{IJ}^{max}} dd (R_i(d) - R_j(d))^2}$$





First Principles

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



Schrödinger

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

variational (deductive)

Feynman

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

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

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

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



correlational (inductive)

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

supervised
learning

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



Vapnik



Thanks for your attention!

First principles view on chemical compound space: Gaining rigorous atomistic control of molecular properties

OAvL, Int J Quant Chem (2013), <http://onlinelibrary.wiley.com/doi/10.1002/qua.24375/abstract>

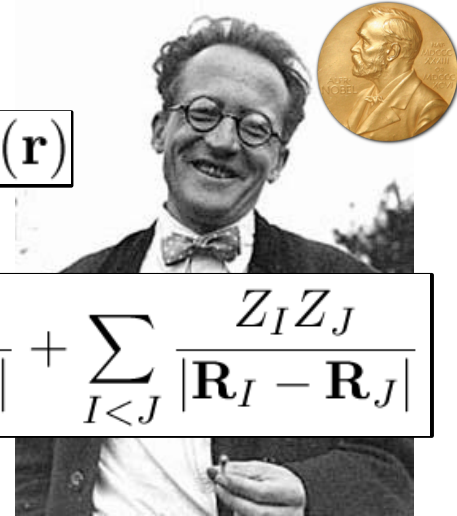
<http://www.quantum-machine.org/>



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

correlational (inductive)

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

supervised
learning

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

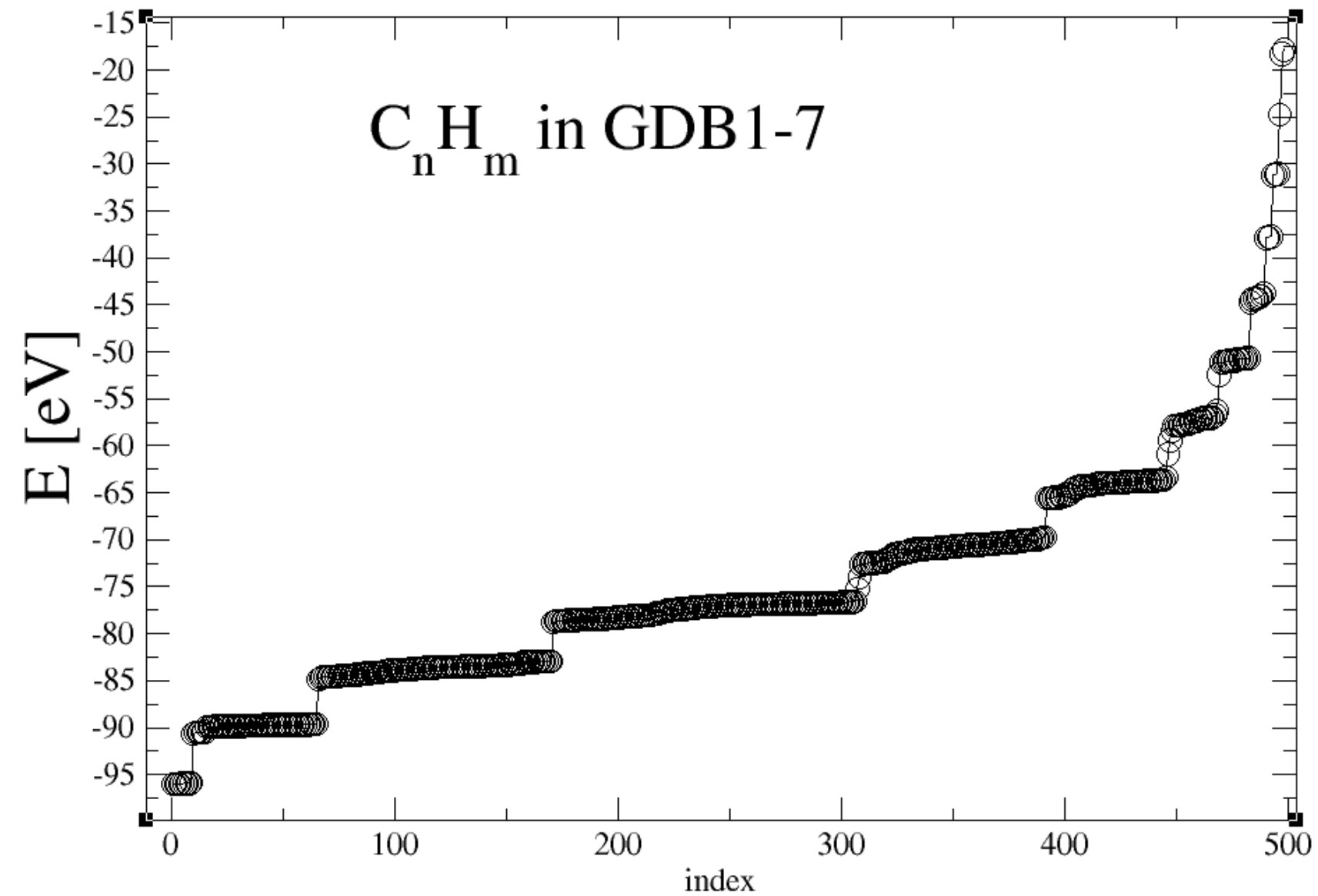
Infer solution by comparison
to previous examples

- Regression method?
- Function?
- Variables?
- Metric?
- **Data?**

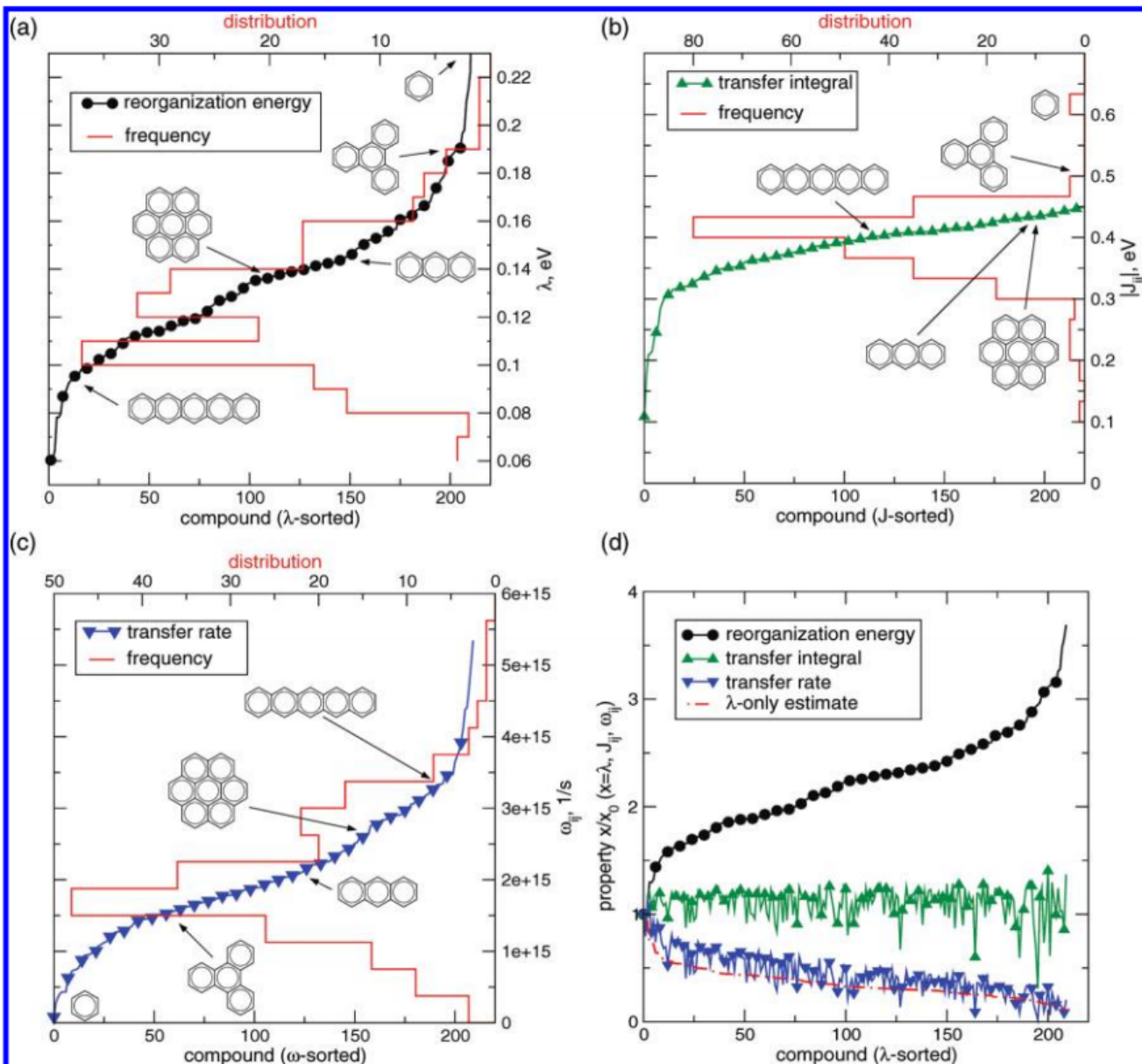
Vapnik



Data stratification



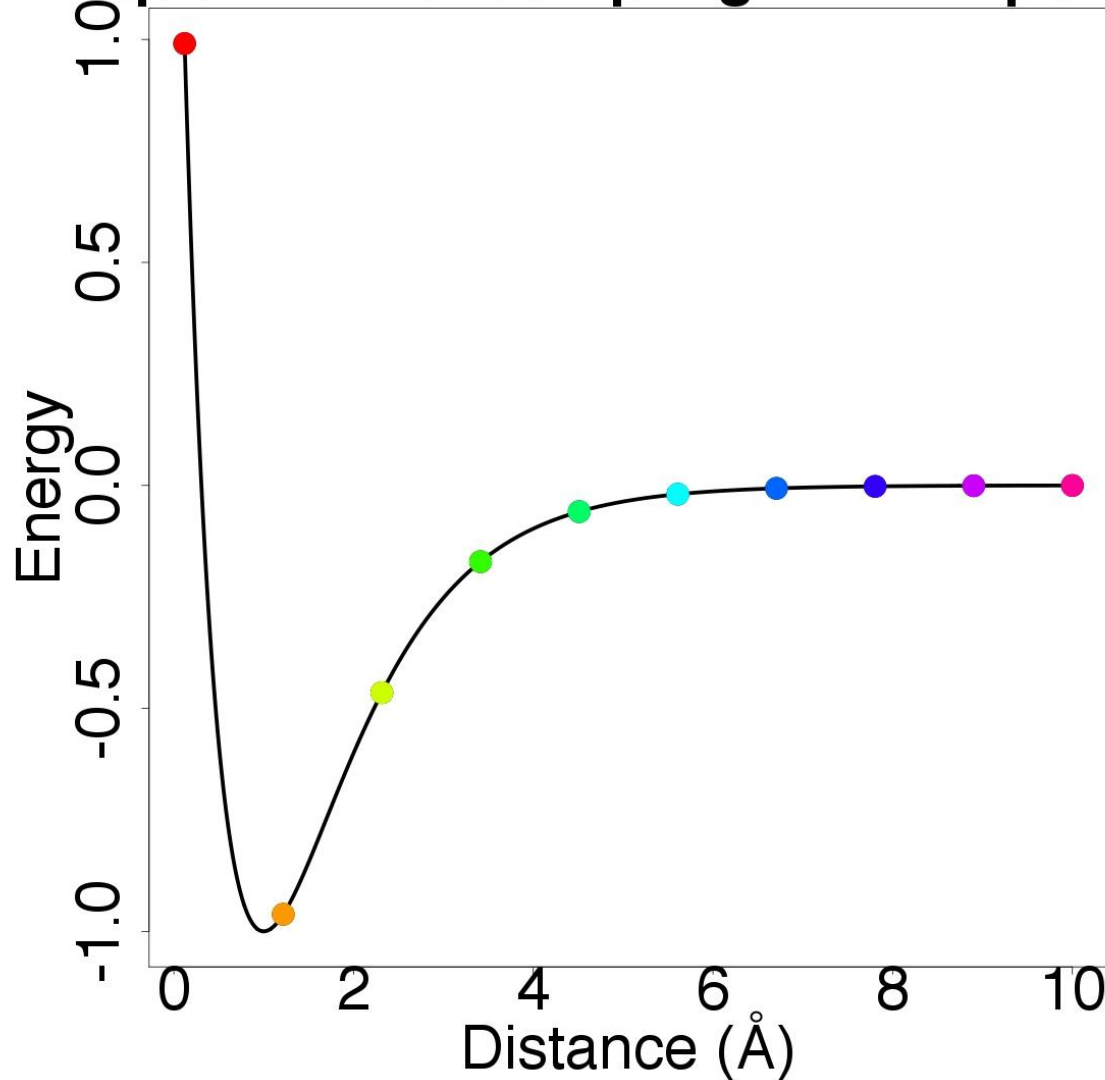
Data stratification



Misra et al JCTC (2011)

Outlook: Selection bias

Equal interval sampling after 10 points



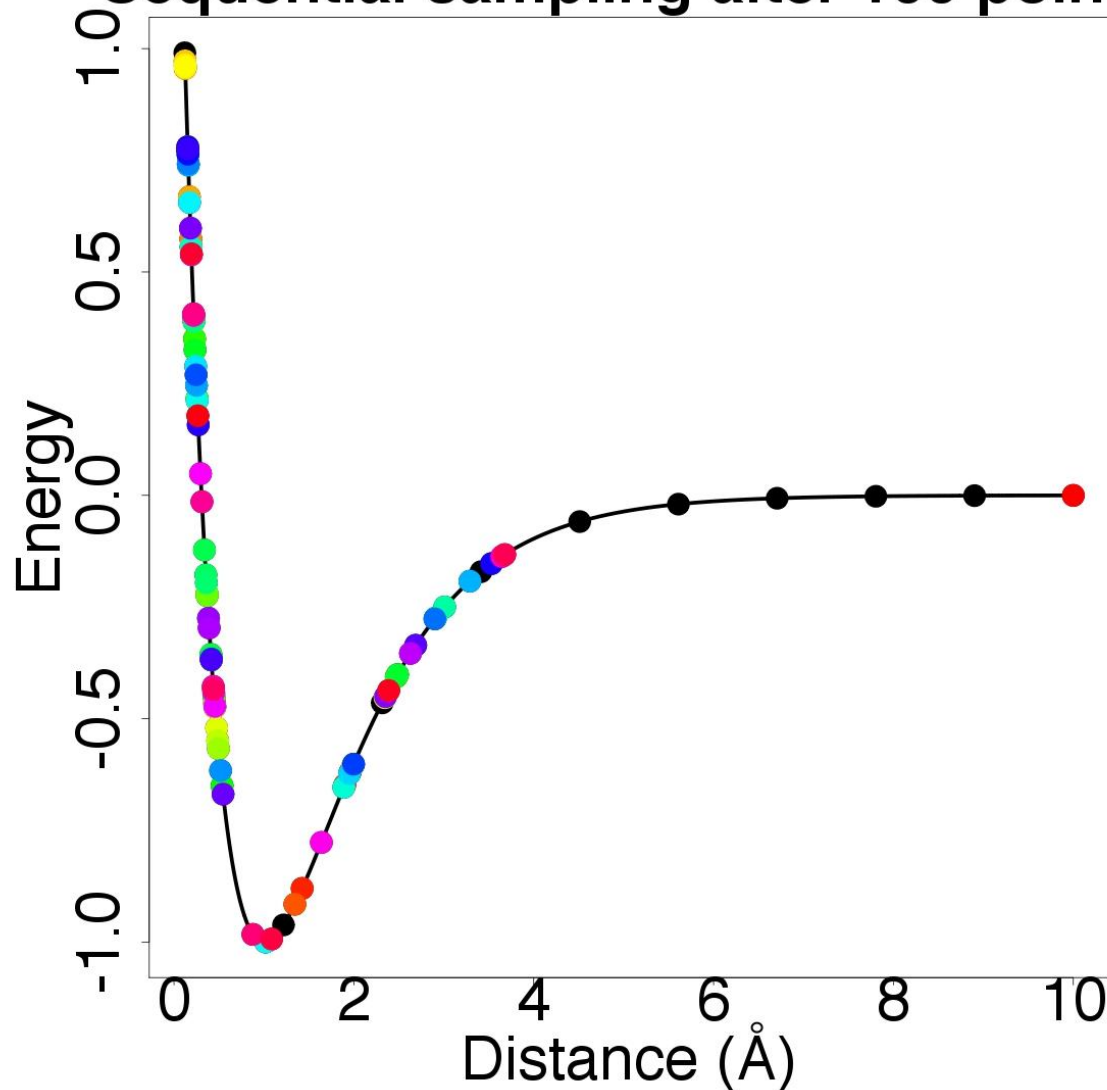
Balaprakash
(ANL)



Vazquez
(ANL)

Outlook: Selection bias

Sequential sampling after 100 points



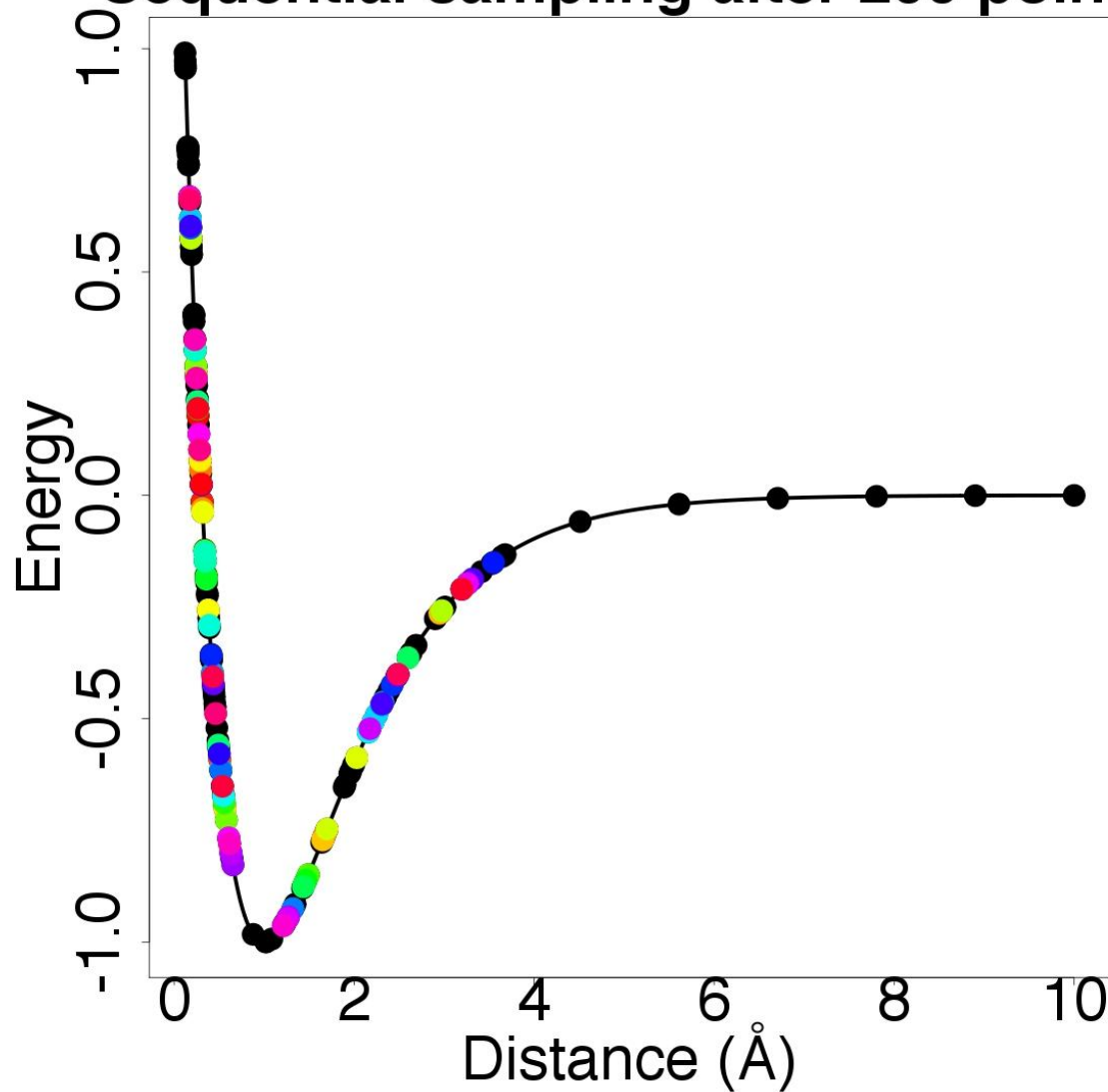
Balaprakash
(ANL)



Vazquez
(ANL)

Outlook: Selection bias

Sequential sampling after 200 points



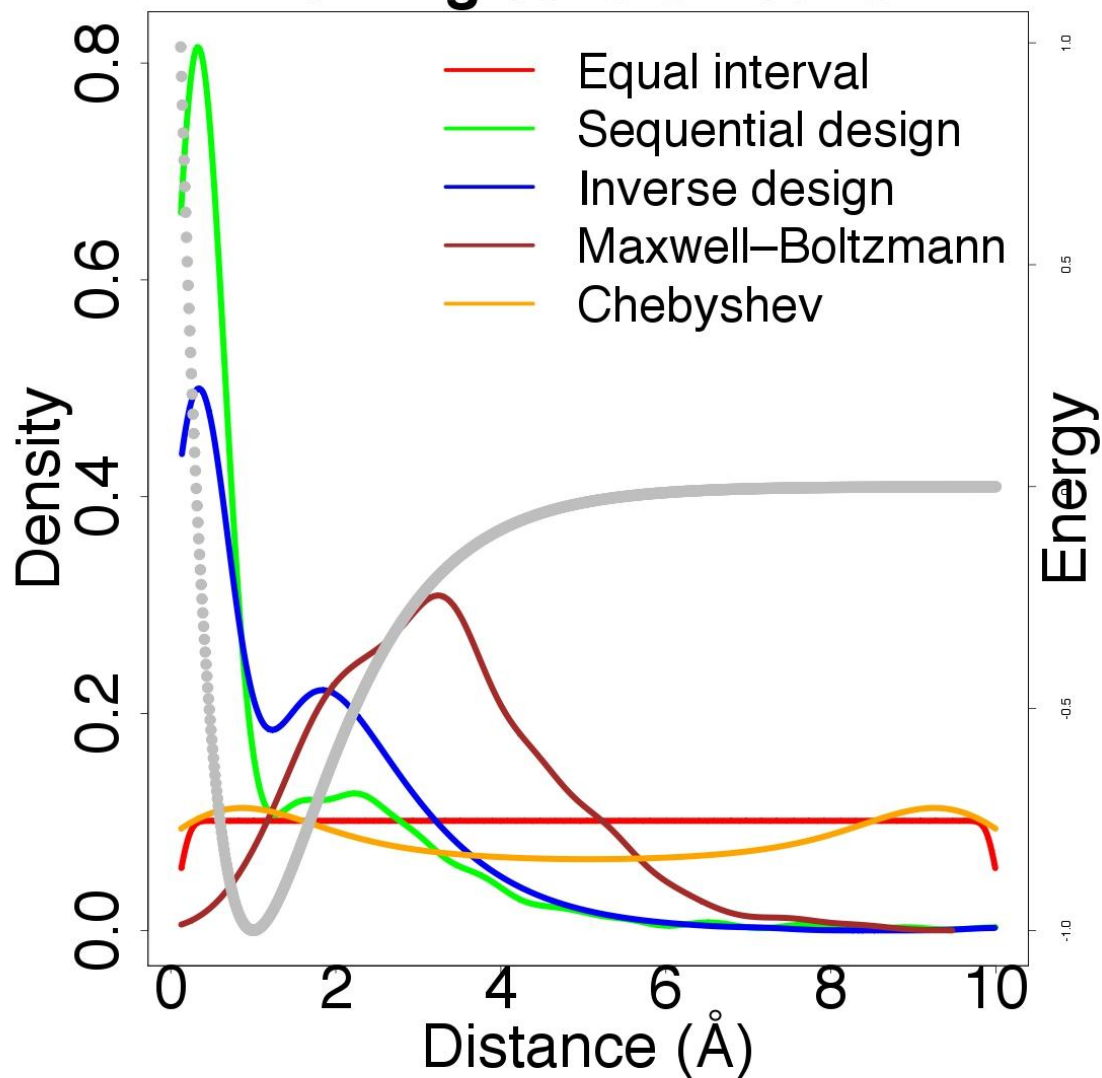
Balaprakash
(ANL)



Vazquez
(ANL)

Outlook: Selection bias

Training set distribution

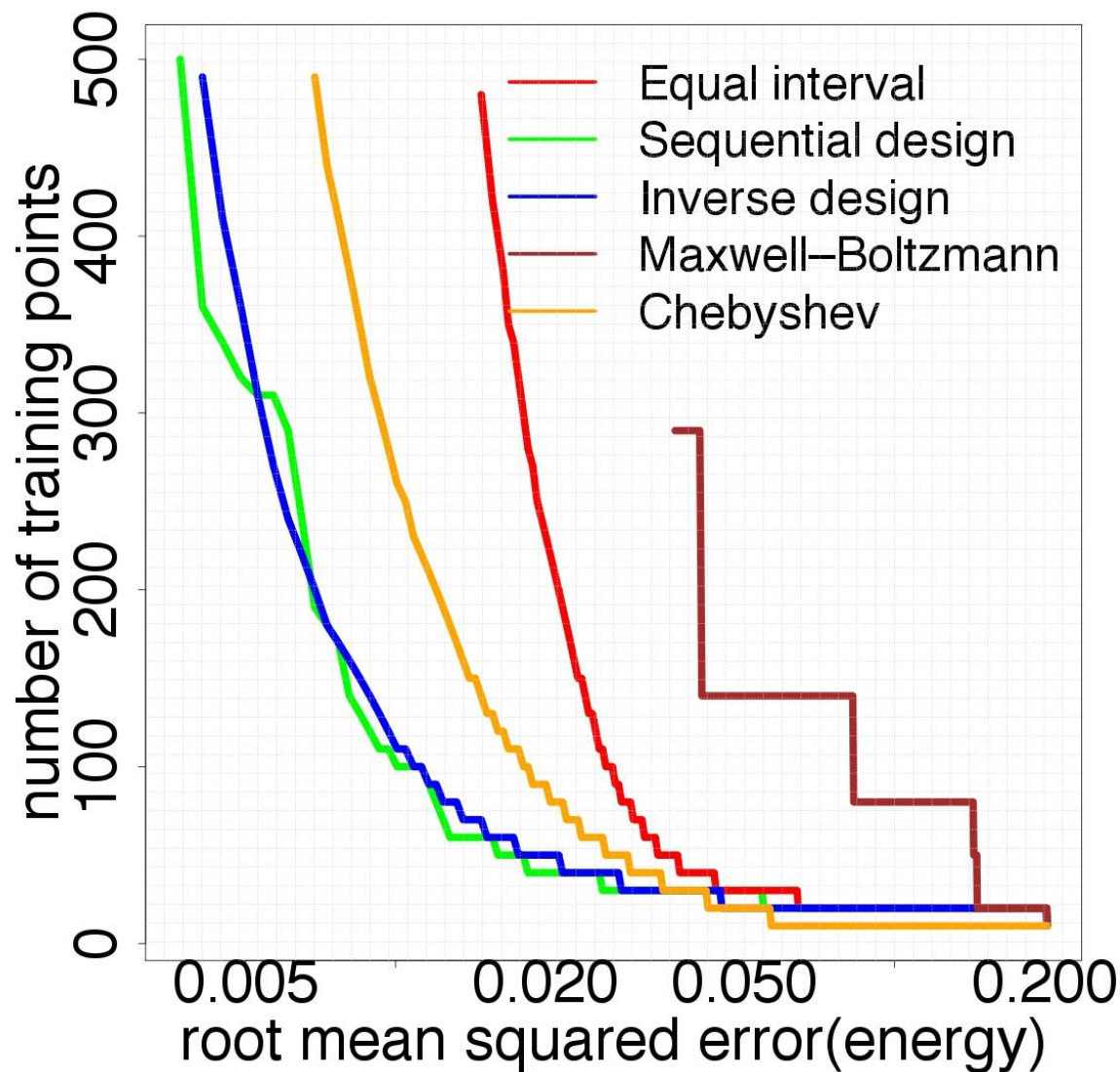


Balaprakash
(ANL)



Vazquez
(ANL)

Outlook: Selection bias



Balaprakash
(ANL)



Vazquez
(ANL)

First Principles

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



Schrödinger

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

variational (deductive)

Feynman

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

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

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

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



correlational (inductive)

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

supervised
learning

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



Vapnik



Thanks for your attention!

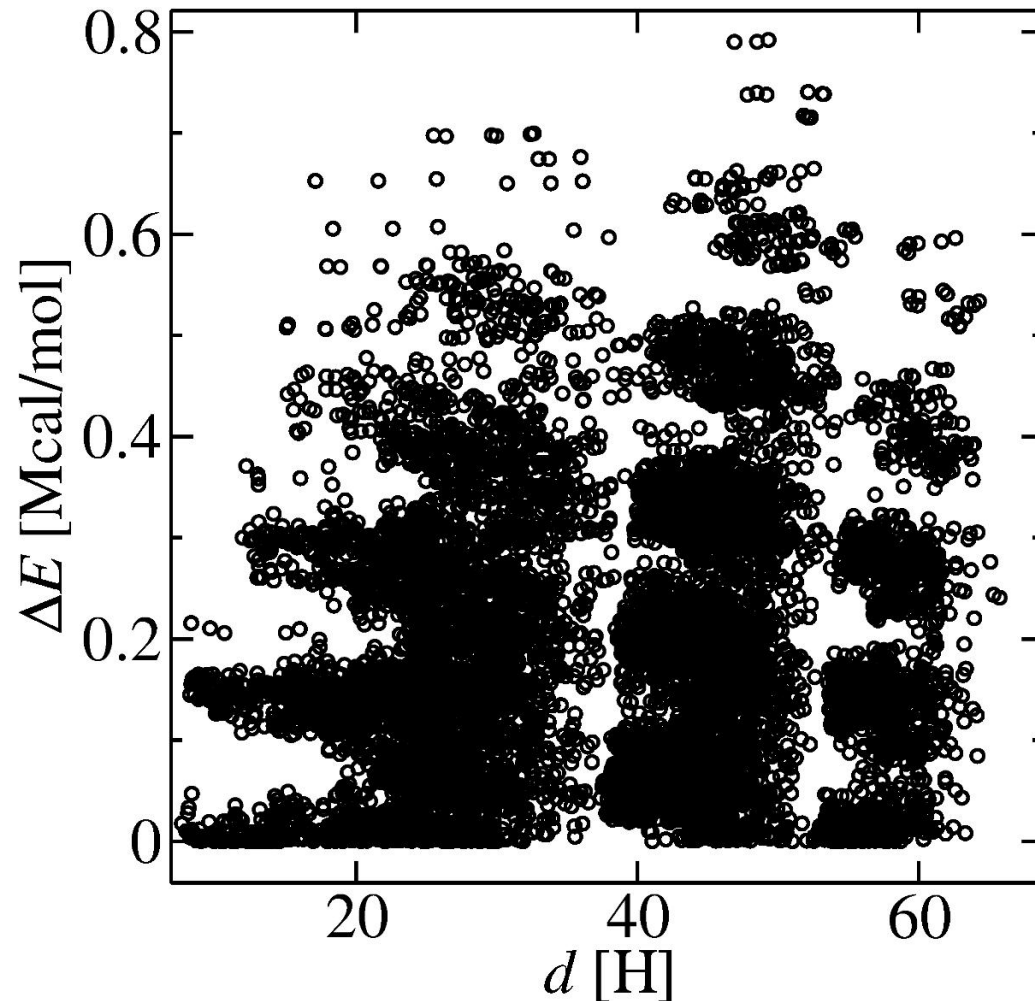
First principles view on chemical compound space: Gaining rigorous atomistic control of molecular properties

OAvL, Int J Quant Chem (2013), <http://onlinelibrary.wiley.com/doi/10.1002/qua.24375/abstract>

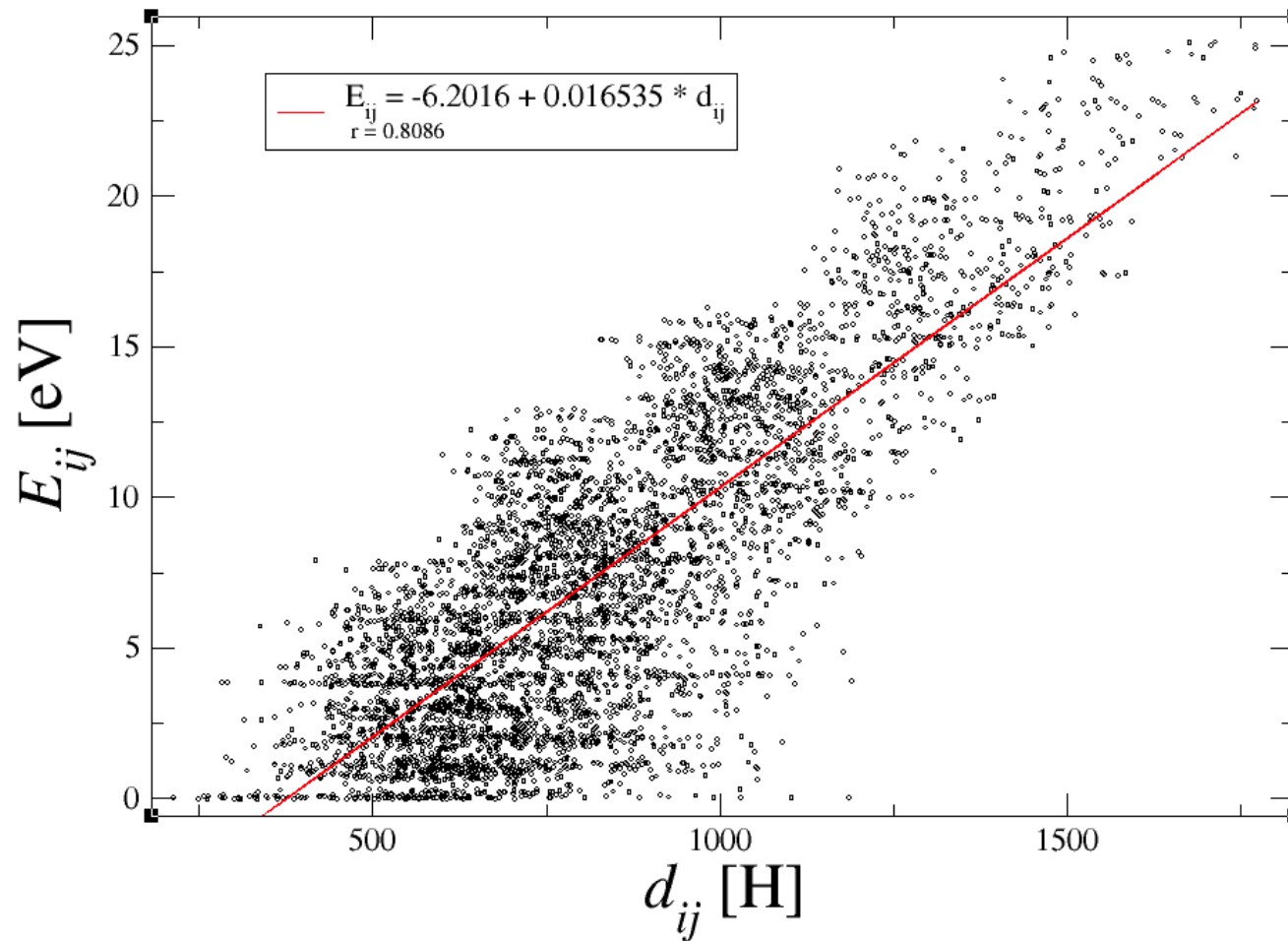
<http://www.quantum-machine.org/>



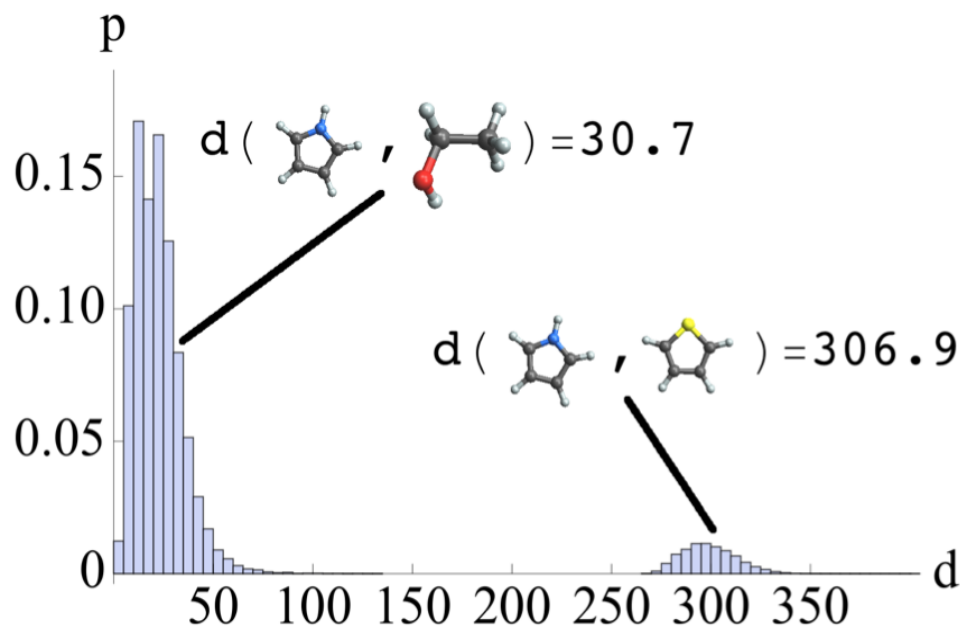
$$d(\mathbf{M}, \mathbf{M}_i) = \sqrt{\sum_{IJ} |M_{IJ} - M_{IJ}^{(i)}|^2}$$



$$D(M_i, M_j) = \sqrt{\int_{d=0}^{d \geq d_{IJ}^{max}} dd (R_i(d) - R_j(d))^2}$$

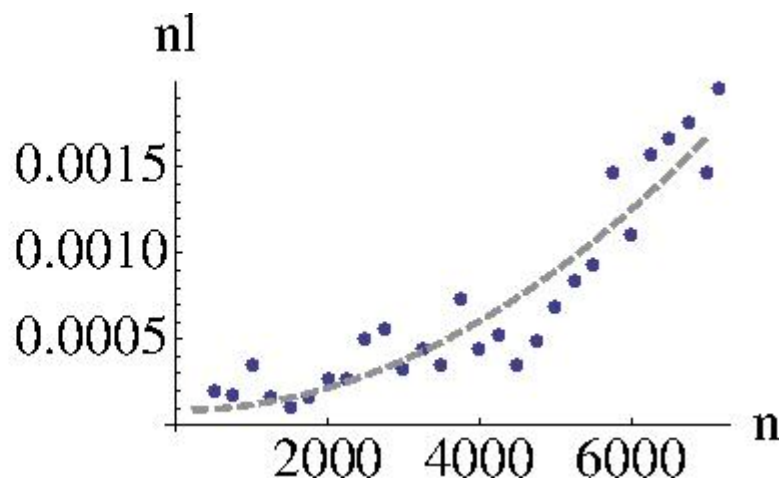
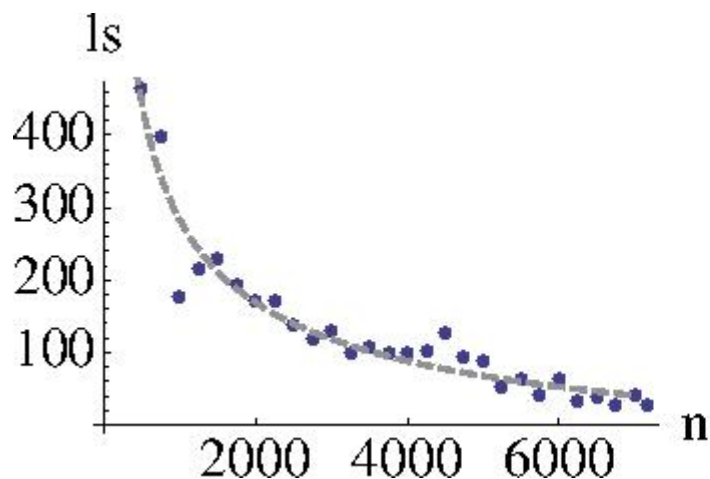


Locality



Model becomes local at ~5k molecules in training set

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



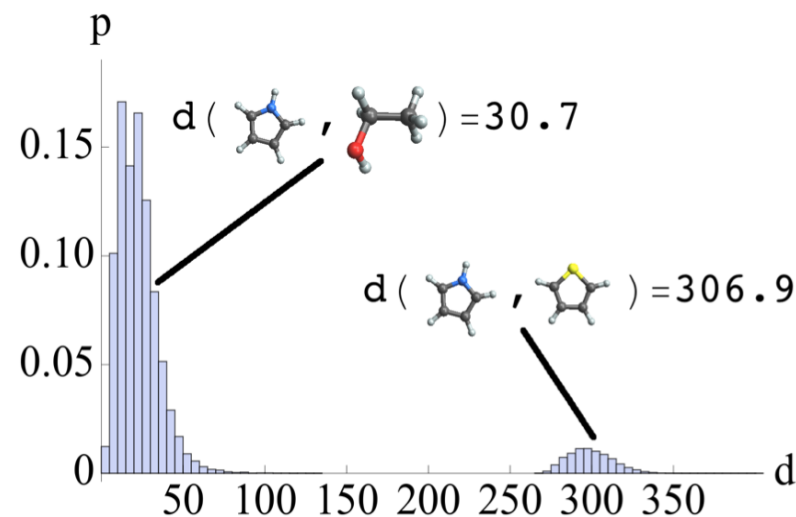
Correlational: Kernel Ridge Regression

$$d(\mathbf{M}, \mathbf{M}') = \sqrt{\sum_{IJ} |M_{IJ} - M'_{IJ}|^2}$$

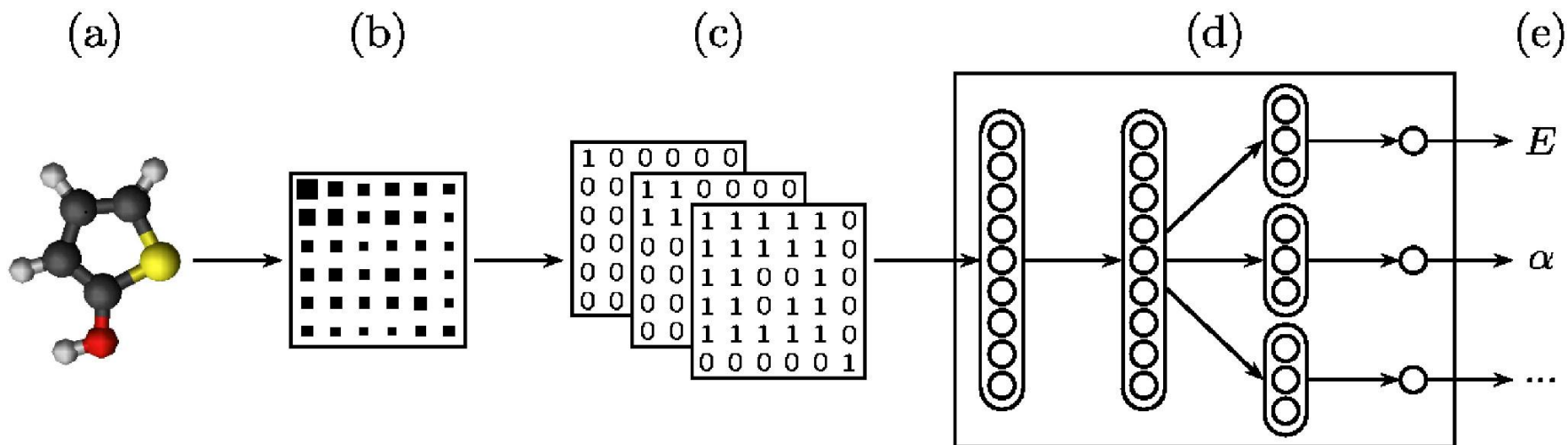
$$\min_{\alpha} \sum_i (E^{est}(\mathbf{M}_i) - E_i^{ref})^2 + \lambda \sum_i \alpha_i^2$$

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

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

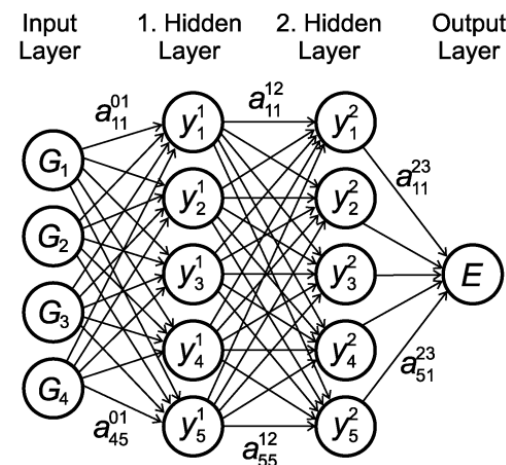


Correlational: Regression 2



$$E = f_1^3 \left(b_1^3 + \sum_{l=1}^5 a_{l1}^{23} \cdot f_l^2 \left(b_l^2 + \sum_{k=1}^5 a_{kl}^{12} \cdot f_k^1 \left(b_k^1 + \sum_{j=1}^4 a_{jk}^{01} \cdot G_j \right) \right) \right)$$

T. B. Blank, S. D. Brown, A. W. Calhoun and D. J. Doren,
J. Chem. Phys., 1995, **103**, 4129.
 J. Behler, *Phys Chem Chem Phys* (2011)



G. Montavon, M. Rupp, V. Gobre, A. Vazquez, K. Hansen, A. Tkatchenko, K.-R. Müller,
 OAvL, *submitted* (2012)

Transferability (no overfitting)

For training set: k -fold **cross-validation**

1. divide data into k blocks
2. predict each block with model trained on remaining blocks
3. average coefficients

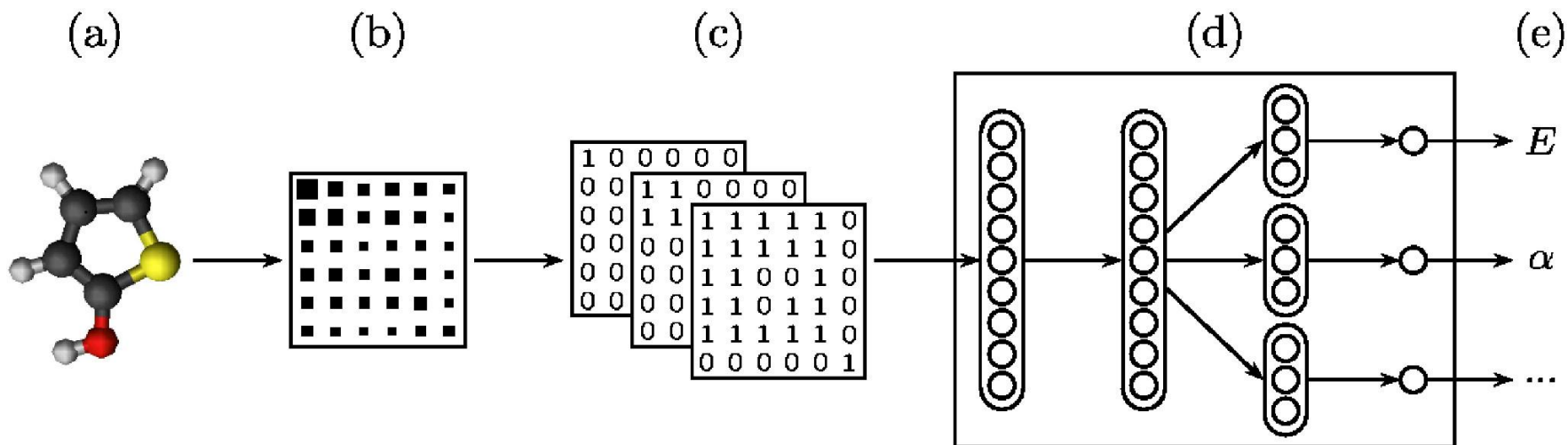


Two nested loops for training *and* hyper parameter optimization

Apply to test set to measure out-of-sample performance



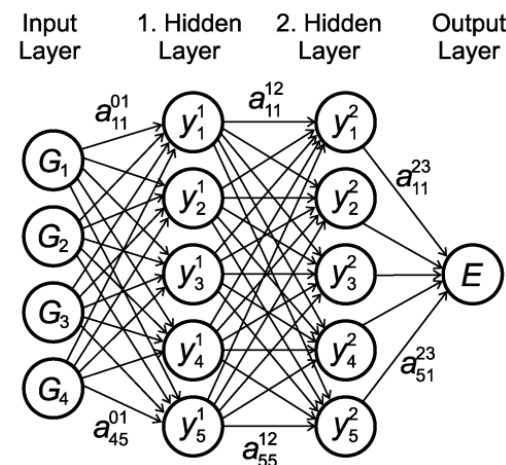
Correlational: Regression 2



$$E = f_1^3 \left(b_1^3 + \sum_{l=1}^5 a_{l1}^{23} \cdot f_l^2 \left(b_l^2 + \sum_{k=1}^5 a_{kl}^{12} \cdot f_k^1 \left(b_k^1 + \sum_{j=1}^4 a_{jk}^{01} \cdot G_j \right) \right) \right)$$

T. B. Blank, S. D. Brown, A. W. Calhoun and D. J. Doren,
J. Chem. Phys., 1995, **103**, 4129.

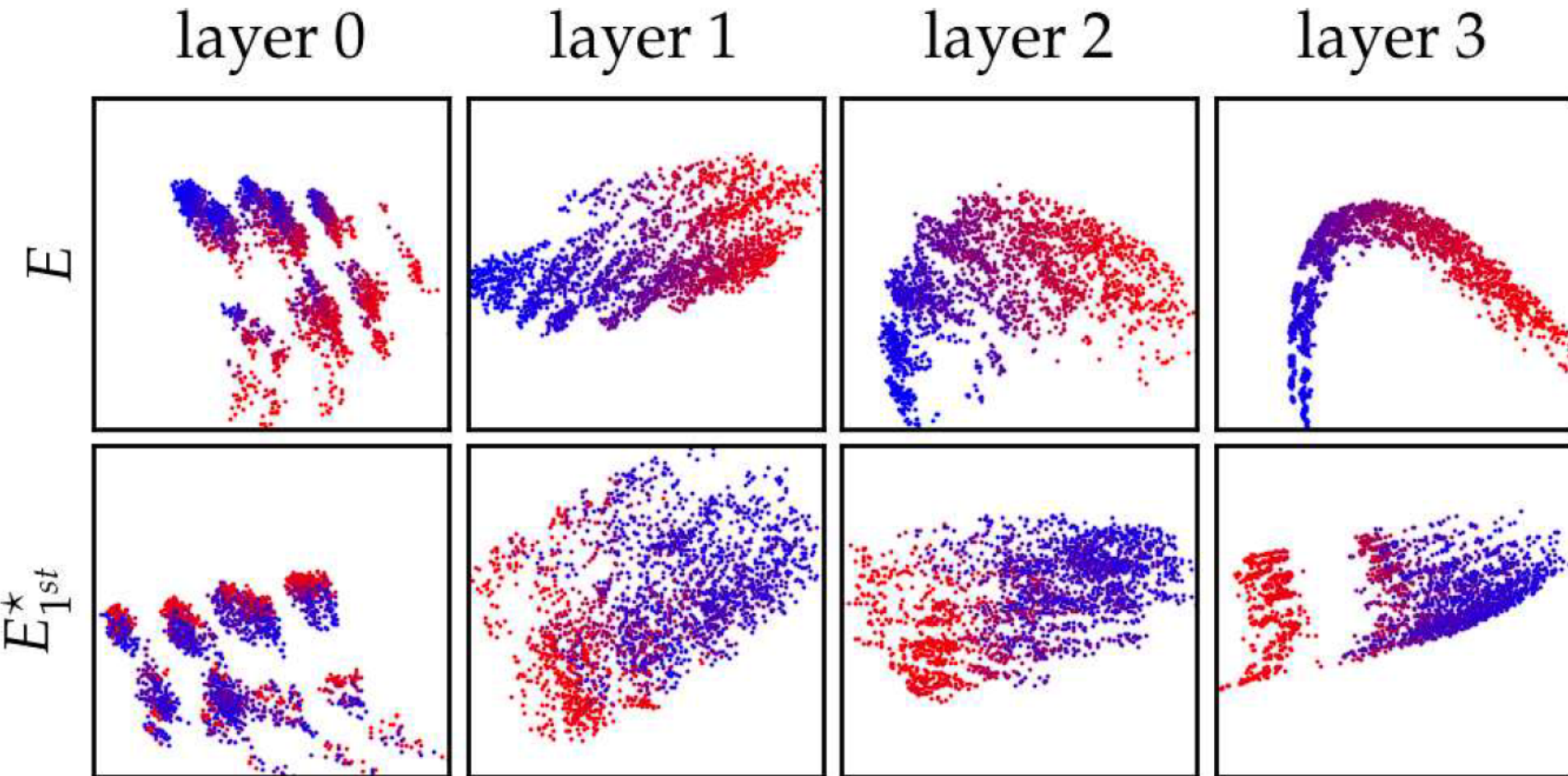
J. Behler, *Phys Chem Chem Phys* (2011)



G. Montavon, M. Rupp, V. Gobre, A. Vazquez, K. Hansen, A. Tkatchenko, K.-R. Müller,
 OAvL, *NJP* accepted (2013)

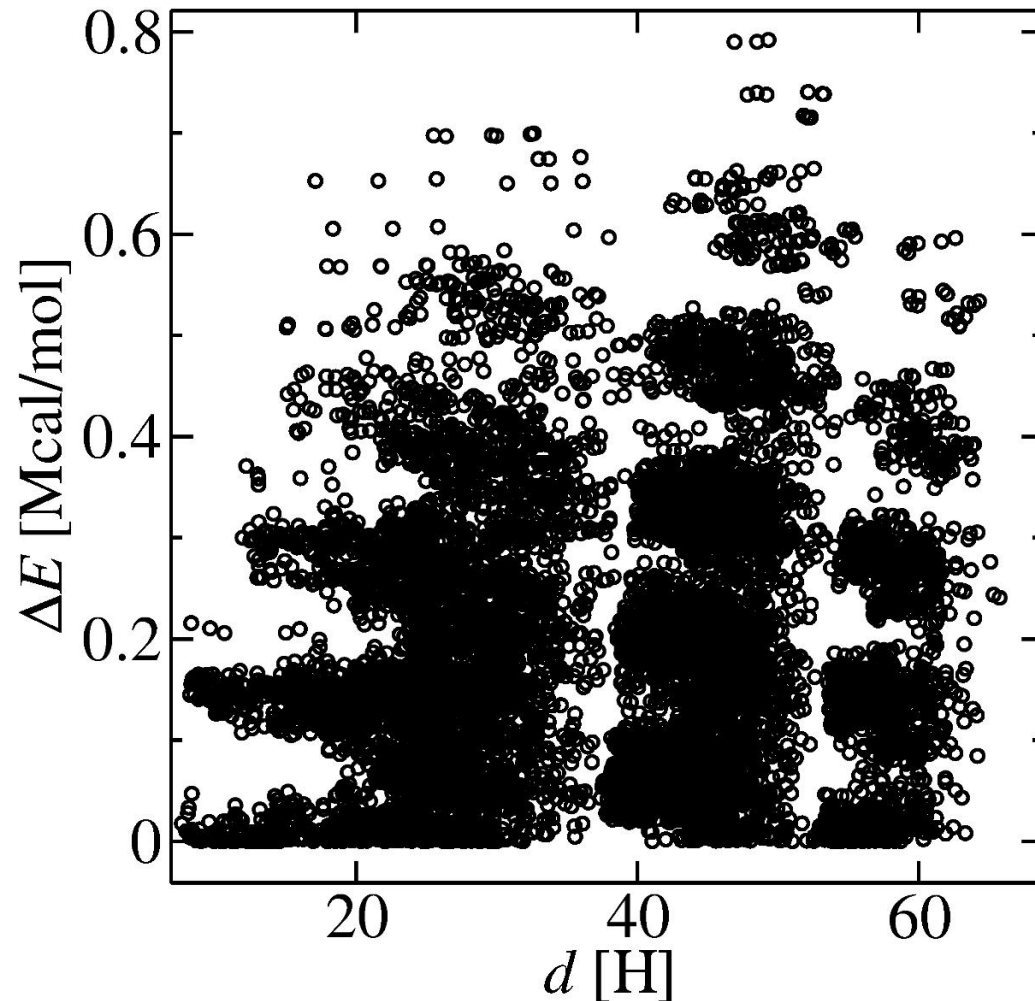
Deep Neural Networks

– PCA on properties for four layers



G. Montavon, M. Rupp, V. Gobre, A. Vazquez, K. Hansen, A. Tkatchenko, K.-R. Müller,
OAvL, *NJP* accepted (2013)

$$d(\mathbf{M}, \mathbf{M}_i) = \sqrt{\sum_{IJ} |M_{IJ} - M_{IJ}^{(i)}|^2}$$

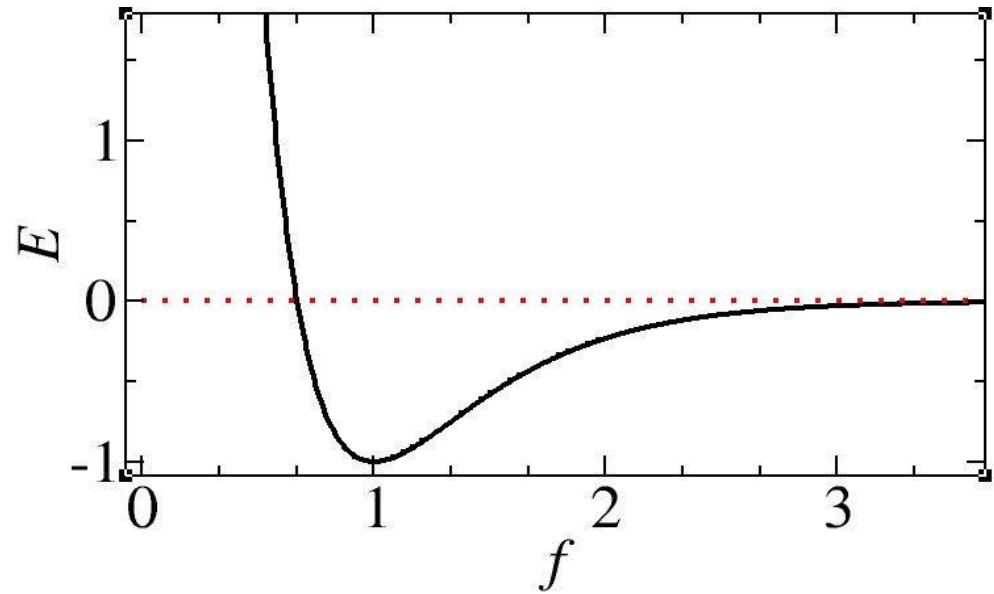


Outlook: Forces

1. Interpolate binding:

- a. $E(f=3) = 0$
- b. $E(f=1) = E(\text{PBE0})$
- c. $dE/df(f=1) = 0$
- d. $E(f=2/3) = 0$

Train on 1k molecules



Outlook: Forces

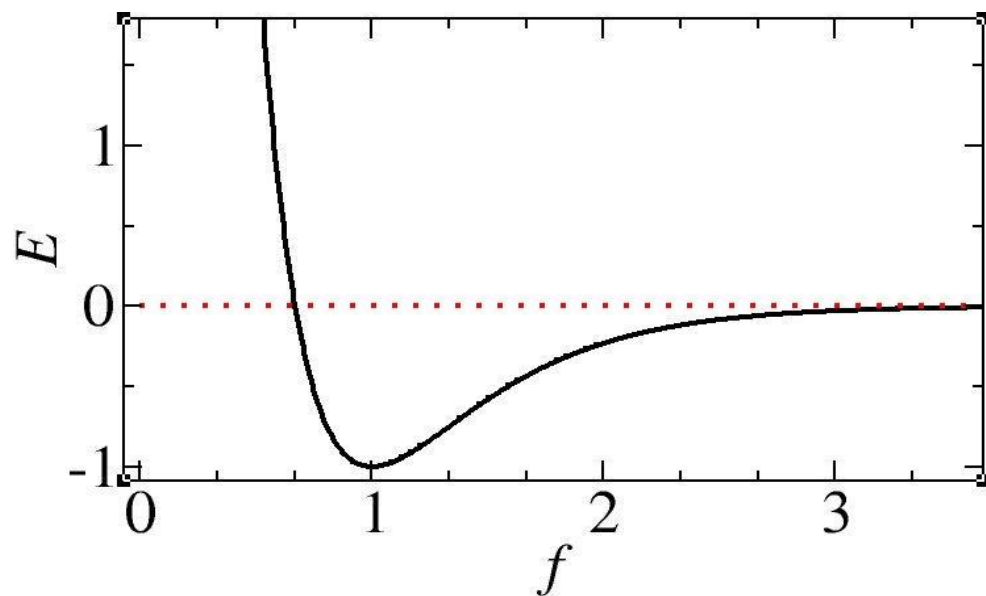
1. Interpolate binding:

- $E(f=3) = 0$
- $E(f=1) = E(\text{PBE0})$
- $dE/df(f=1) = 0$
- $E(f=2/3) = 0$

Train on 1k molecules

2. Test on remaining 6k molecules

MAE $\sim 15\text{kcal/mol}$

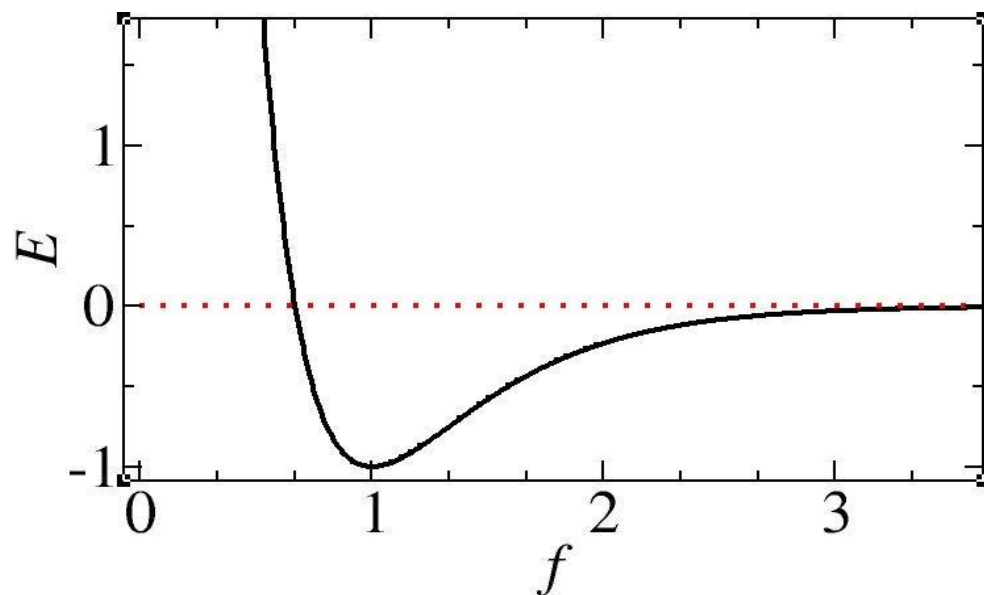


Outlook: Forces

1. Interpolate binding:

- a. $E(f=3) = 0$
- b. $E(f=1) = E(\text{PBE0})$
- c. $dE/df(f=1) = 0$
- d. $E(f=2/3) = 0$

Train on 1k molecules



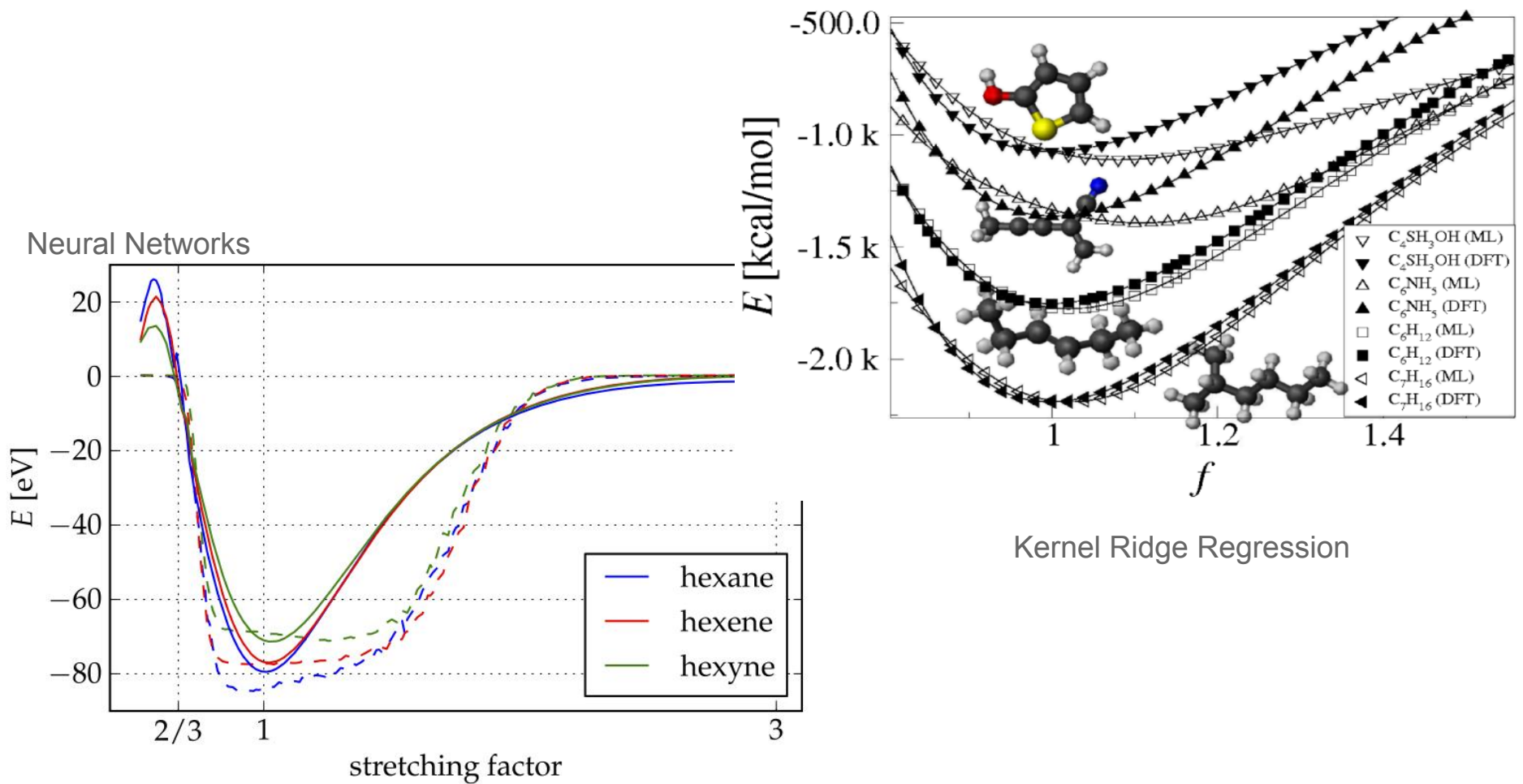
2. Test on remaining 6k molecules

MAE $\sim 15\text{kcal/mol}$

3. Experiment: Predict binding curve for some molecules



Outlook: Forces



M. Rupp, A. Tkatchenko, K.-R. Müller, OAvL, *Phys Rev Lett* (2012)

G. Montavon, M. Rupp, V. Gobre, A. Vazquez, K. Hansen, A. Tkatchenko, K.-R. Müller, OAvL, *submitted* (2012)

