Machine Learning of Quantum Chemical Space

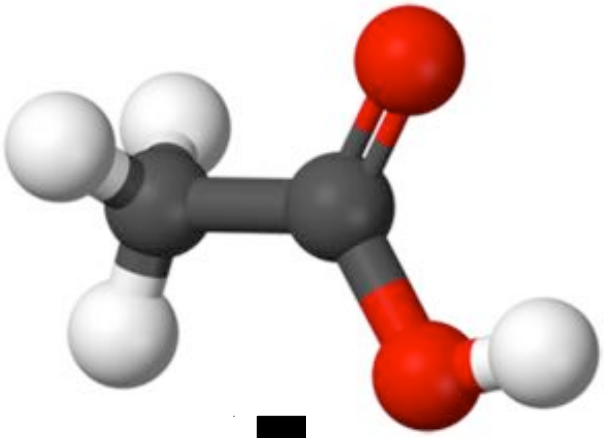
Alexandre Tkatchenko

Physics and Materials Science (PhyMS), University of Luxembourg

alexandre.tkatchenko@uni.lu

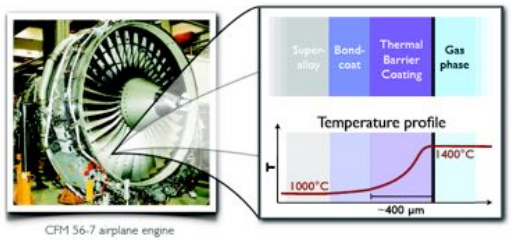
IPAM, Sept 23, 2019

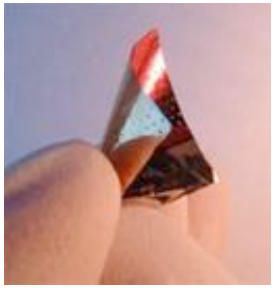
Quantum physics/chemistry today



DFT
MP2
CCSD(T)
...

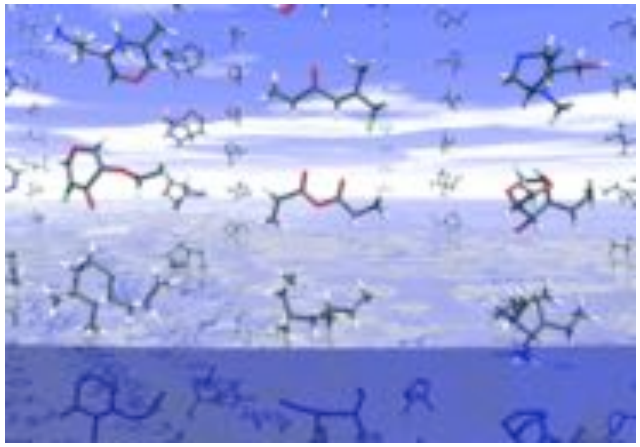
$$\hat{\mathcal{H}}(R_1, Z_1, ..., R_N, Z_N)\tilde{\Psi} = E\tilde{\Psi}$$





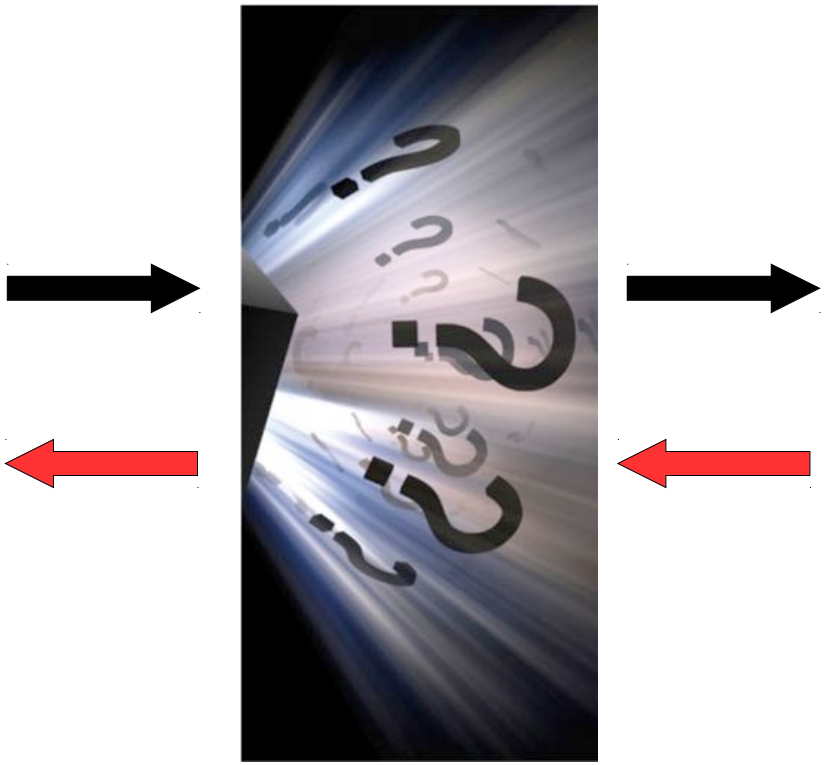
Properties: Energy, polarizability, HOMO, LUMO, …
Dynamics: Thermal properties, spectroscopy, ...

Quantum physics/chemistry tomorrow?



Training data:
molecular properties

ML

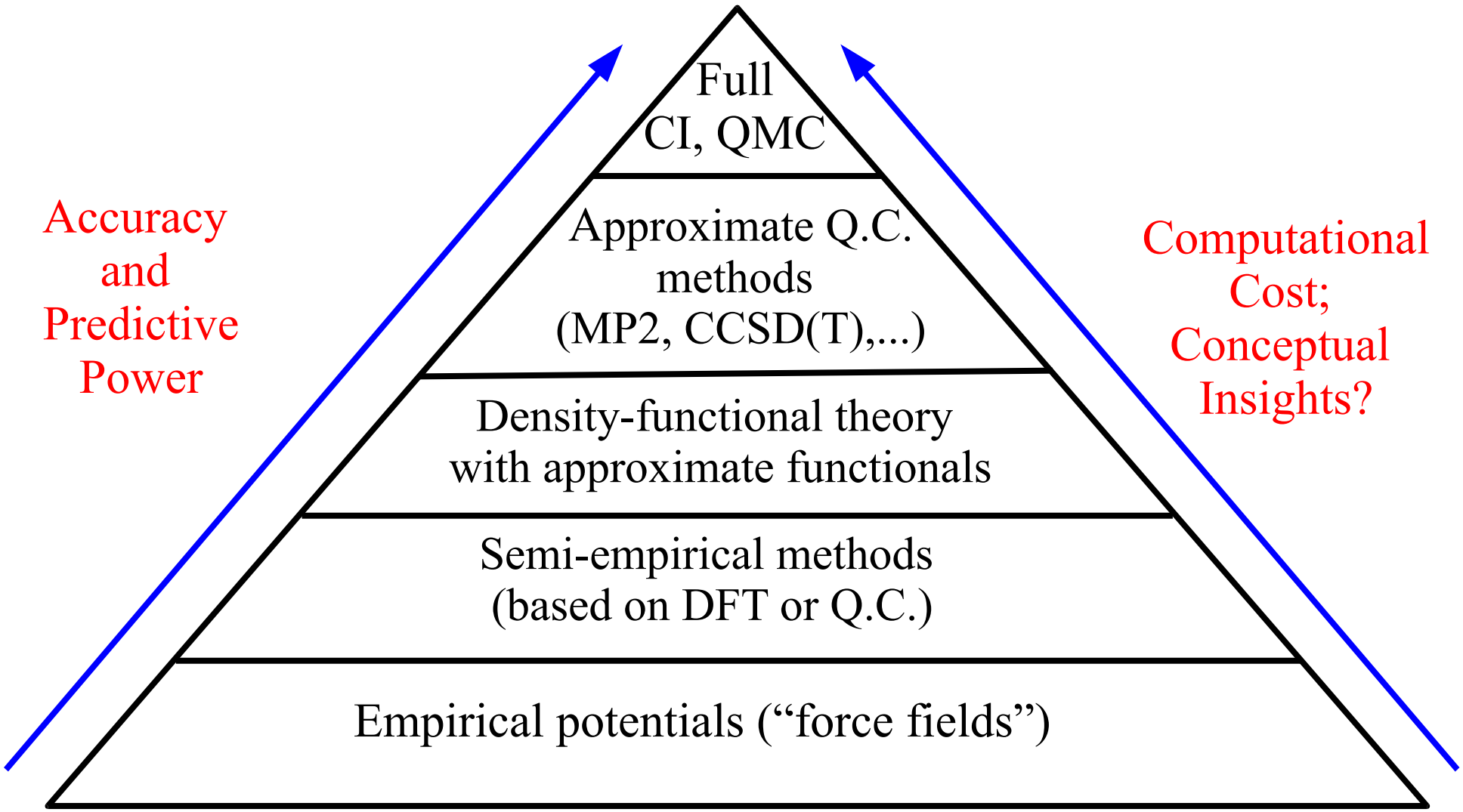


Insights:

- Structure of chemical space
- Reactivity trends, aromaticity, “new” chemistry
- Molecular design through multi-property optimization
- ...

Current state of the art of atomistic modeling

$$\hat{\mathcal{H}}\Psi = E\Psi$$



Computational Complexity of $\hat{\mathcal{H}}\Psi = E\Psi$

Hierarchy of **numerical approximations** to Schrödinger's equation:

Abrv.	Method	Runtime
CISDTQ	Configuration Interaction (up to quadruples)	$O(N^{10})$
CCSD(T)	Coupled Cluster (up to perturbative triples)	$O(N^{7})$
CISD	Configuration Interaction (up to doubles)	$O(N^{6})$
MP2	Møller-Plesset second order perturbation theory	$O(N^{5})$
HF	Hartree-Fock	$O(N^{4})$
DFT	Density Functional Theory (Kohn-Sham)	$O(N^{3-4})$
TB	Tight Binding	$O(N^{3})$
MM	Molecular Mechanics	$O(N^{1-2})$

N = system size

Generalized gradient approximation: ~ 100k citations
Perdew, Burke, Ernzerhof, ***Phys. Rev. Lett.*** (1996)

Density-functional approximations (DFA): *Solid starting point*

$$\Psi(\vec{r}_1, \vec{r}_2, \vec{r}_3, \ldots, \vec{r}_n)$$

Exchange and **Correlation** functionals

LDA **PBE**

$$n(\vec{r}); \vec{\nabla} n(\vec{r}); \nabla^2 n(\vec{r}); \ldots$$

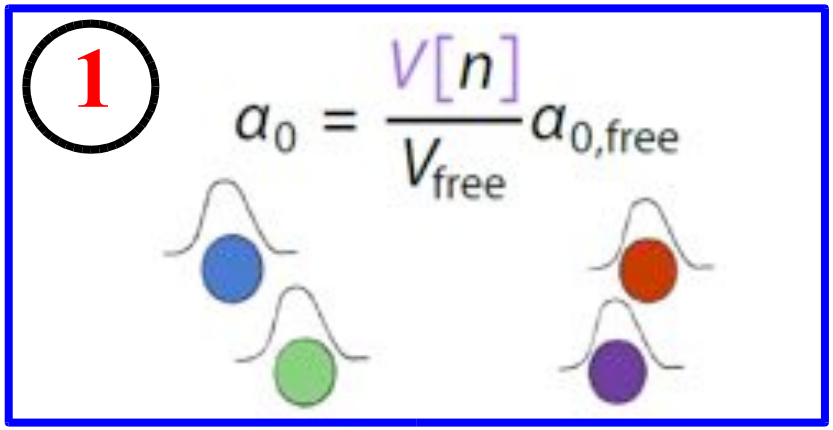
TPSS, SCAN

- ✗ **Self-interaction error**
- ✗ **Lack of long-range correlation (van der Waals interactions)**

$$n(\vec{r})$$

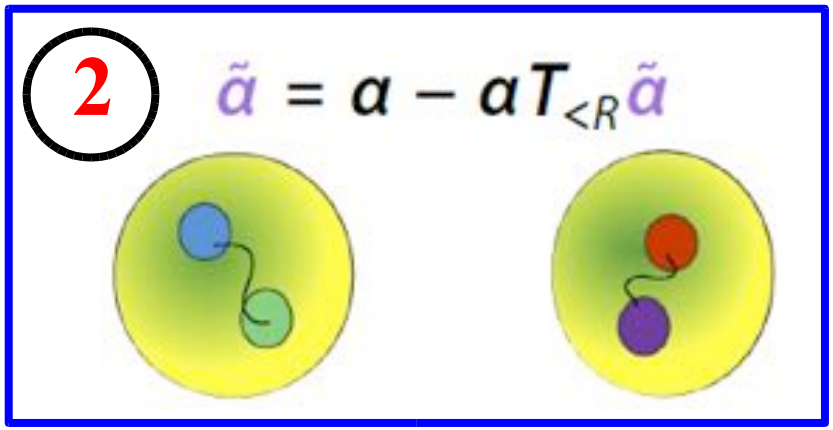
Modeling Real Materials: DFT+MBD Method

A. Tkatchenko and M. Scheffler, ***Phys. Rev. Lett.*** (2009)



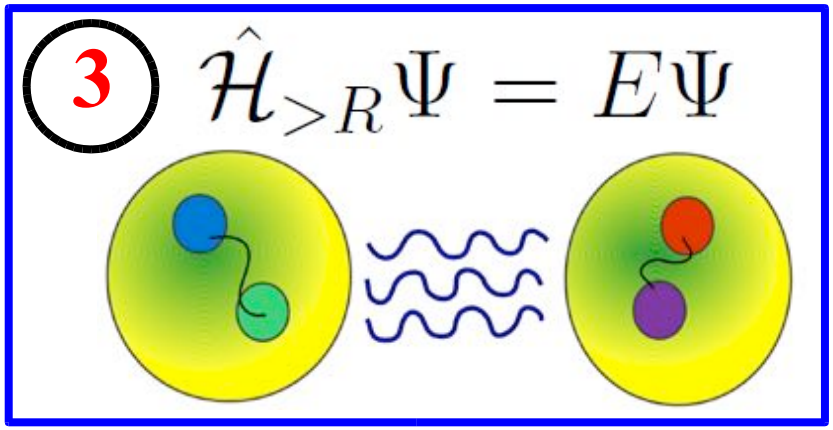
Valence electrons projected to oscillators *(Tkatchenko-Scheffler)*

A. Tkatchenko, R. A. DiStasio Jr., R. Car, M. Scheffler, ***Phys. Rev. Lett.*** (2012)

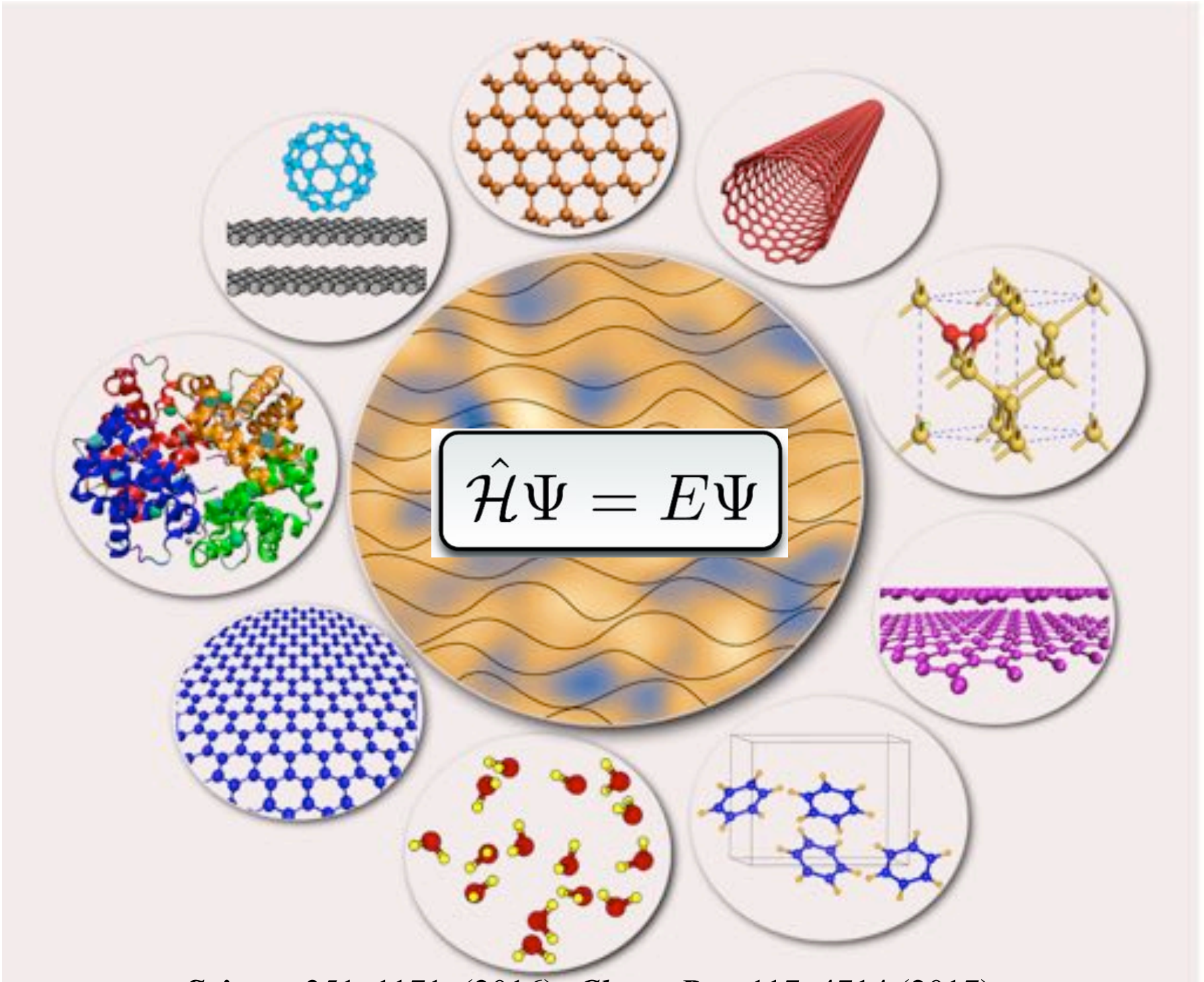


Dyson-like short-range electrodynamic screening

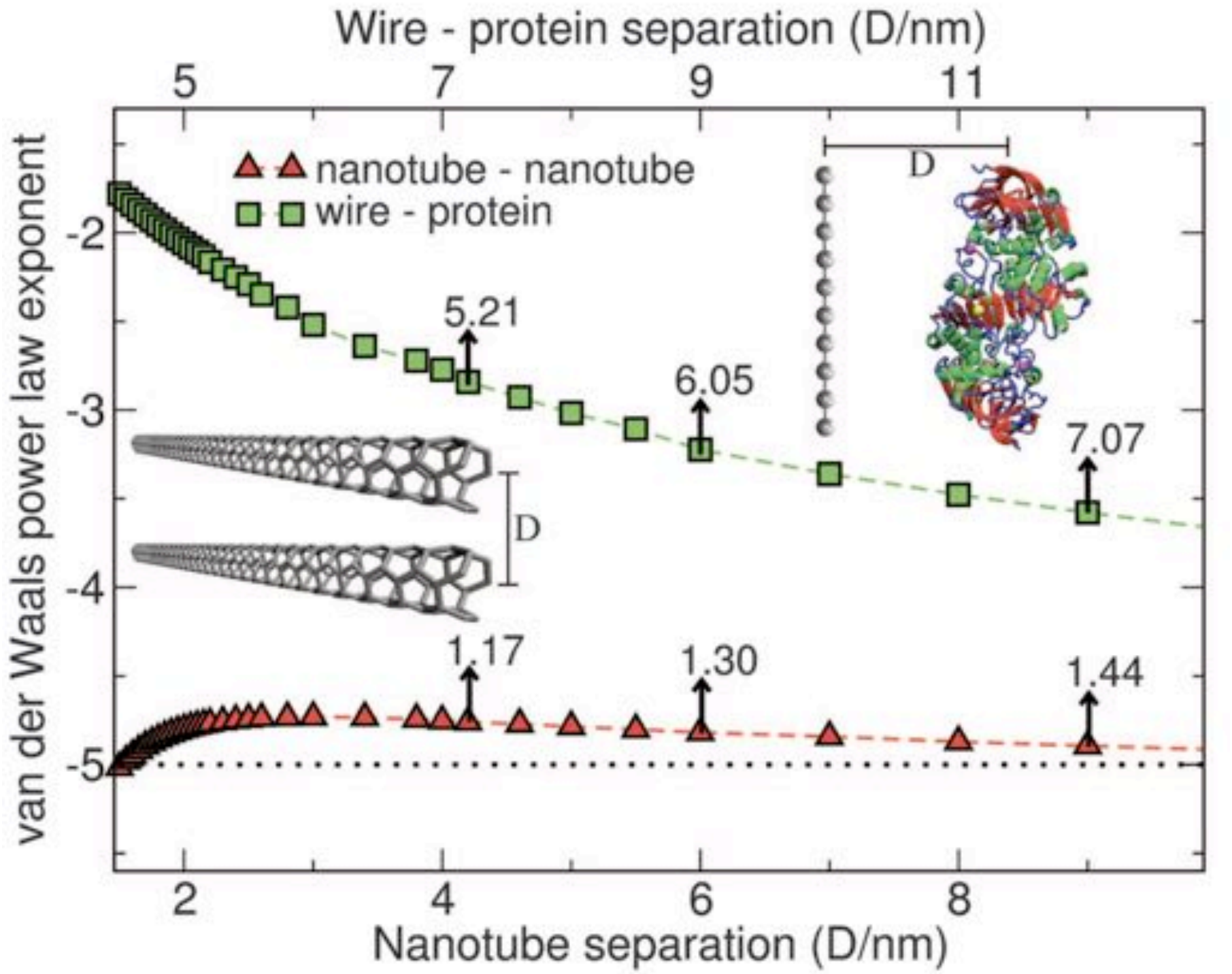
A. Ambrosetti, R. A. Distasio Jr., A. M. Reilly, A. Tkatchenko, ***J. Chem. Phys.*** (2014)



Long-range correlation energy calculated using SE

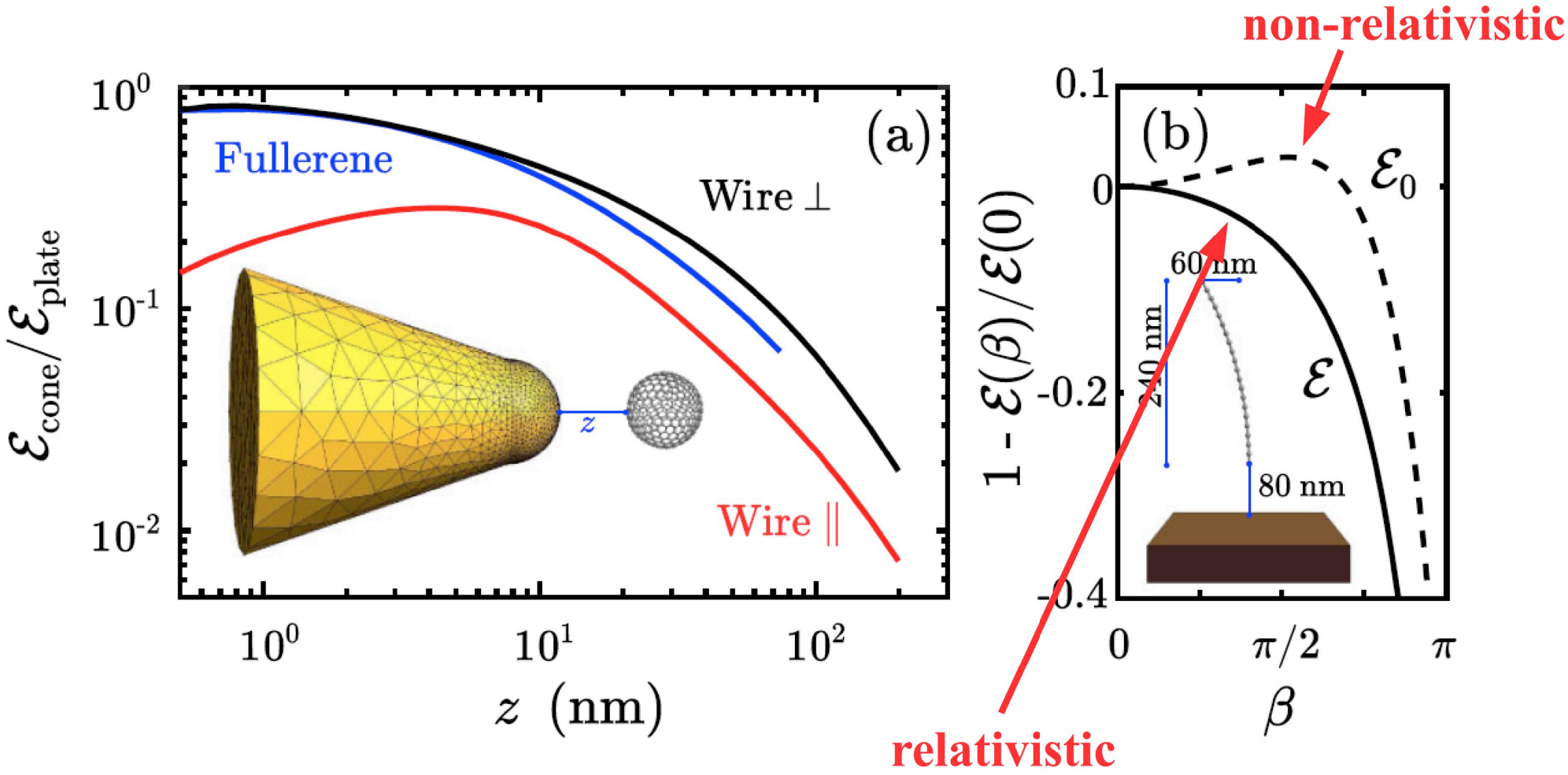


Science 351, 1171 (2016); ***Chem. Rev.*** 117, 4714 (2017).



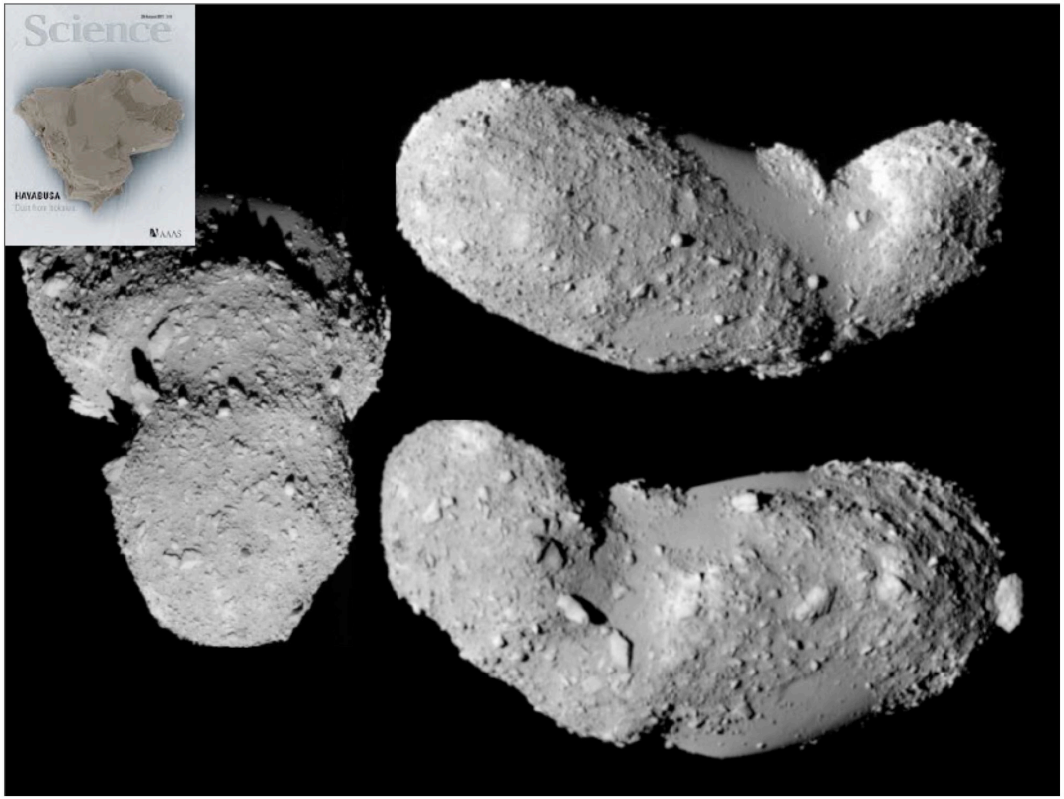
Ambrosetti, Ferri, DiStasio Jr., and Tkatchenko, ***Science*** (2016).

Relativistic and Quantum Effects at Mesoscopic Scales



P. S. Venkataram, J. Hermann, A. Tkatchenko, and A. W. Rodriguez,
Phys. Rev. Lett. 118, 266802 (2017).

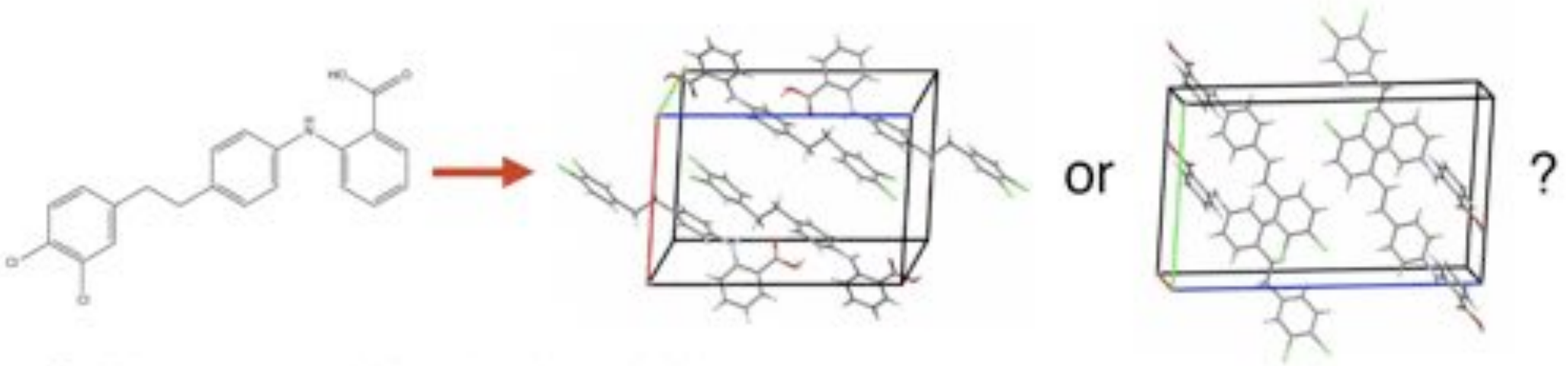
QM effects at macroscopic scales:



B. Rozitis *et al.*, ***Nature* 512,** 174 (2014).

Predictive Modeling of Real Materials: What state-of-the-art QM can do for us today?

Predictive Modeling of Real Materials



- Solids composed of molecular moieties
- Held together by intermolecular interactions
- Different crystal-packing motifs (polymorphs) possible
- Energy difference between polymorphs **~ 1 - 4 kJ/mol**
  - ↳ **~ 1 - 2 % of lattice energy**

Polymorphs can exhibit completely different

- Kinetic stabilities
- Solubilities
- Densities
- Vibrational Spectra (THz)
- NMR chemical shifts
- Melting Points
- Conductivities
- Refractive Indices
- Vapor pressure
- Elastic constants
- Heat capacities
- …

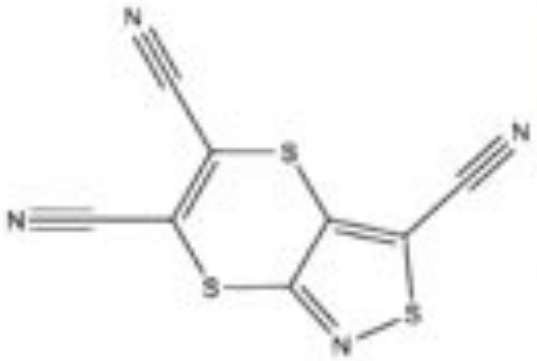
Pharmaceuticals

Organic electronics

Explosives

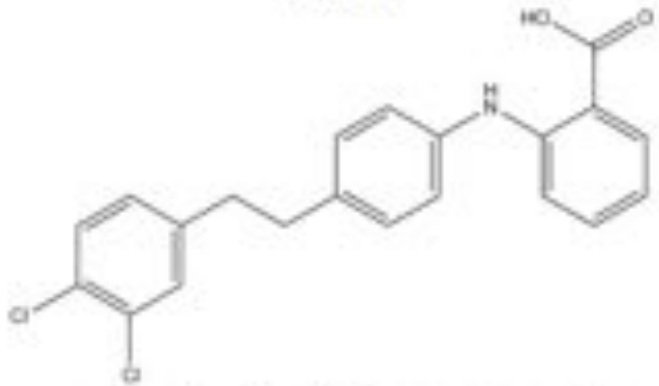
Targets of Cambridge Blind Test 2016

XXII



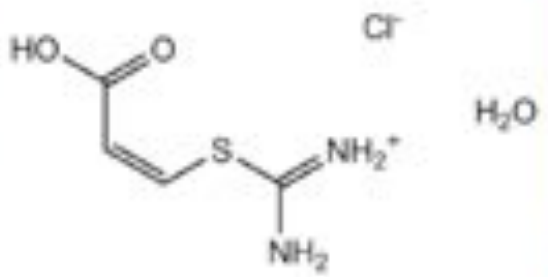
rigid molecule

XXIII



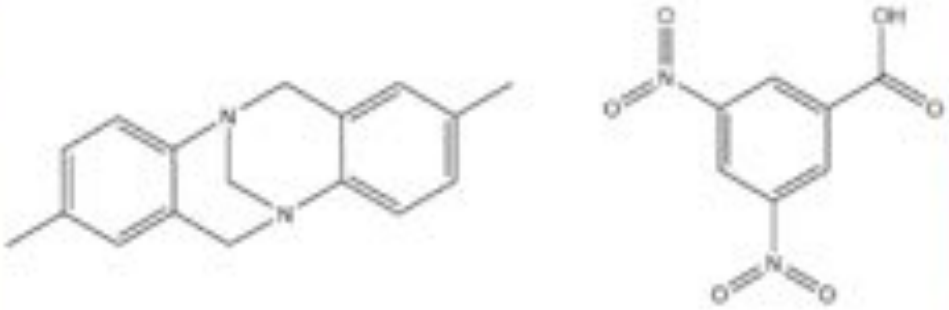
partially flexible molecule,
polymorphic system

XXIV



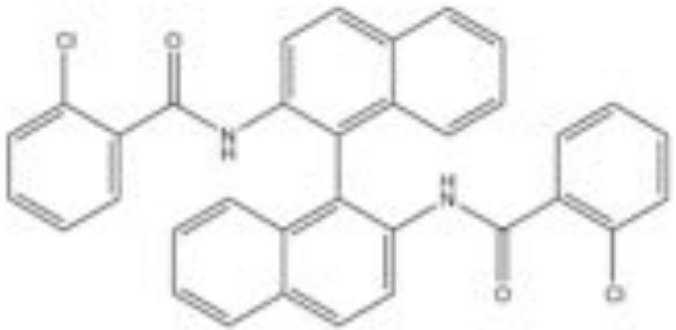
salt

XXV



multiple partially flexible molecules
as co-crystal

XXVI

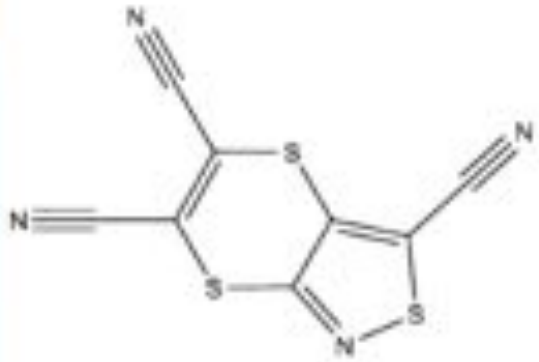


molecule with 4-8 internal
degrees of freedom

A. M. Reilly *et al*., ***Acta Cryst. B*72**, 439 (2016).

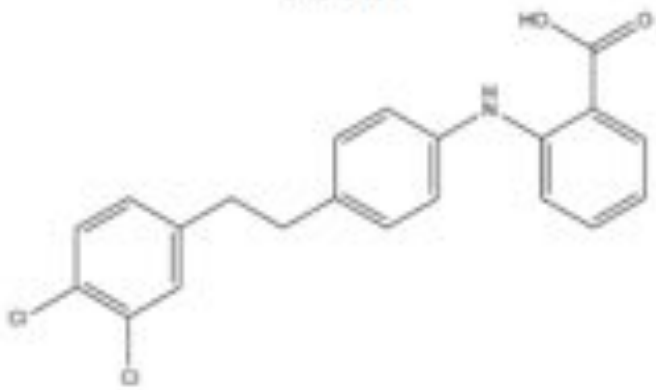
Targets of Cambridge Blind Test 2016

XXII



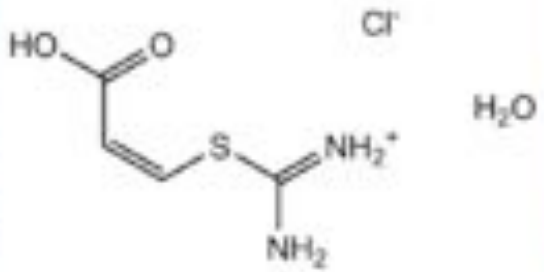
57 %

XXIII



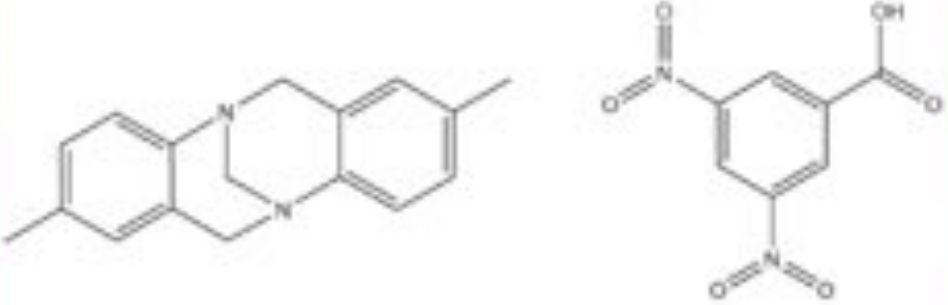
28 %

XXIV



13 %

XXV



36 %

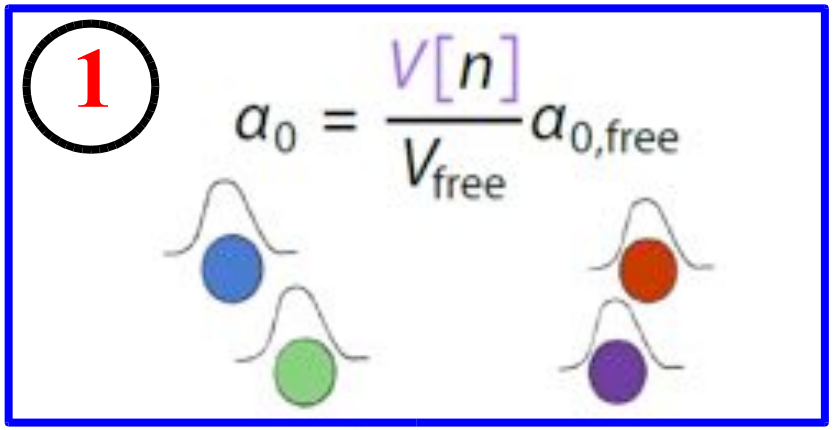
XXVI

25 %

A. M. Reilly *et al.*, ***Acta Cryst. B*72**, 439 (2016).

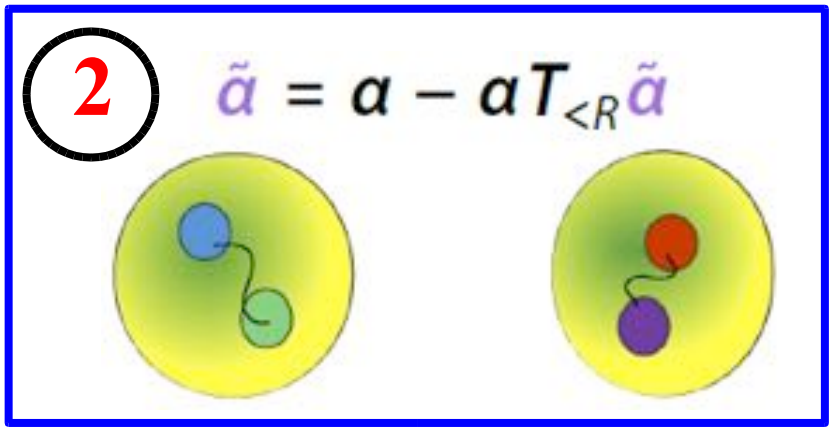
Modeling Real Materials: DFT+MBD Method

A. Tkatchenko and M. Scheffler, ***Phys. Rev. Lett.*** (2009)



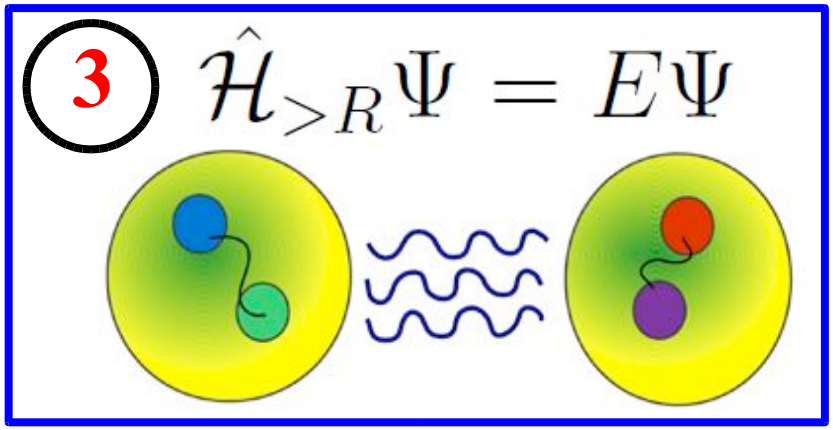
Valence electrons projected to oscillators *(Tkatchenko-Scheffler)*

A. Tkatchenko, R. A. DiStasio Jr., R. Car, M. Scheffler, ***Phys. Rev. Lett.*** (2012)



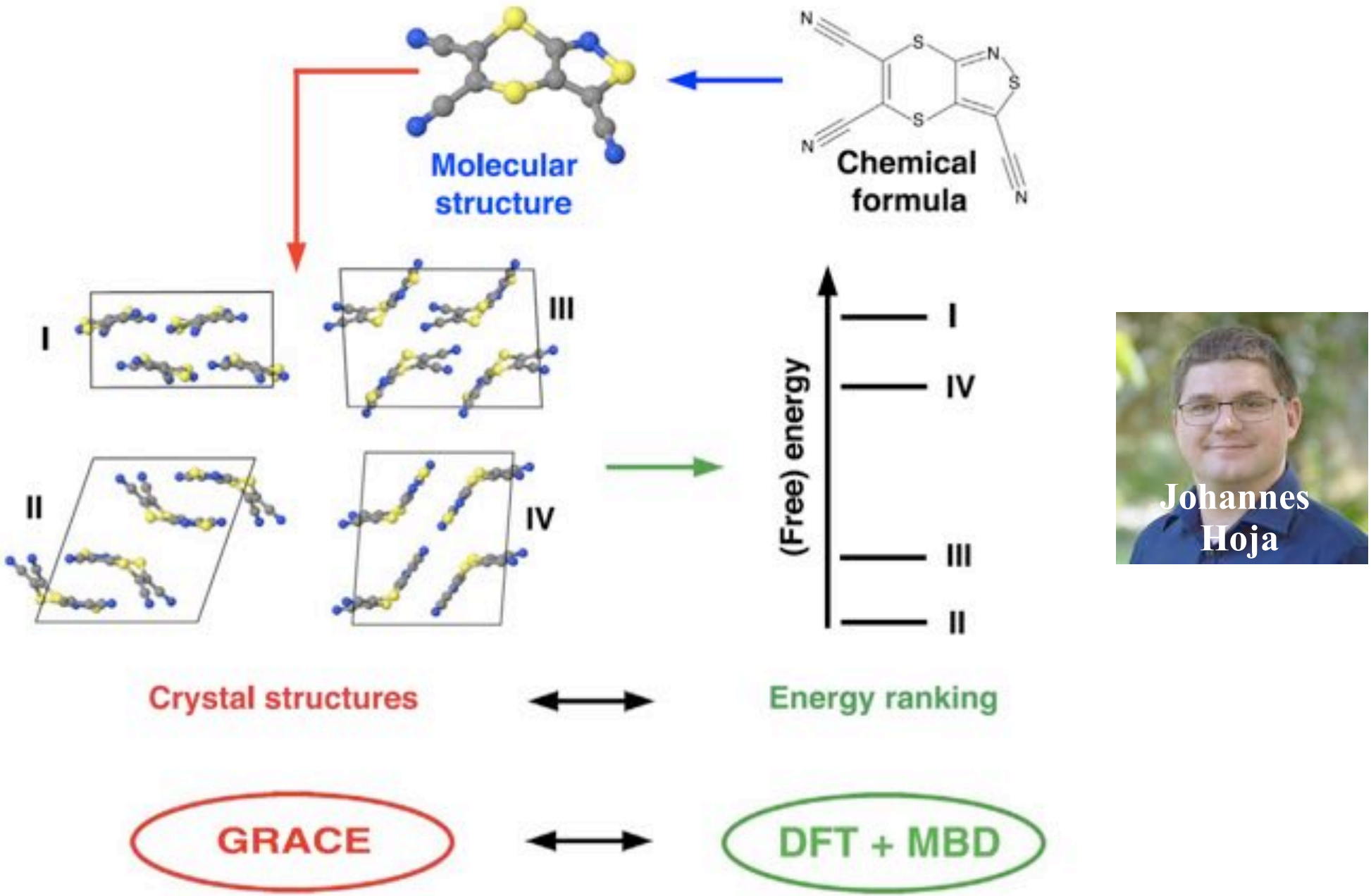
Dyson-like short-range electrodynamic screening

A. Ambrosetti, R. A. Distasio Jr., A. M. Reilly, A. Tkatchenko, ***J. Chem. Phys.*** (2014)



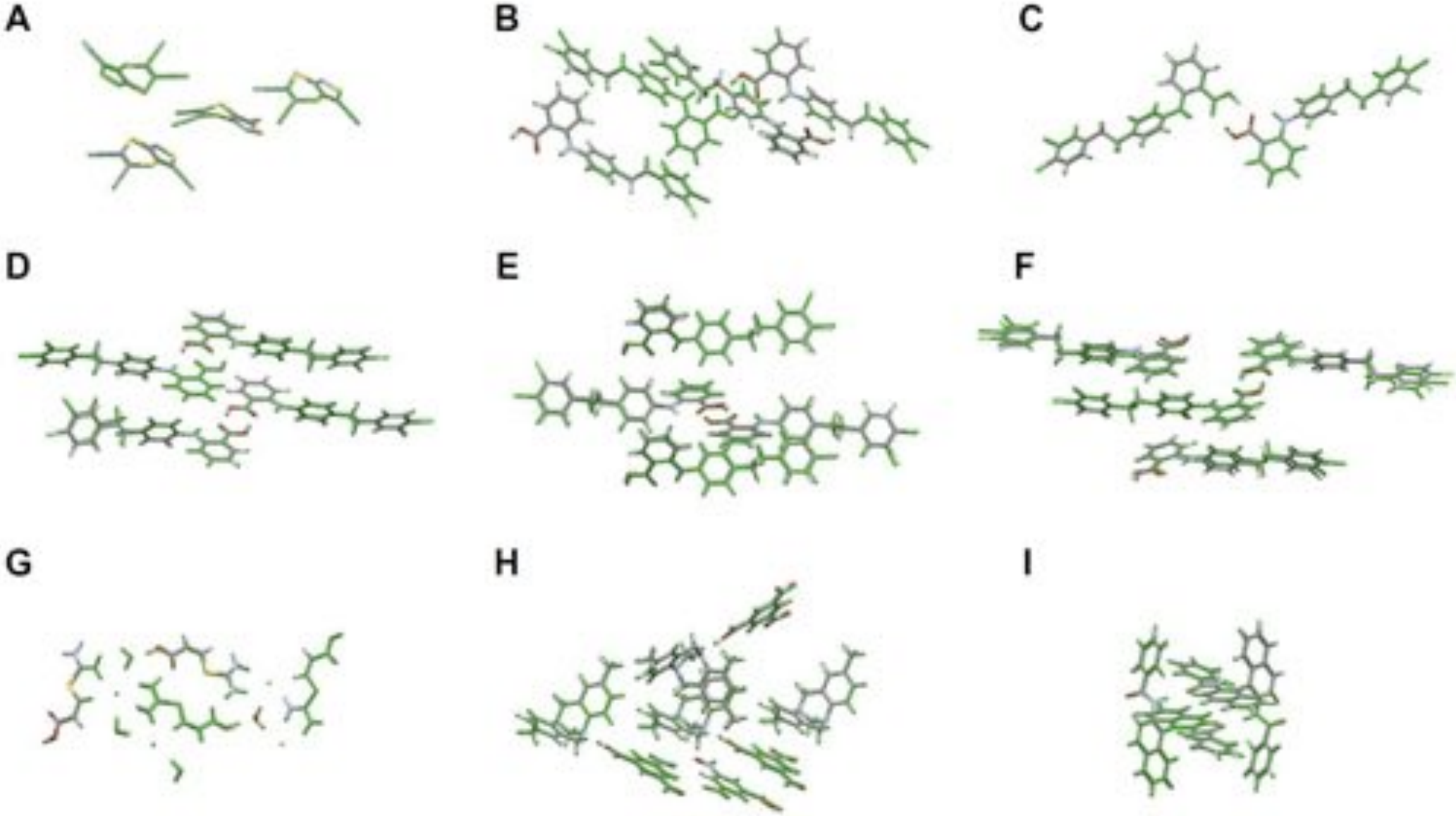
Long-range correlation energy calculated using SE

Blind Crystal Structure Prediction with DFT+MBD



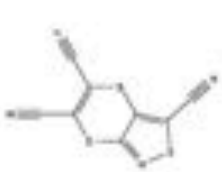
J. Hoja, H.Y. Ko, M.A. Neumann, R. Car, R.A. DiStasio Jr., and A. Tkatchenko, ***Science Adv.*** 5, eaau3338 (2019).

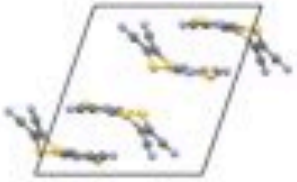
DFT+MBD yields quantitative agreement with experiment and new predictions

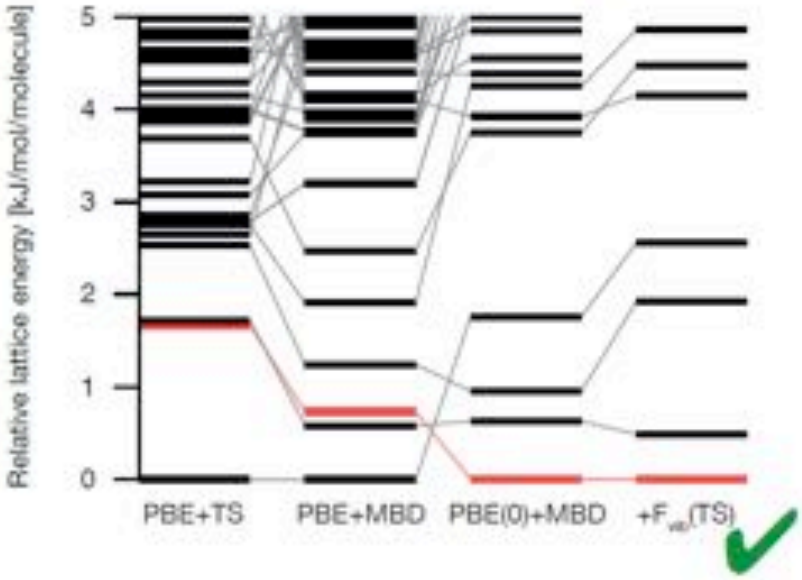


J. Hoja, H.Y. Ko, M.A. Neumann, R. Car, R.A. DiStasio Jr., and A. Tkatchenko, ***Science Adv.*** 5, eaau3338 (2019).

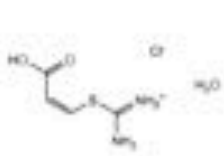
XXII

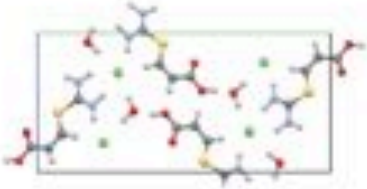


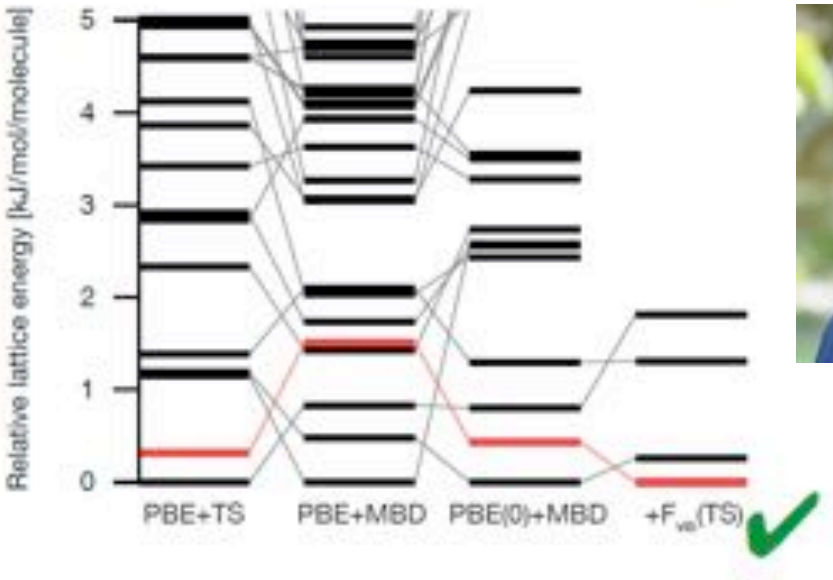




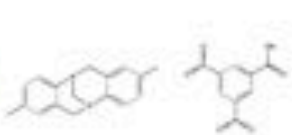
XXIV

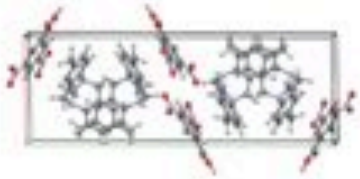


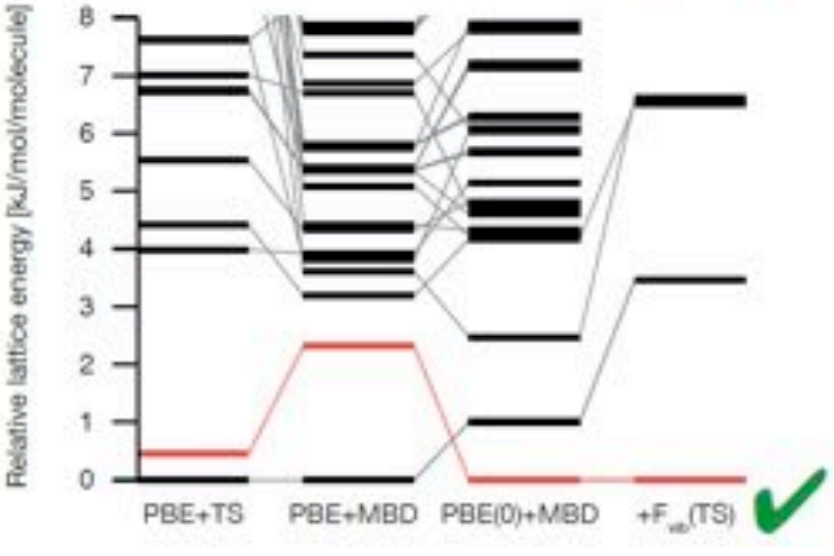




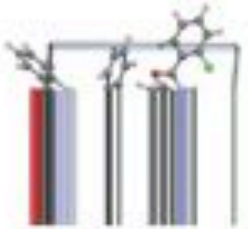
XXV

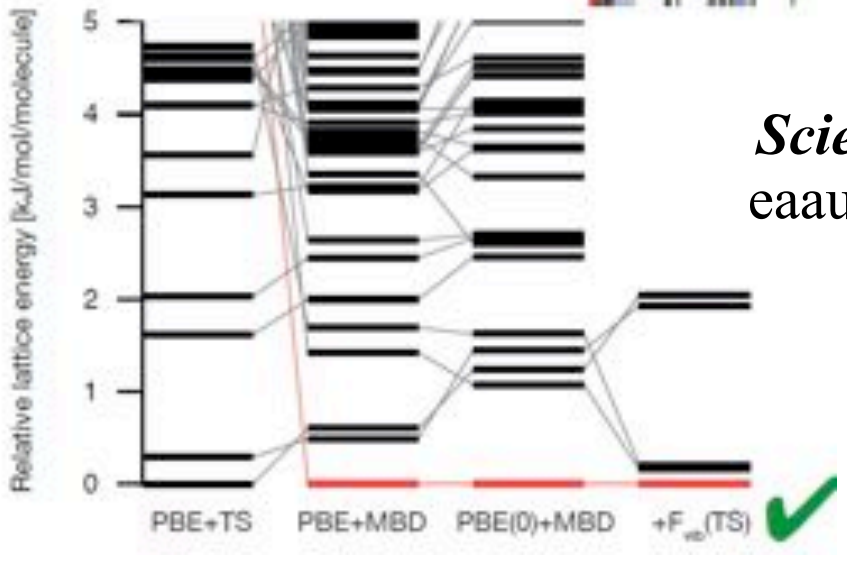






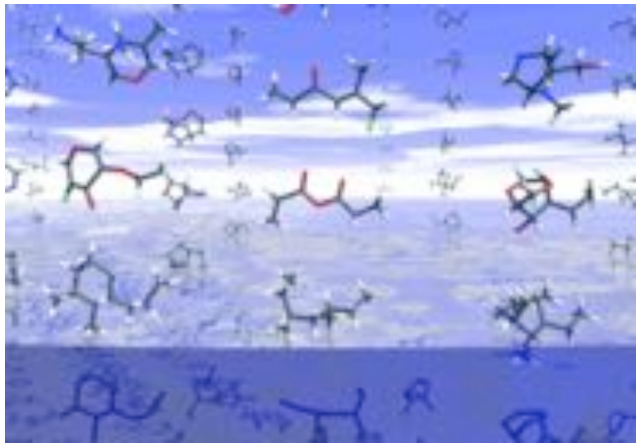
XXVI





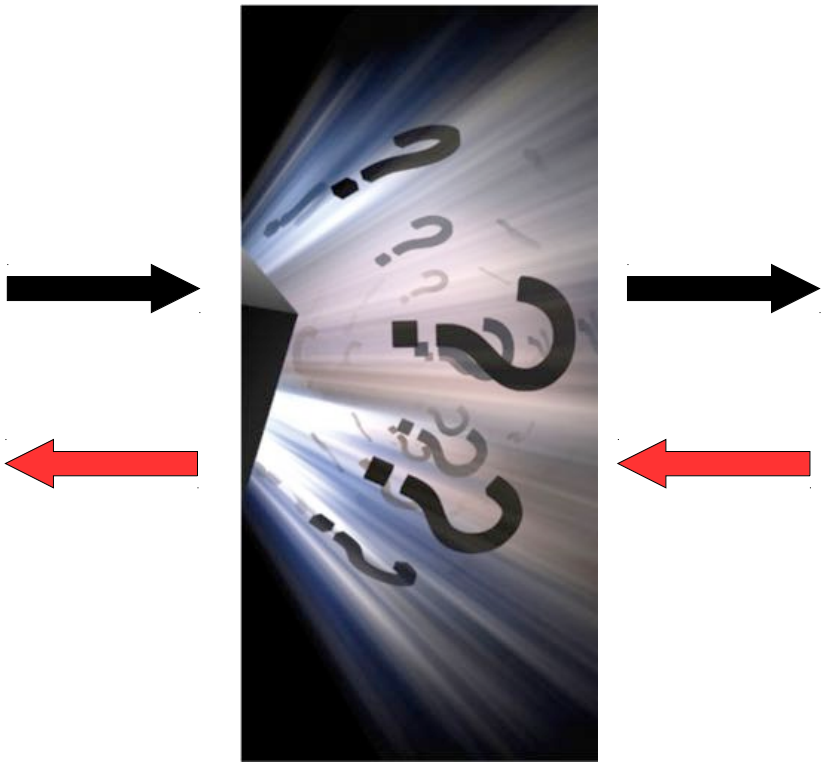
Science Adv. 5, eaau3338 (2019)

Quantum physics/chemistry tomorrow?



Training data:
molecular properties

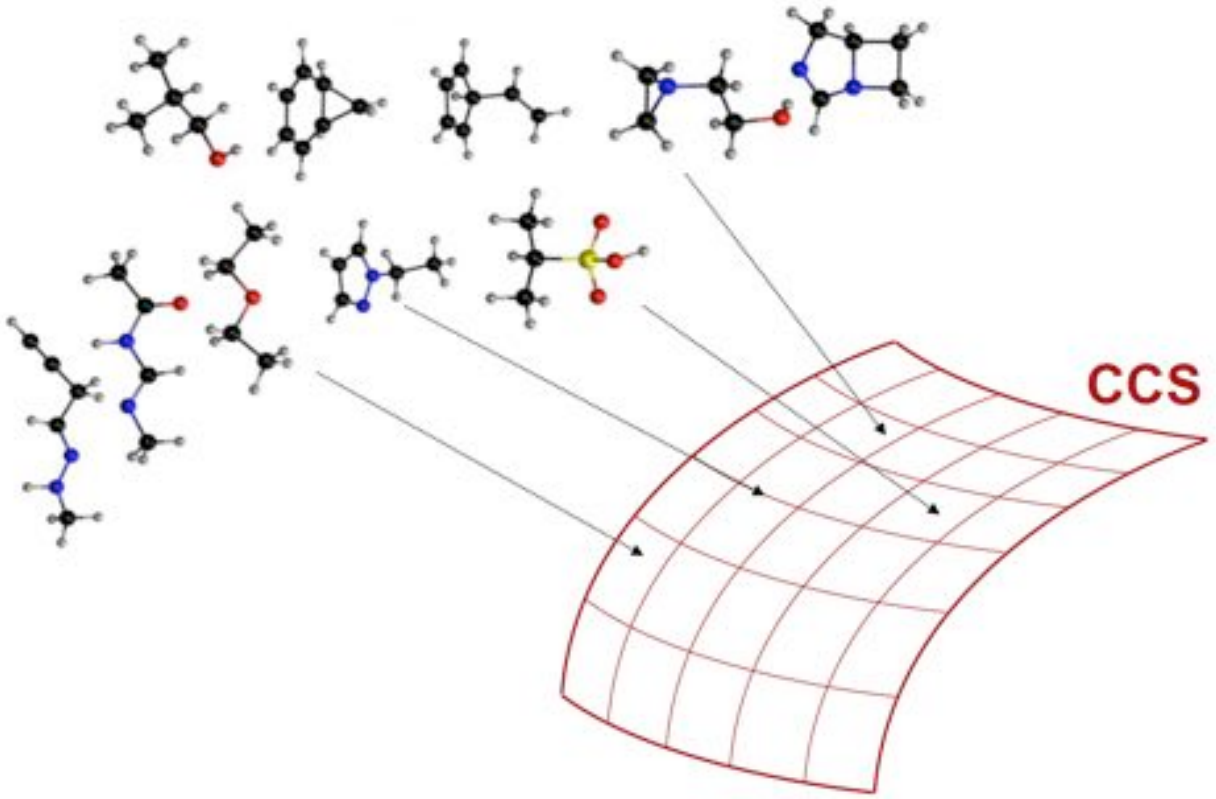
ML



Insights:

- Structure of chemical space
- Reactivity trends, aromaticity, “new” chemistry
- Molecular design through multi-property optimization
- ...

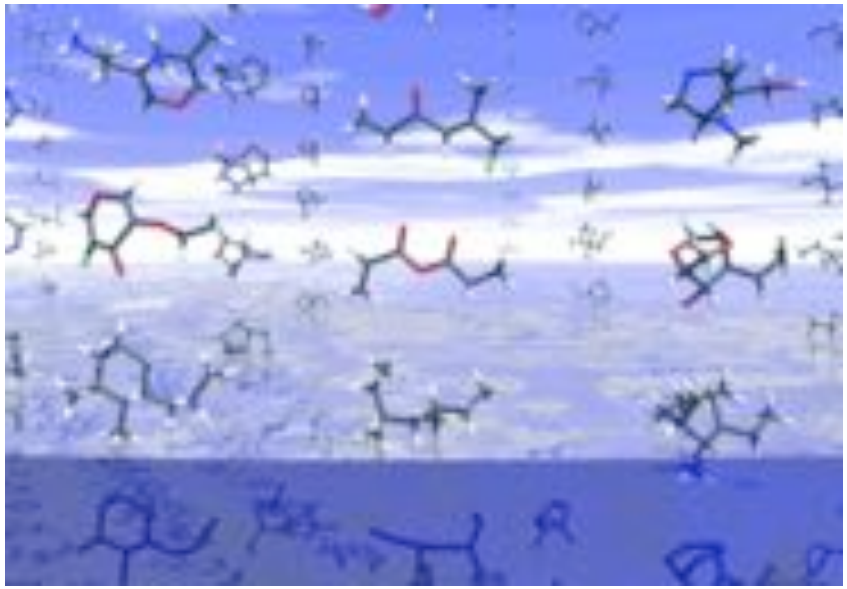
Molecular Quantum Chemical Space



- Graph theory: combinatorial explosion
- At least 10^{60} small drug candidate molecules
- Finding needles in a haystack

$\{\boldsymbol{R}_i, Z_i\}$ maps to $\{P_1, P_2, P_3, P_4, \ldots\}$

Machine learning for molecular big data

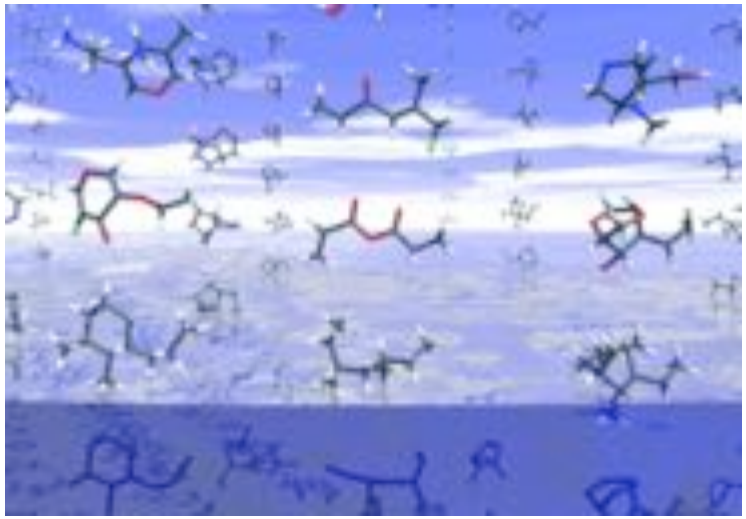


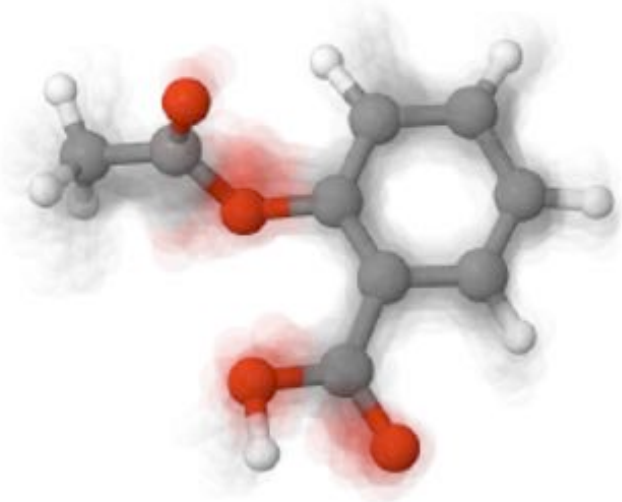
$\{\boldsymbol{R}_i, Z_i\}$ maps to $\{P_1, P_2, P_3, P_4, \ldots\}$

- Descriptor: what's a good representation of a molecule?
- Metric: how to define distance between two molecules?
- Data selection: Which molecules to use for training?
- Properties: which set of properties uniquely defines a molecule?
- Degrees of freedom: composition vs. conformation

Can we obtain insights into Chemical Compound Space (CCS) ?

Molecular Data in this Talk





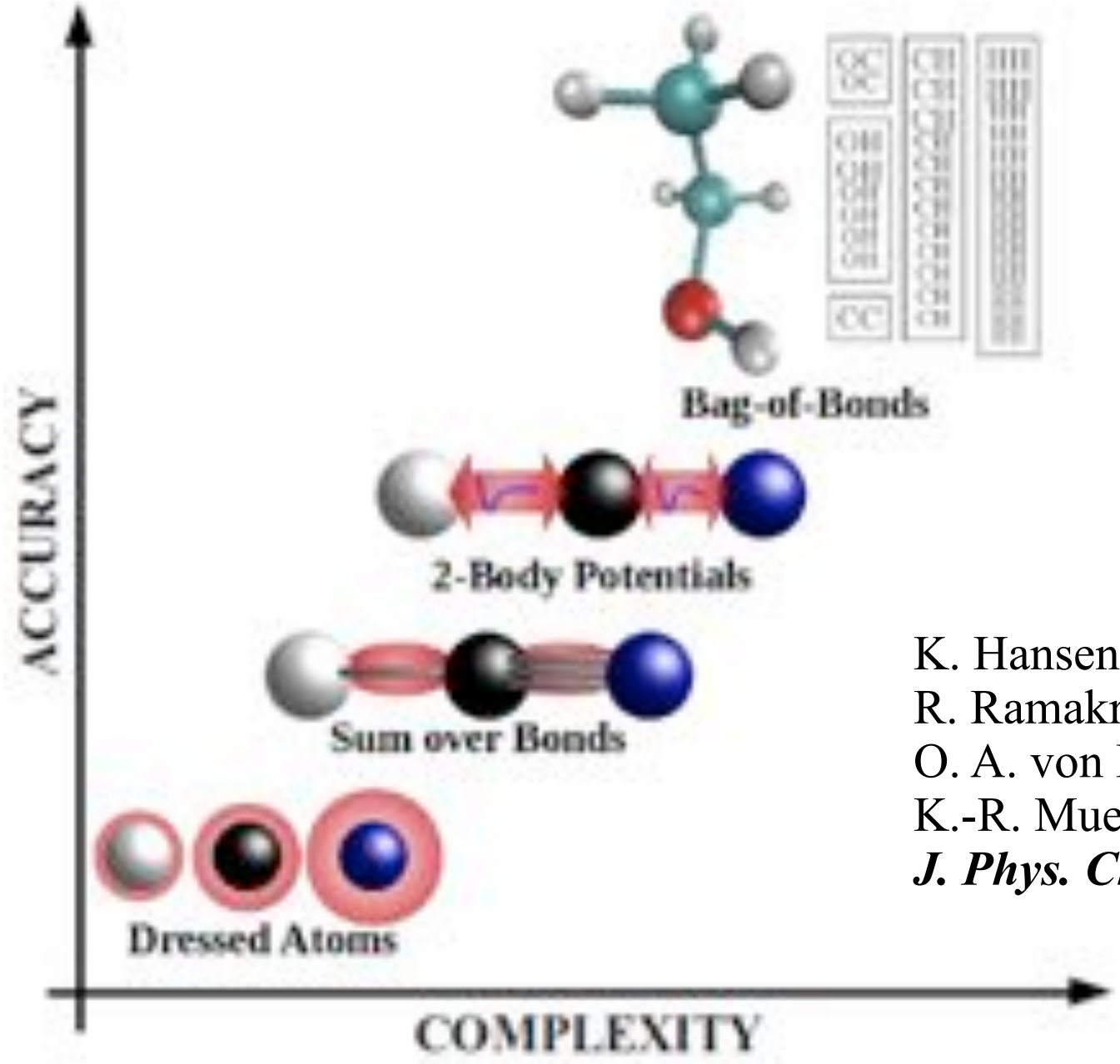
GDB mol graphs: J. L. Reymond (U. Bern)
http://gdb.unibe.ch/downloads/

QM7/QM9 datasets: Hybrid DFT calculations by von Lilienfeld's group (Sci. Data 2014) and my group (PRL 2012).

MD17/ISO17 datasets: Molecular dynamics trajectories from my group (DFT and CCSD(T) levels)

www.quantum-machine.org

Predicting Molecular Properties: The Importance of *Physical Baselines*



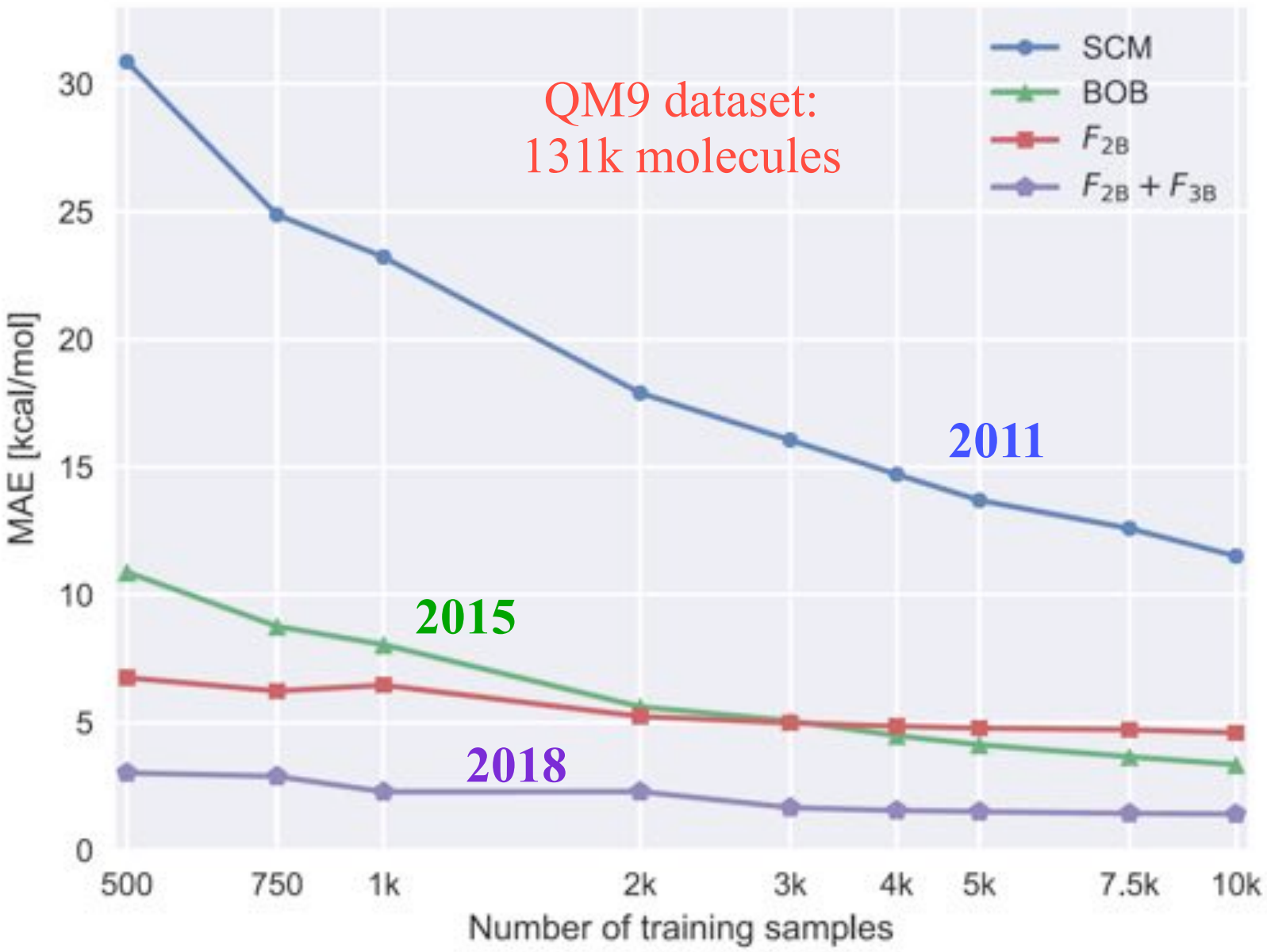
K. Hansen, F. Biegler,
R. Ramakrishnan, W. Pronobis,
O. A. von Lilienfeld,
K.-R. Mueller, and A. Tkatchenko,
J. Phys. Chem. Lett. 6, 2326 (2015).

Predicting Molecular Properties: QM7 dataset

model	MAE [kcal/mol]
dressed atoms	15.1
sum-overbonds	9.9
Lennard-Jones potential	8.7
polynomial pot. ($n = 6$)	5.6
polynomial pot. ($n = 10$)	3.9
polynomial pot. ($n = 18$)	3.0
Bag of Bonds ($p = 2$, Gaussian)	4.5
Bag of Bonds ($p = 1$, Laplacian)	1.5
Coulomb matrix ($p = 2$, Gaussian)[17]	10.0
Coulomb matrix ($p = 1$, Laplacian)[16]	4.3
2+3body many-body expansion	**0.8**

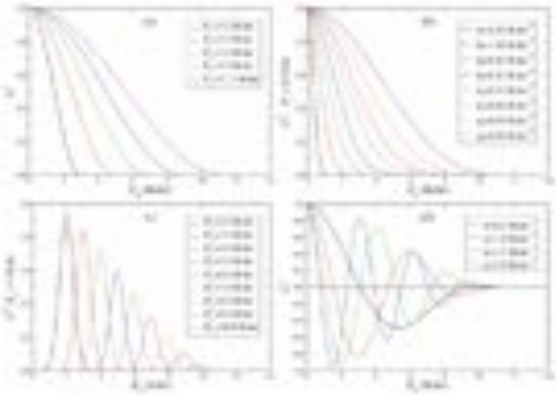
W. Pronobis, A. Tkatchenko, and K.-R. Mueller, ***J. Chem. Theory Comput.*** (2018).

QM9 dataset: Evolution from Coulomb Matrix to Many-Body Representation



W. Pronobis, A. Tkatchenko, and K.-R. Mueller, ***J. Chem. Theory Comput.*** (2018).

Zoo of Descriptors for Molecules and Solids



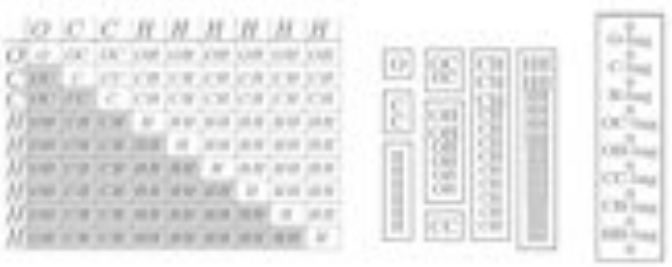
Atom-centered
symmetry functions
(Behler et al. 2007)

$$M_{ij} = \begin{cases} 0.5 Z_i^{2.4} & \text{for } i = j \\ \frac{Z_i Z_j}{d_{ij}} & \text{for } i \neq j \end{cases}$$

Coulomb matrix
(Rupp et al. 2012)

$$\{Z_i, \mathbf{R}_i\}$$

$$\{Z_i, d_{ij}\}$$



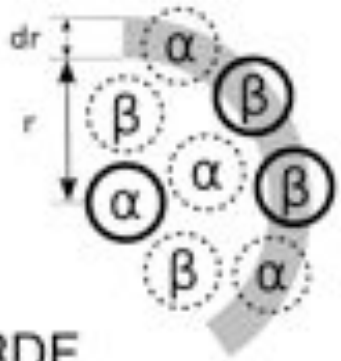
Bag of bonds
(Hansen et al. 2015)

$$k(\rho, \rho') = \int d\hat{R} \left| \rho(\mathbf{r}) \rho'(\hat{R}\mathbf{r}) \right|^n$$

SOAP
(Bartók et al. 2013)

$$x_{ij} = \begin{cases} 0.5 Z_i^{2.4} & \text{if } i = j \\ \frac{Z_i Z_j}{\phi(r_i, r_j)} & \text{if } i \neq j \end{cases}$$

Sine matrix
(Faber et al. 2015)

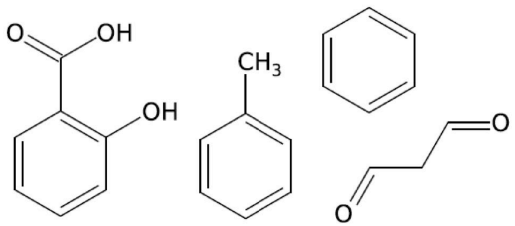


PRDF
(Schütt et al, 2014)

Slide from K. Schuett

Learning the Representation: Deep Tensor Neural Networks (DTNN)

Input: Atomic numbers and interatomic distances



$$Z = [\, Z_1 \quad Z_2 \quad \cdots \quad Z_n \,]$$

$$D = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{12} & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nn} \end{bmatrix}$$

Embedding of based on atom types

$$\mathbf{x}_i^{(0)} = \mathbf{x}_{Z_i} \in \mathbb{R}^d$$

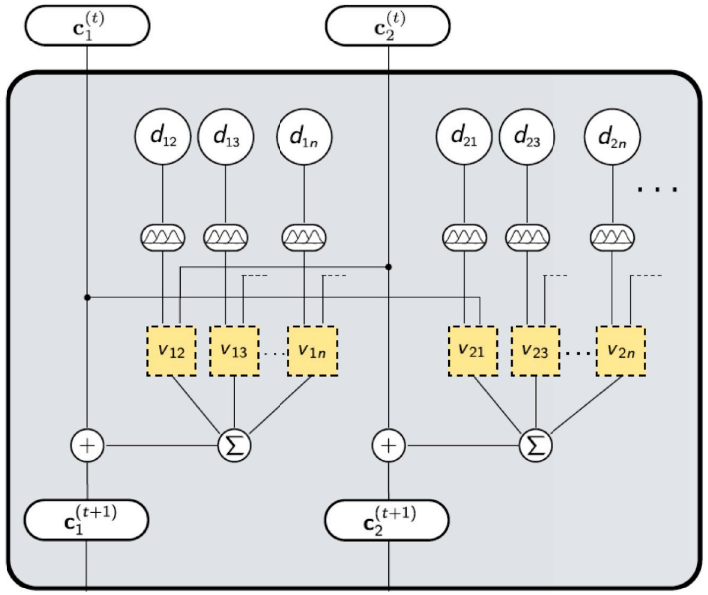
Add interaction with environment using $t = 1 \ldots T$ sequential refinements $\mathbf{v}_i^{(t)}$

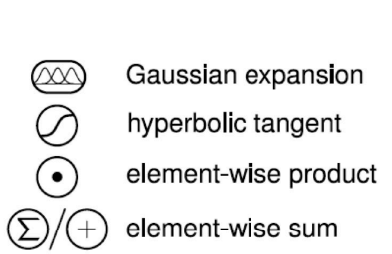
$$\mathbf{x}_i^{(t+1)} = \mathbf{x}_i^{(t)} + \mathbf{v}_i^{(t)} \left(\mathbf{x}_1^{(t)}, \ldots \mathbf{x}_{n_{\text{atoms}}}^{(t)}, d_{i1}, \ldots, d_{in_{\text{atoms}}} \right)$$

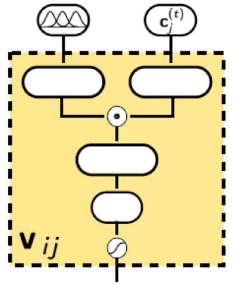
Prediction via atom-wise contributions:

$$\hat{E} = \sum_{i=1}^{n_{\text{atoms}}} f_{\text{out}}(\mathbf{x}_i^{(T)})$$

$$\hat{\mathcal{H}}\Psi = E\Psi$$





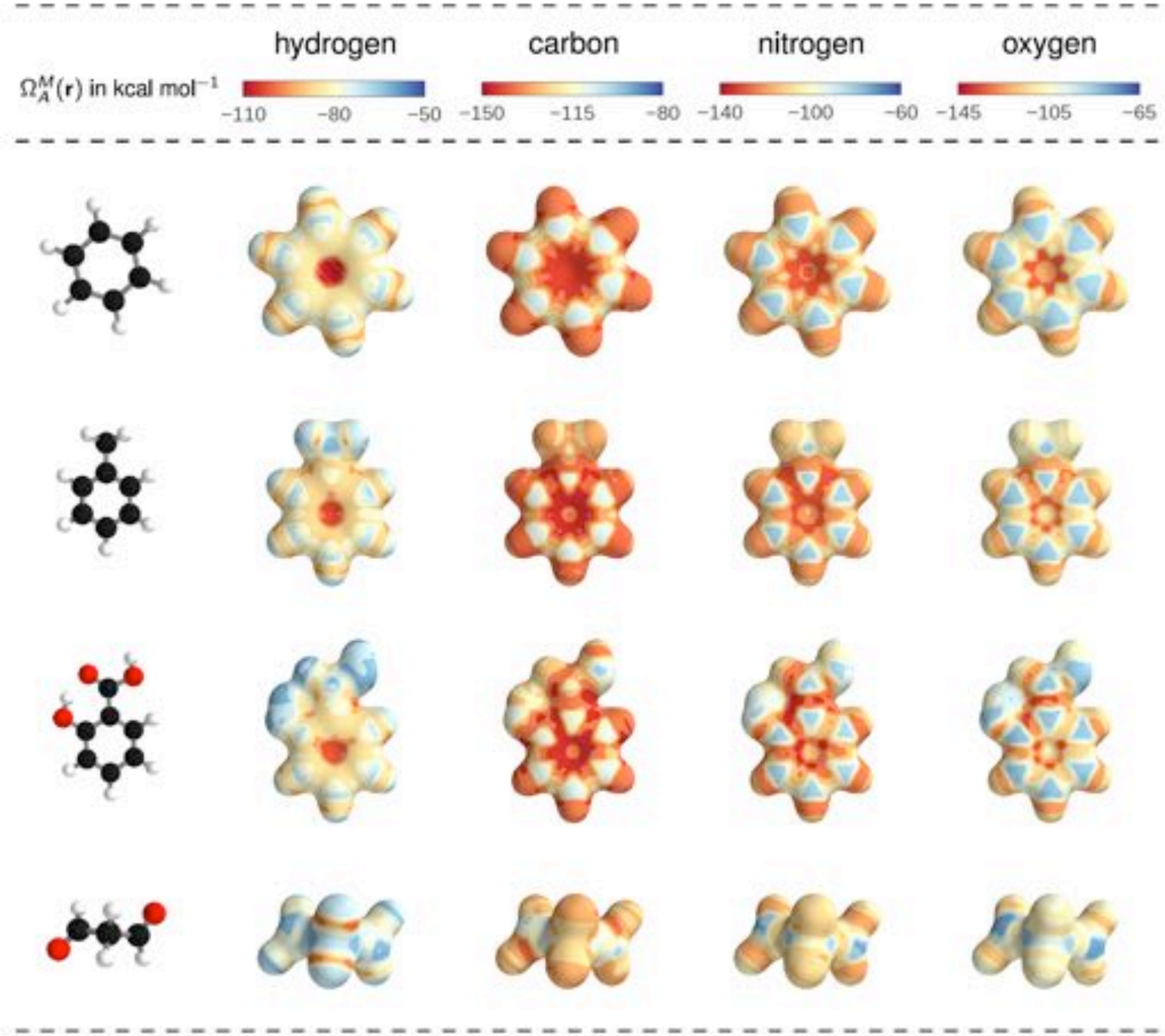


$$\tanh\left(W^{fc}((W^{cf}\mathbf{c}_j + \mathbf{b}^{f_1}) \circ (W^{df}\hat{\mathbf{d}}_{ij} + \mathbf{b}^{f_2}))\right)$$

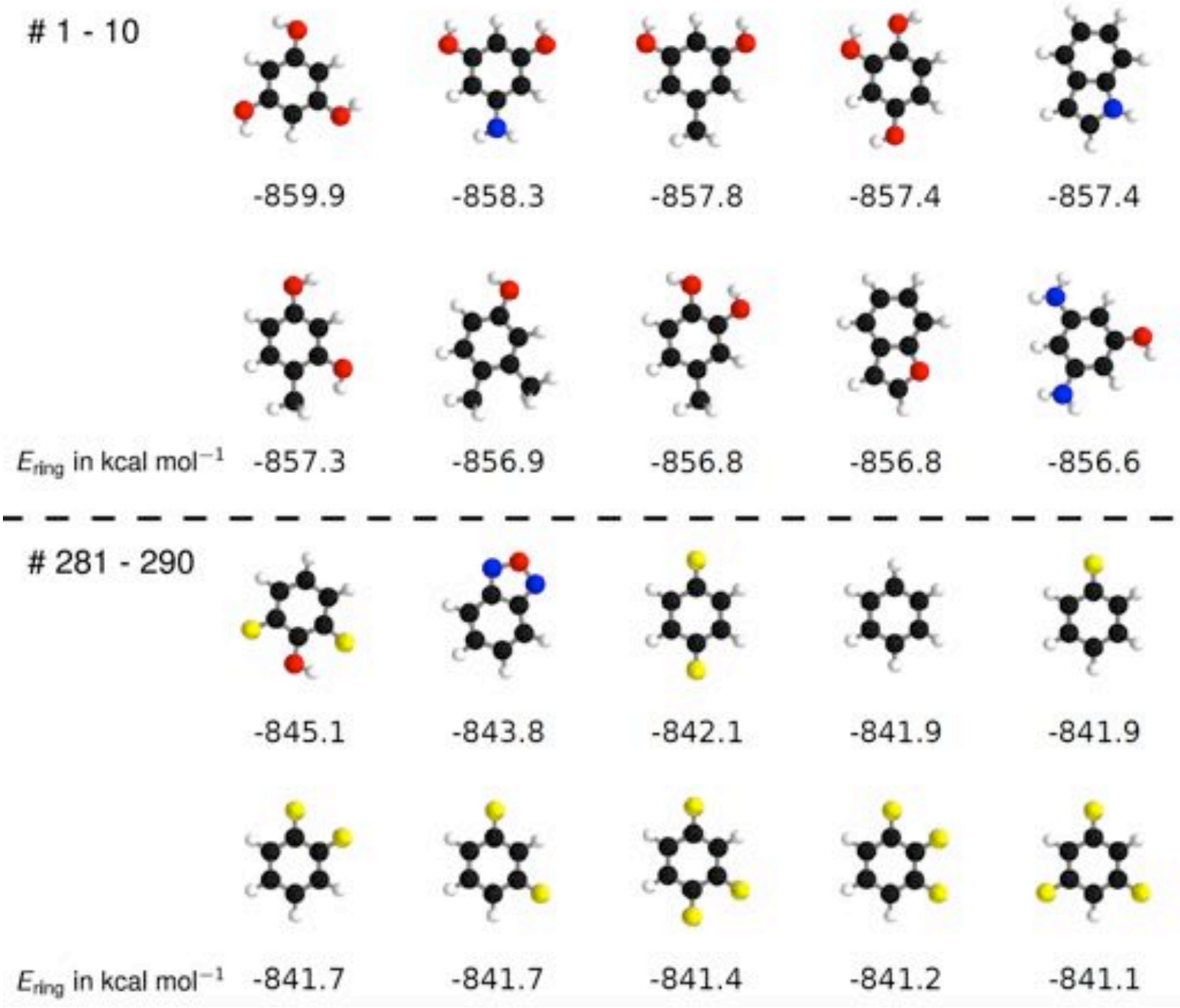
Mean absolute error on QM9: **0.2 kcal/mol**

K. T. Schuett, F. Arbabzadah, S. Chmiela, K.-R. Mueller, and A. Tkatchenko, ***Nature Commun.*** 8, 13890 (2017).

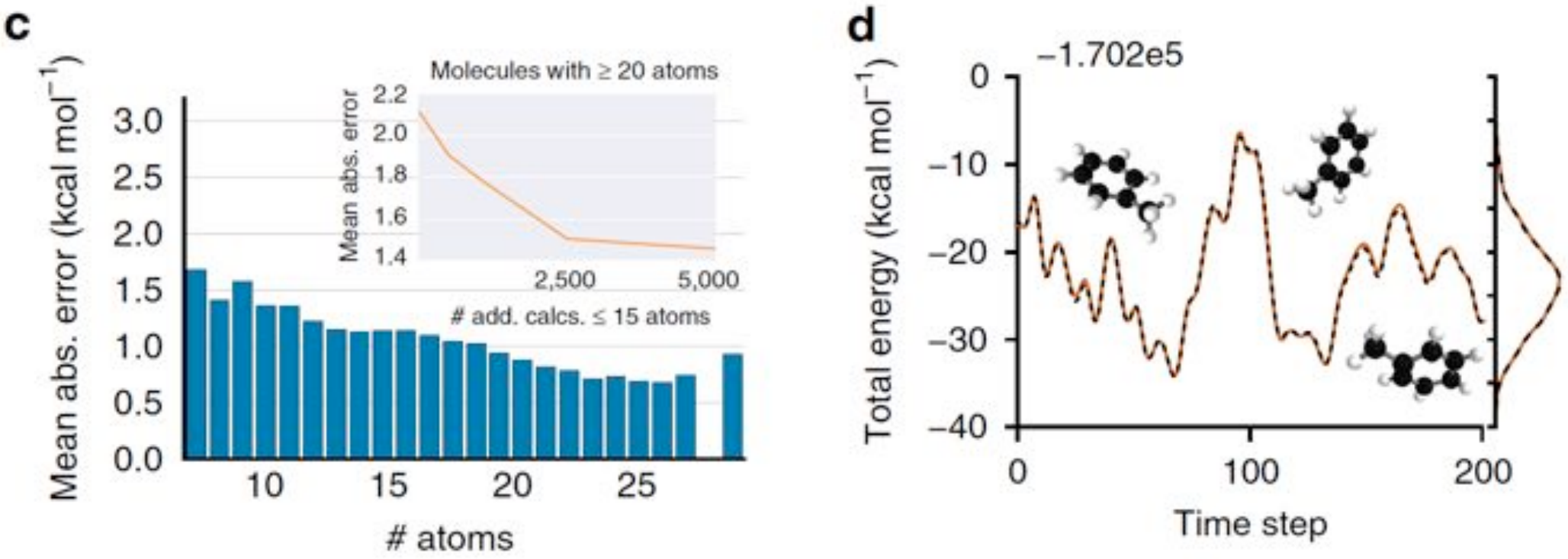
Molecular DTNN: What Did it Learn ?



Quantum Chemical Insights: Aromaticity



Learning Full Chemical Space with DTNN?

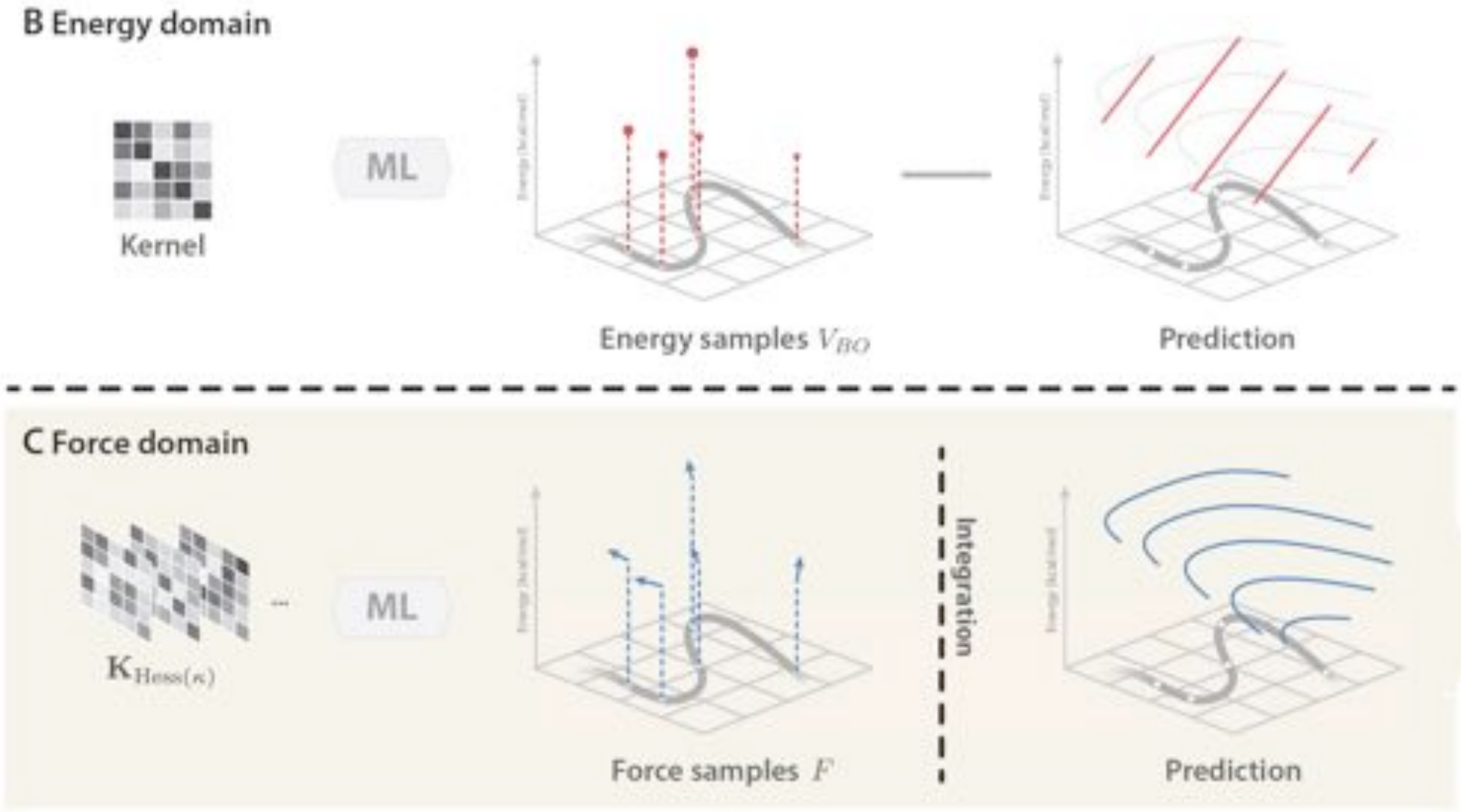


Accurately representing BOTH compositional and conformational degrees of freedom is difficult.

For C7O2H10 isomer and MD data, the error grows to > 1.0 kcal/mol

K. T. Schuett, F. Arbabzadah, S. Chmiela, K.-R. Mueller, and A. Tkatchenko, ***Nature Commun.*** 8, 13890 (2017).

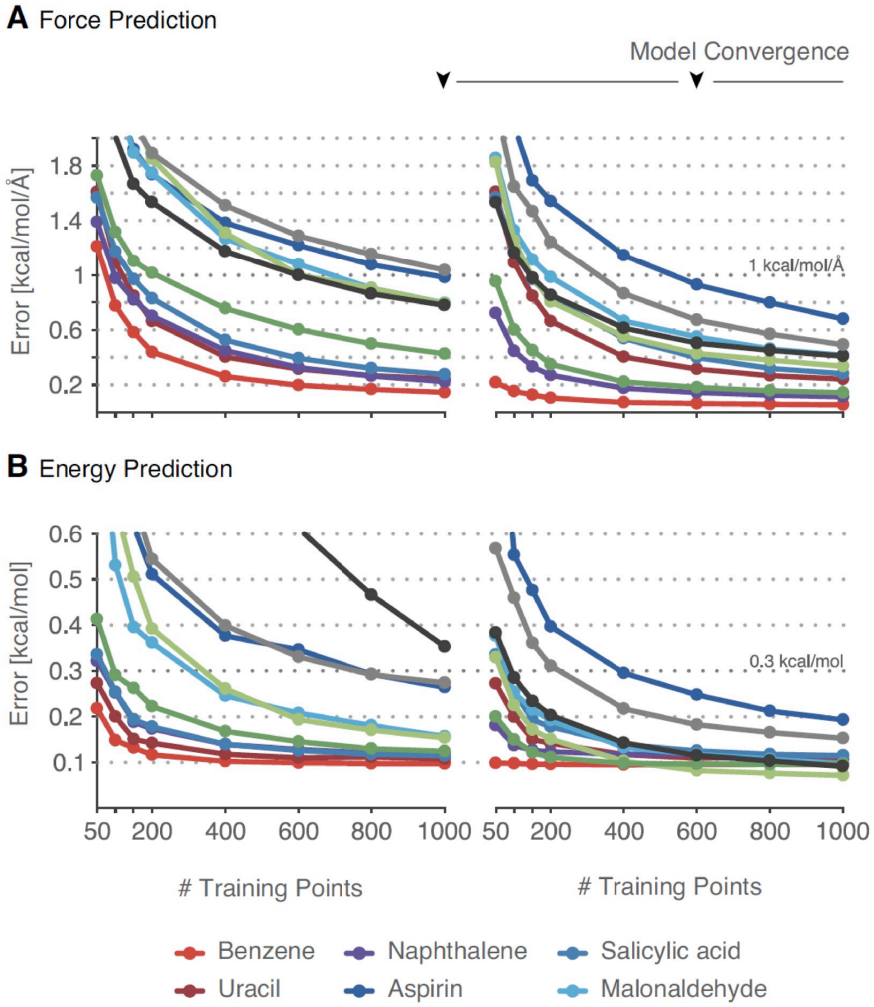
Beating the Hell out of Data:
Gradient-Domain Machine Learning (GDML)



S. Chmiela, A. Tkatchenko, H. Sauceda, I. Poltavsky, K. T. Schuett, K.-R. Mueller,
Science Adv. 3, e1603015 (2017).

Symmetrized Gradient-Domain Machine Learning: Towards Exact Molecular Force Fields

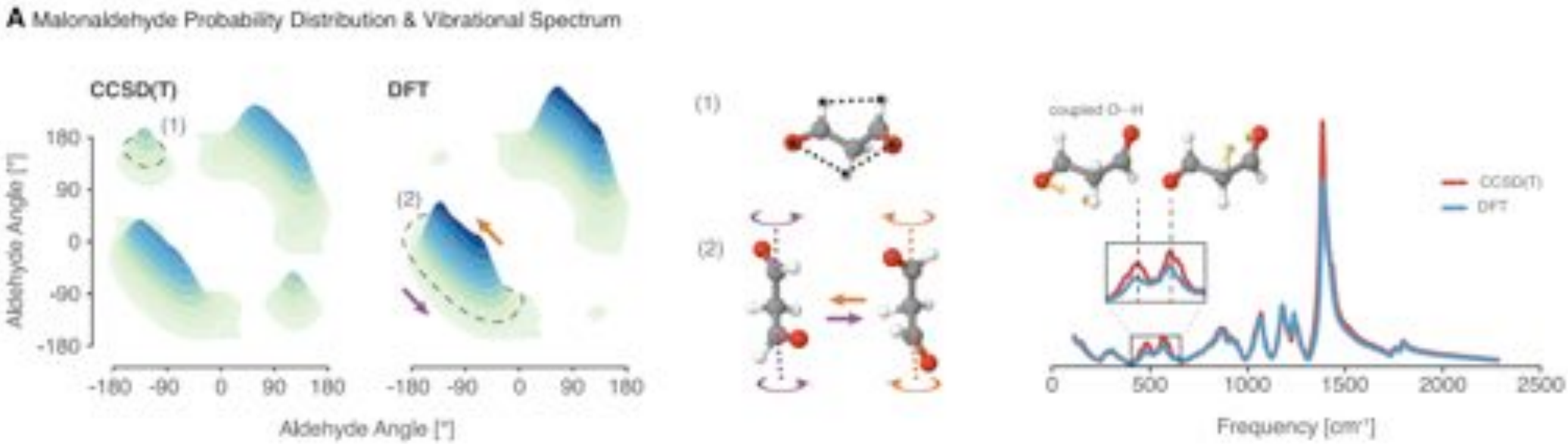
$$\hat{\mathbf{f}}_{\mathbf{F}}(\mathbf{x}) = \sum_{i}^{M} \sum_{l}^{3N} \sum_{q}^{S} (\mathbf{P}_q \alpha_i)_l \frac{\partial}{\partial x_l} \nabla \kappa\left(\mathbf{x}, \mathbf{P}_q \mathbf{x}_i\right)$$



Globally accurate force field from only 100s of conformations

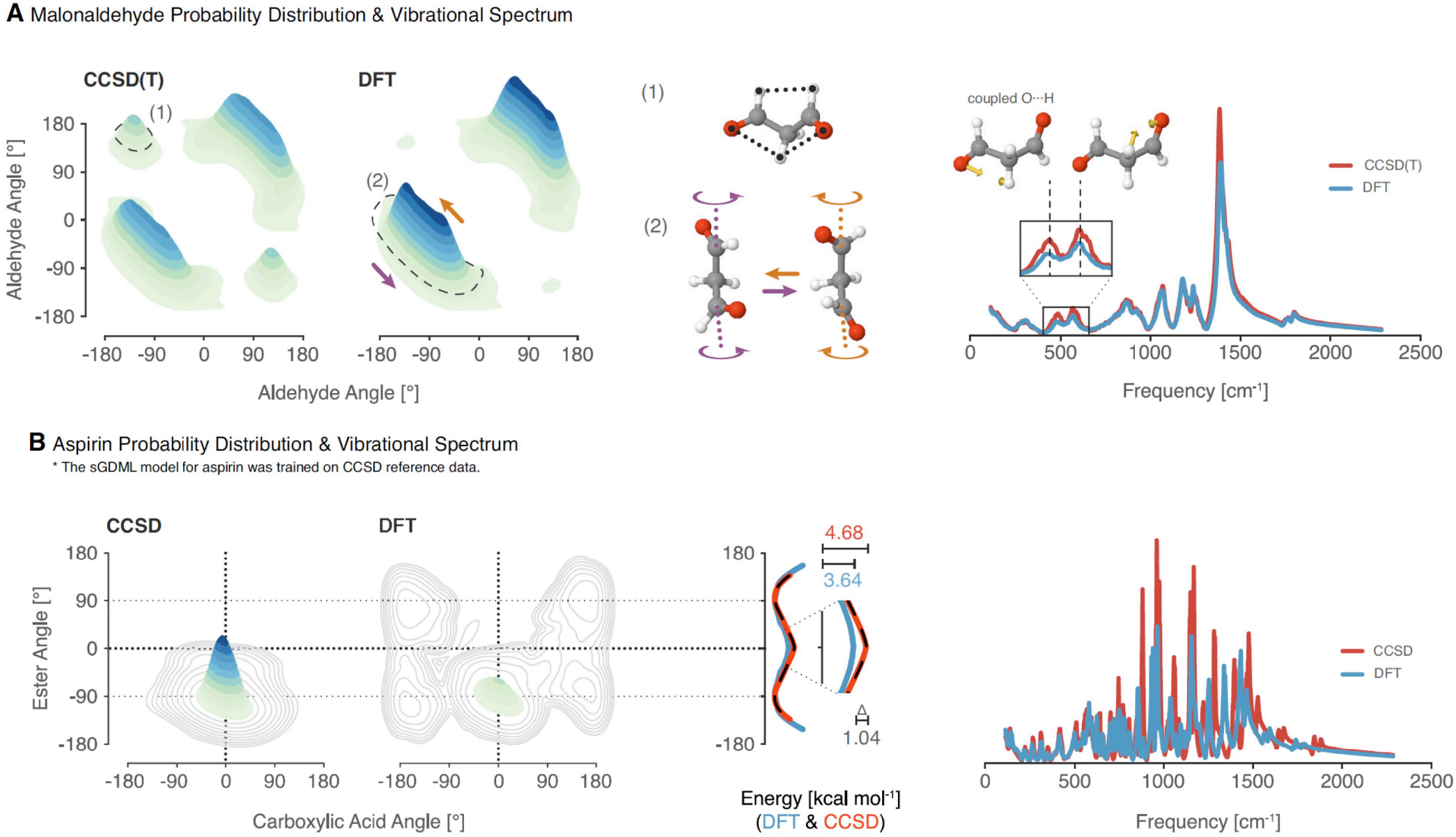
S. Chmiela, H. Sauceda, K.-R. Mueller, and A. Tkatchenko, ***Nature Commun.*** 9, 3887 (2018).

Embarrassingly Quantum MD for Molecules: Quantized Electrons [CCSD(T)] and Nuclei [PIMD]

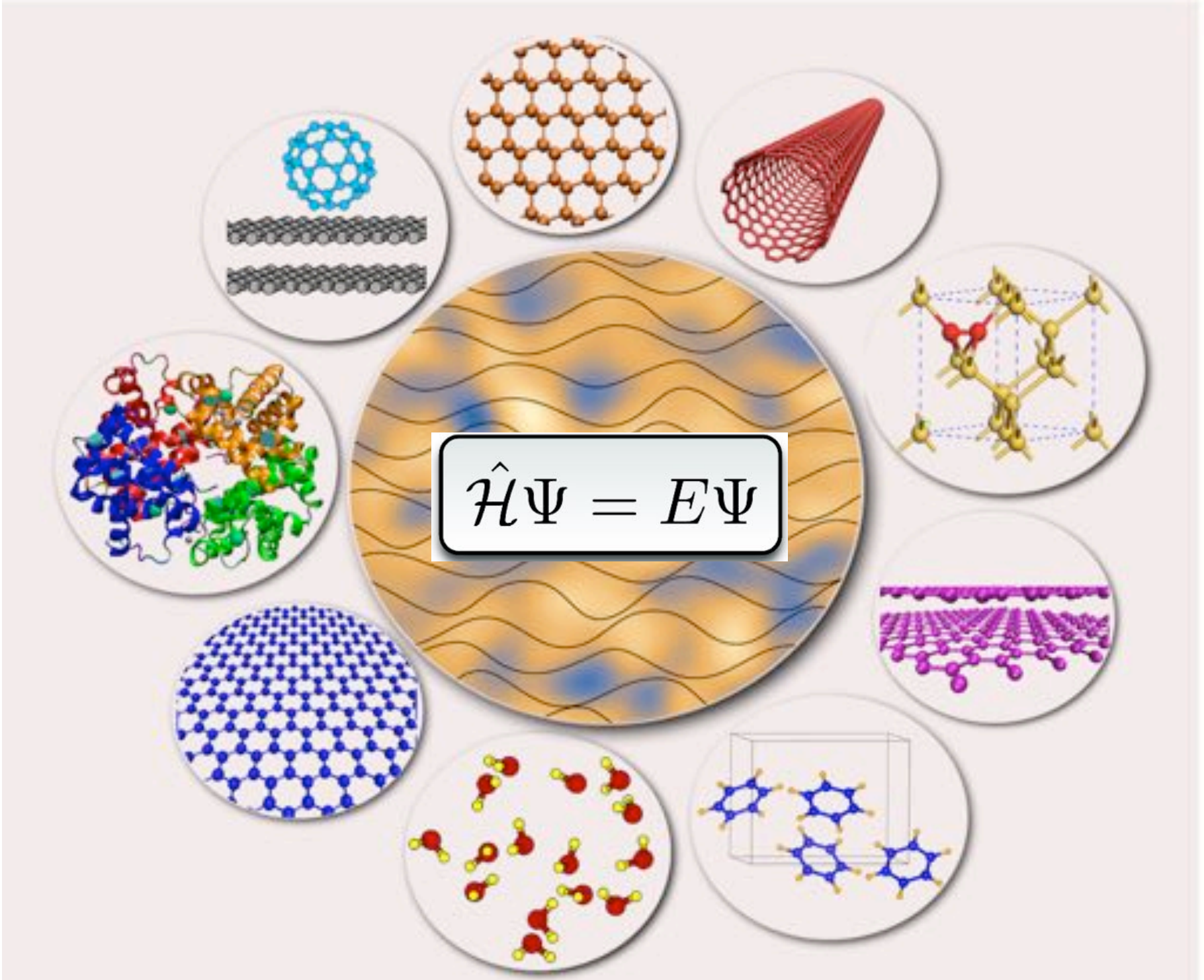


S. Chmiela, H. Sauceda, K.-R. Mueller, and A. Tkatchenko
Nature Commun. 9, 3887 (2018).

Embarrassingly Quantum MD for Molecules: Quantized Electrons [CCSD(T)] and Nuclei [PIMD]



S. Chmiela, H. Sauceda, K.-R. Mueller, and A. Tkatchenko
Nature Commun. 9, 3887 (2018).



Science 351, 1171 (2016); ***Chem. Rev.*** 117, 4714 (2017).

Grand Challenges for Machine Learning in Physics/Chemistry

- *What is chemical space*: descriptors of molecules and materials, metric?
- *How to learn intensive properties*: energy levels, excited states, spectra?
- How to combine ML with physical laws (symmetries) and interaction models?
- Can we learn (approximate) Hamiltonians?
- Can ML suggest better approximations for $\hat{\mathcal{H}}\Psi = E\Psi$?
- More and better (big) data

Physics + Chemistry + ML = *Rational design of molecules and materials in chemical space*