Luxembourg National Research Fund

DFG Deutsche Forschungsgemeinschaft

Machine Learning of Quantum Chemical Space

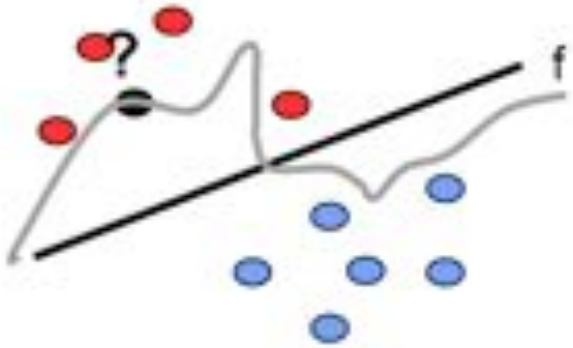
Alexandre Tkatchenko

Physics and Materials Science (PhyMS), University of Luxembourg

alexandre.tkatchenko@uni.lu

CECAM/CSM/IRTG School 2018

Machine Learning in a nutshell



Typical scenario: learning from data

- given data set **X** and labels **Y** (generated by some joint probabilty distribution p(x,y))
- **LEARN/INFER** underlying **unknown** mapping

$$Y = f(X)$$

fit

Example: ~~understand~~ chemical compound space, distinguish brain states …

BUT: how to do this optimally with good performance on **unseen** data?

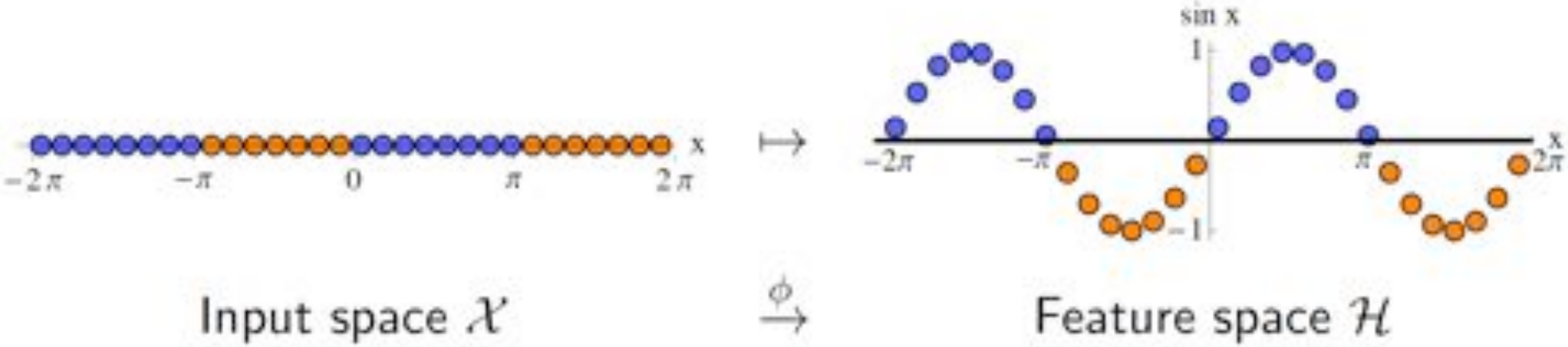
Most popular techniques: **kernel methods** and (deep) **neural networks**

Slide by Klaus Mueller – TU Berlin

Kernel Learning

Idea:

- Transform samples into higher-dimensional space
- *Implicitly* compute inner products there
- Rewrite linear algorithm to use only inner products



Input space $\mathcal{X}$ $\xrightarrow{\phi}$ Feature space $\mathcal{H}$

$$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}, \qquad k(x, z) = \langle \phi(x), \phi(z) \rangle$$

Regularized Kernel Ridge Regression

- Regularized form of ordinary regression
- Regularization prevents over-fitting by penalizing large coefficients
- Use of kernels for non-linearity

Solution has form

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i k(\mathbf{x}_i, \mathbf{x})$$

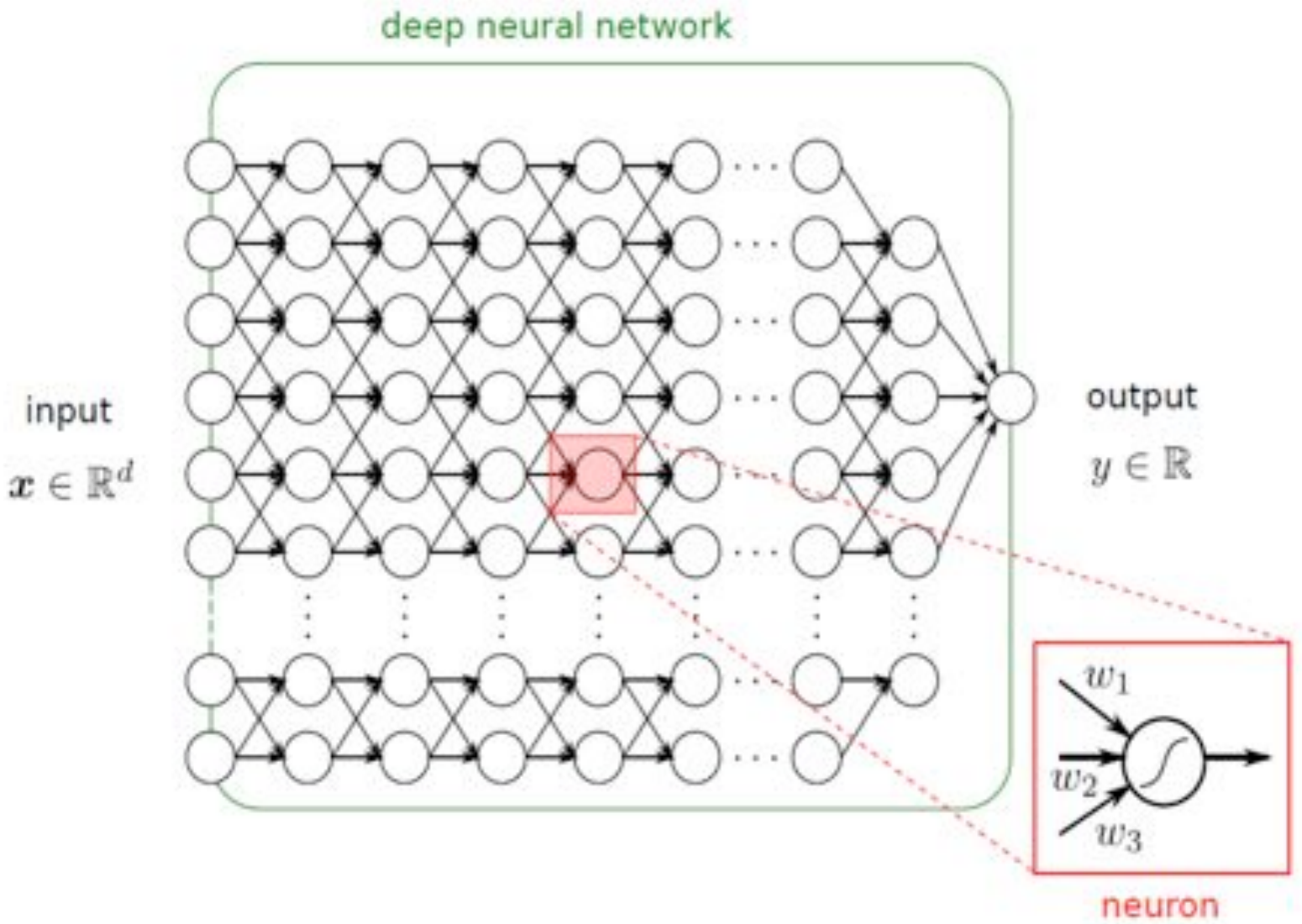
Coefficients α are obtained by solving

$$\sum_{i=1}^{n} (f(\mathbf{x}_i) - y_i)^2 + \lambda \boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha},$$

which has solution

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

Neural Networks



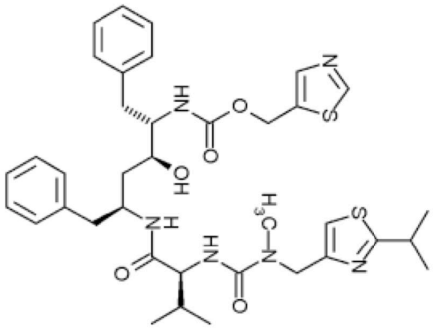
- Neuron applies a nonlinear function to its input.
- Examples of functions: hyperbolic tangent, rectification.

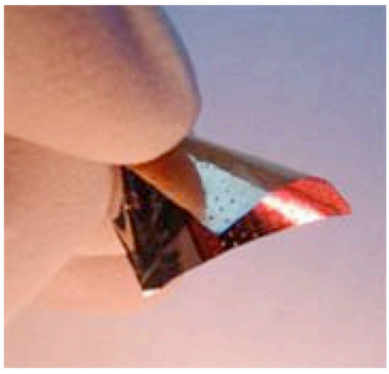
Physics and Chemistry (and Biology?)

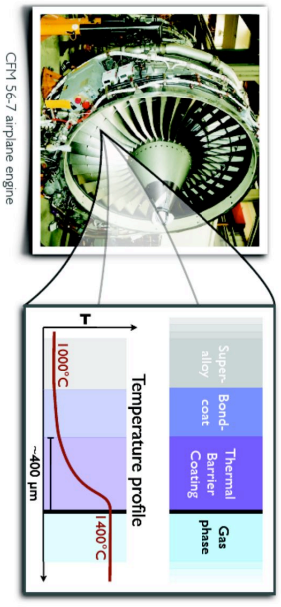
$$\hat{\mathcal{H}}\Psi = E\Psi$$

Density-functional theory, perturbation theory, coupled cluster, configuration interaction, ...

$$E_v[n] = T_s[n] + \int v(\mathbf{r})n(\mathbf{r})d^3\mathbf{r} + E^{\text{Hartree}}[n] + E^{\text{xc}}[n]$$



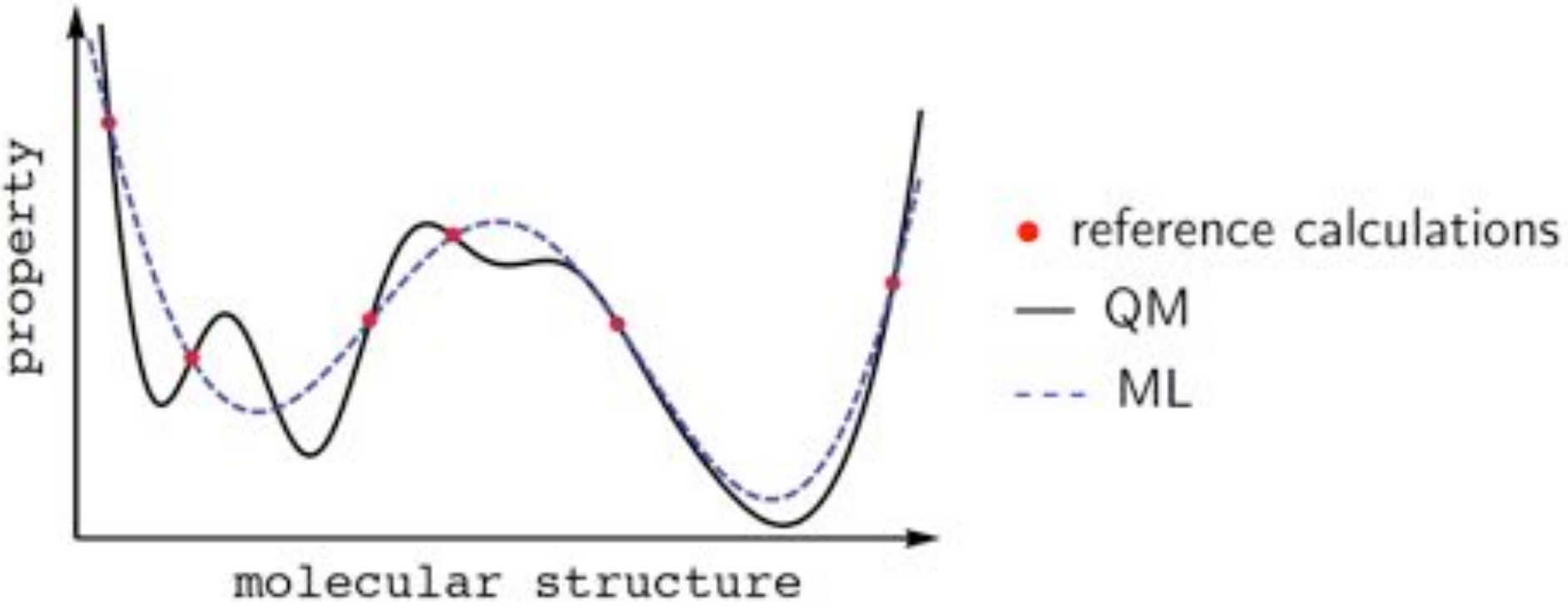




Quantum Mechanics / ML models

Exploit redundancy in a series of QM calculations

- QM/ML = quantum mechanics + machine learning
- Interpolate between QM calculations using ML
- Smoothness assumption (regularization)



Big Data for Molecules and Materials

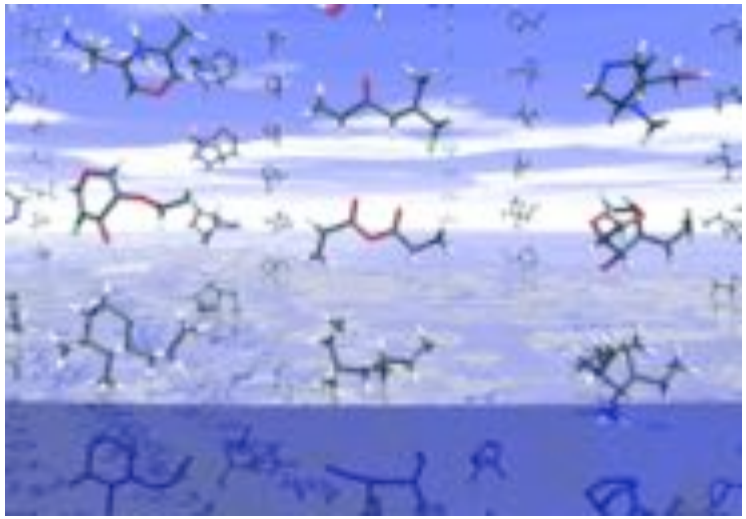
nomad-coe.eu

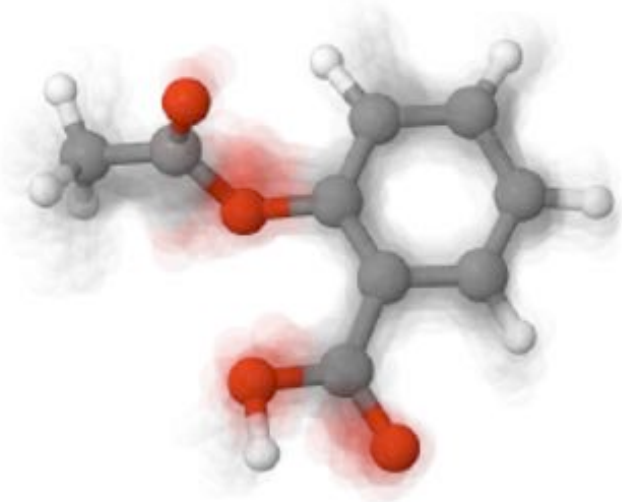
max-centre.eu

e-cam2020.eu

nccr-marvel.ch

Molecular Data in this Talk





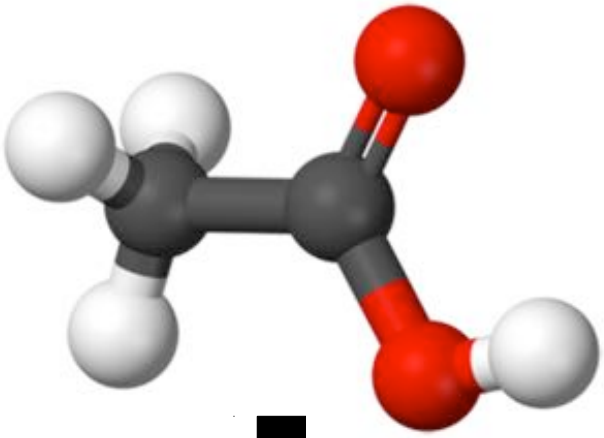
GDB mol graphs: J. L. Reymond (U. Bern)
http://gdb.unibe.ch/downloads/

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

Quantum physics/chemistry today

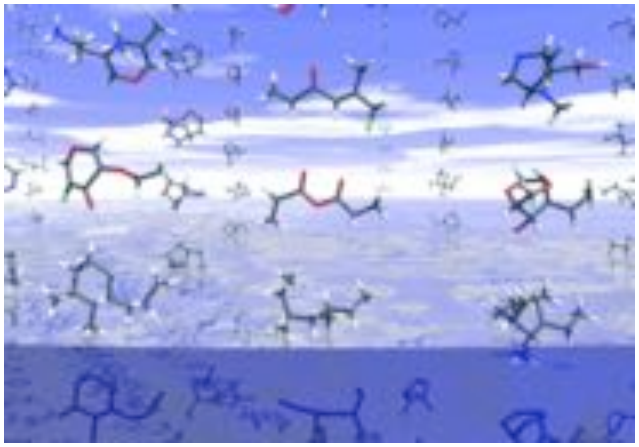


DFT
MP2
CCSD(T)
...

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

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

Quantum physics/chemistry tomorrow?



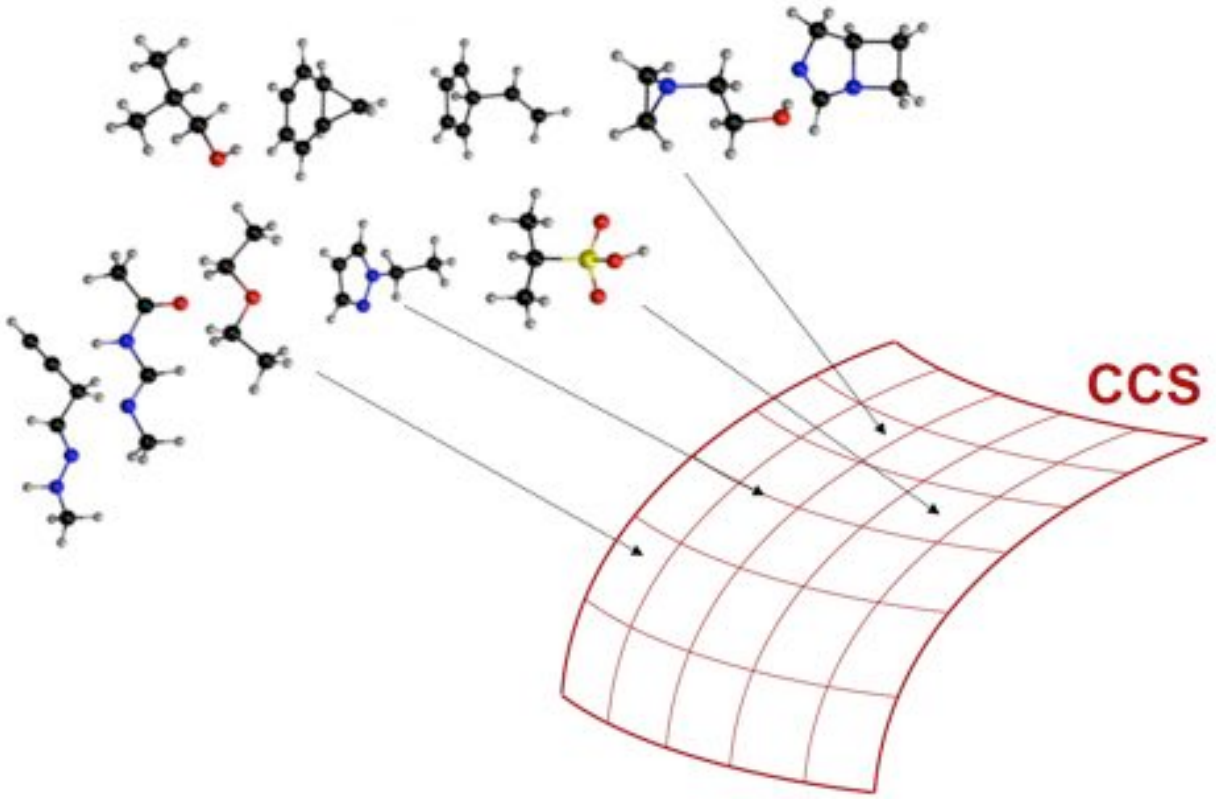
Training data:
molecular properties

ML

Insights:

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

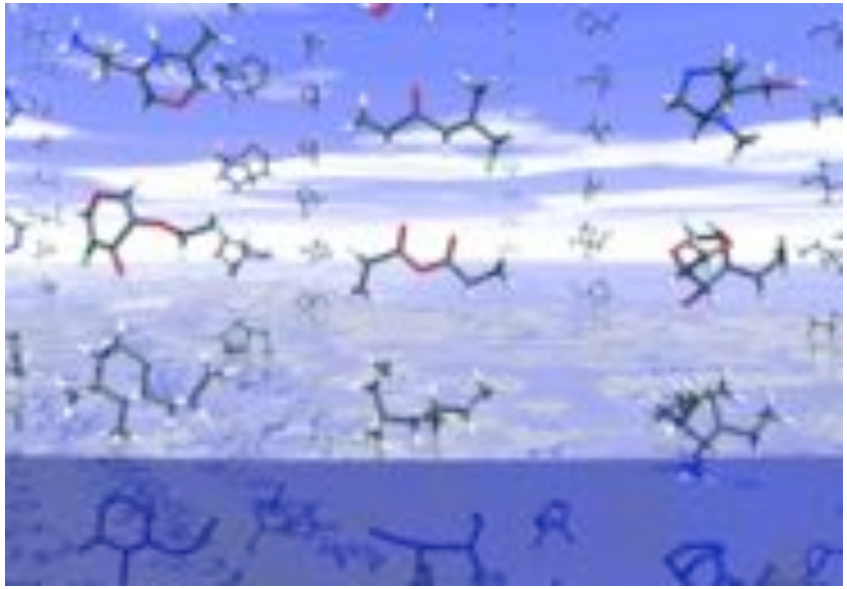
Molecular big data



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

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

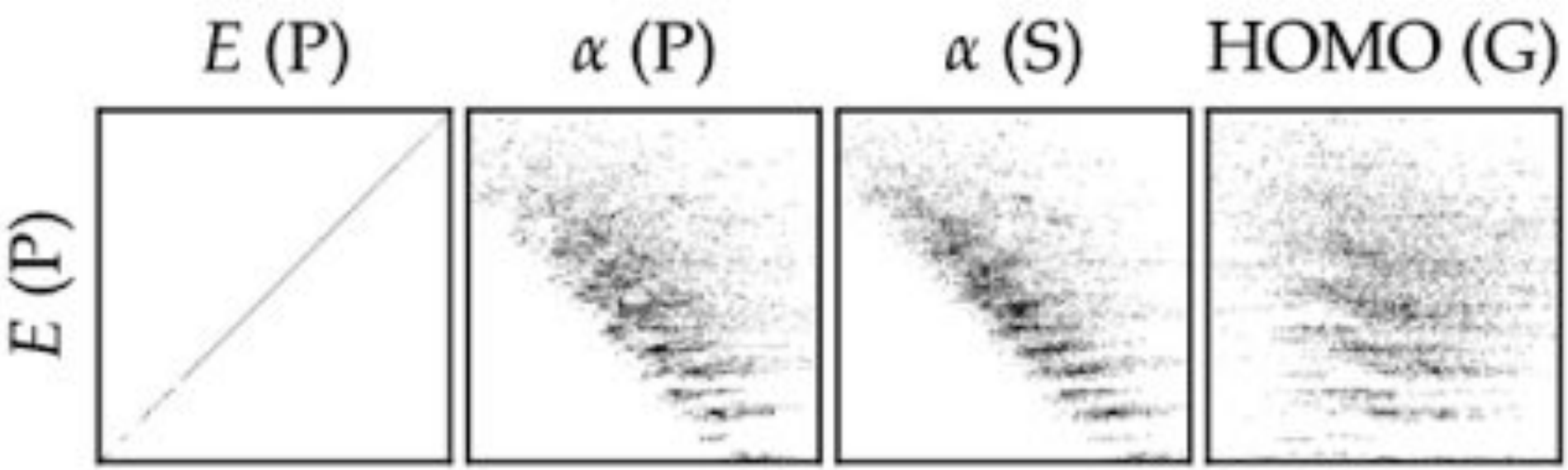
Machine learning for molecular big data

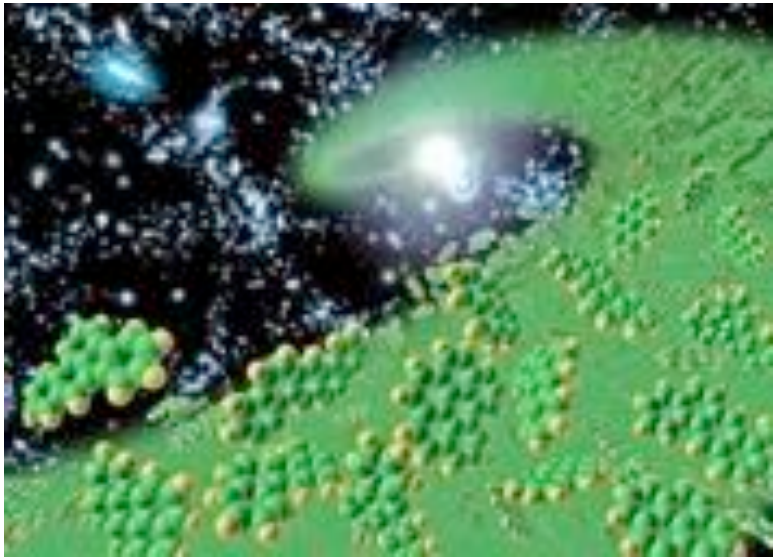


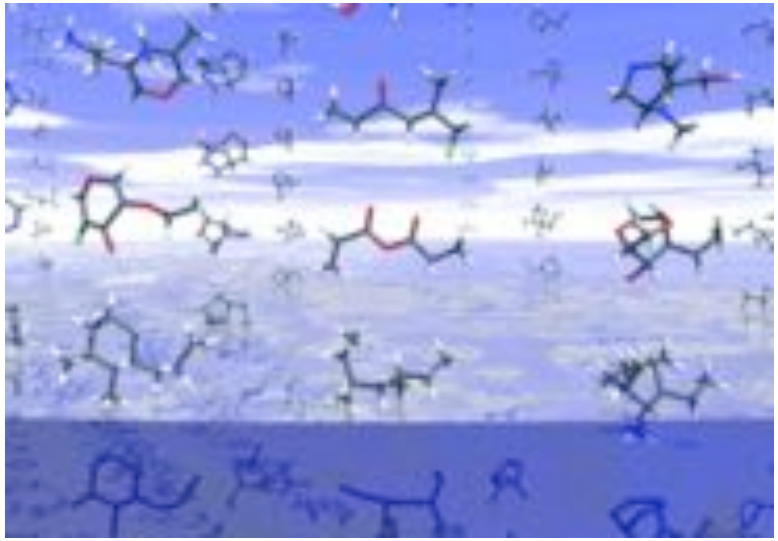
$\{\boldsymbol{R}_i, Z_i\}$ maps to $\{P_1, P_2, P_3, P_4, \ldots\}$

- 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?

Chemical Compound Space: Freedom of design

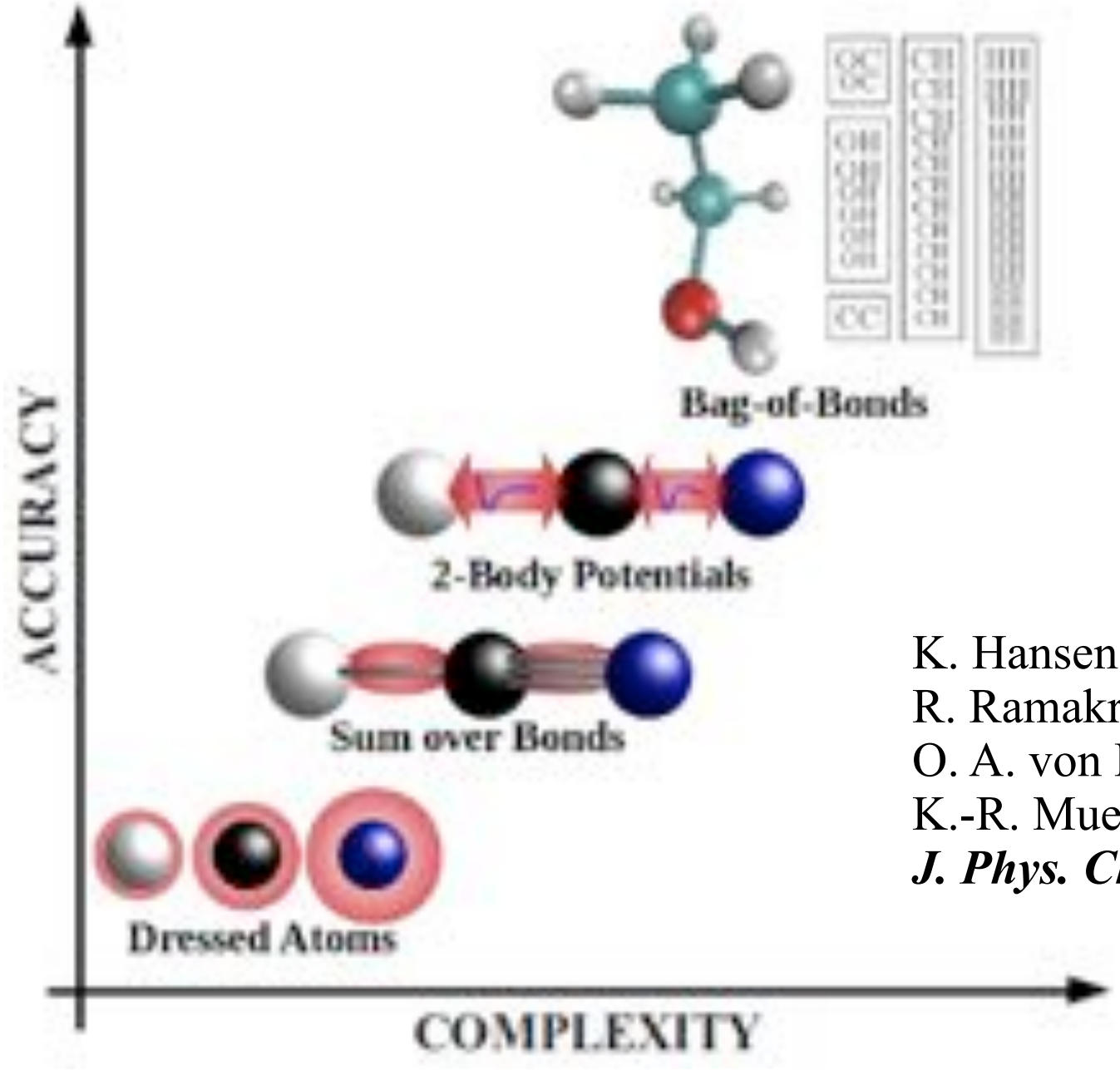






G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Mueller, A. von Lilienfeld, ***New J. Phys.*** 15, 095003 (2013).

Predicting Molecular Properties:
Descriptors From "Dressed Atoms" to Bag-of-Bonds



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

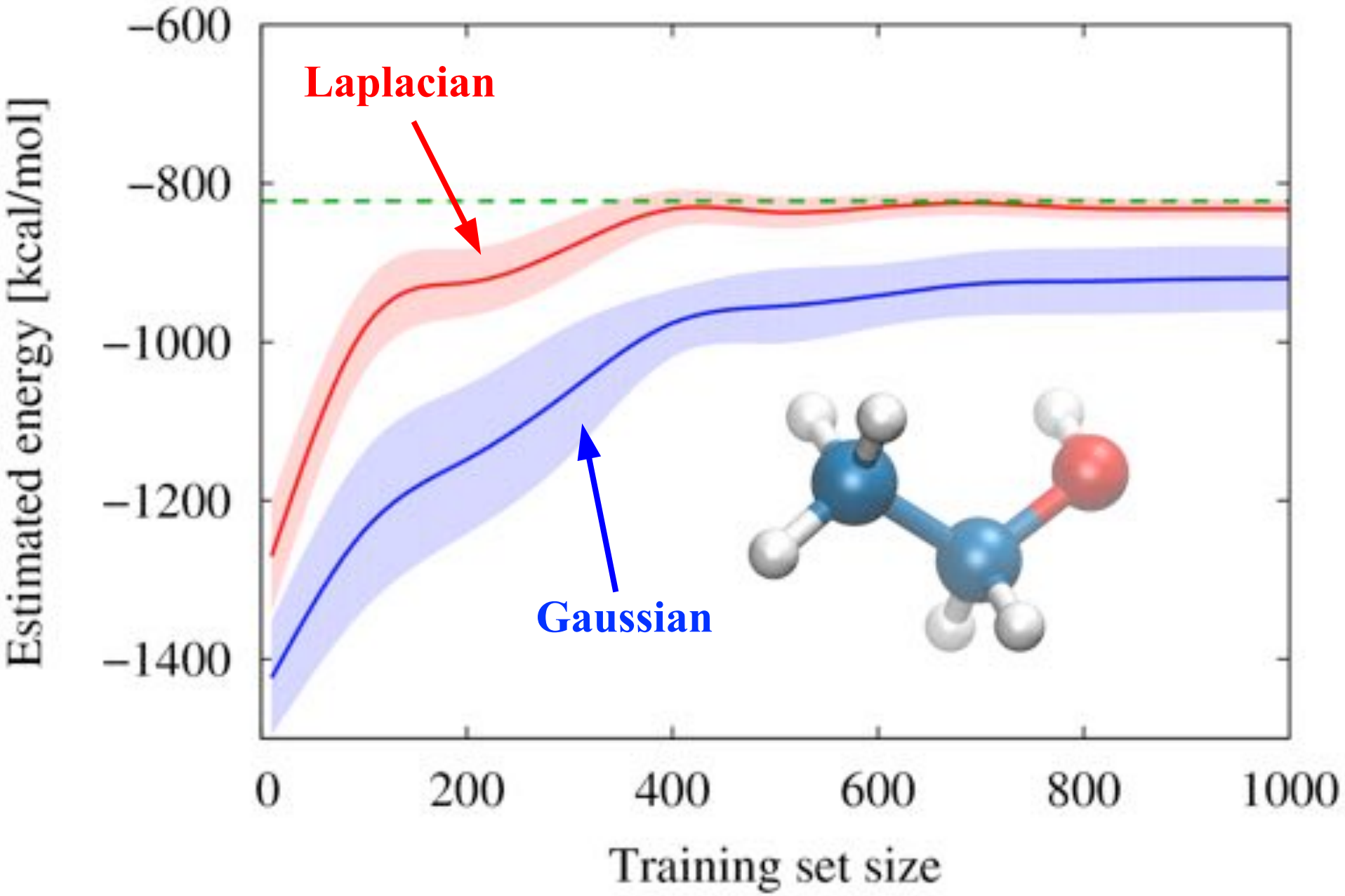
K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. von Lilienfeld, K.-R. Mueller, and A. Tkatchenko, ***J. Phys. Chem. Lett.*** 6, 2326 (2015).

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).

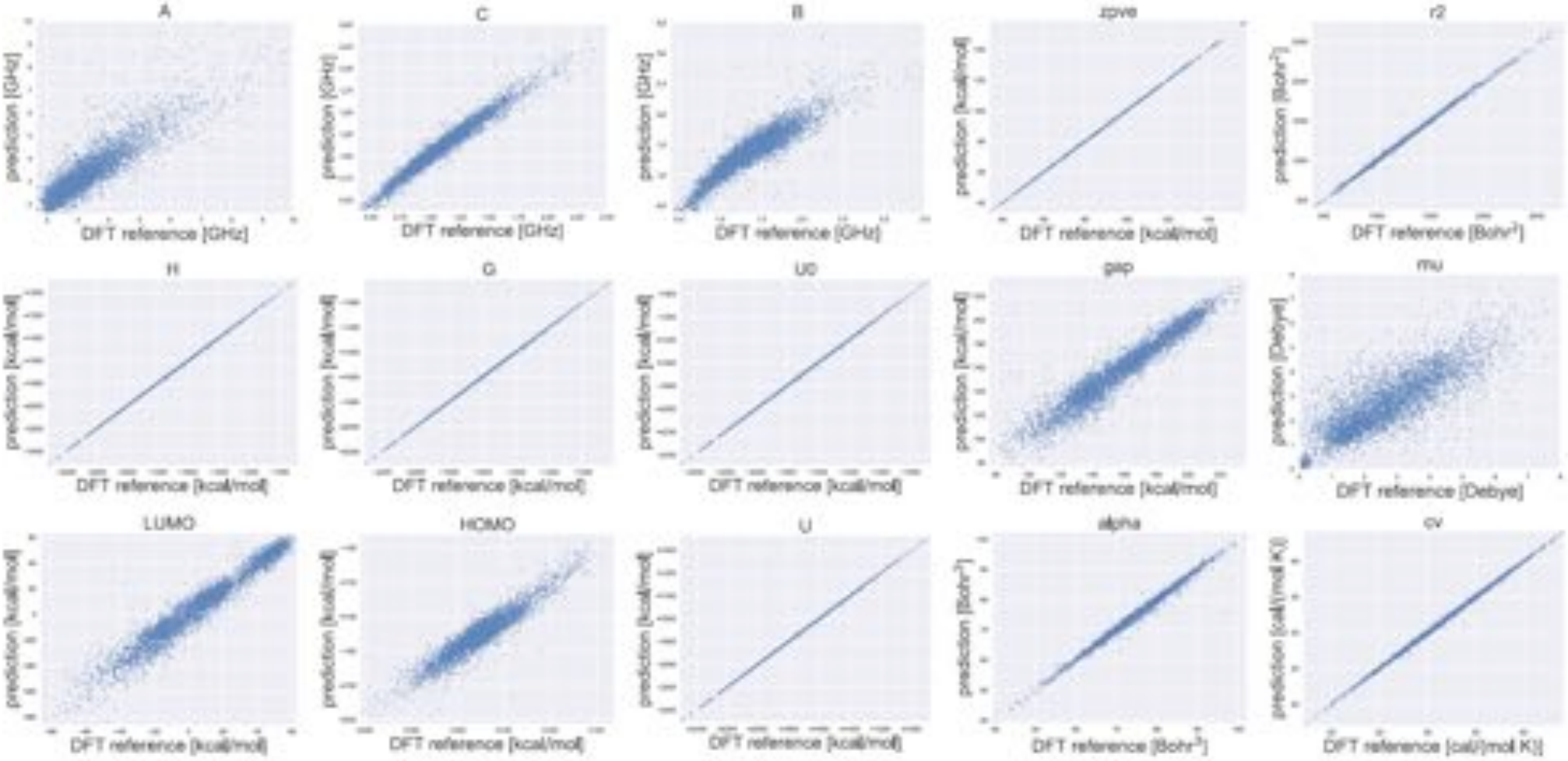
Bag-of-Bonds (BoB): Non-Locality in Chemical Space



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

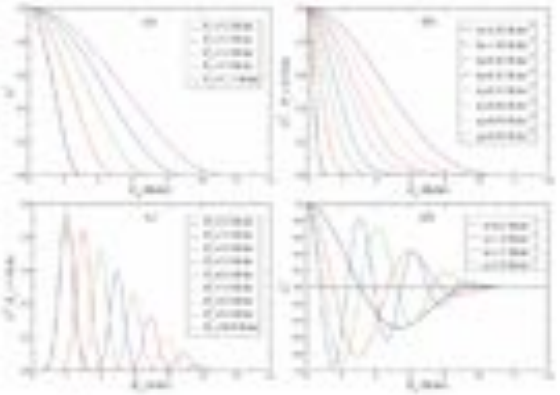
W. Pronobis, A. Tkatchenko, and K.-R. Mueller, ***J. Chem. Theory Comput.*** (2018).

QM9 dataset: Extensive and Intensive Properties



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)

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

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

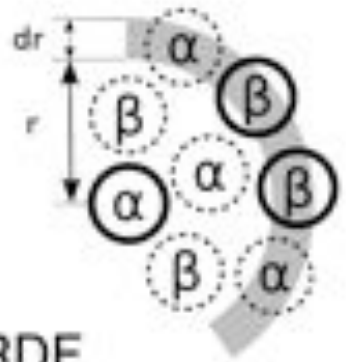
Bag of bonds
(Hansen et al. 2015)

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

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)

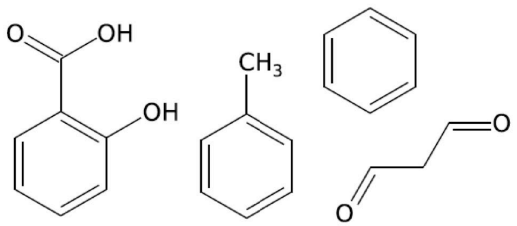


PRDF
(Schütt et al, 2014)

Learning the Representation: Deep Tensor Neural Networks (DTNN)

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

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

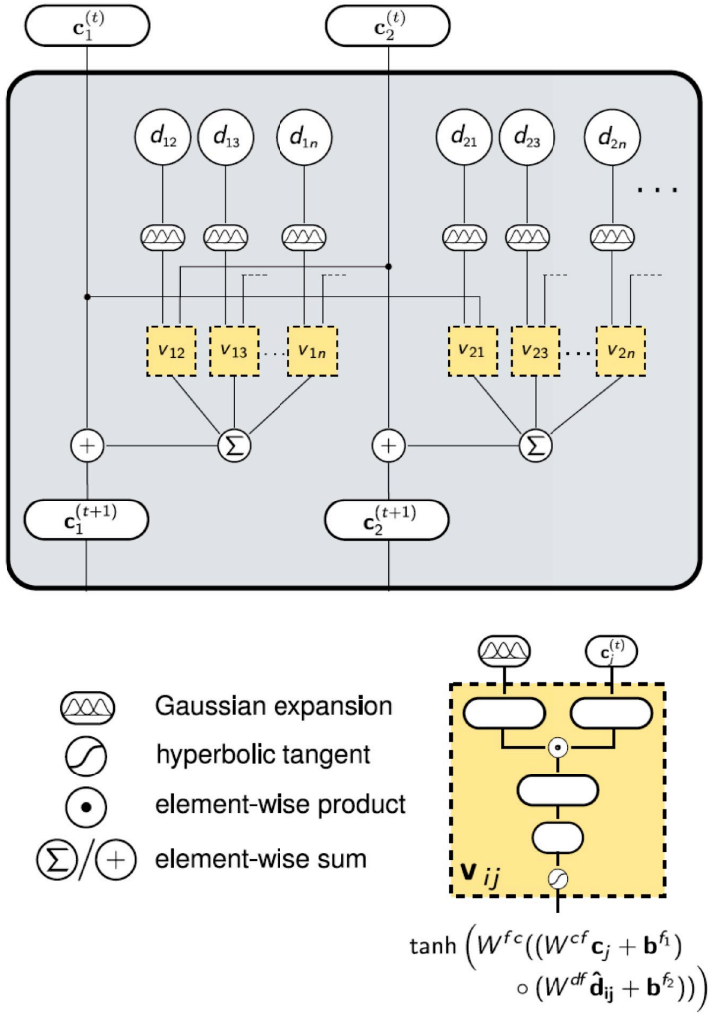
Add interaction with environment using $t = 1 \ldots T$ sequential refinements $\mathbf{v}_i^{(t)}$

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

Prediction via atom-wise contributions:

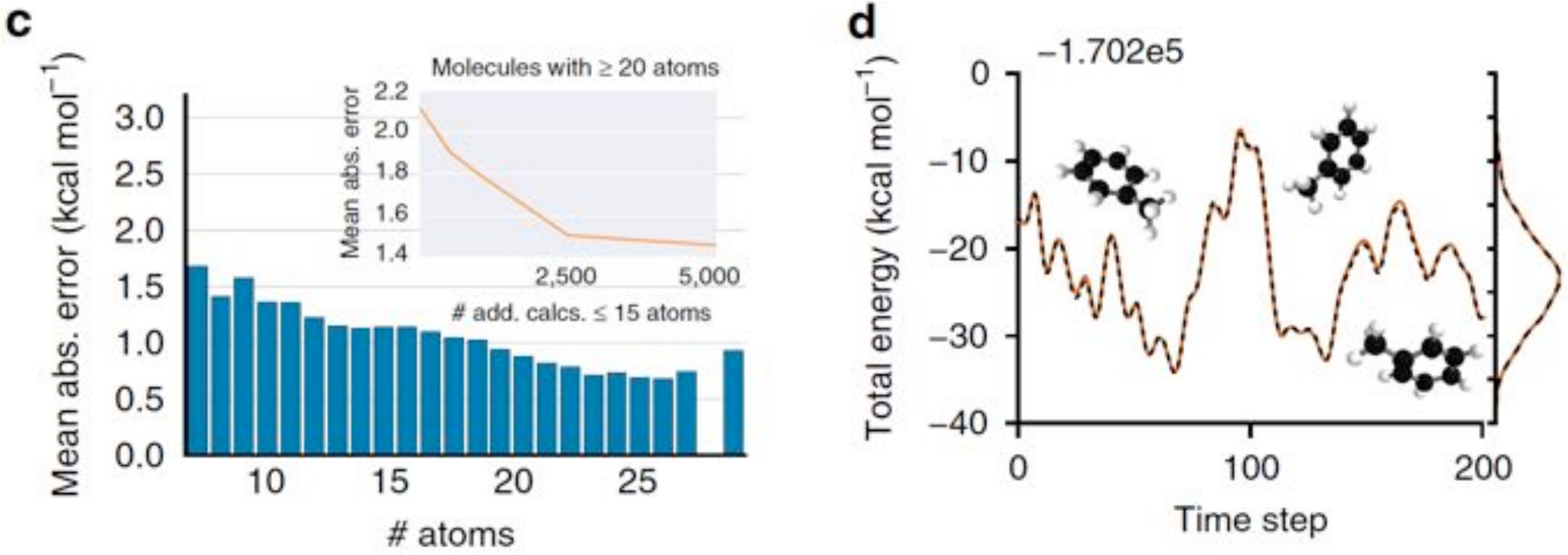
$$\hat{E} = \sum_{i=1}^{n_{\text{atoms}}} f_{\text{out}}(\mathbf{x}_i^{(T)})$$

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



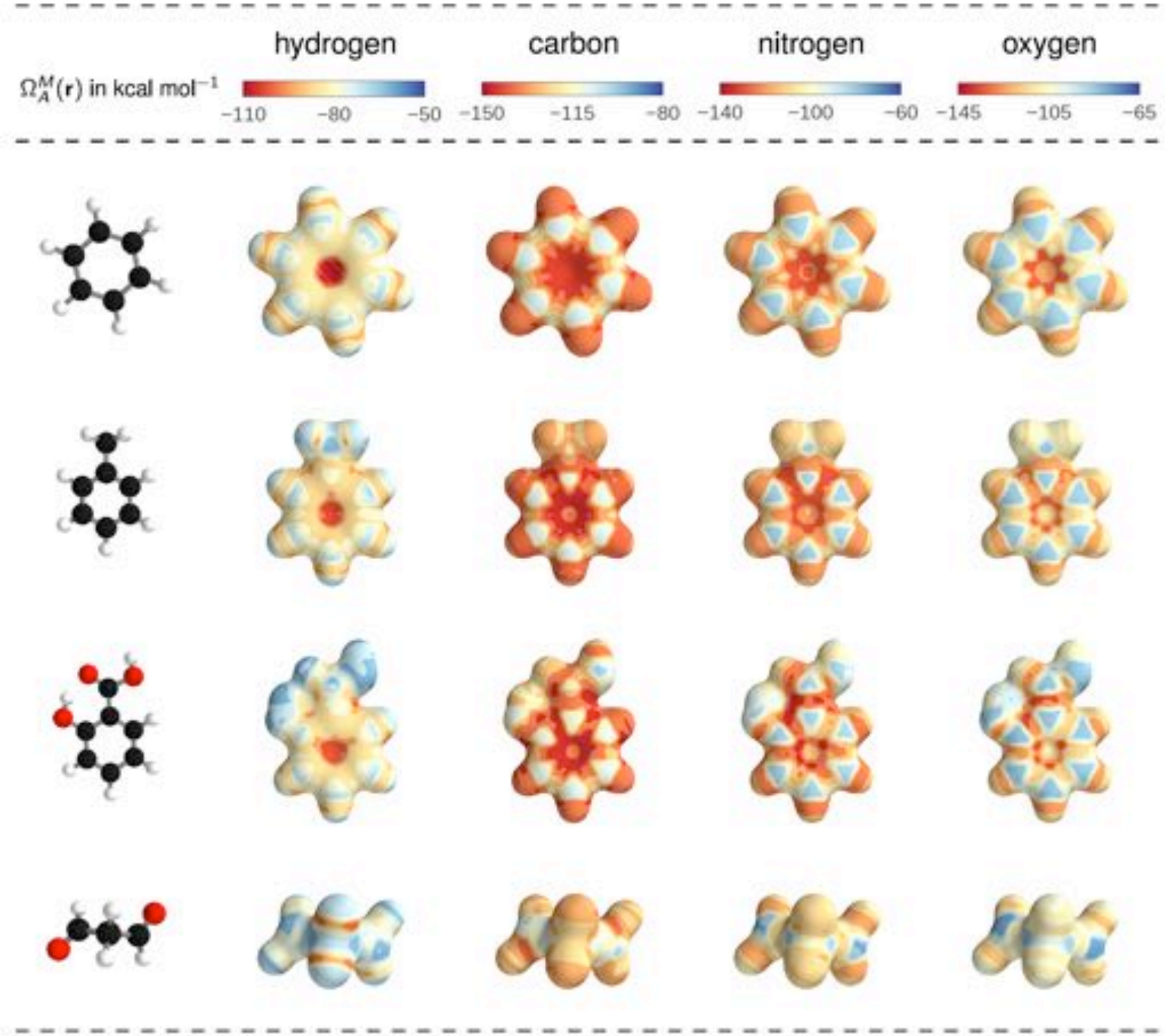
K. T. Schuett, F. Arbabzadah, S. Chmiela, K.-R. Mueller, and A. Tkatchenko, ***Nature Commun.*** 8, 13890 (2017).

Molecular DTNN: Performance on QM9 and MD

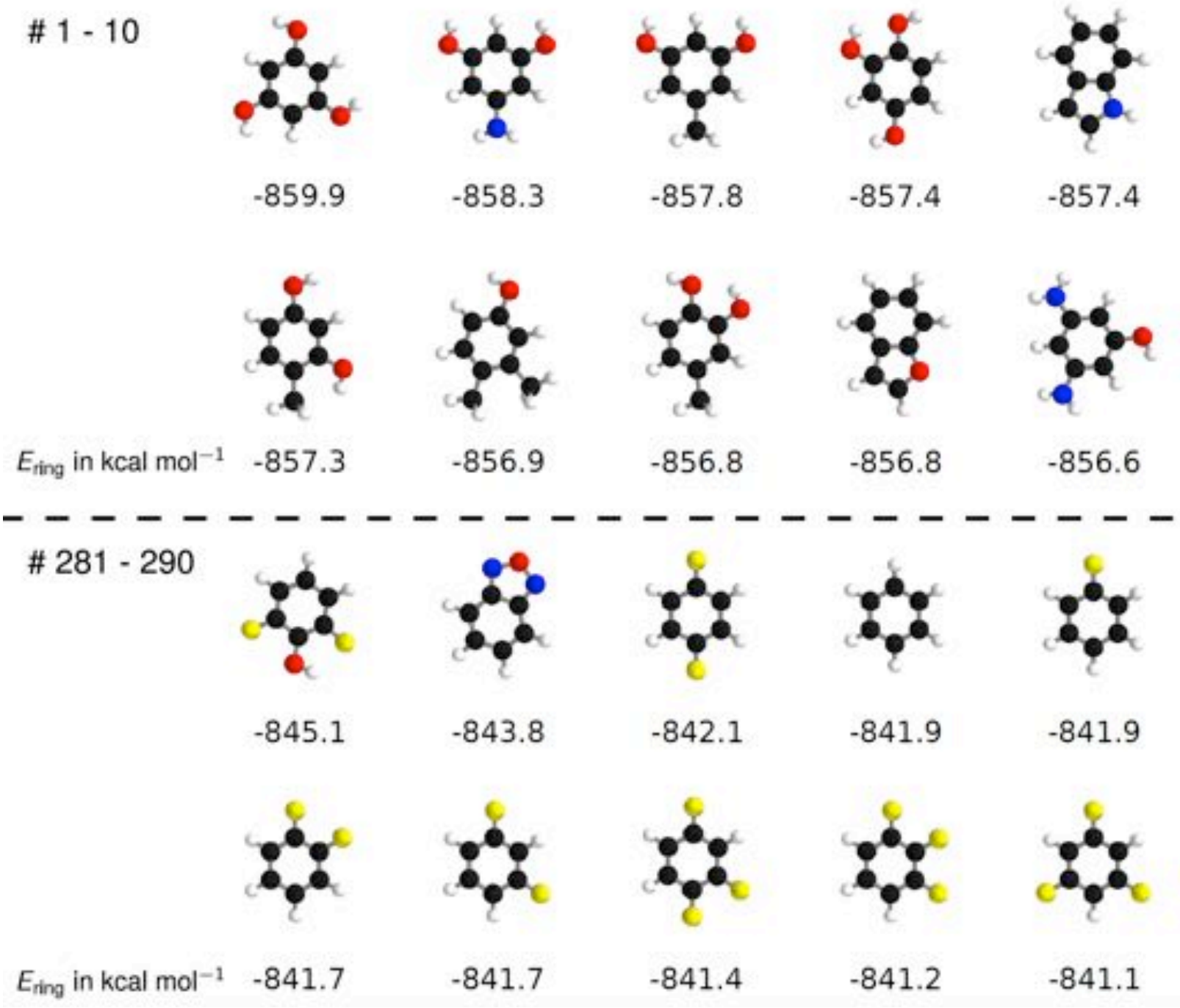


K. T. Schuett, F. Arbabzadah, S. Chmiela, K.-R. Mueller, and A. Tkatchenko, ***Nature Commun.*** 8, 13890 (2017).

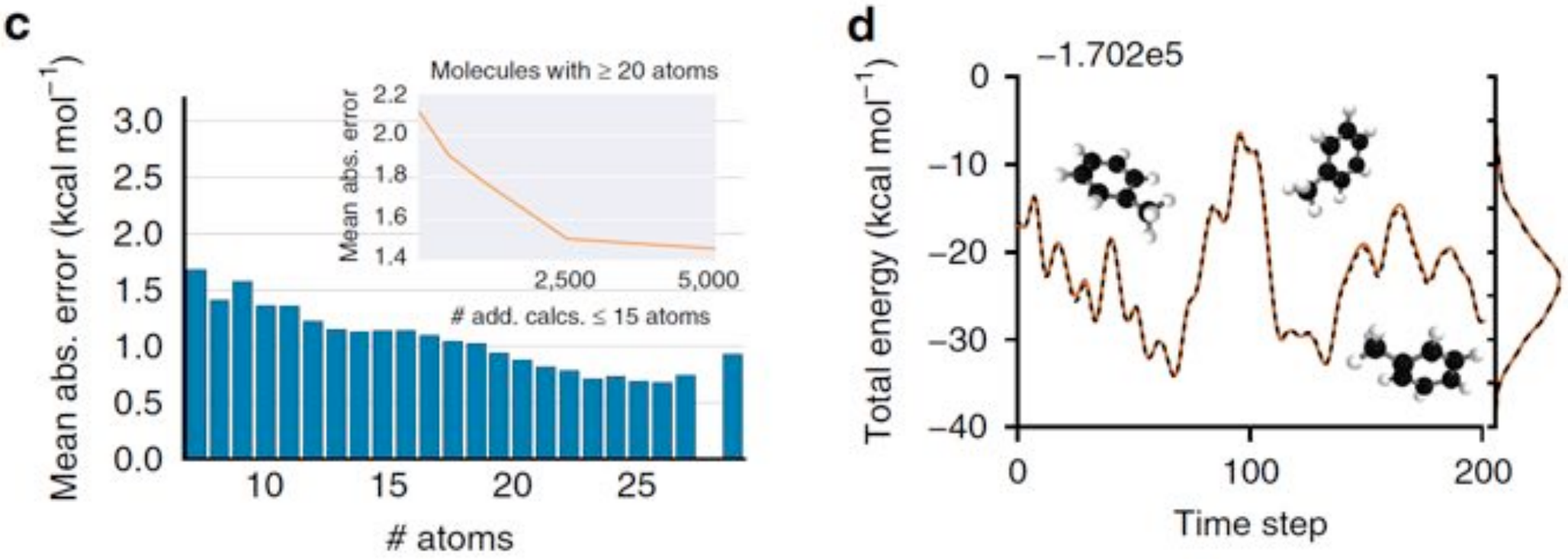
Molecular DTNN: What Did it Learn ?



Quantum Chemical Insights: Aromaticity



Learning Full Chemical Space with DTNN?

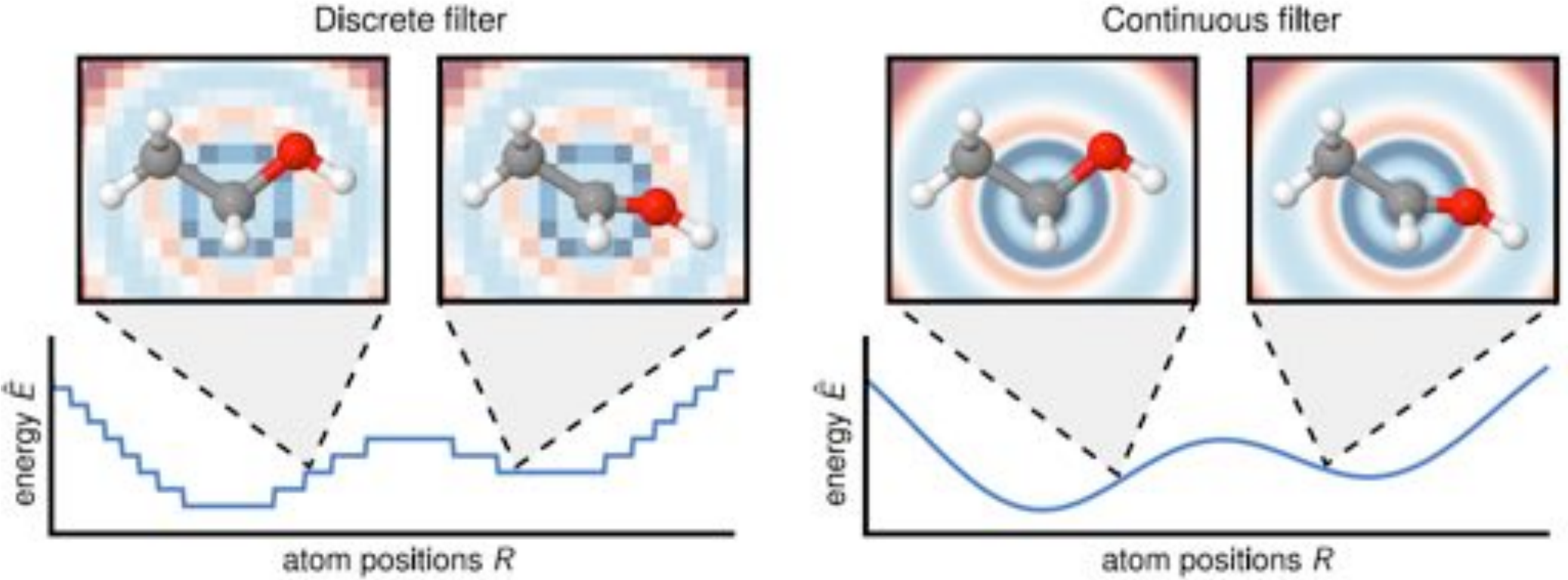


Accurately representing BOTH compositional and conformational degrees of freedom is difficult.

For C7O2H10 isomer and MD data, the error grows to 1.7 kcal/mol.

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

From DTNN to SchNet architecture

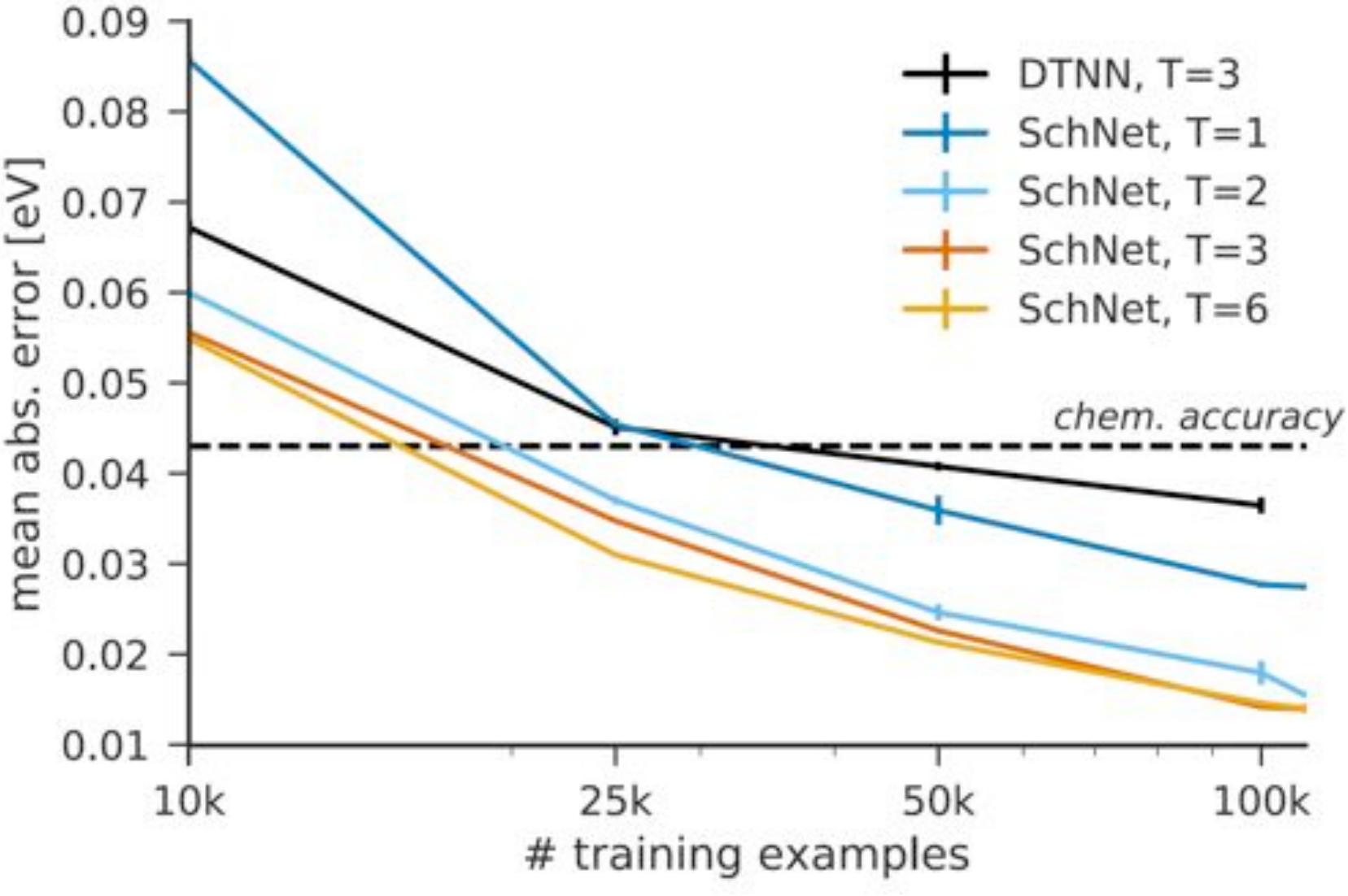


$$\mathbf{v}_i^{(t)} = \sum_{j=1}^{N_{\text{atom}}} \mathbf{x}_j^{(t)} \circ \underbrace{W_{[d_{ij}]}^{(t)}}_{\text{parameter tensor}}$$

$$\mathbf{v}_i^{(t)} = \sum_{j=1}^{N_{\text{atom}}} \mathbf{x}_j^{(t)} \circ \underbrace{W^{(t)}(d_{ij})}_{\text{neural network}}$$

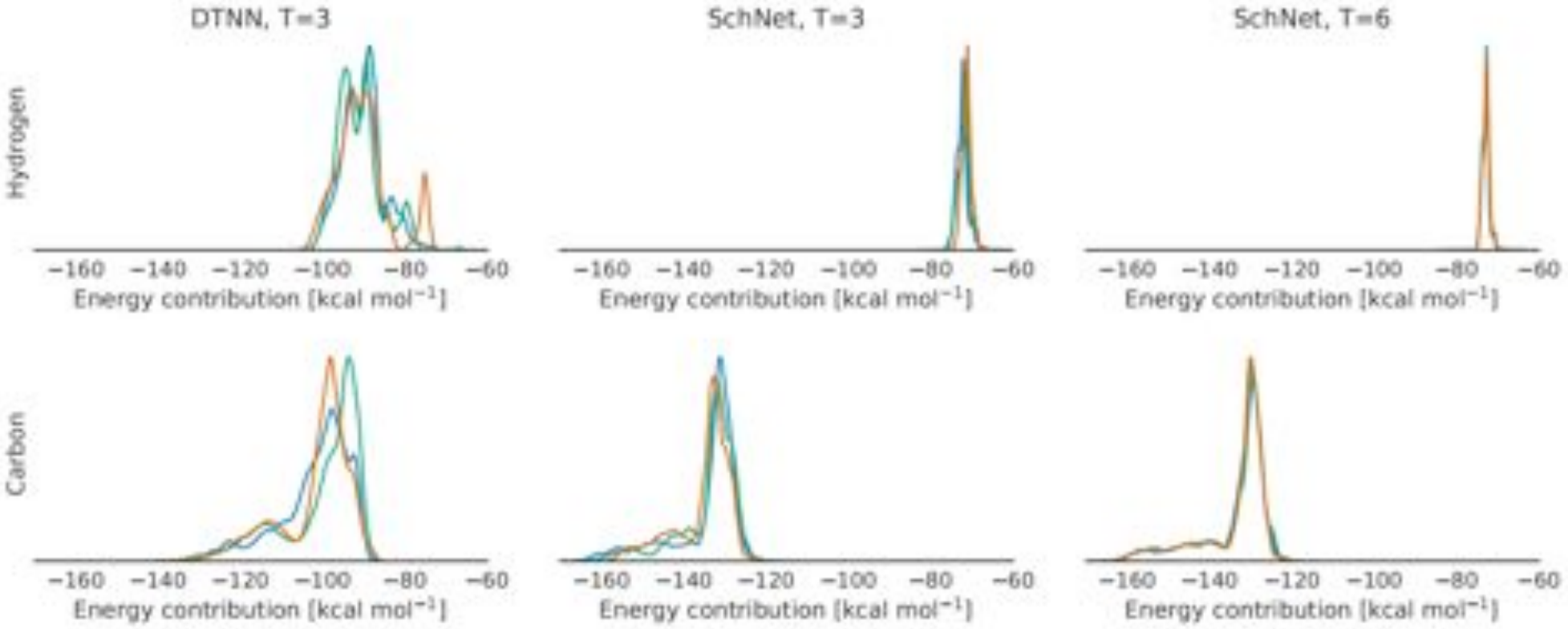
K.T. Schuett, P.J. Kindermans, H.E. Sauceda, S. Chmiela, A. Tkatchenko, K.-R. Mueller (2017).
SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. NIPS.

Data Efficiency and Robustness of Deep Networks



K.T. Schuett, H.E. Sauceda, P. J. Kindermans, S. Chmiela, A. Tkatchenko, K.-R. Mueller,
***J. Chem. Phys.* 148**, 241722 (2018).

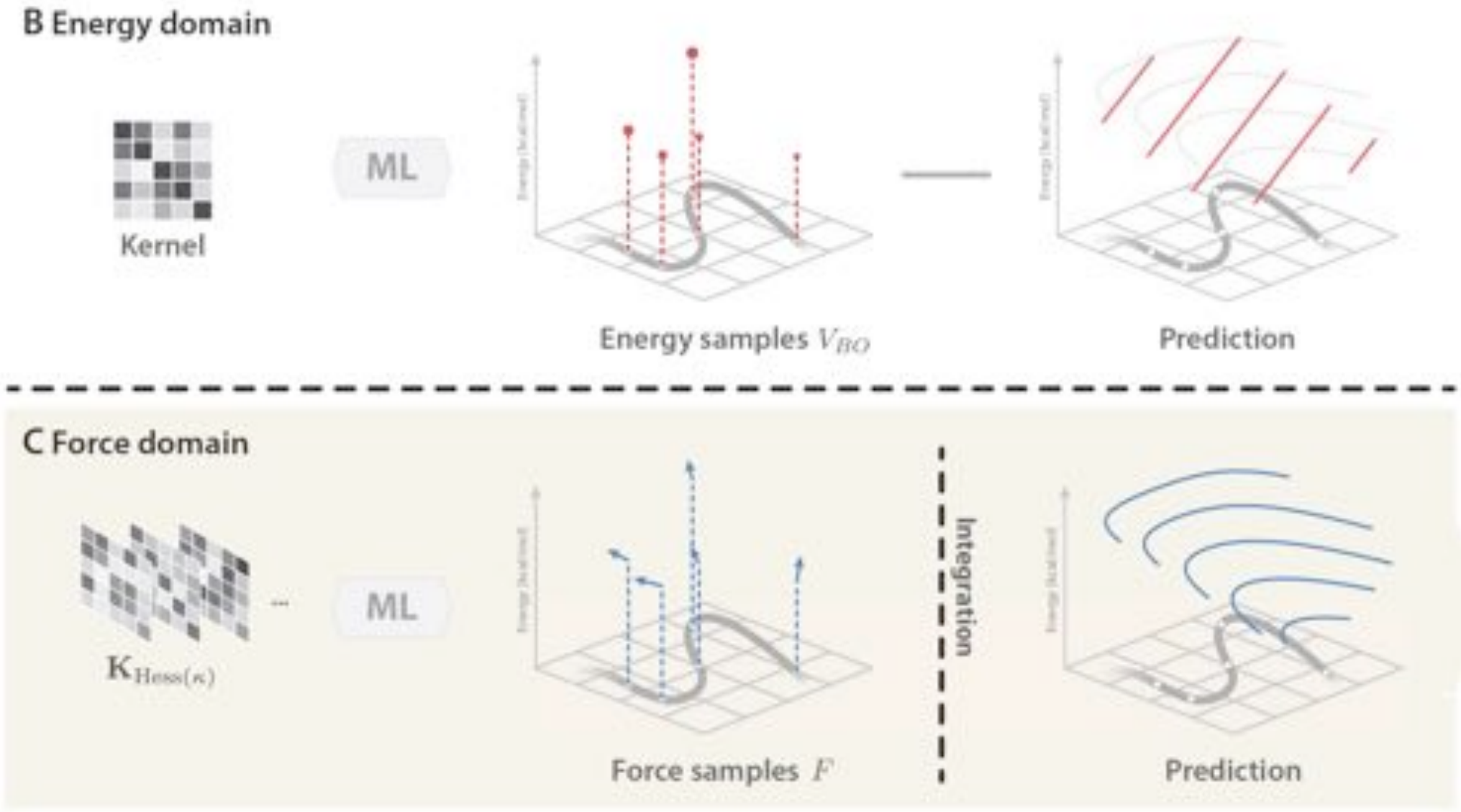
Data Efficiency and Robustness of Deep Networks



K.T. Schuett, *Ph.D. thesis*, TU Berlin, 2018.

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

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).

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

$$\{\mathbf{x}_i, \mathbf{F}_i\}_{i=1}^{N}$$

$$\left(\mathbf{K}_{\mathrm{Hess}(\kappa)} + \lambda \mathbb{I}\right) \vec{\alpha} = \nabla V_{BO} = -\mathbf{F}$$

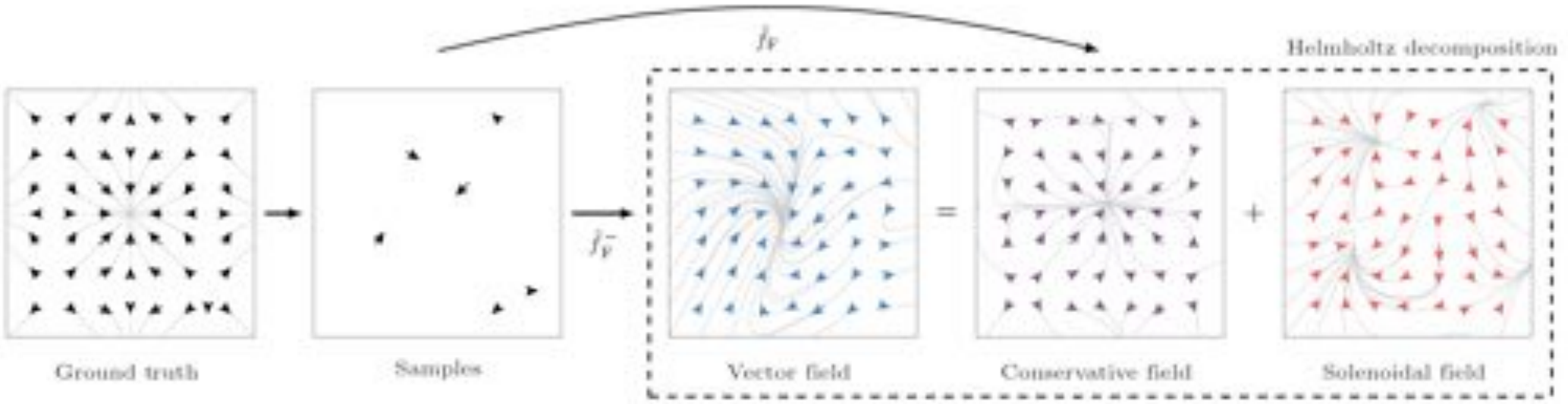
$$\{\boldsymbol{\alpha}_i\}_{i=1}^{N}$$

$$\mathbf{F}(\mathbf{x}) = \sum_{i=1}^{N} (\boldsymbol{\alpha}_i \cdot \nabla) \nabla \kappa(\mathbf{x}, \mathbf{x}_i)$$

$$E(\mathbf{x}) = \sum_{i=1}^{N} \boldsymbol{\alpha}_i \cdot \nabla \kappa(\mathbf{x}, \mathbf{x}_i)$$

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

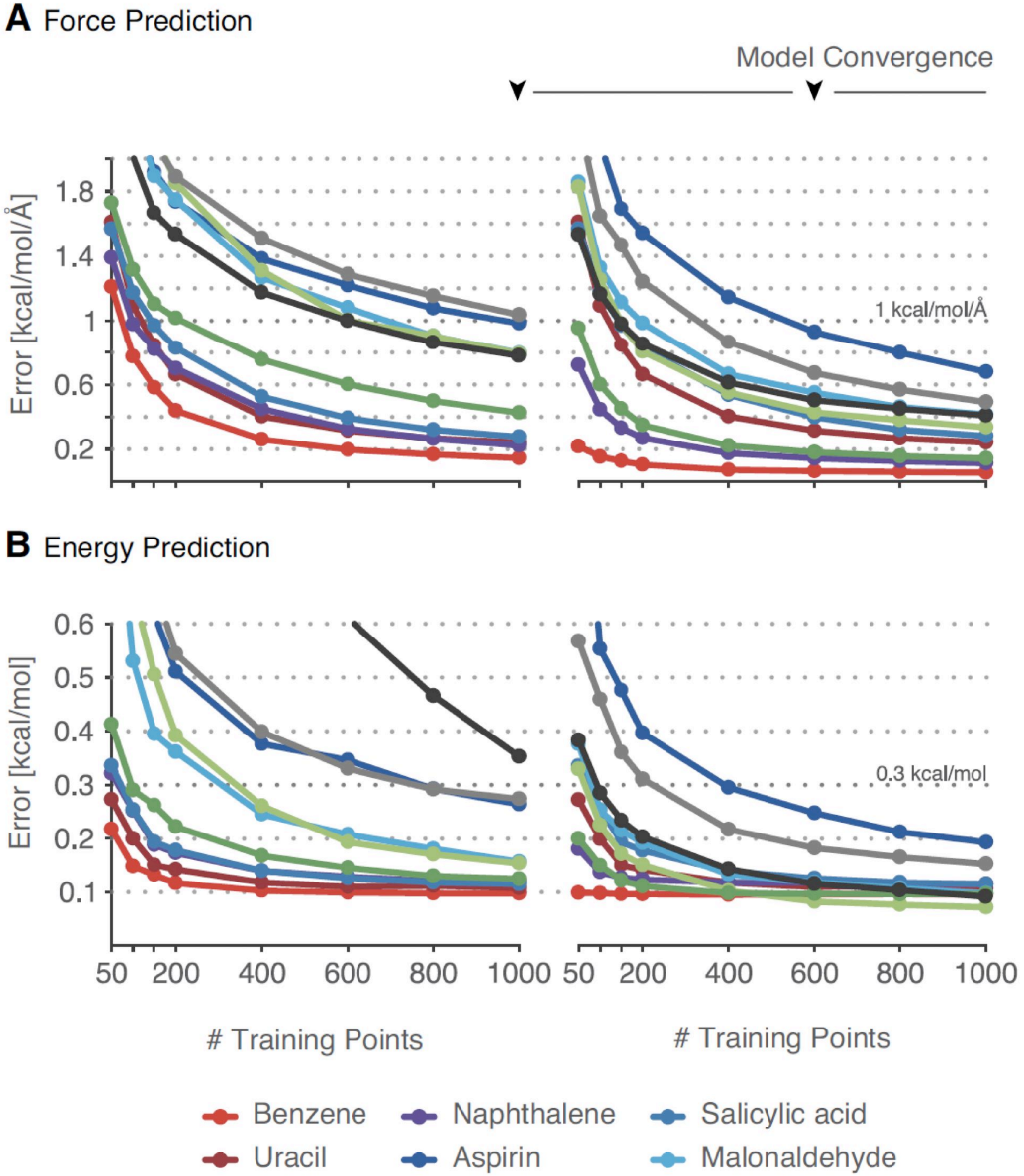
Beating the Hell out of Data:
Gradient-Domain Machine Learning (GDML)



Wrong Better Error

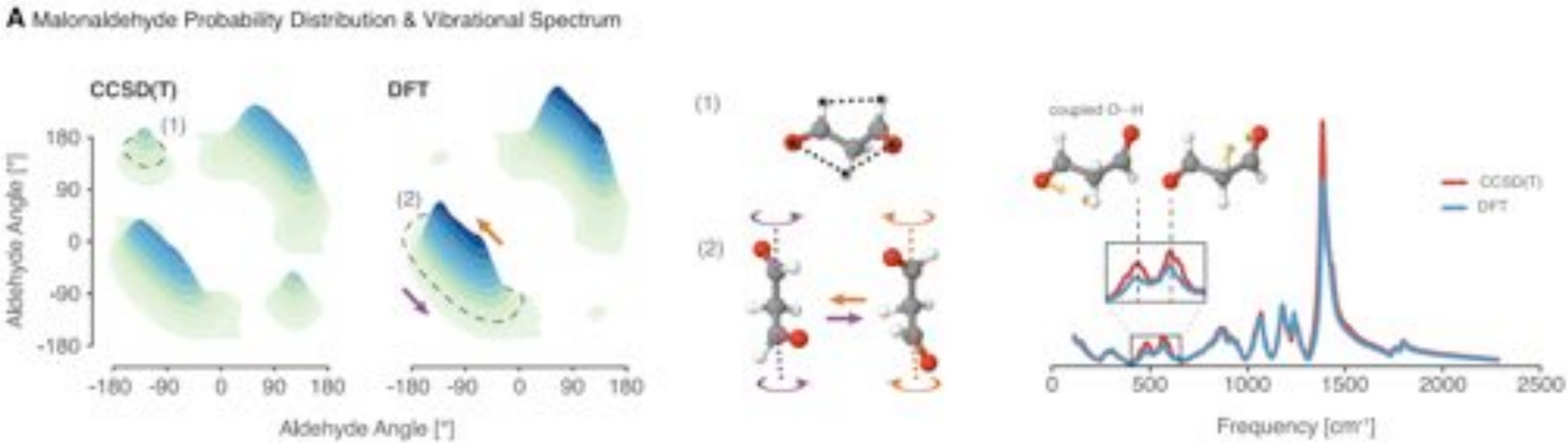
S. Chmiela, A. Tkatchenko, H. Sauceda, I. Poltavsky, K. T. Schuett, K.-R. Mueller,
Science Adv. 3, e1603015 (2017).

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



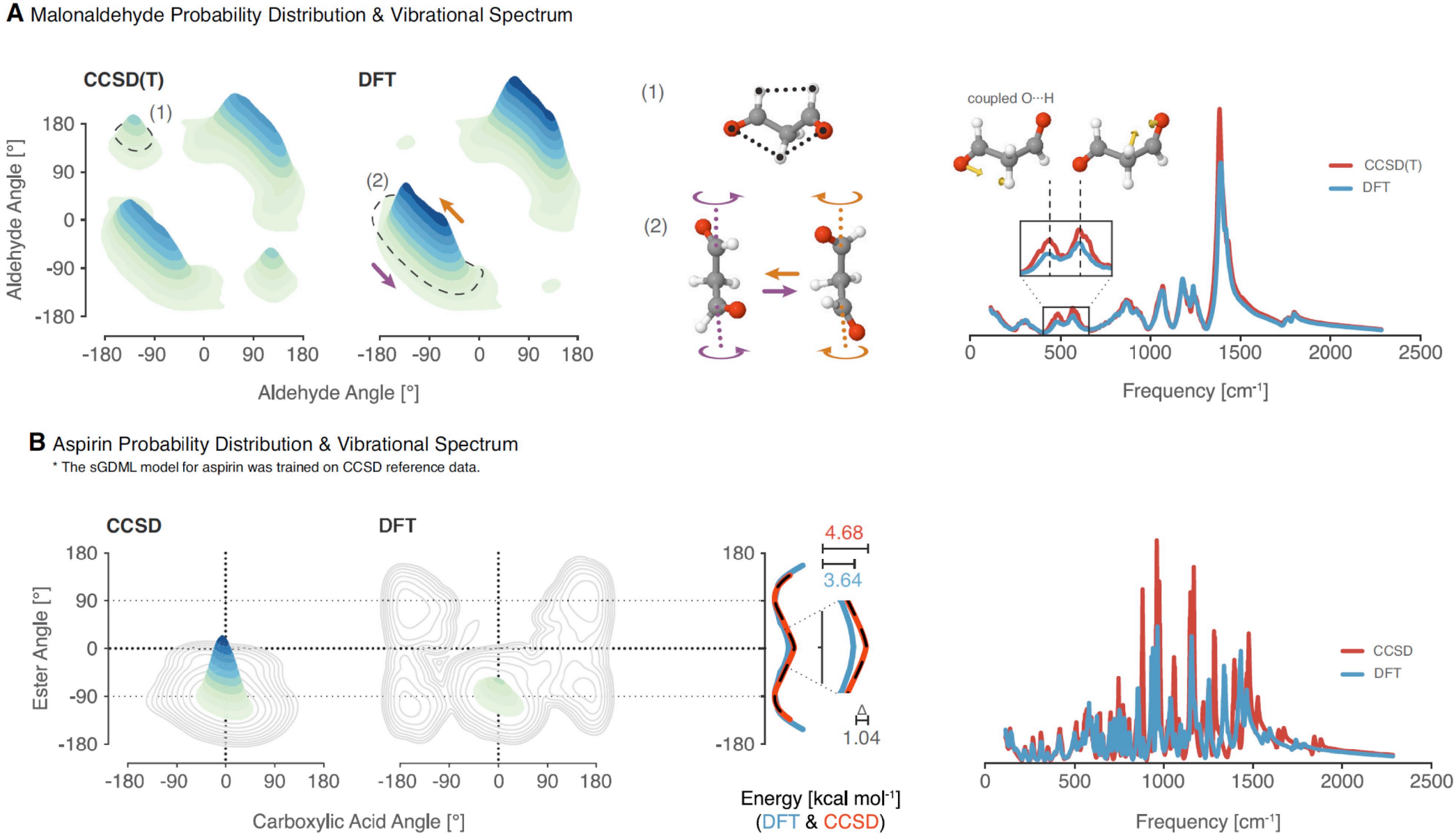
S. Chmiela, H. Sauceda, K.-R. Mueller, and A. Tkatchenko
Nature Commun., *in press* (2018).

Towards Exact Quantum Dynamics for Molecules: Quantized Electrons [CCSD(T)] and Nuclei [PIMD]

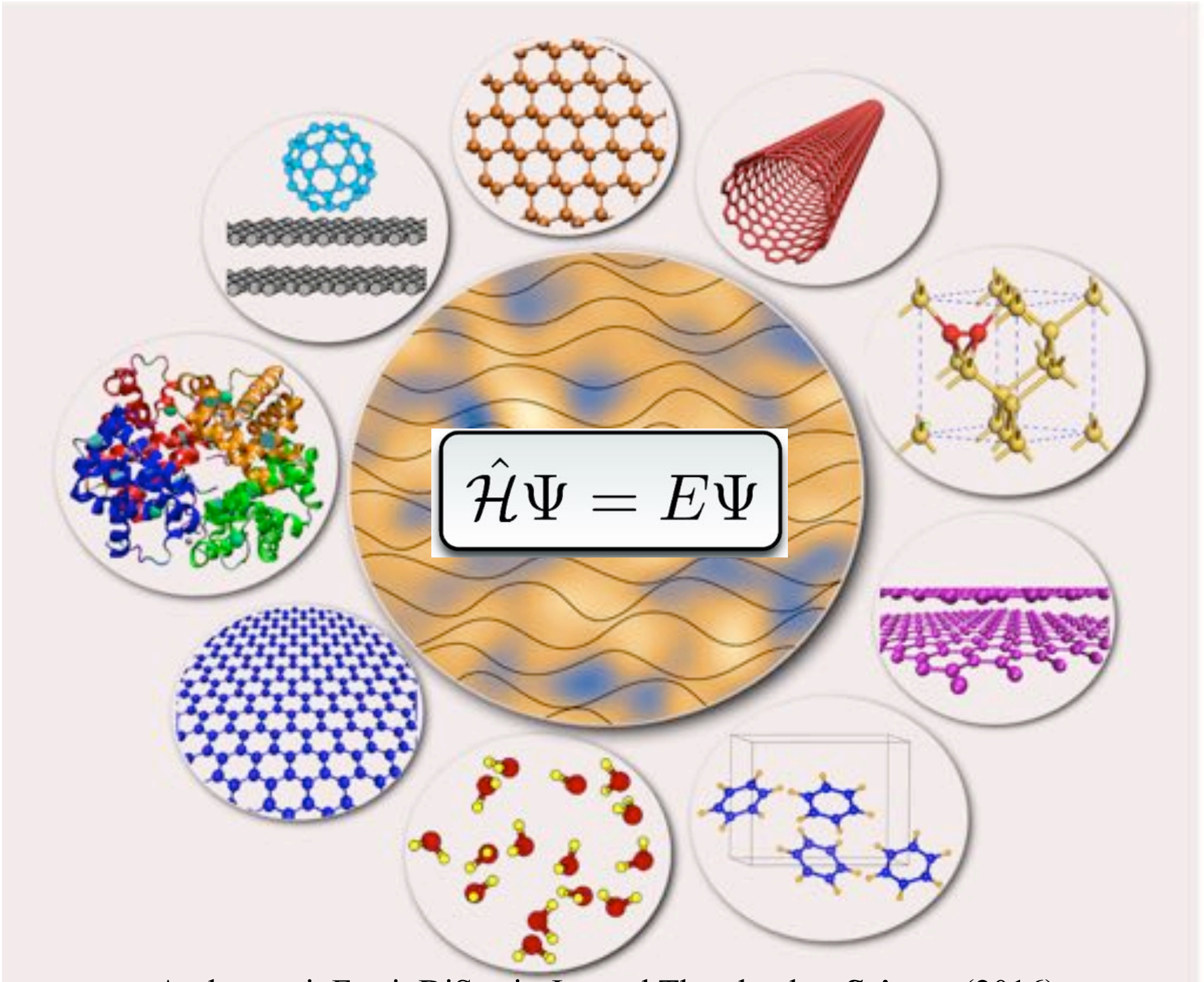


S. Chmiela, H. Sauceda, K.-R. Mueller, and A. Tkatchenko
Nature Commun., *in press* (2018).

Towards Exact Quantum Dynamics for Molecules: Quantized Electrons [CCSD(T)] and Nuclei [PIMD]



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



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

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 SE?
- More and better (big) data

Towards rational design of molecules and materials in chemical space