UNC ESHELMAN SCHOOL OF PHARMACY

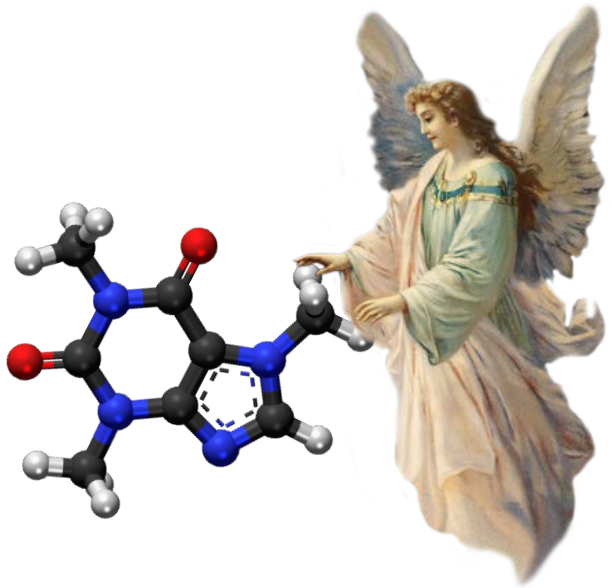
ML4MS Workshop @ Aalto
May 10, 2019

Learning Quantum Chemistry with Neural Networks

@olexandr

Olexandr Isayev
University of North Carolina at Chapel Hill
olexandr@unc.edu
http://olexandrisayev.com

The Ultimate Dream of a Computational Chemist



Challenges:

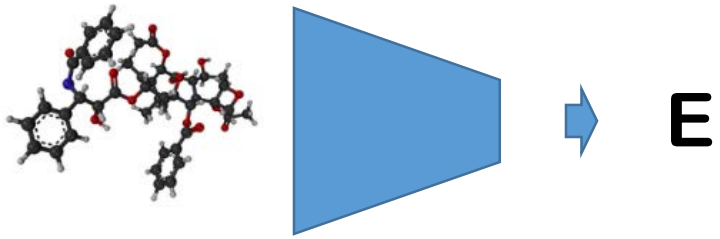
- system complexity
- length scale
- time scale
- model accuracy

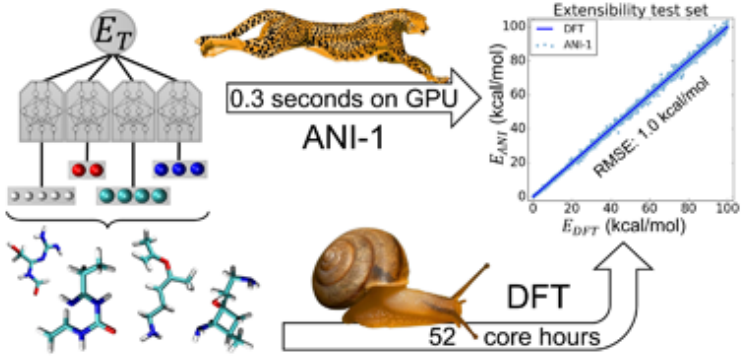
Quantum Mechanics 101

$$\hat{H}\psi = E\psi$$

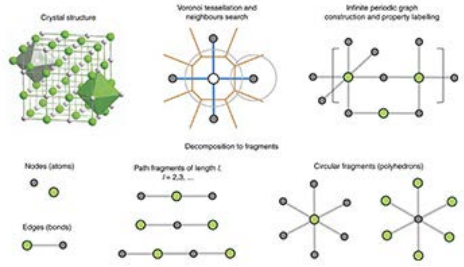
The Schrodinger equation was discovered in 1926 by Erwin Schrodinger, an Austrian theoretical physicist.
It is an important equation that is fundamental to quantum mechanics.

$$E = f(R_{vector})$$





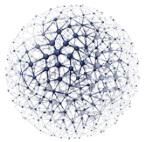
J. Chem. Phys. **2018**, 148, 241733



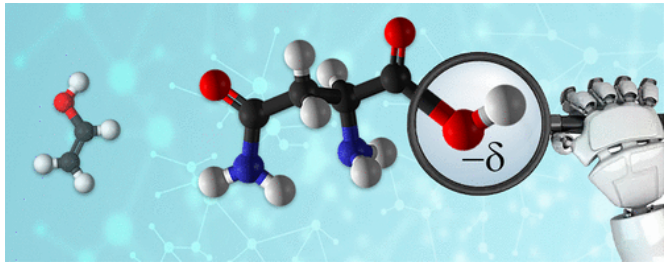
Nature Commun. **2017**, 8, 15679

Comp. Mater. Sci., 152, **2018**, 134-145

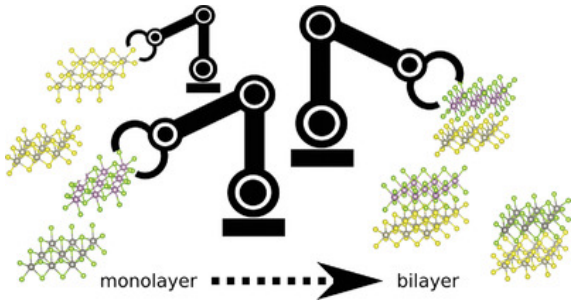
Chem. Sci., **2017**, 8, 3192-3203



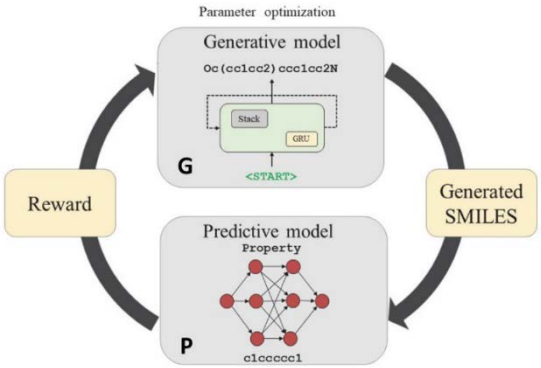
AFLOW
Automatic - FLOW for Materials Discovery



J. Phys. Chem. Lett., **2018**, 9 (16), pp 4495–4501

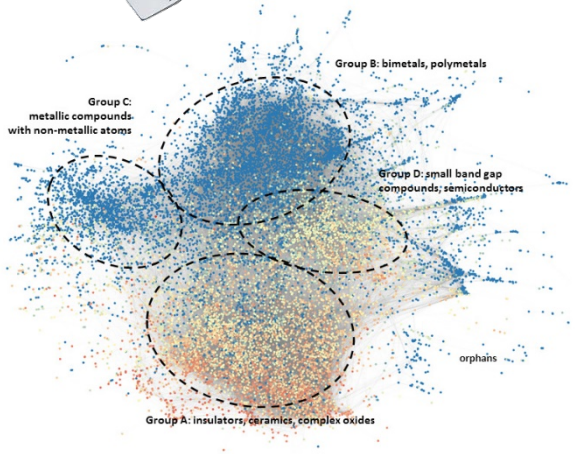


Adv. Theory Simul., **2019**, 2: 1800128

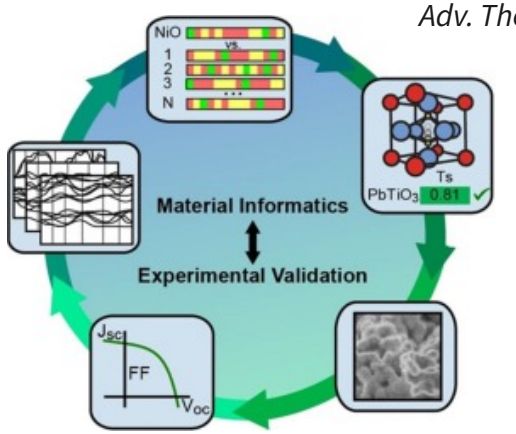


ACS Med. Chem. Lett. **2018**. 9, 1065–1069

Science Advances, **2018**, 4 (7) ,eaap7885

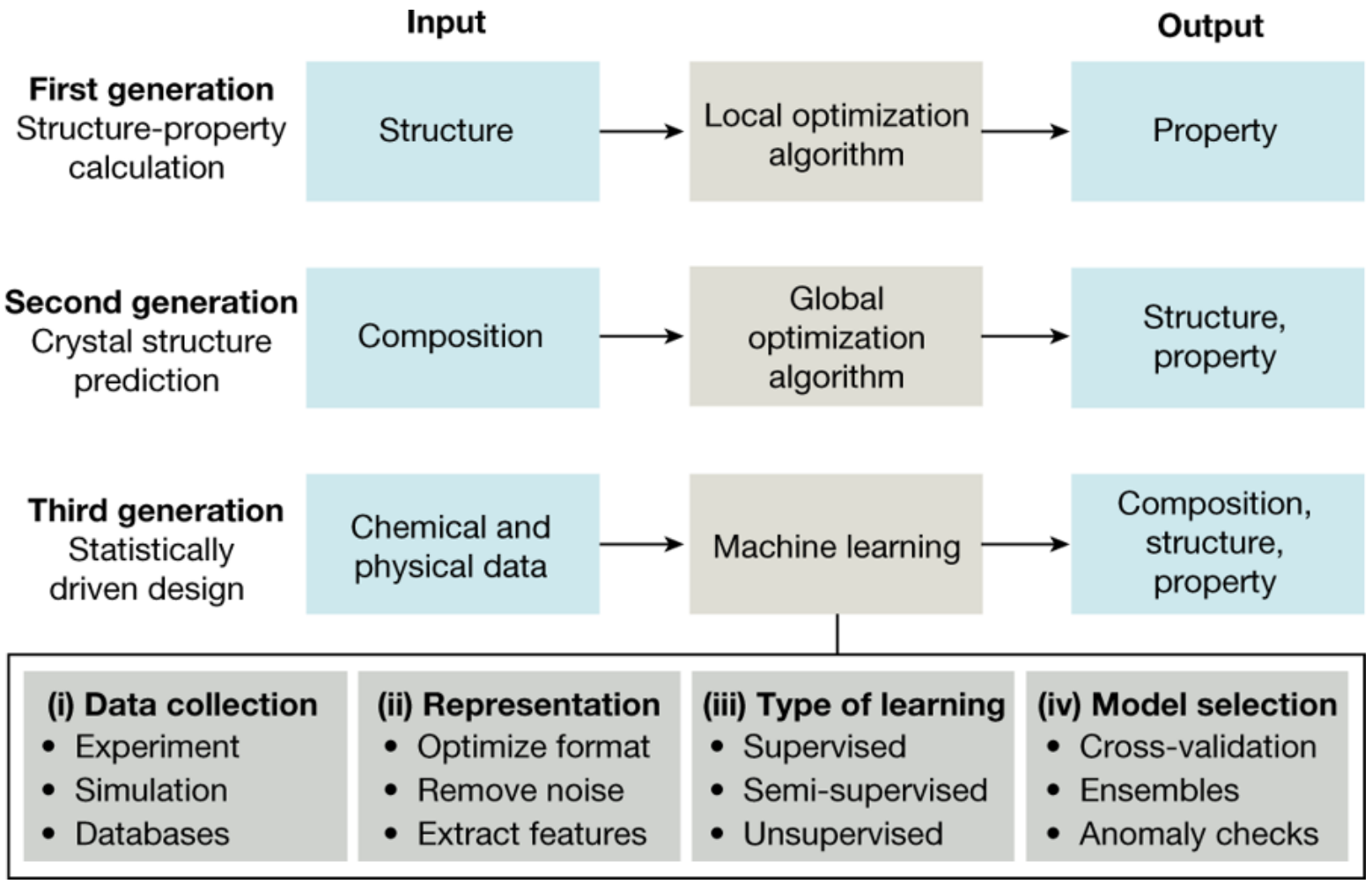


Chem. Mater., **2015**, 27, 735-742.



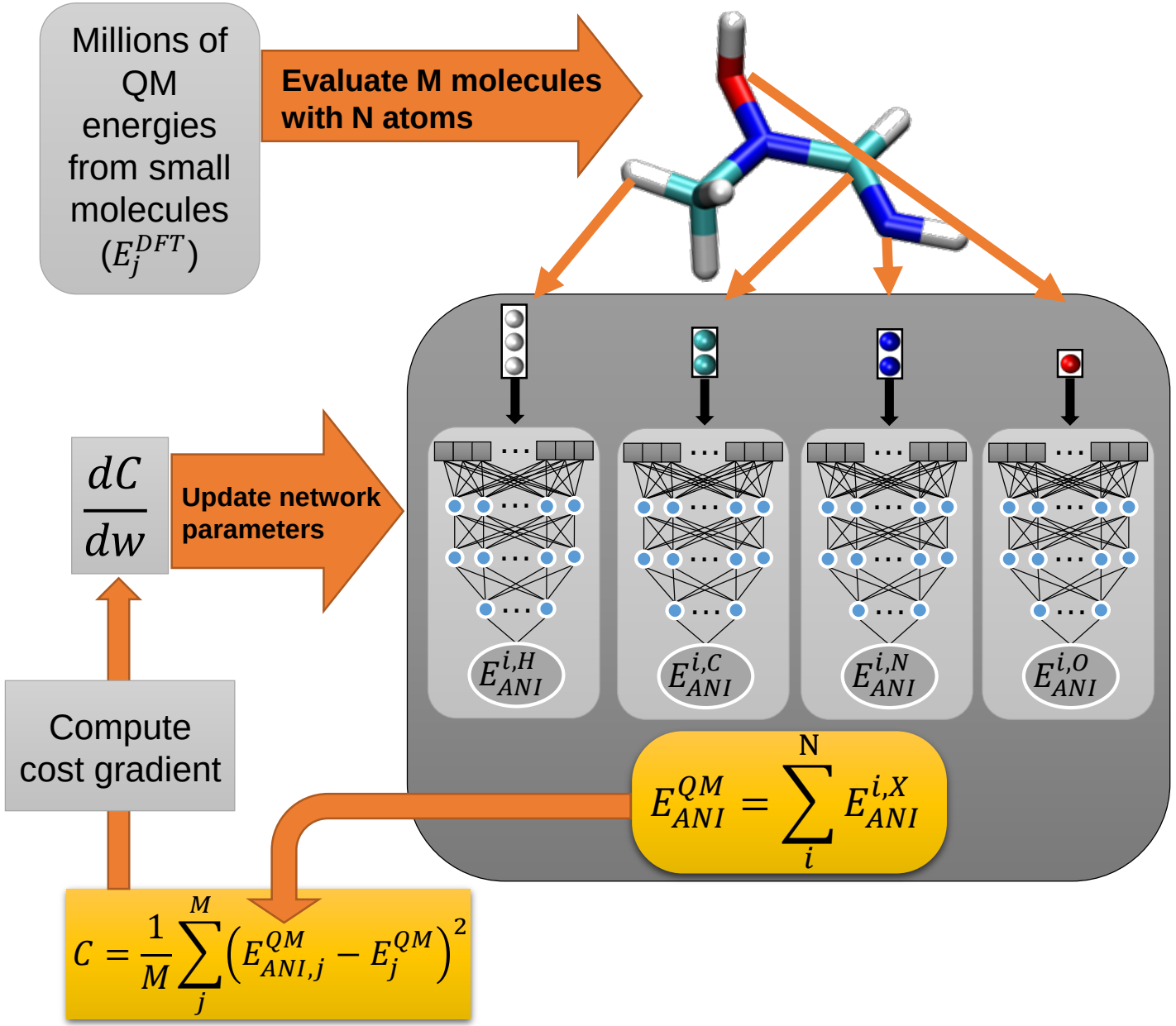
Materials Discovery. **2017**, 6, *9-16*

Evolution of statistical modeling applications in computational chemistry



Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A Machine learning for molecular and materials science. *Nature,* **559**, 547–555 (2018). DOI: 10.1038/s41586-018-0337-2

Neural Network molecular potential - training



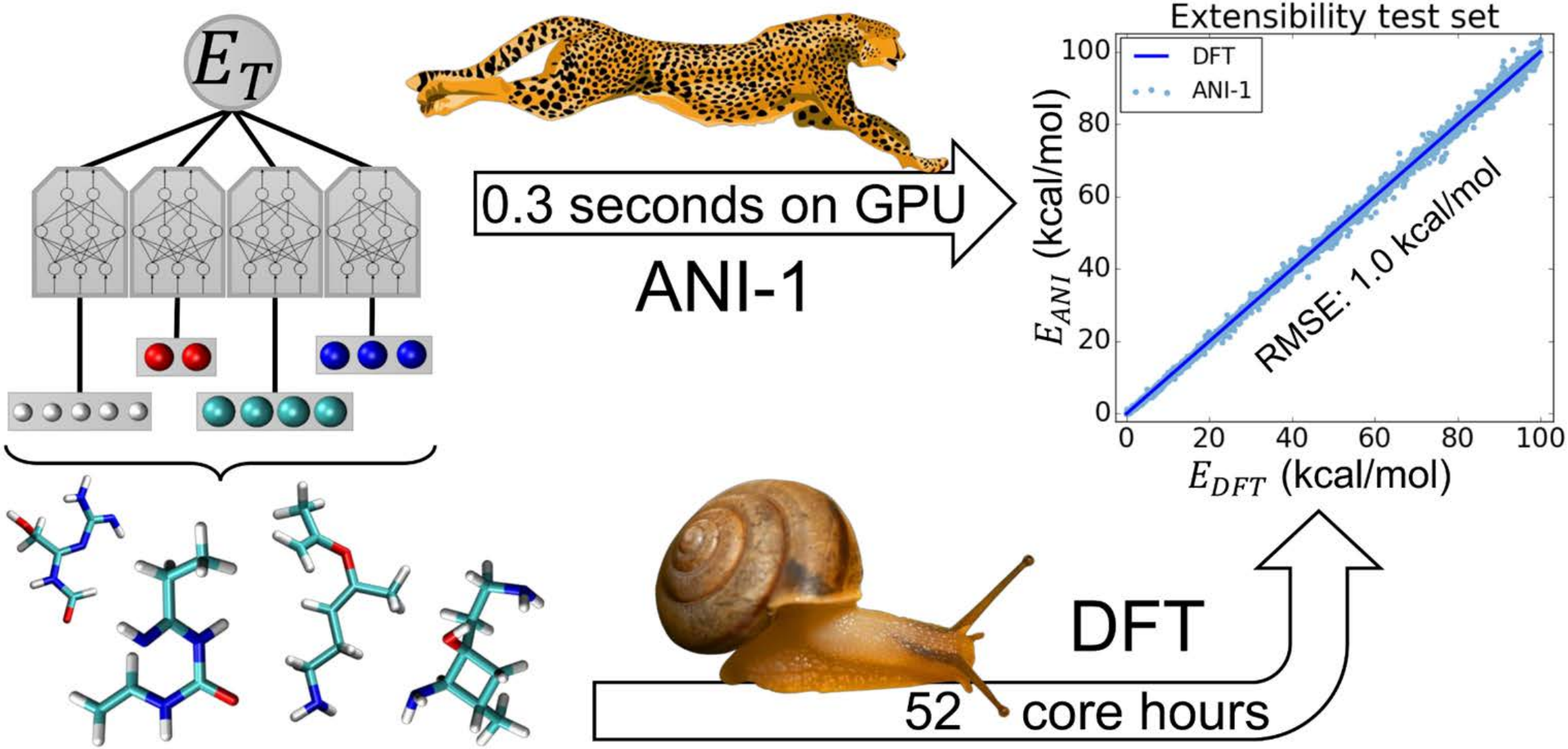
Neural Networks are not Magic!

Currently available:
CHNOSFCl

P, Si, Br, I, …
in progress

ωB97x/DZ (TZ soon)

ANI Deep Neural Network



Chem. Sci., 2017, **8**, 3192-3203

ANAKIN-ME

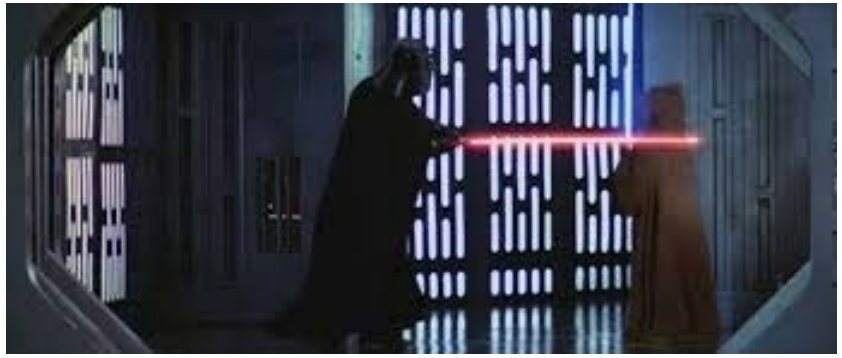
Accurate NeurAl networK engINe for Molecular Energies

We want to train a padawan network to become a DFT jedi master

ANI

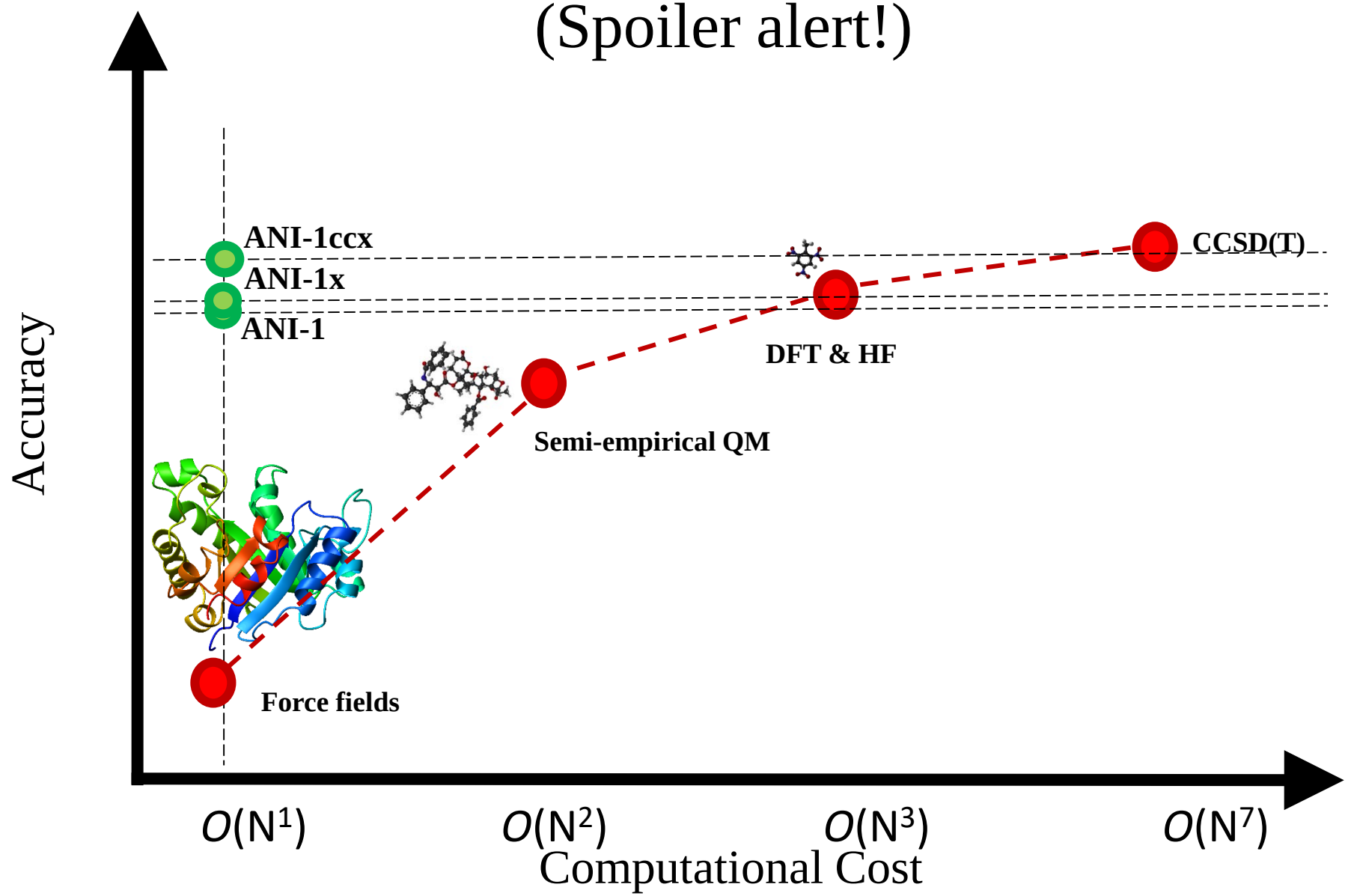
+

=



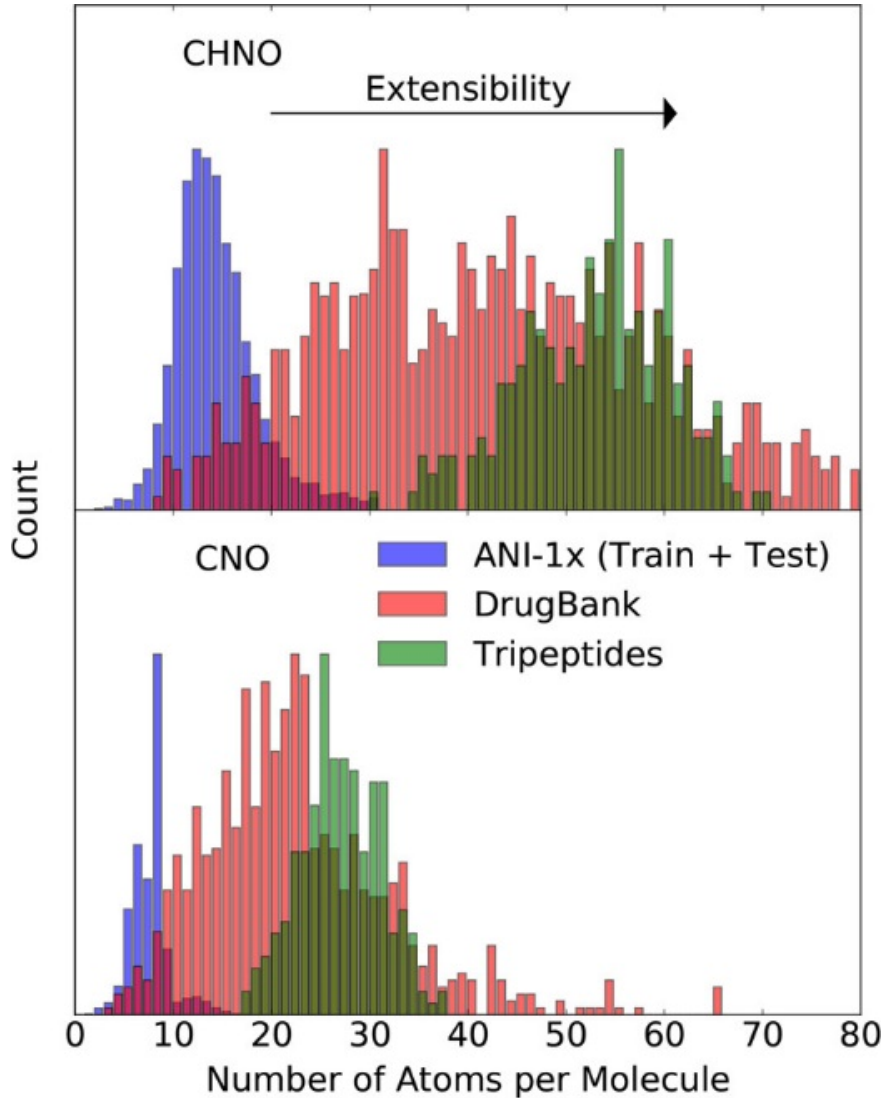
Where do we fit?

(Spoiler alert!)



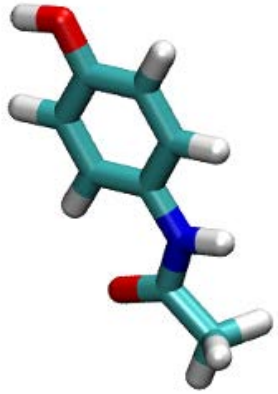
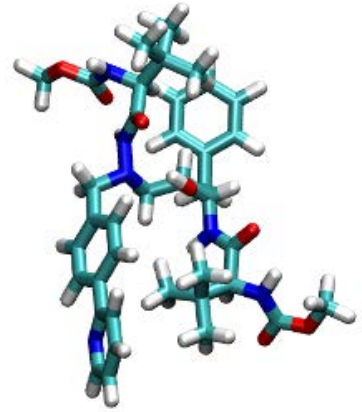
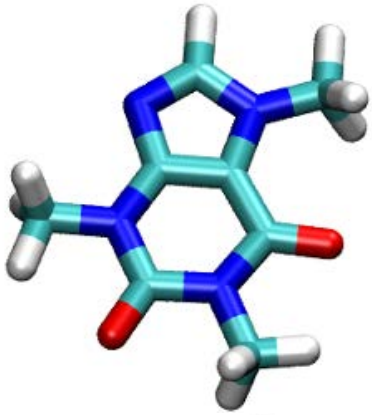
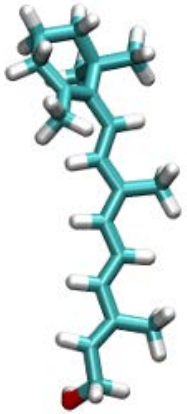
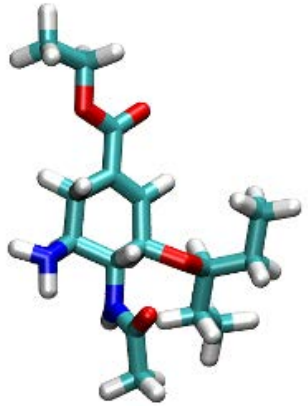
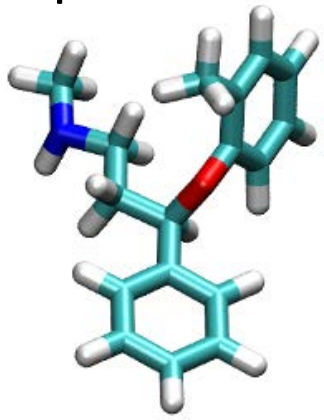
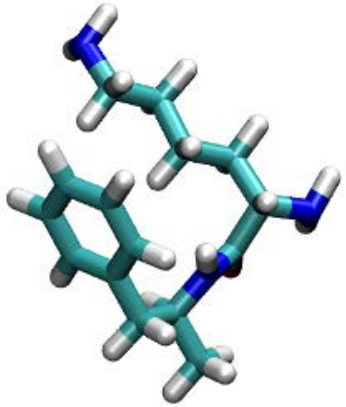
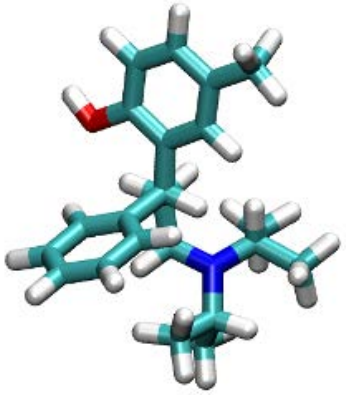
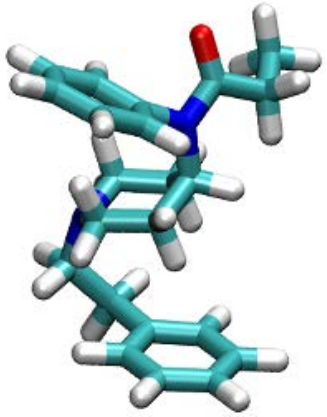
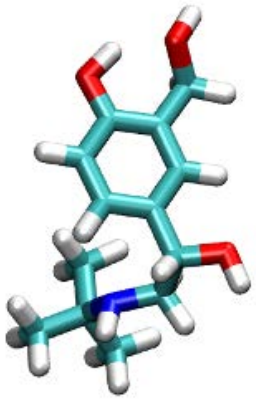
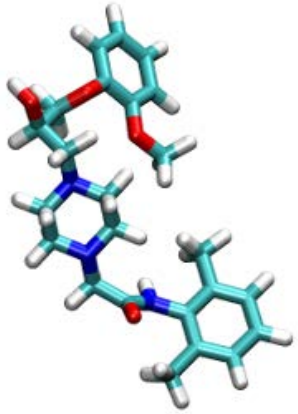
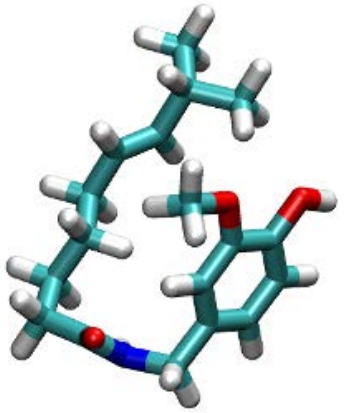
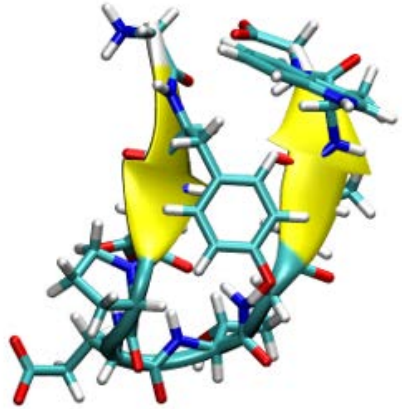
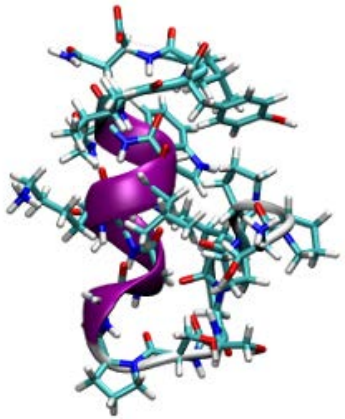
What do you need?

- ANI requires TONS of data
  - For ANI-1 we run ~20M DFT data points @ wB97x/DZ.
  - Available to anyone!
  - Molecules with 1 to 8 heavy atoms from the GDB database
  - Out-of-equilibrium geometry sampling with NMS, MD
- Train network on a fraction of available data, validate on independent data
- Test on ‘known sizes’ (Molecules with <= # max heavy atoms per molecule in training set)
  - Interpolation
- Test on ‘unknown sizes’ (Molecules larger than any in the training set)
  - Extrapolation



ANI-MD Benchmark // COMP6

- 12 drug molecules and 2 proteins
- Mean size 75 atoms (max 312 atoms)
- 1ns of molecular dynamics (MD)
- Dynamics at 300K
- MD ran on ANI-1x potential
- 128 randomly sampled frames

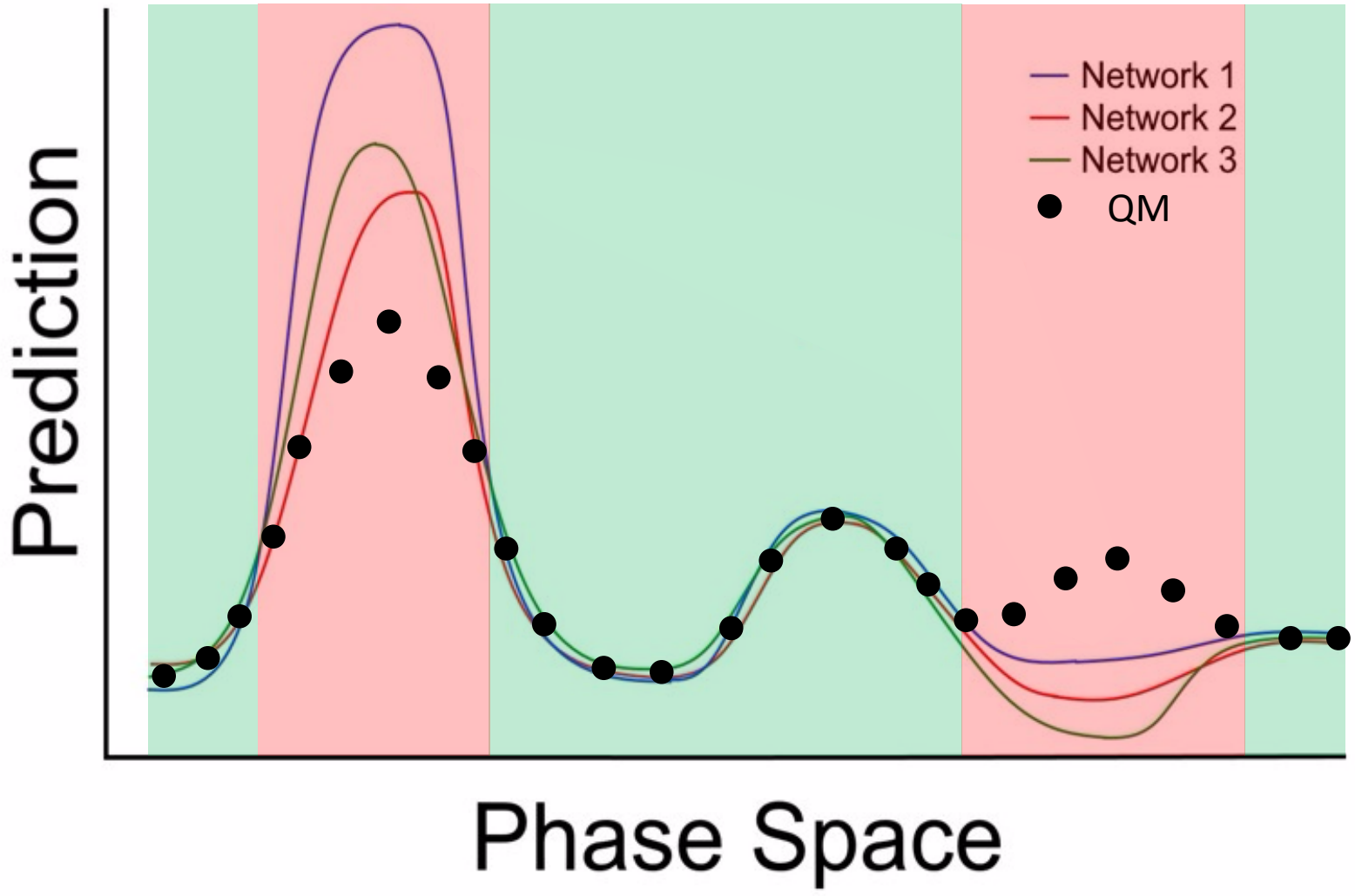















Can we predict when the model is wrong?

Ensemble disagreement can drive data generation

Good data coverage

Bad data coverage

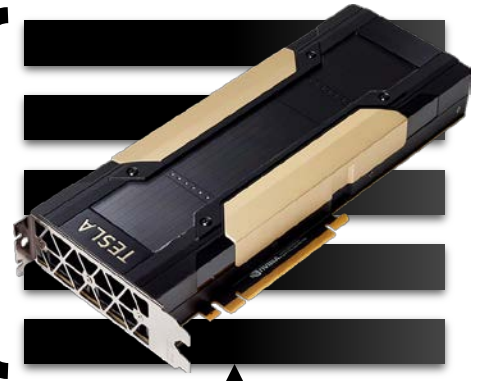


JS Smith, O.Isayev, A. Roitberg; *Journal of Chemical Physics*, (**2018**), 148 (24), 241733

Active Learning - The Big Picture

An automated and self-consistent data generation framework

Ensemble of ANI networks



Non-equilibrium Conformational sampler

New test data

Check ensemble disagreement

Train network ensemble

Molecule Sampling

(e.g. GDB small molecule database, small peptides, drug like molecules)

Structure Pools

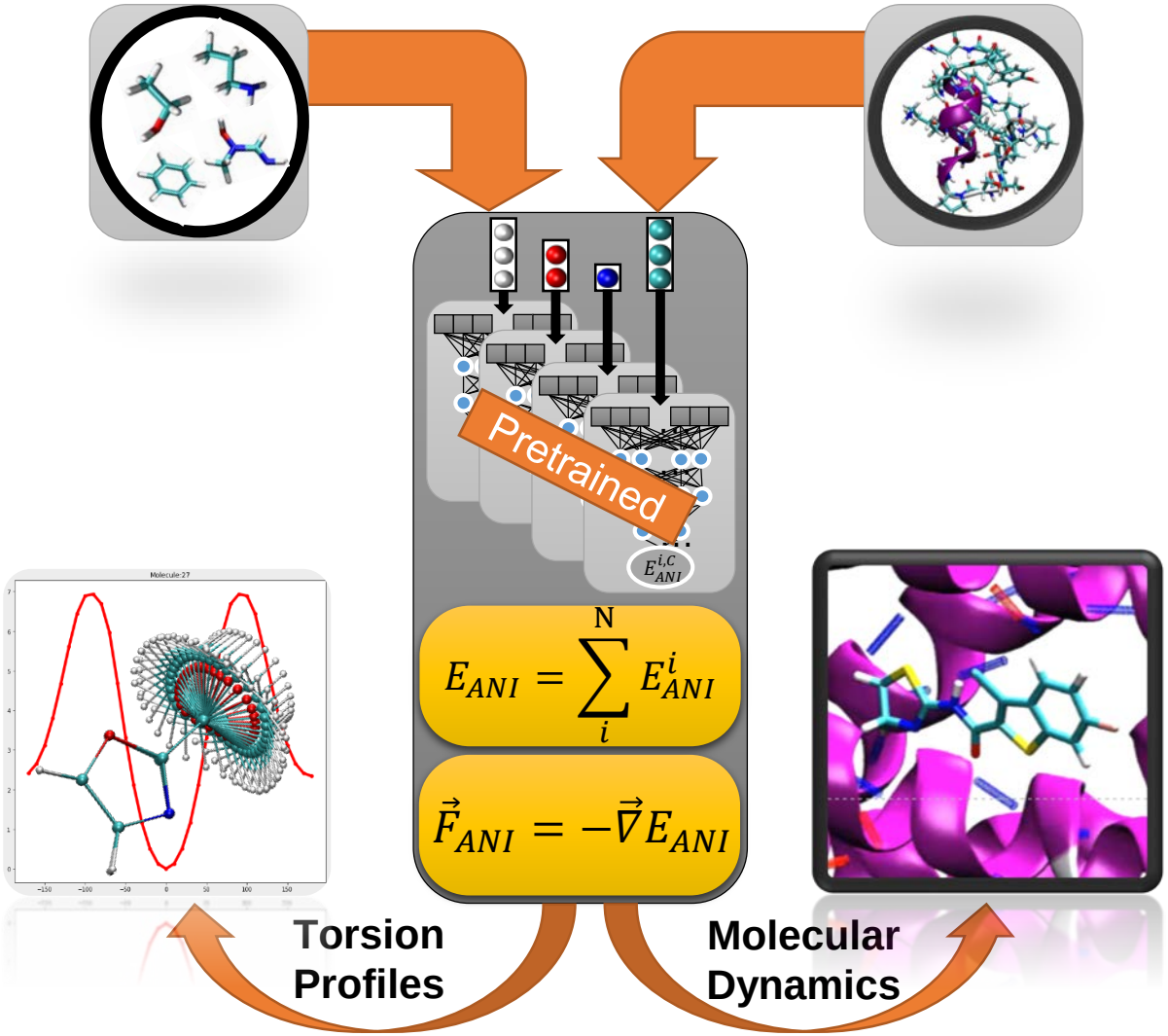
Compute Cluster

Computations with QM

ANI-1x Dataset (i.e. energies, forces, dipoles)

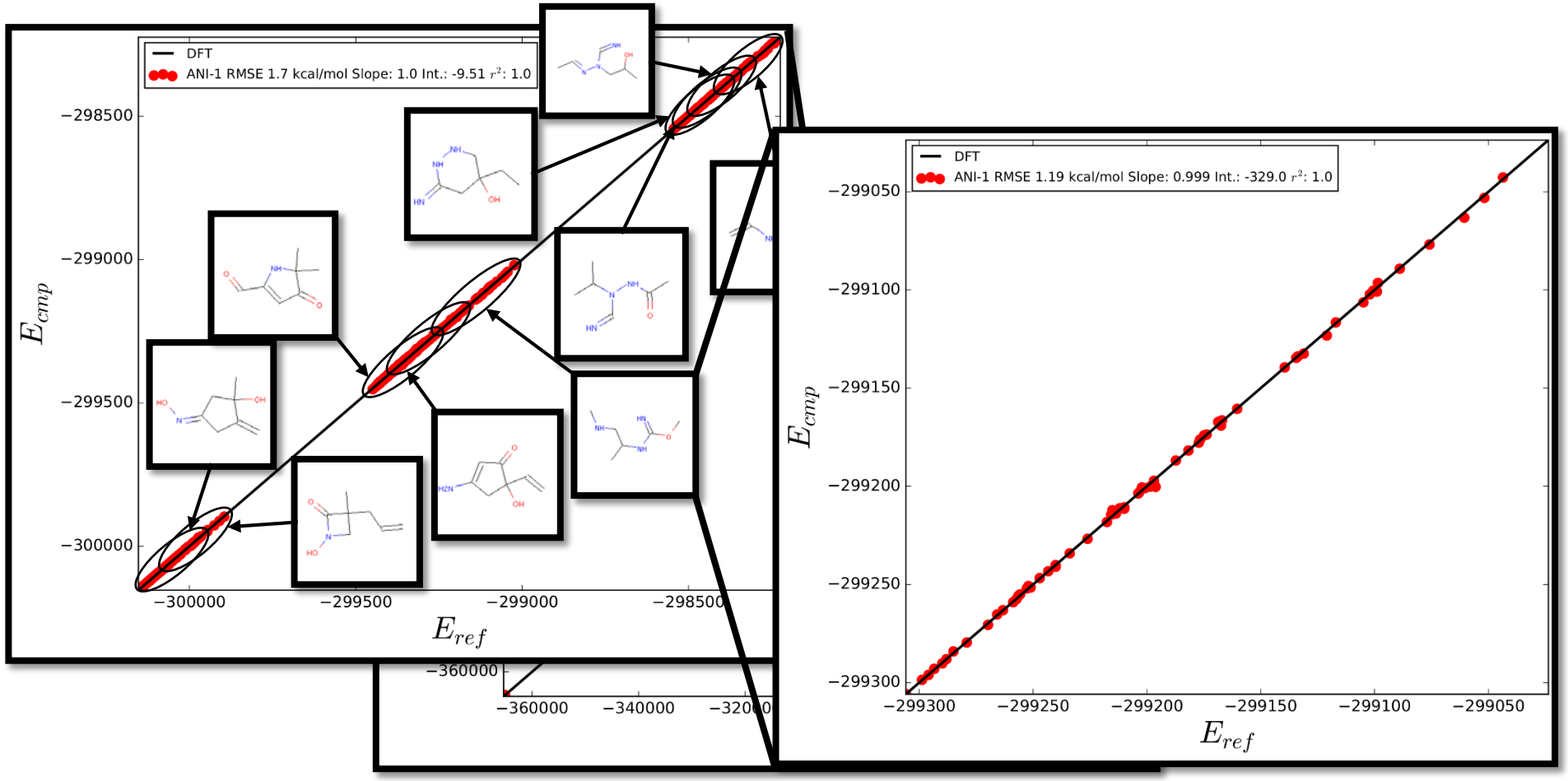
JS Smith, O.Isayev, A. Roitberg;
Journal of Chemical Physics, (**2018**), 148 (24), 241733

ANI molecular potential - application

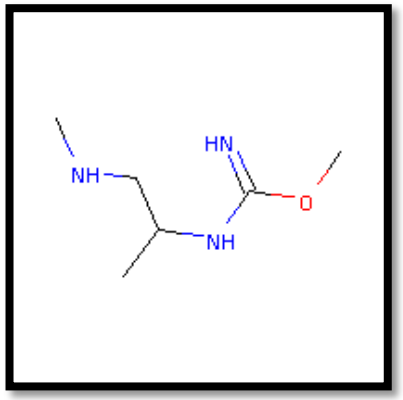


Total energy correlation for ANI-1 vs. DFT

(External test of 131 molecules with 10 heavy atoms, 8200 total molecules + conformations) [units: kcal/mol]

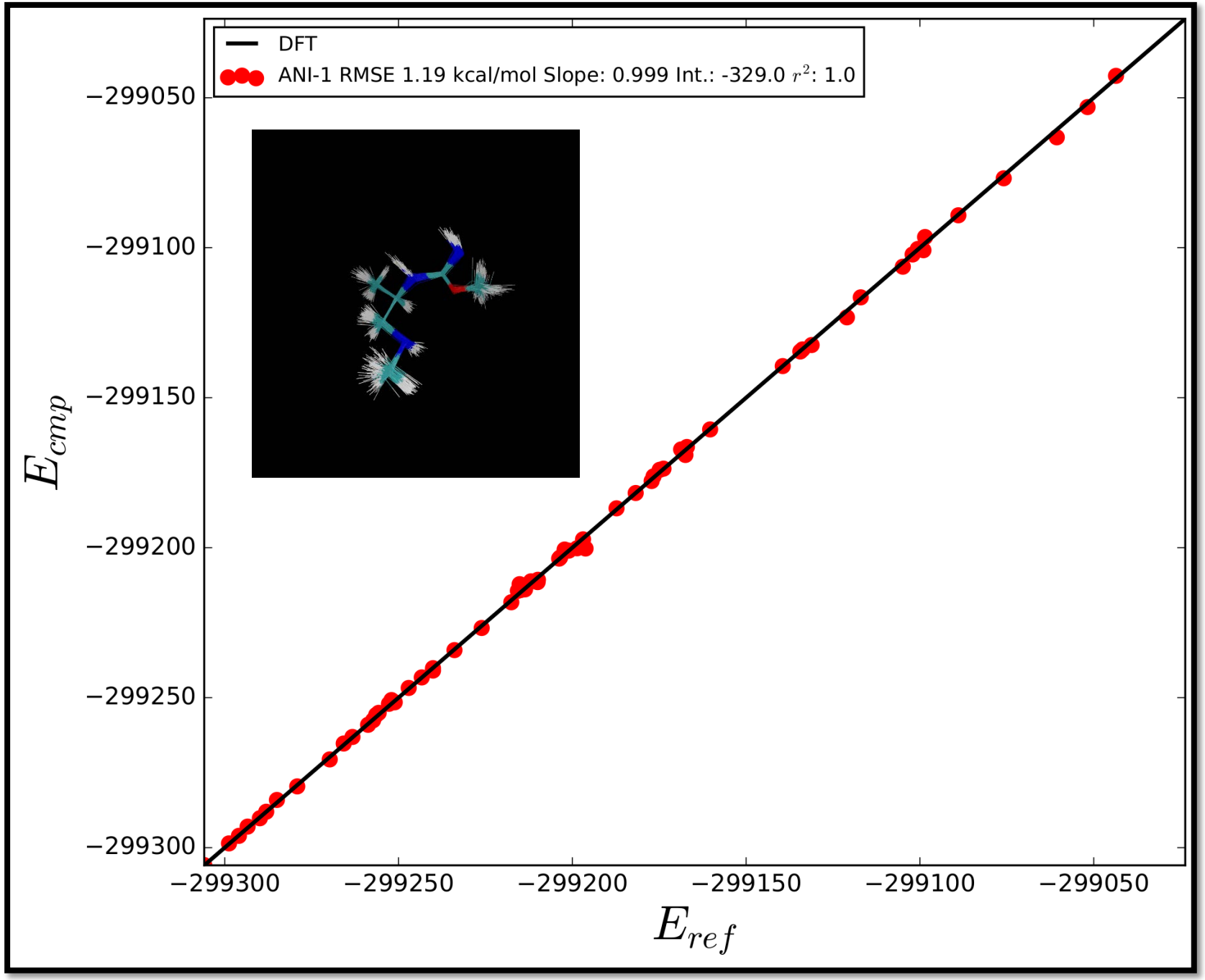


J. Smith, O. Isayev, A. Roitberg. ***Chem. Sci.***, 2017, **8**, 3192-3203

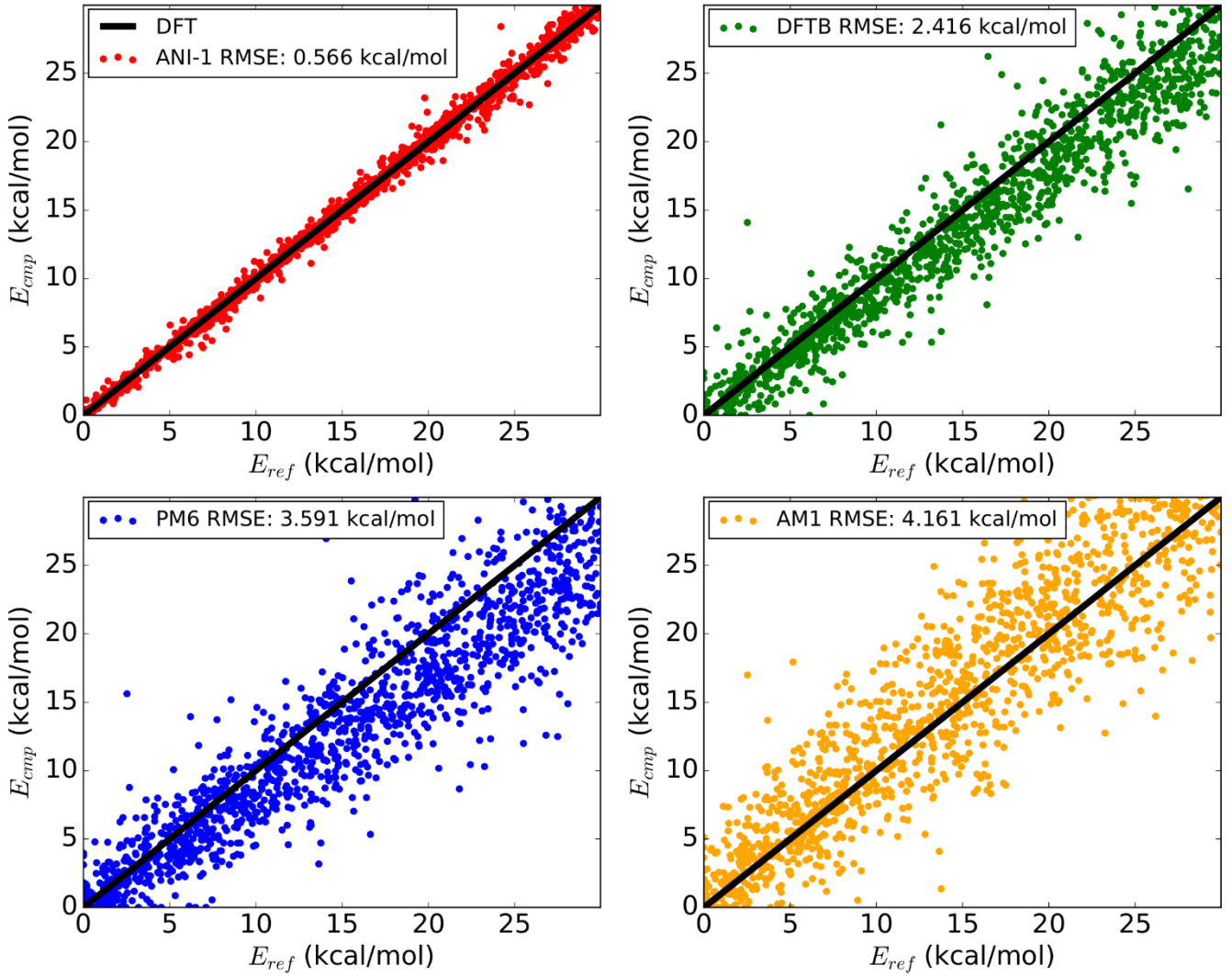


73 total test structures
10 Heavy atoms
25 Total atoms
RMSE: 1.2 kcal/mol
(0.048 kcal/mol/atom)
DFT time: 1143.11s
ANI time: 0.0032s

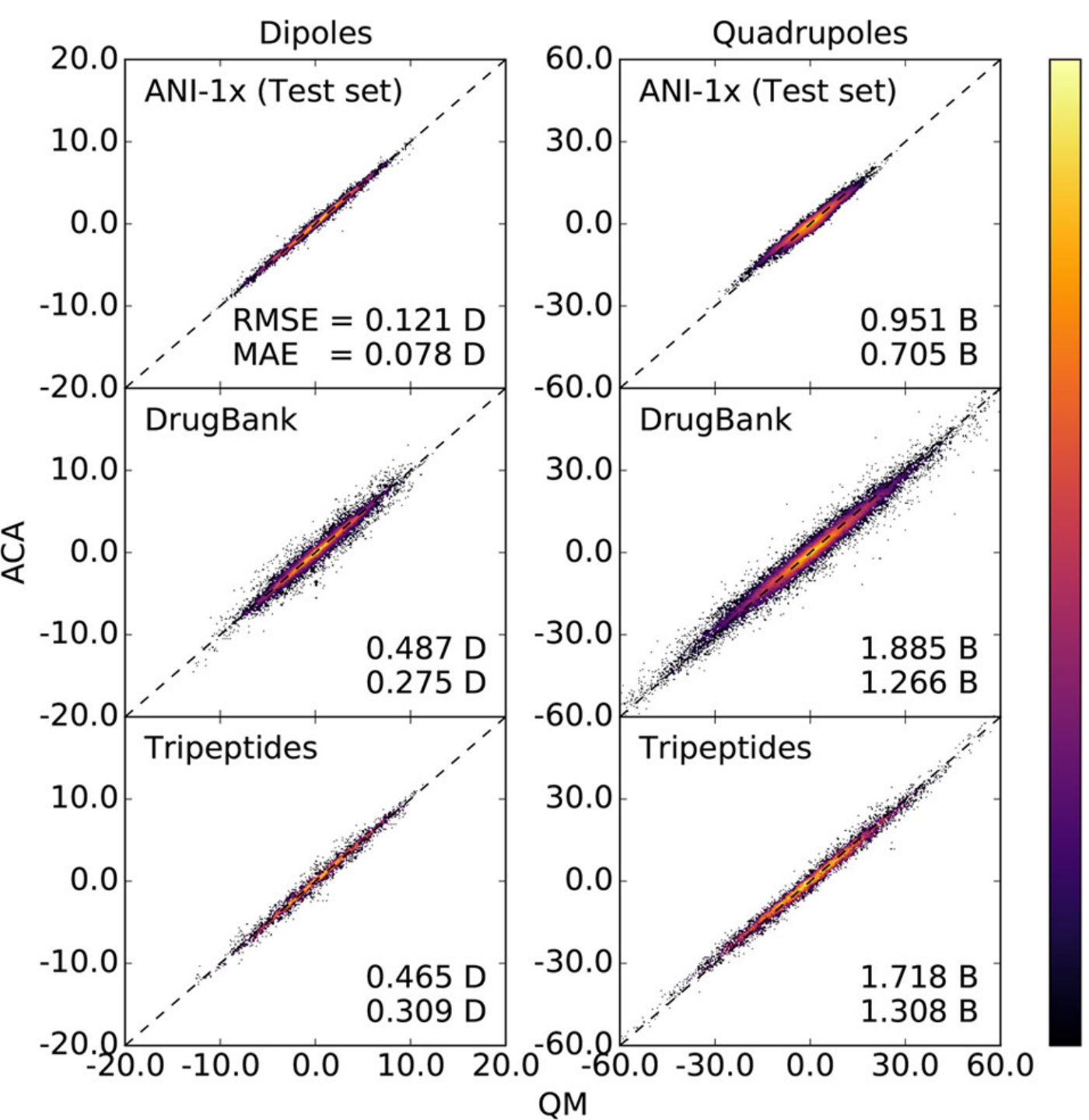
357,000x speedup!

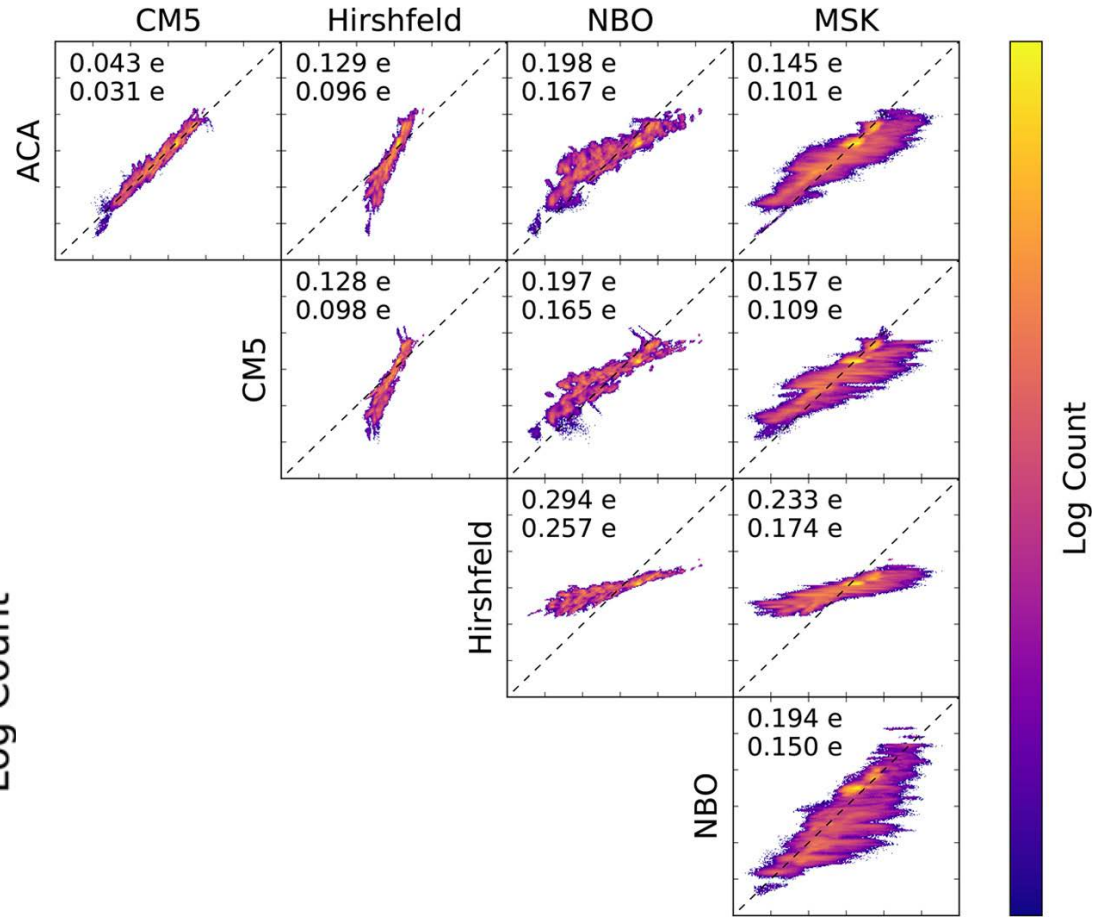


Relative Energy correlation (30kcal/mol)



J. Smith, O. Isayev, A. Roitberg. ***Chem. Sci.***, 2017, **8**, 3192-3203



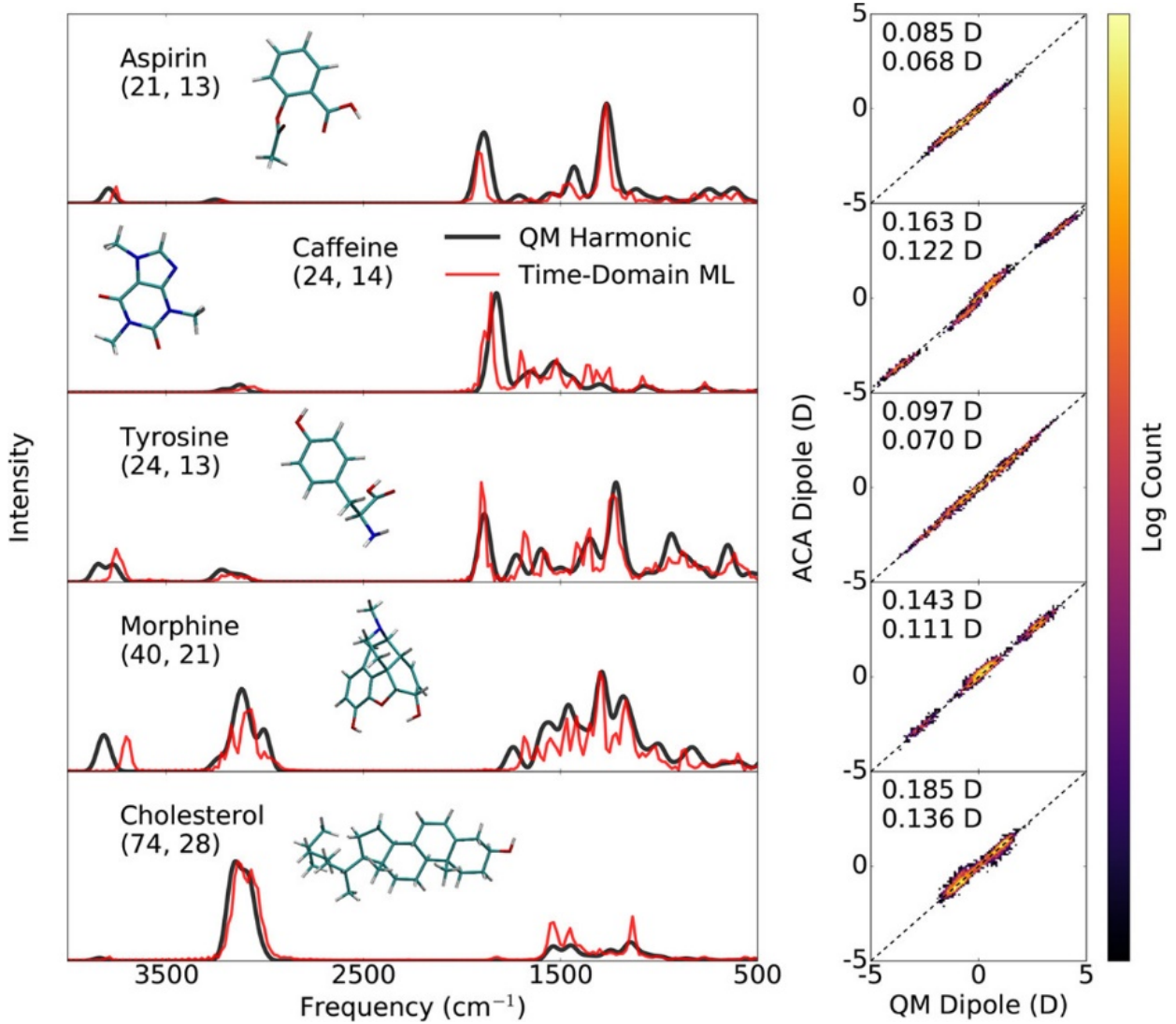


Discovering a Transferable Charge Assignment Model Using Machine Learning

A.E. Sifain, N. Lubbers, B.T. Nebgen, J.S. Smith, A.Y. Lokhov, O. Isayev, A. E. Roitberg, K. Barros, S. Tretiak.

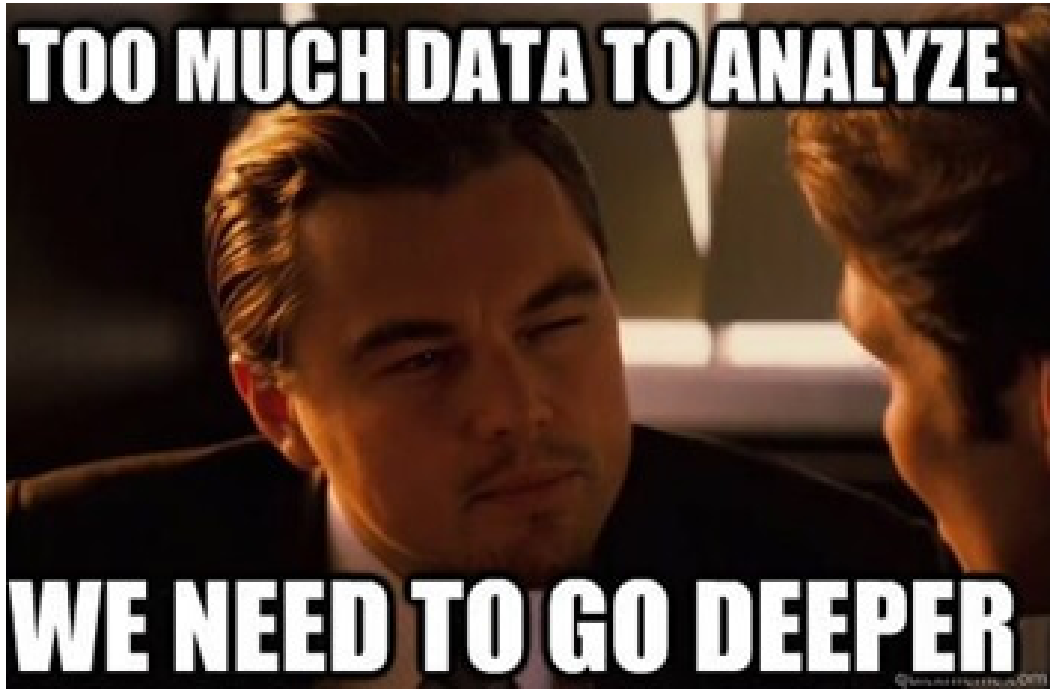
J. Phys. Chem. Lett. 9, **2018**, 4495-4501

Accurate IR spectra simulation with time-domain ML



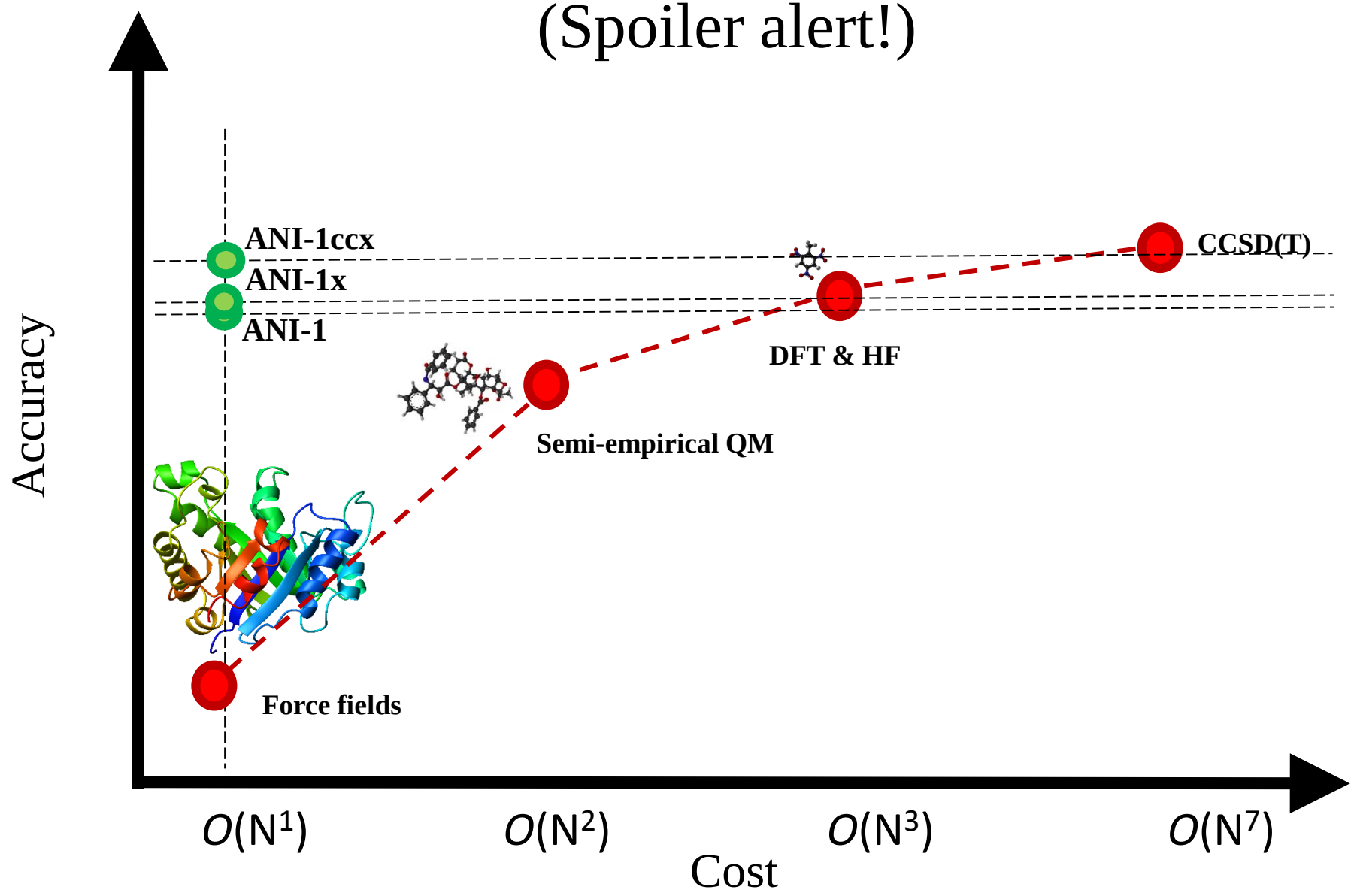
A.E. Sifain, N. Lubbers, B.T. Nebgen, J.S. Smith, A.Y. Lokhov, O. Isayev, A. E. Roitberg, K. Barros, S. Tretiak.
J. Phys. Chem. Lett. 2018. DOI: 10.1021/acs.jpclett.8b01939

Can we go beyond DFT?



Where do we fit?

(Spoiler alert!)



High Throughput CSDT(T)/CBS

	CPU-core hours		**Mean absolute deviation from CCSD(T)-F12 (kcal/mol)**	
	Alanine (13 atoms)	Aspirin (21 atoms)	S66	W4-11
CCSD(T)/CBS	9.13	427.00	0.03	1.31
CCSD(T)*/CBS (this work)	**1.44**	**7.44**	**0.09**	**1.46**

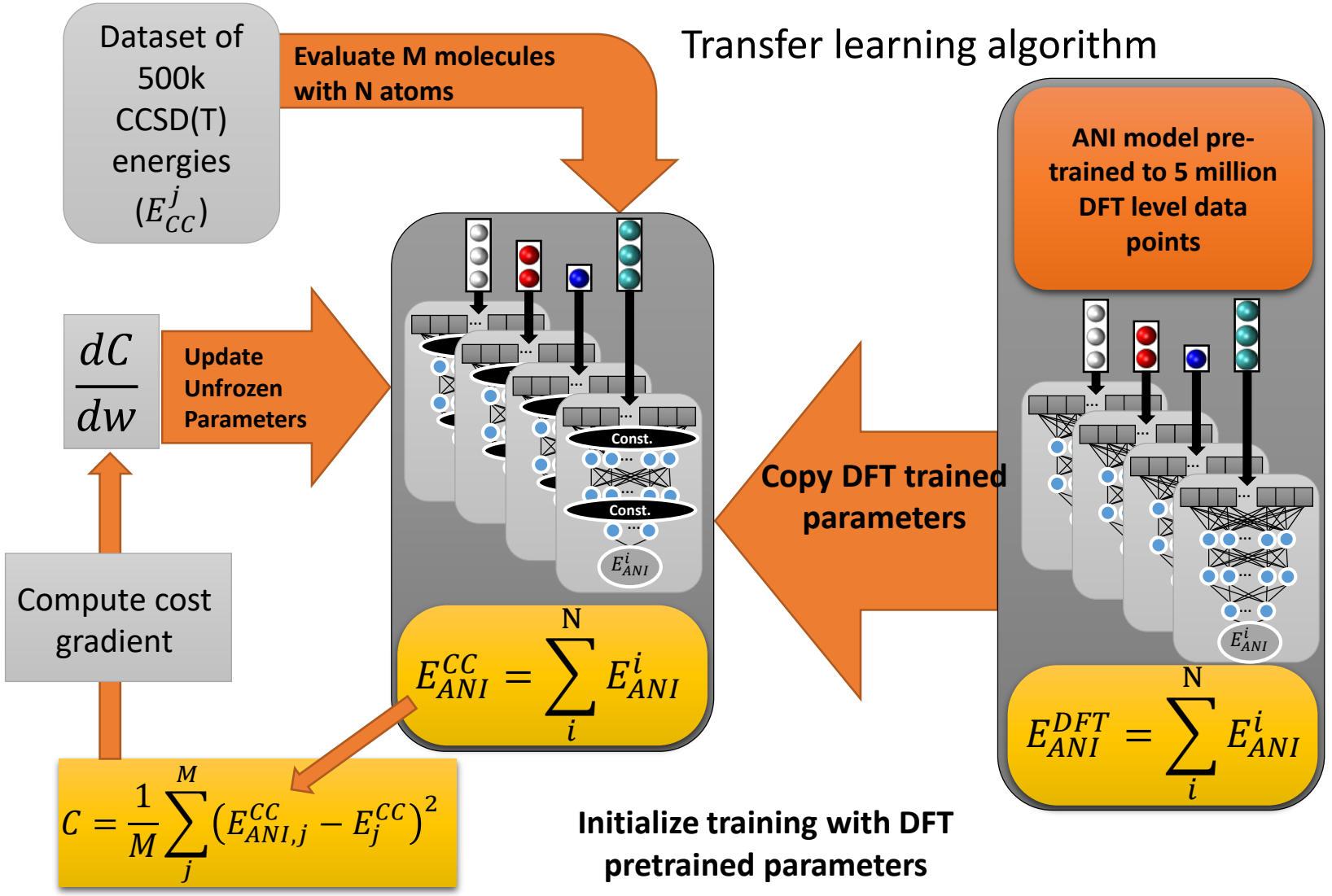
$$E_{total}^{CBS} \approx E_{HF}^{CBS} + E_{MP2}^{CBS} + \left(E_{CCSD(T)}^{cc-pVTZ} - E_{MP2}^{cc-pVTZ}\right)$$

$$E_{CCSD(T)}^{cc-pVTZ} \approx E_{Normal-DPLNO-CCSD(T)}^{cc-pVTZ} + (E_{Tight-DPLNO-CCSD(T)}^{cc-pVDZ} - E_{Normal-DPLNO-CCSD(T)}^{cc-pVDZ})$$

J.S. Smith et al. Outsmarting Quantum Chemistry Through Transfer Learning. *Nature Comm.* **2019**. Accepted. 10.26434/chemrxiv.6744440

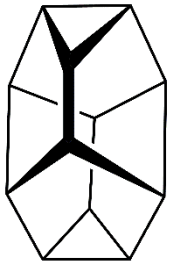
Transferring knowledge of CCSD(T)/CBS

- Regenerate 10% of ANI-1x training data (0.5M of 5M)
- For high-level reference we use CCSD(T)/CBS accurate QM model
- We only retrain 60k of 400k neural network parameters
- Results show clear improvement over DFT trained model
- New models are exceeding the DFT in accuracy

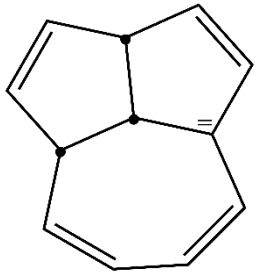


J.S. Smith et al. Outsmarting Quantum Chemistry Through Transfer Learning. *Nature Comm.* **2019**. Accepted. 10.26434/chemrxiv.6744440

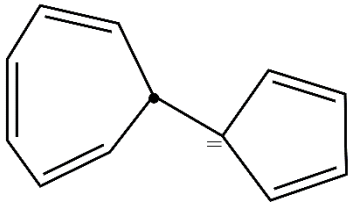
Hydrocarbon reaction energy benchmark, DFT vs CCSD(T)



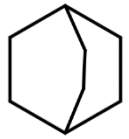
E1 (1)



E2 (22)



E3 (31)

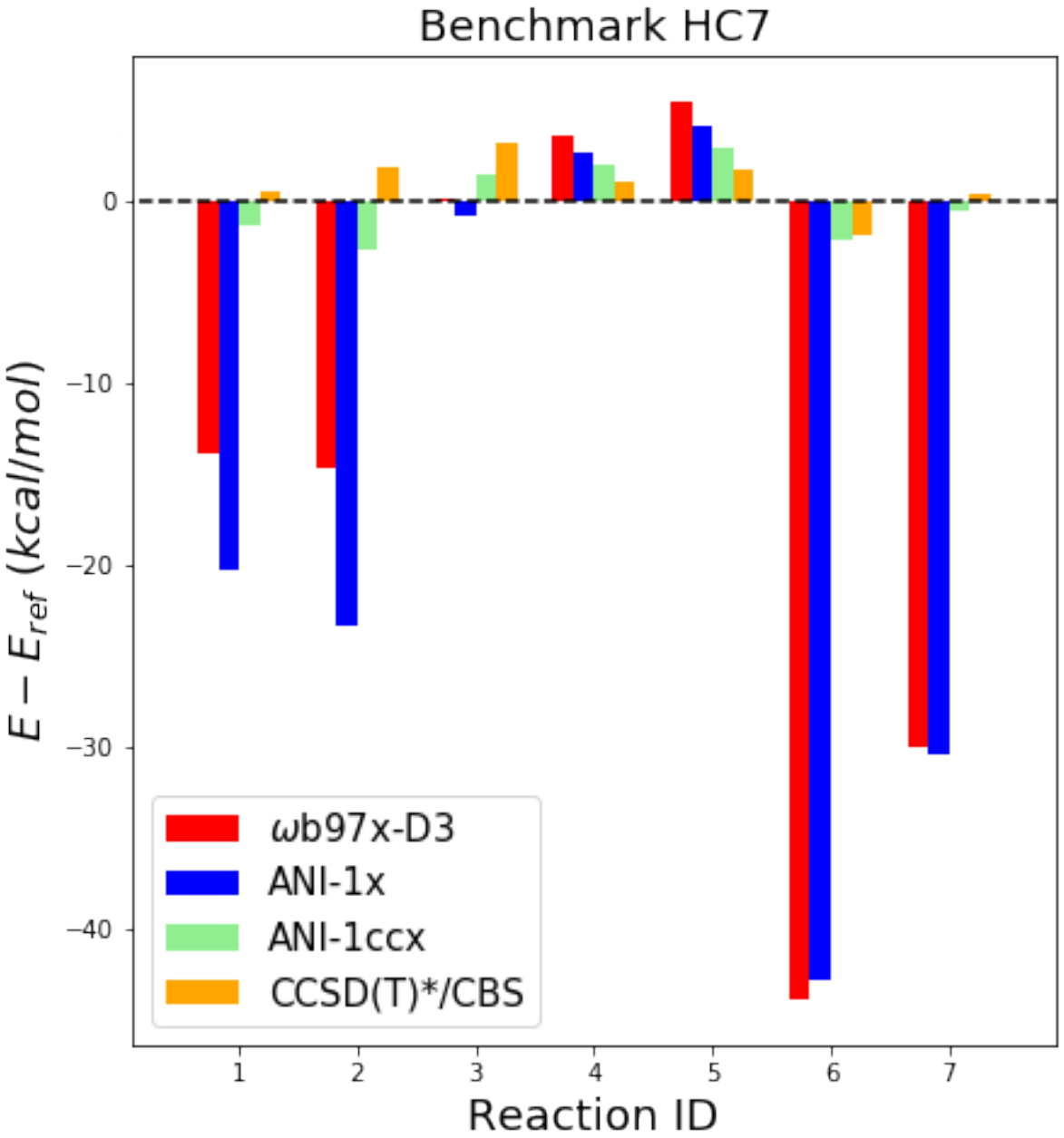


E4
(Bicyclo[2.2.2]octane)

Units: kcal/mol

Reaction	Ref.	ANI-1ccx	CCSD(T)*/CBS	ANI-1x	ωb97x
1) $\mathrm{E1} \rightarrow \mathrm{E2}$	14.3	15.6	13.8	34.6	28.2
2) $\mathrm{E1} \rightarrow \mathrm{E3}$	25.0	27.7	23.1	48.3	39.7
3) Octane-a → Octane-b	1.9	0.4	-1.3	2.7	1.7
4) $4CH_4 + C_6H_{14} \rightarrow 5C_2H_6$	9.8	7.9	8.7	7.2	6.2
5) $6CH_4 + C_8H_{18} \rightarrow 7C_2H_6$	14.8	11.9	13.1	10.8	9.3
6) Adamantane → $3CH_4 + 2C_2H_2$	194.0	196.2	195.9	236.8	238.0
7) E4 → $3CH_4 + 2C_2H_2$	127.2	127.8	126.9	157.7	158.0

Reference data: Peverati, R.; Zhao, Y.; Truhlar, D. G., *J. Phys. Chem. Lett.* **2011**, *2* (16), 1991–1997.



ChemRxiv: 10.26434/chemrxiv.6744440

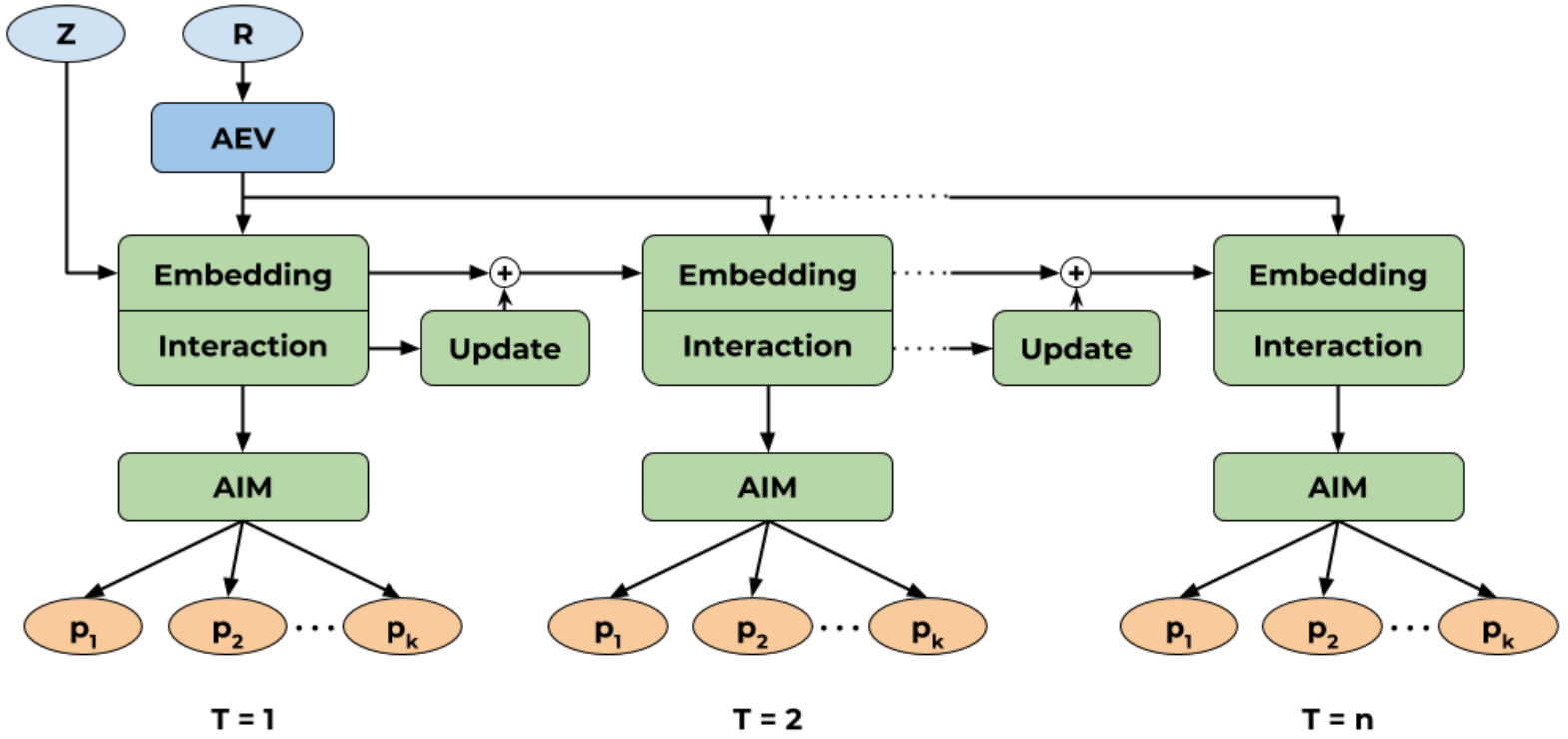
Can we go beyond simple energies?

Rethinking Network Architecture: AIMNet

Atoms-in-molecules neural net

Iterative “SCF-like” update for better accuracy and
Long range interactions

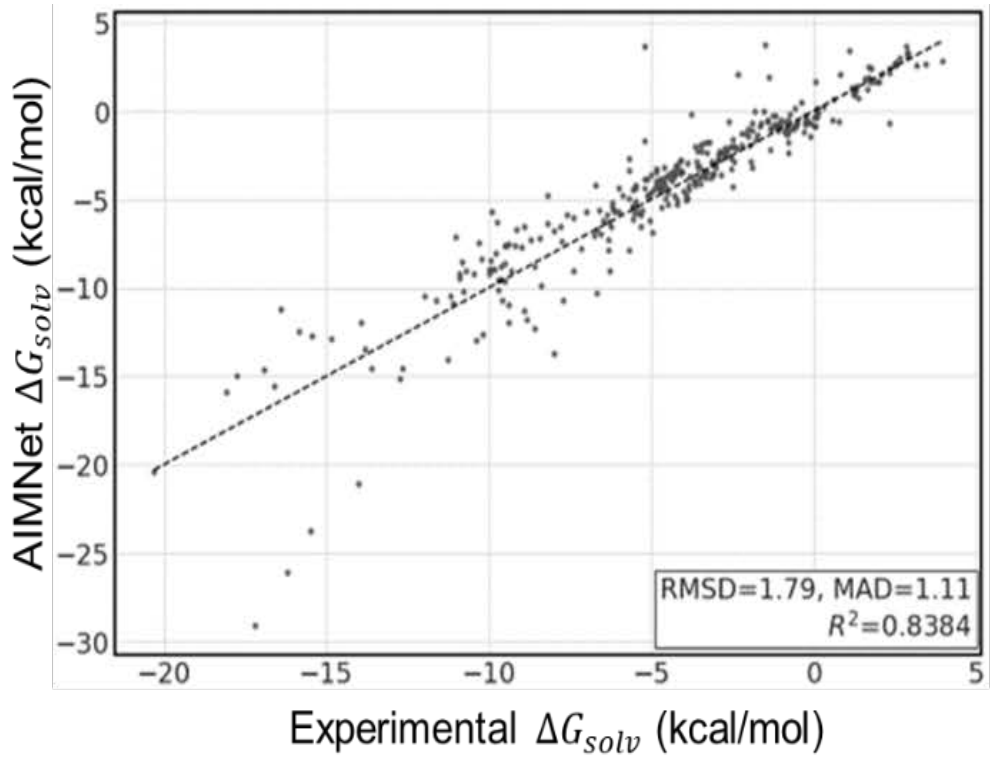
Multimodal and
multi-task learning: gas phase energy, charges, atomic volumes, continuum solvent (SMD) Correction

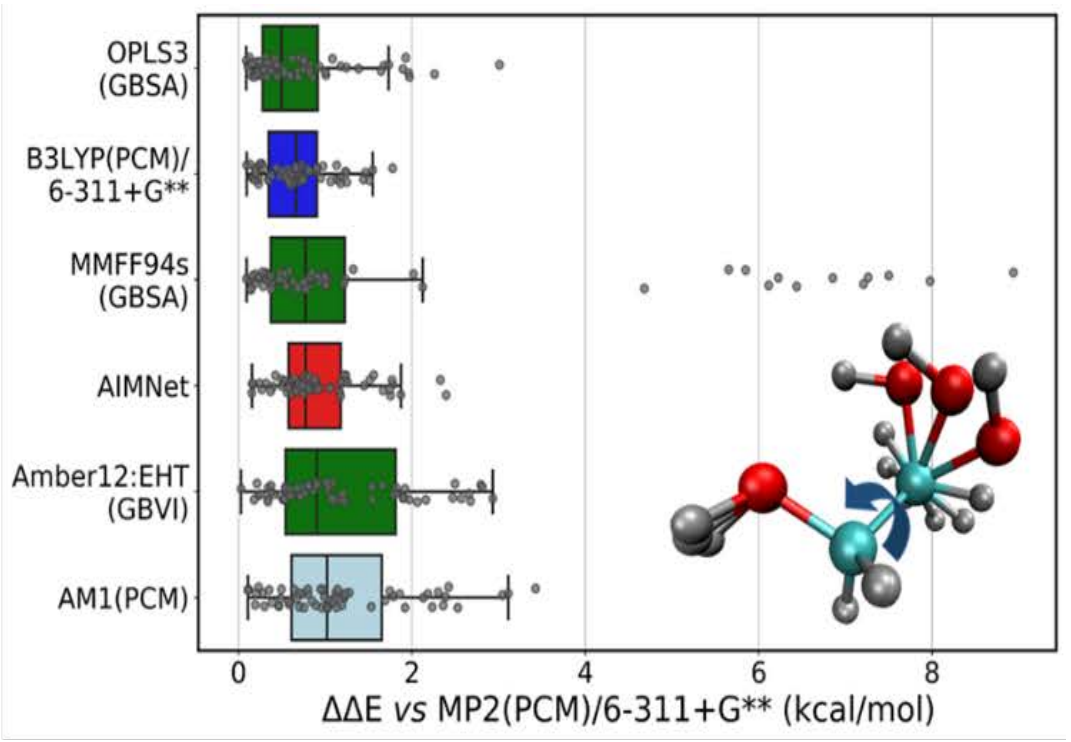


Deep NN network, AIMNet with T=3:
33 hidden layers, ~1M parameters

R. Zubatyuk, J.S. Smith, J. Leszczynski, O. Isayev, Accurate and Transferable Multitask Prediction of Chemical Properties with an Atoms-in-Molecule Neural Network. ChemRxiv, 2018. Submitted.

Fast & Accurate Solvation Free Energies with AIMNet

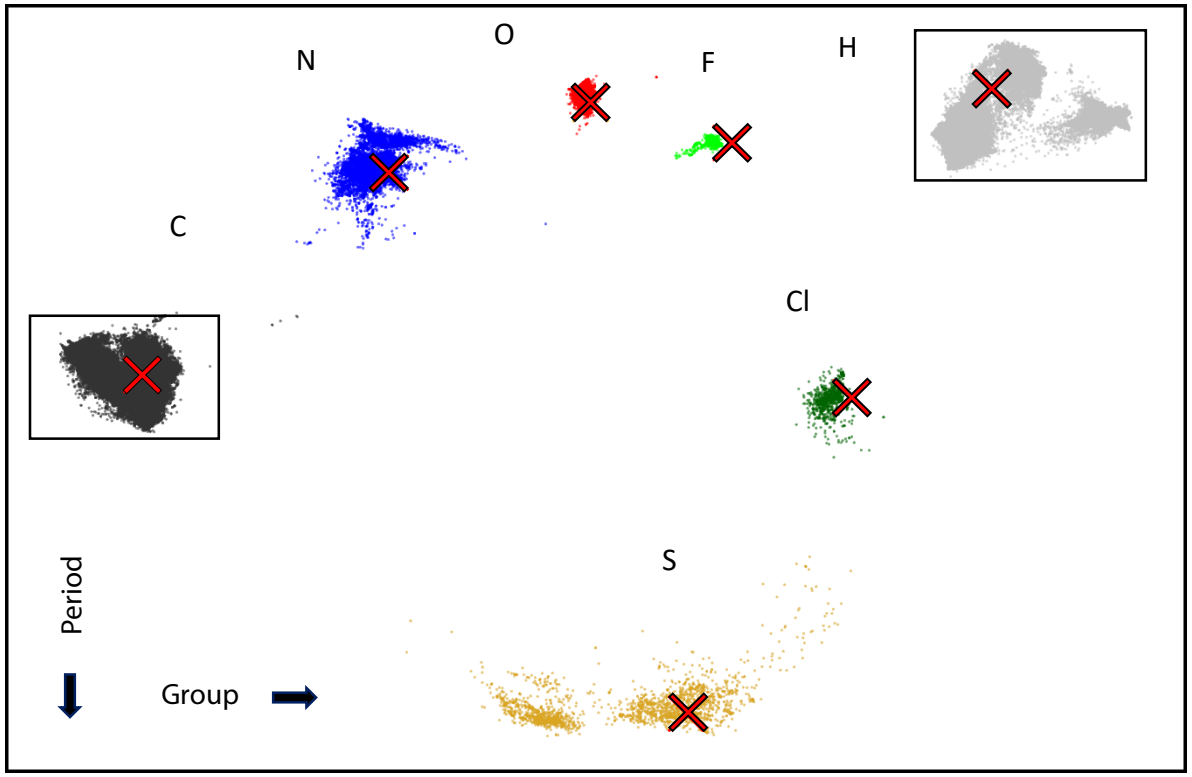




a) Experimental versus predicted with AIMNet solvation free energies (kcal/mol) for 414 neutral molecules from MNSol database. b) performance of AIMNet and other solvation models on torsion benchmark of Sellers et al.

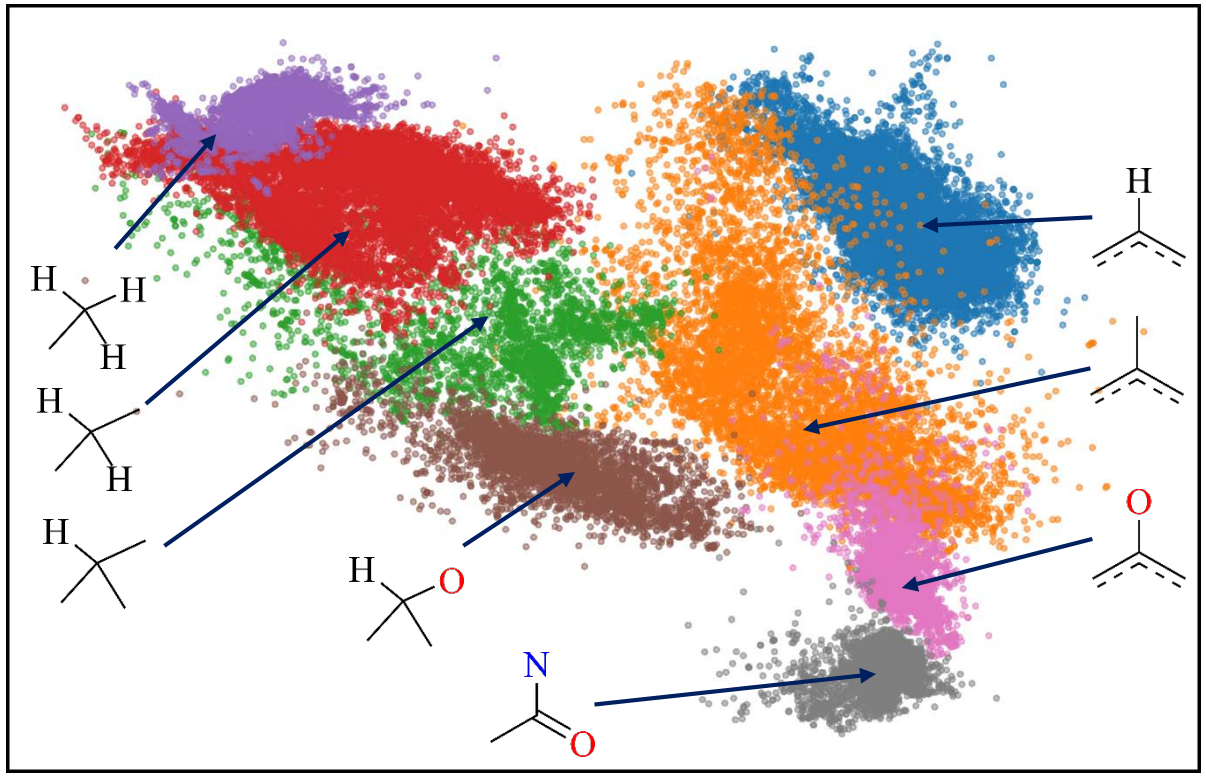
R. Zubatyuk, J.S. Smith, J. Leszczynski, O. Isayev, Accurate and Transferable Multitask Prediction of Chemical Properties with an Atoms-in-Molecule Neural Network. ChemRxiv, 2018. Submitted.

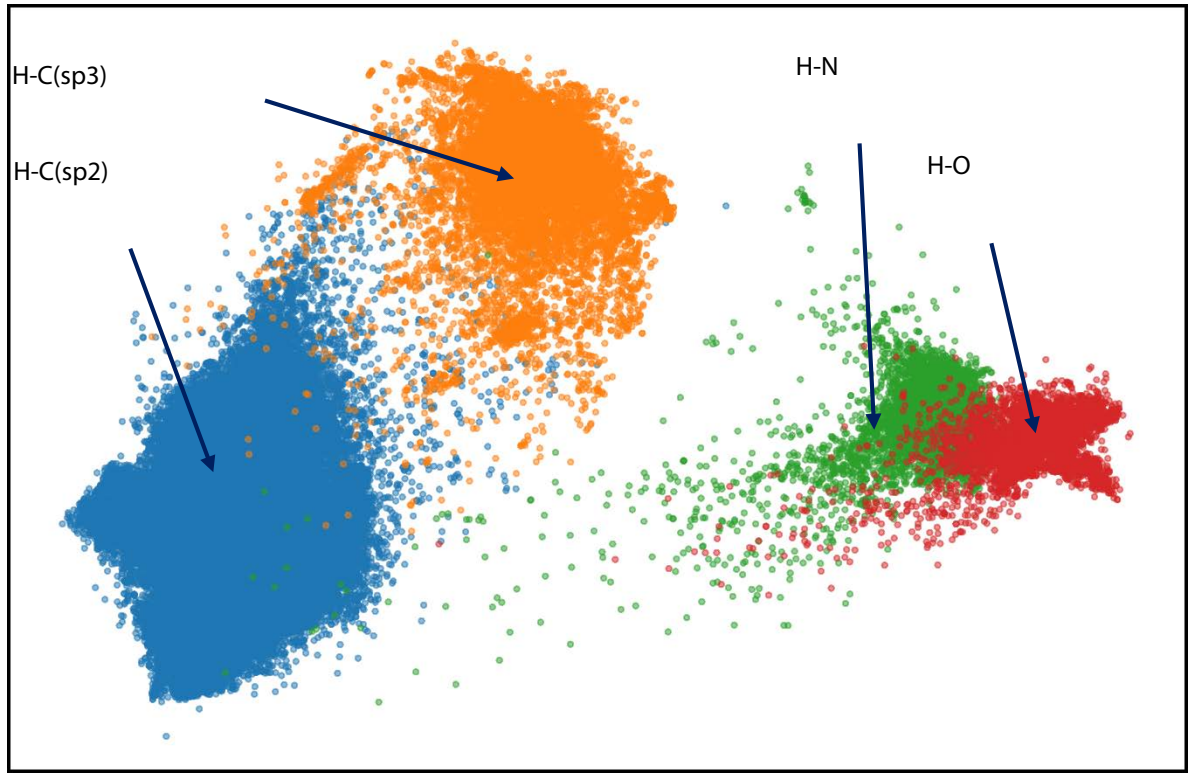
Nature of Learned Atomic Embeddings



R. Zubatyuk, J.S. Smith, J. Leszczynski, O. Isayev, Accurate and Transferable Multitask Prediction of Chemical Properties with an Atoms-in-Molecule Neural Network. ChemRxiv, 2018. Submitted.

Example of Carbon and Hydrogen

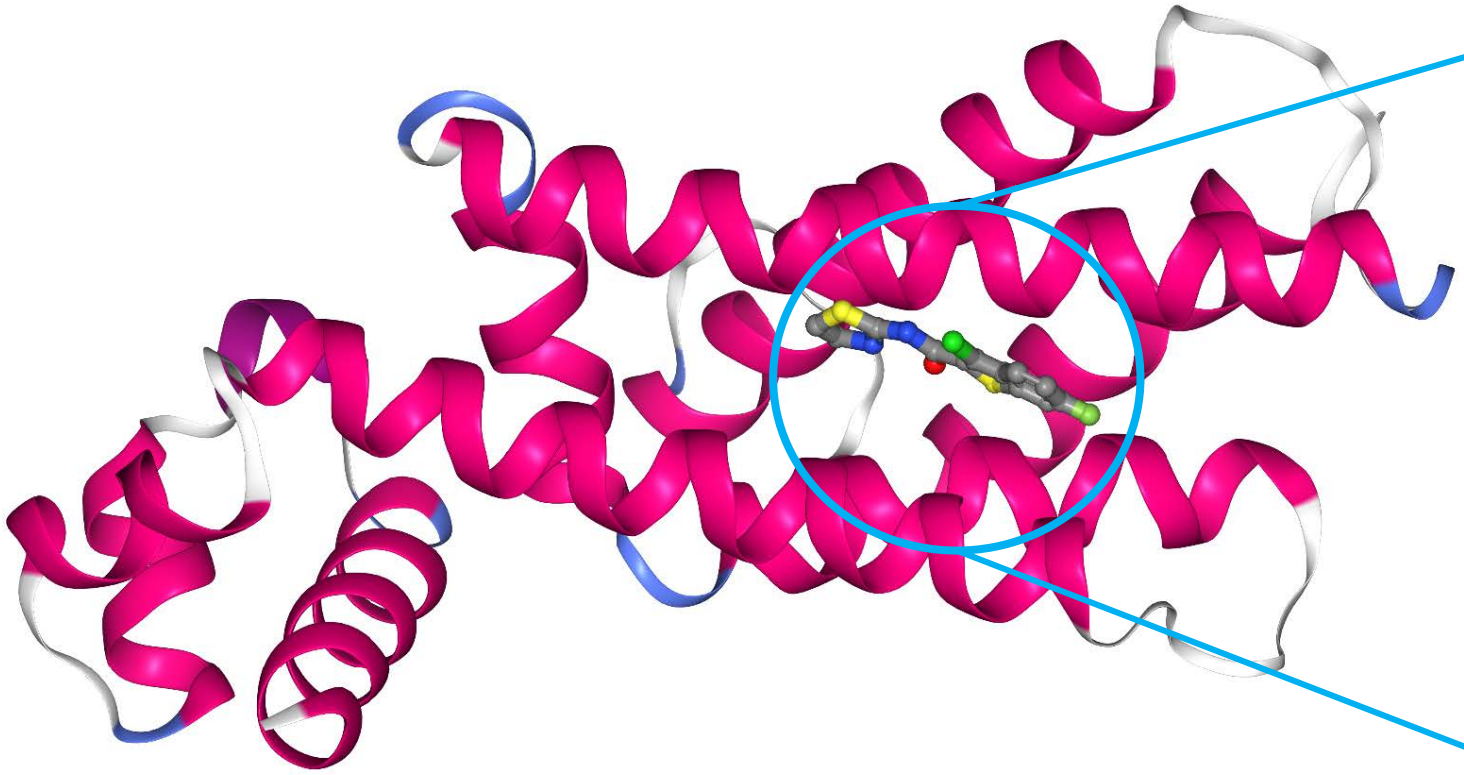


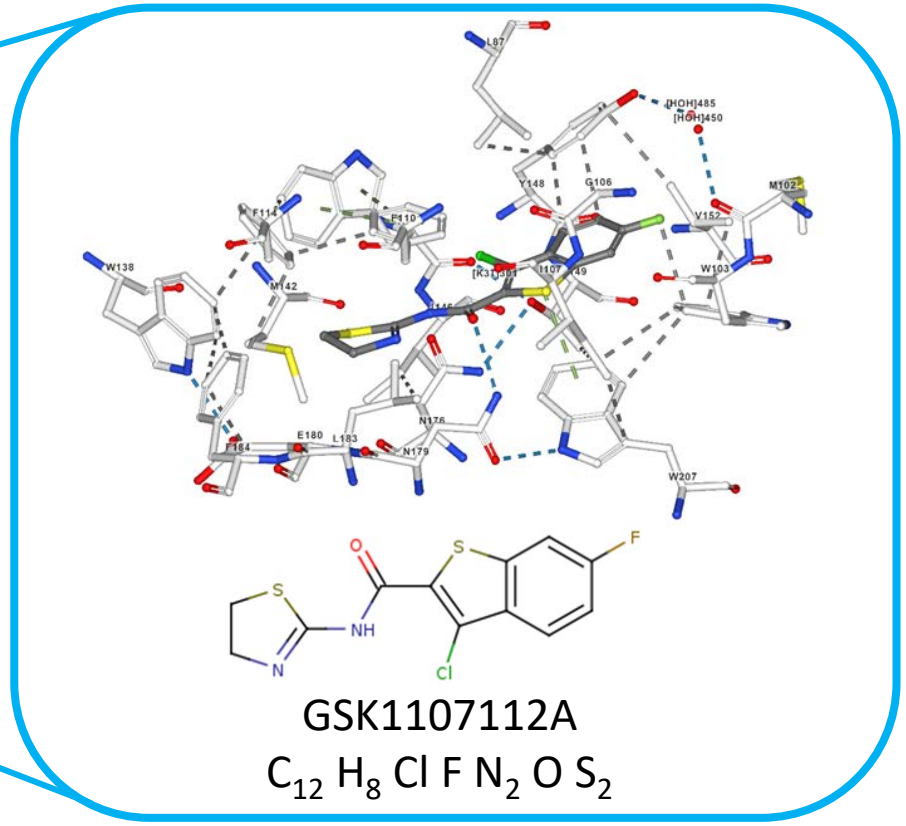


R. Zubatyuk, J.S. Smith, J. Leszczynski, O. Isayev, Accurate and Transferable Multitask Prediction of Chemical Properties with an Atoms-in-Molecule Neural Network. ChemRxiv, 2018. Submitted.

Major future developments

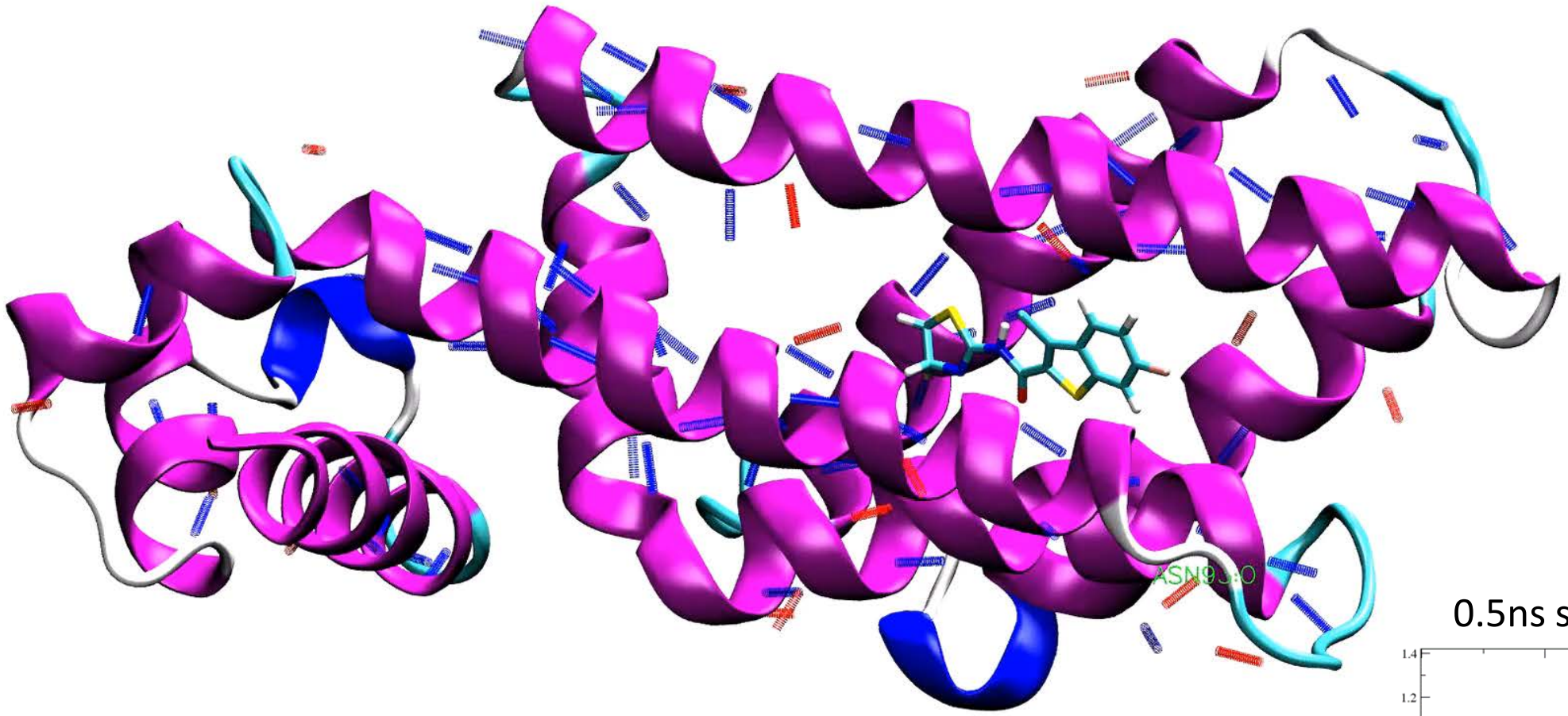
Toward Realistic Macromolecular Simulations



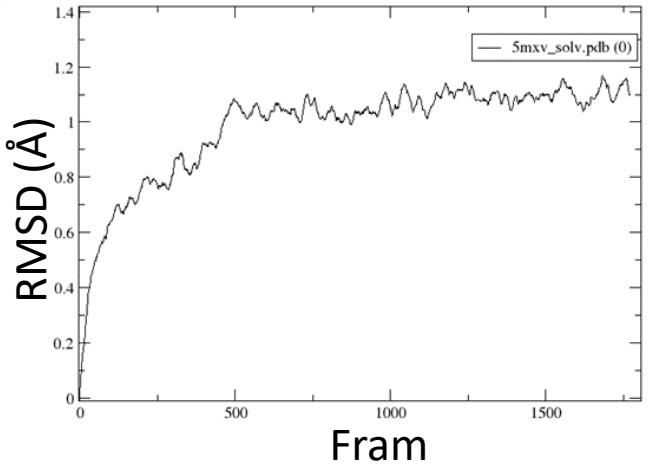


Mycobacterium tuberculosis (5MXV) in explicit water
Simulated with ANI-2 (CHNOSFCl)

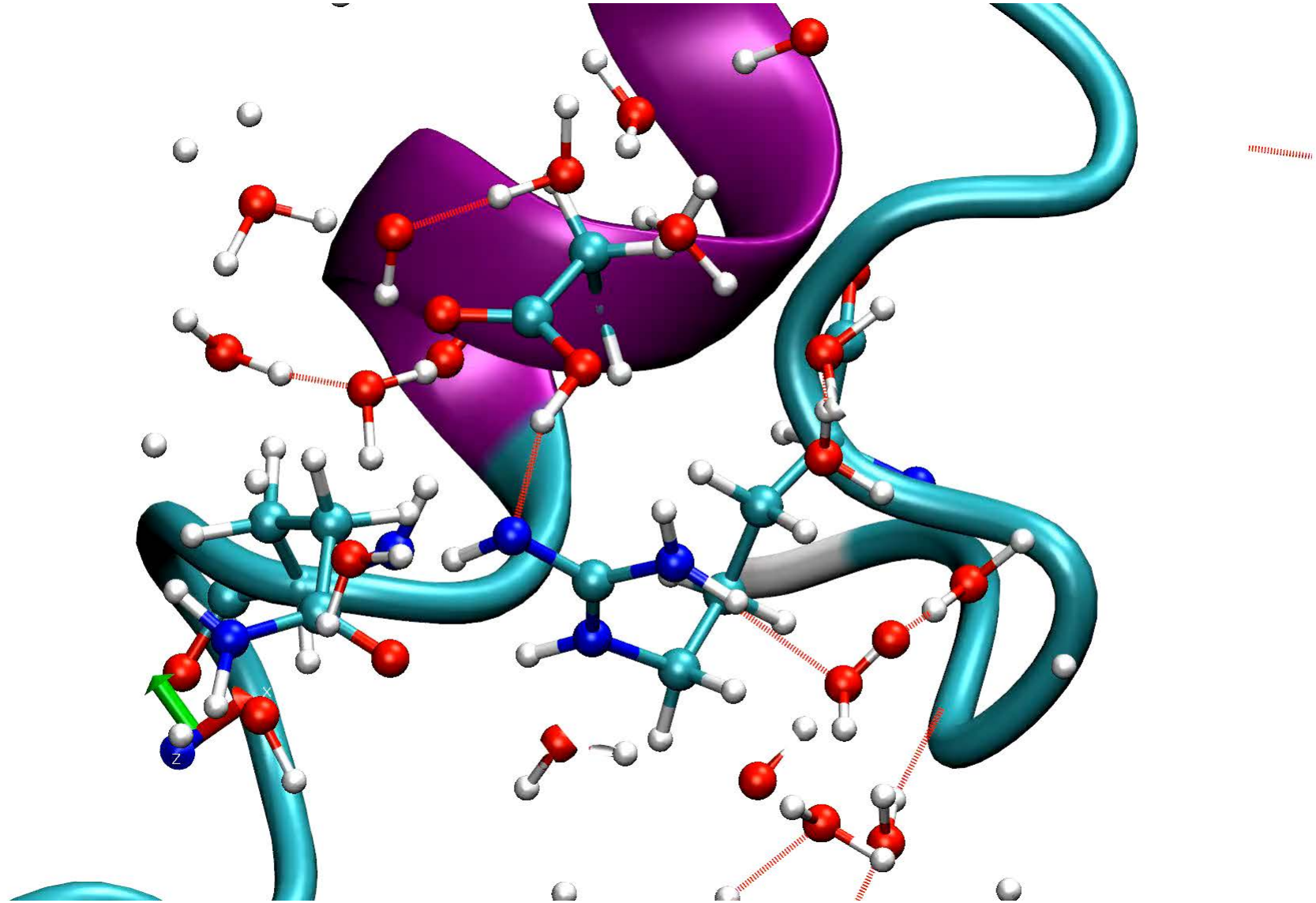
- ~35K atoms
- Explicit water
- No ions
- S, F and Cl in ligand



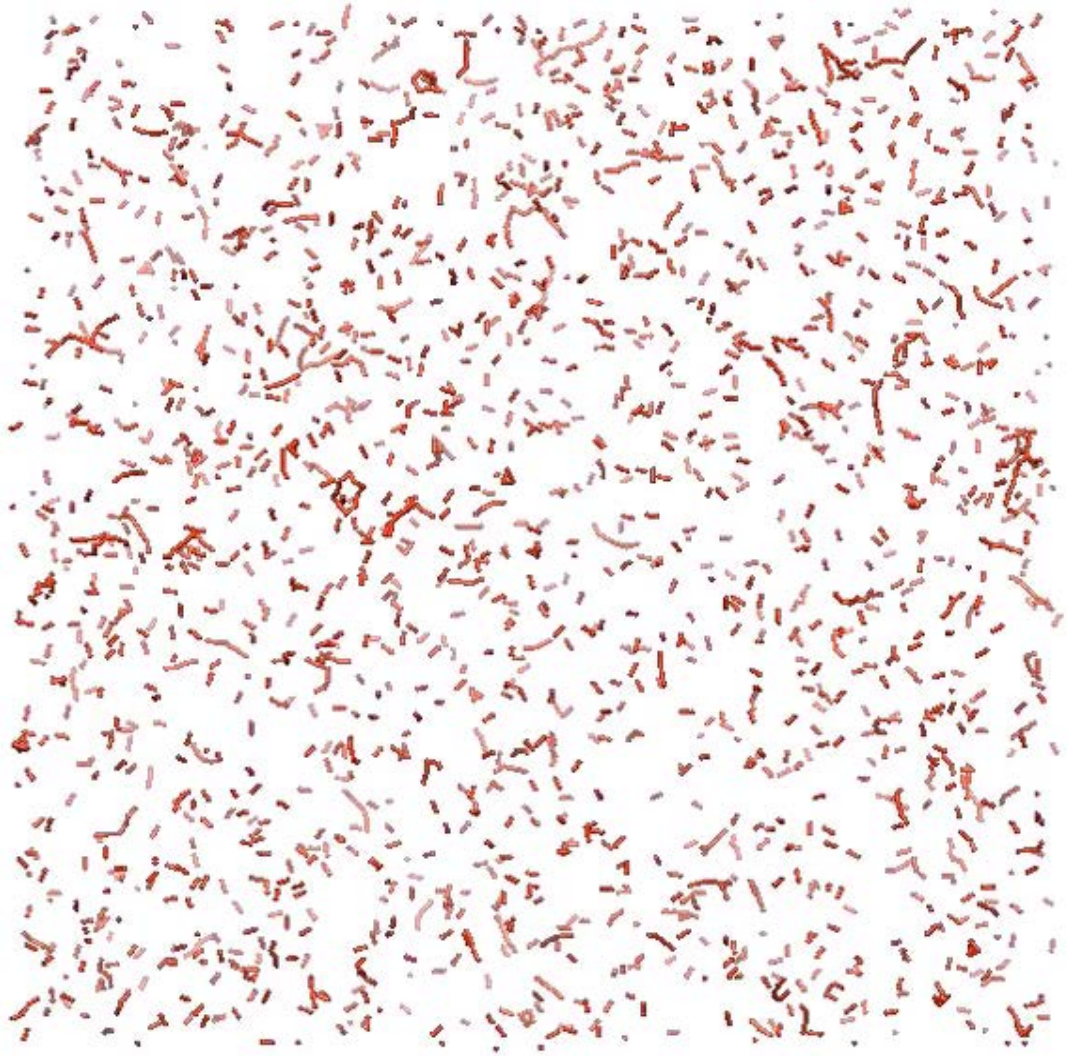
0.5ns simulation time



Timings for a 5x ensemble prediction for ANI-2x			
GPU	ANI-2x time per step	Total time per step	Steps per day
Tesla V100	297ms	317ms	**272k**



Simulation of Complex Chemical Reactions



https://youtu.be/DRVMH5u8EA0

Carbon nanoparticles/sheets nucleation [4000 atoms in 60A box at 2500K, 5ns MD simulation]

Use the ANI-1x potential:

ANI-1x interfaced to ASE Python library
Available at: https://github.com/isayev/ASE_ANI

ANI-1x implementation in PyTorch
Available at: https://github.com/aiqm/torchani

Coming soon to AMBER, OpenMM & LAMMPS

Use the ANI-1 dataset:

ANI-1: A data set of 20M off-equilibrium DFT calculations for organic molecules
Sci. Data, 2017, **4**, 170193 DOI: 10.1038/sdata.2017.193

ANI Data set Python library
Available at: https://github.com/isayev/ANI1_dataset

Users:

academic labs:

- Stanford (Vijay Pande)
- U Pitt (Geoff Hutchison)
- CMU MSE (Noa Marom)
- USF
- NCSU
- Barcelona
- Helsinki
- Tel Aviv

Government labs, companies etc.

Los Alamos NATIONAL LABORATORY EST. 1943

BERKELEY LAB Lawrence Berkeley National Laboratory

NIH National Institutes of Health

BASF We create chemistry

SCHRÖDINGER.

Pfizer

Genentech

OpenEye Scientific Software

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

Mariya Popova
Roman Zubatyuk
Daniel Korn
Kyle Bowler
Hatice Gockan

Adrian Roitberg
Justin S. Smith
Christian Devereux
Kavindri Ranasinghe

UF UNIVERSITY of FLORIDA

Funding:

CHE-1802789

Collaborators:

Nicholas Lubbers
Ben Nebgen
Andrew Sifain
Kipton Barros
Sergei Tretiak

HPC Computing: