Los Alamos
NATIONAL LABORATORY
EST.1943

LA-UR-18-26999
Approved for public release; distribution is unlimited.

Title: Machines Learning Quantum Chemistry: Potentials, Properties, and P-Orbitals

Author(s): Nebgen, Benjamin Tyler
Smith, Justin Steven
Lubbers, Nicholas Edward
Sifain, Andrew E.
Barros, Kipton Marcos
Tretiak, Sergei

Intended for: Telluride workshop: Multi-scale quantum mechanical analysis of condensed phase systems: methods and applications, 2018-07-23/2018-07-27 (Telluride, Colorado, United States)

Issued: 2018-07-26

Disclaimer:
Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by the Los Alamos National Security, LLC for the National Nuclear Security Administration of the U.S. Department of Energy under contract DE-AC52-06NA25396. By approving this article, the publisher recognizes that the U.S. Government retains nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy. Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

Machines Learning Quantum Chemistry: Potentials, Properties, P-Orbitals

Ben Nebgen

Justin Smith, Nicholas Lubbers, Andrew Sifain, Kipton Barros, Sergei Tretiak

Acknowledgements

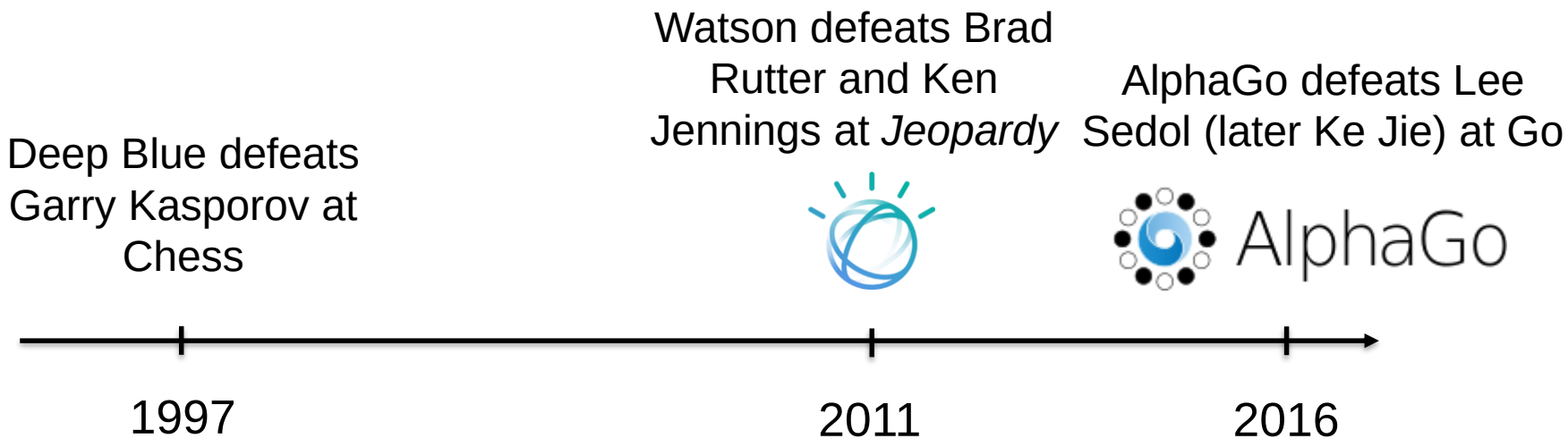
- Nickolas Lubbers: Built HIP-NN
- Justin Smith (University of Florida): Built ANI
- Andrew Sifain
- Andrey Lokhov
- Kipton Barros
- Adrian Roitberg (University of Florida)
- Olexandr Isayev (University of North Carolina at Chapel Hill)
- Sergei Tretiak

- Funding: LDRD-ER Started in Oct-2017

Computers are Becoming Smarter

In the last two decades, every major game has fallen to some form of artificial intelligence.

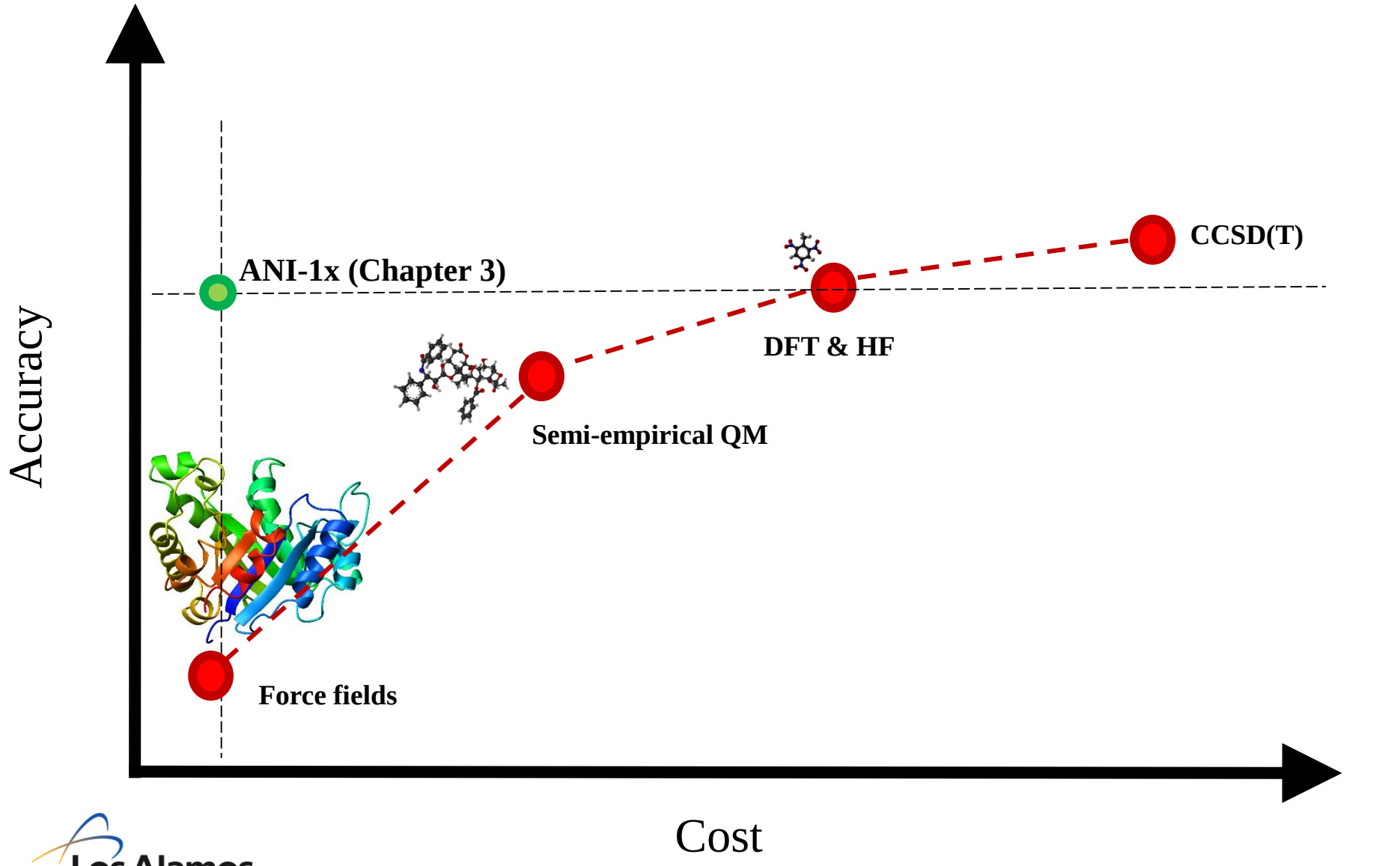
While Deep Blue was a played by computation, AlphaGo can be characterized as an intelligence. AlphaGo plays games against itself to get better at Go.



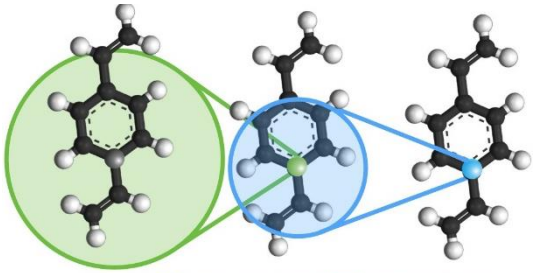
Overview

- **ML for Energy and Force Prediction**
  - Active Learning
  - Molecular Dynamics
- **Singlet Triplet**
  - Charges
  - IR spectrum
- **Machine Learned Effective Hamiltonians**
  - Effective Hamiltonian Form
  - Learning Parameters

Speed and Cost

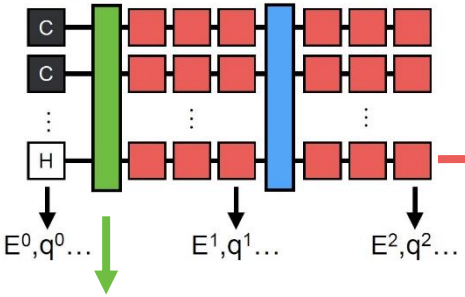


Structure of HIP-NN Neural Network



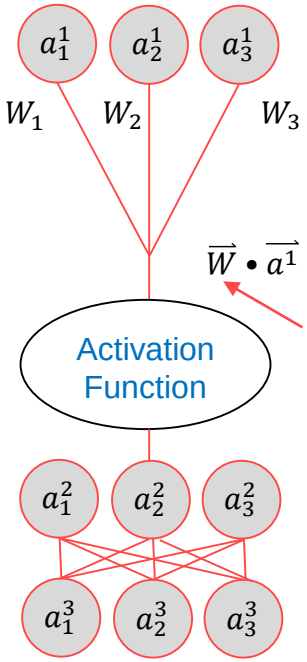
Typically results in ~10,000 different parameters.

Steepest gradient descent algorithms optimize W, V, s



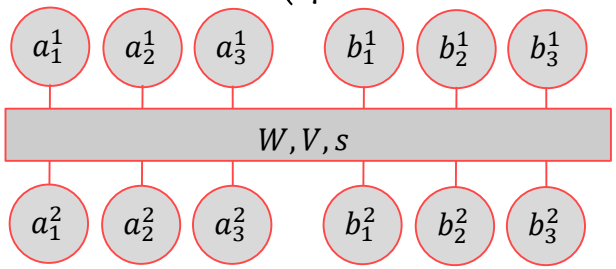
Nicholas Lubbers Justin S. Smith, Kipton Barros “Hierarchical modeling of molecular energies using a deep neural network” arXiv:1710.00017, 2017.

Atomic Layer

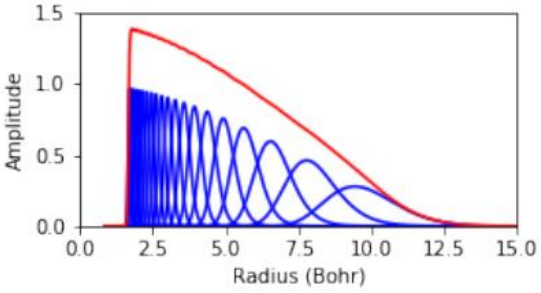


Interaction Layer

$$a_i^2 = f\left(\sum_\mu s^\mu(r) V_b \bullet b^1 + W_a \bullet a^1 + W_b \bullet b^1 + C_a\right)$$



$s^\mu(r)$



UNCLASSIFIED

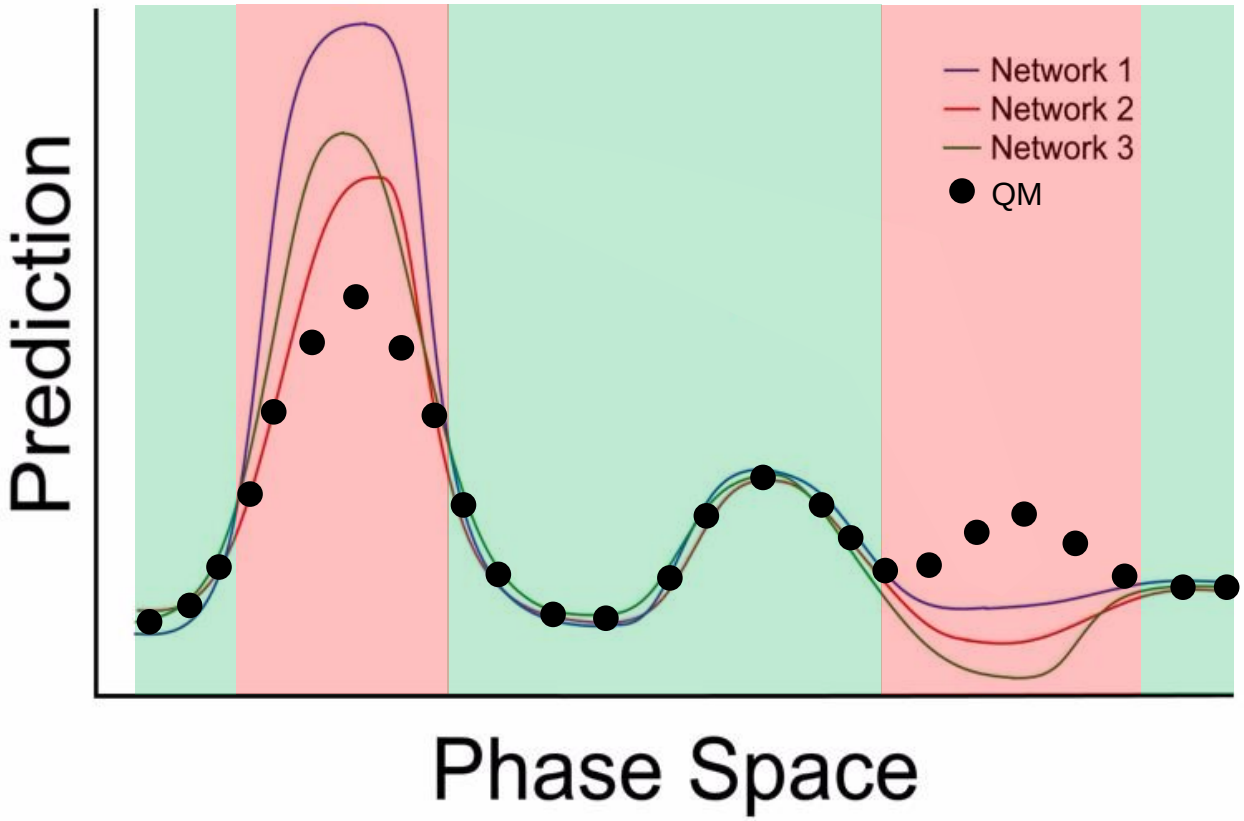
Los Alamos NATIONAL LABORATORY EST.1943
Operated by Los Alamos National Security, LLC for NNSA
NNSA

Optimal Training Selection

Ensemble disagreement can drive data generation

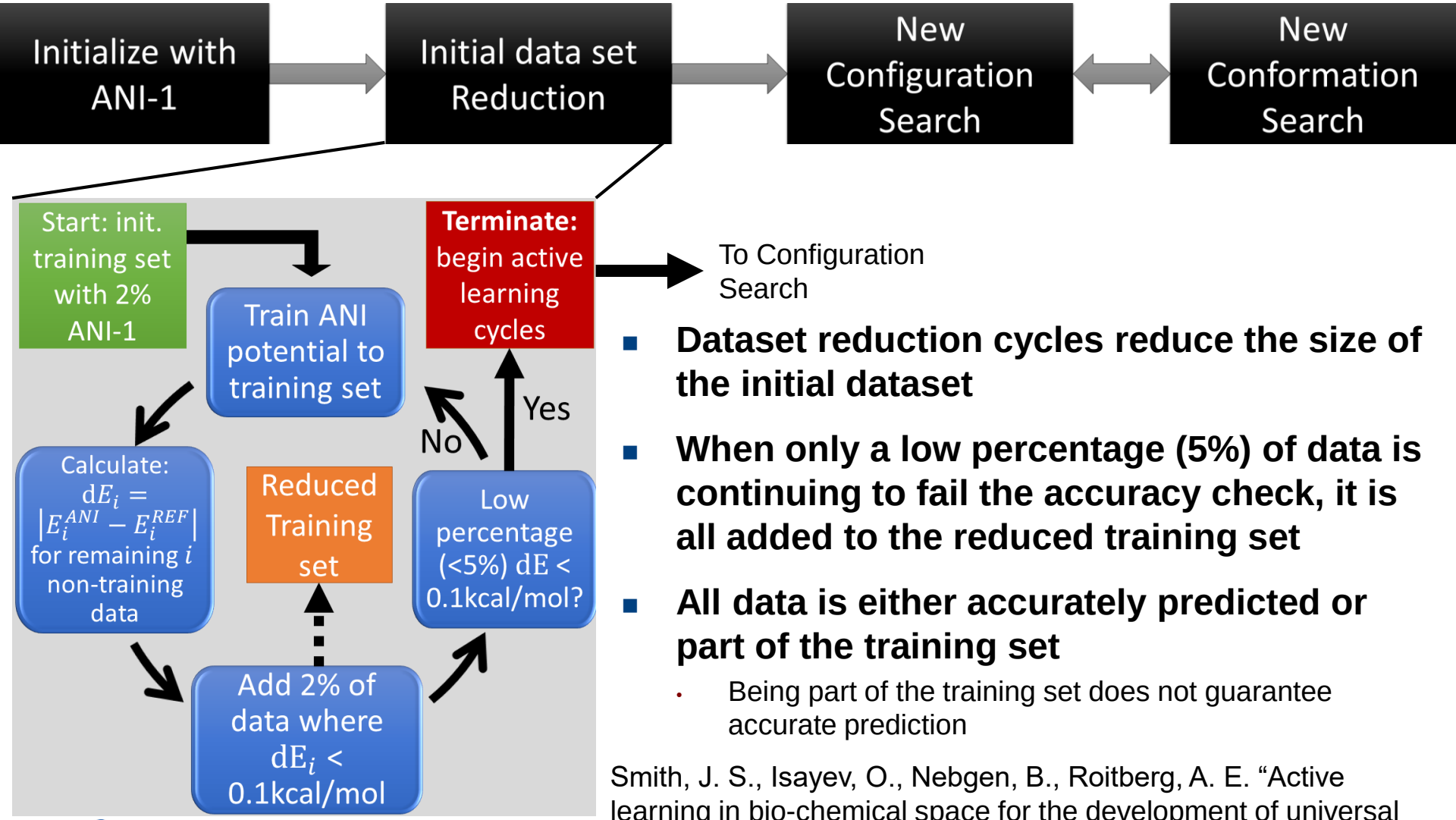
Good data coverage

Bad data coverage



Smith, J. S., Isayev, O., Nebgen, B., Roitberg, A. E. “Active learning in bio-chemical space for the development of universal machine learned molecular potentials” **148**, 241733 (2018)

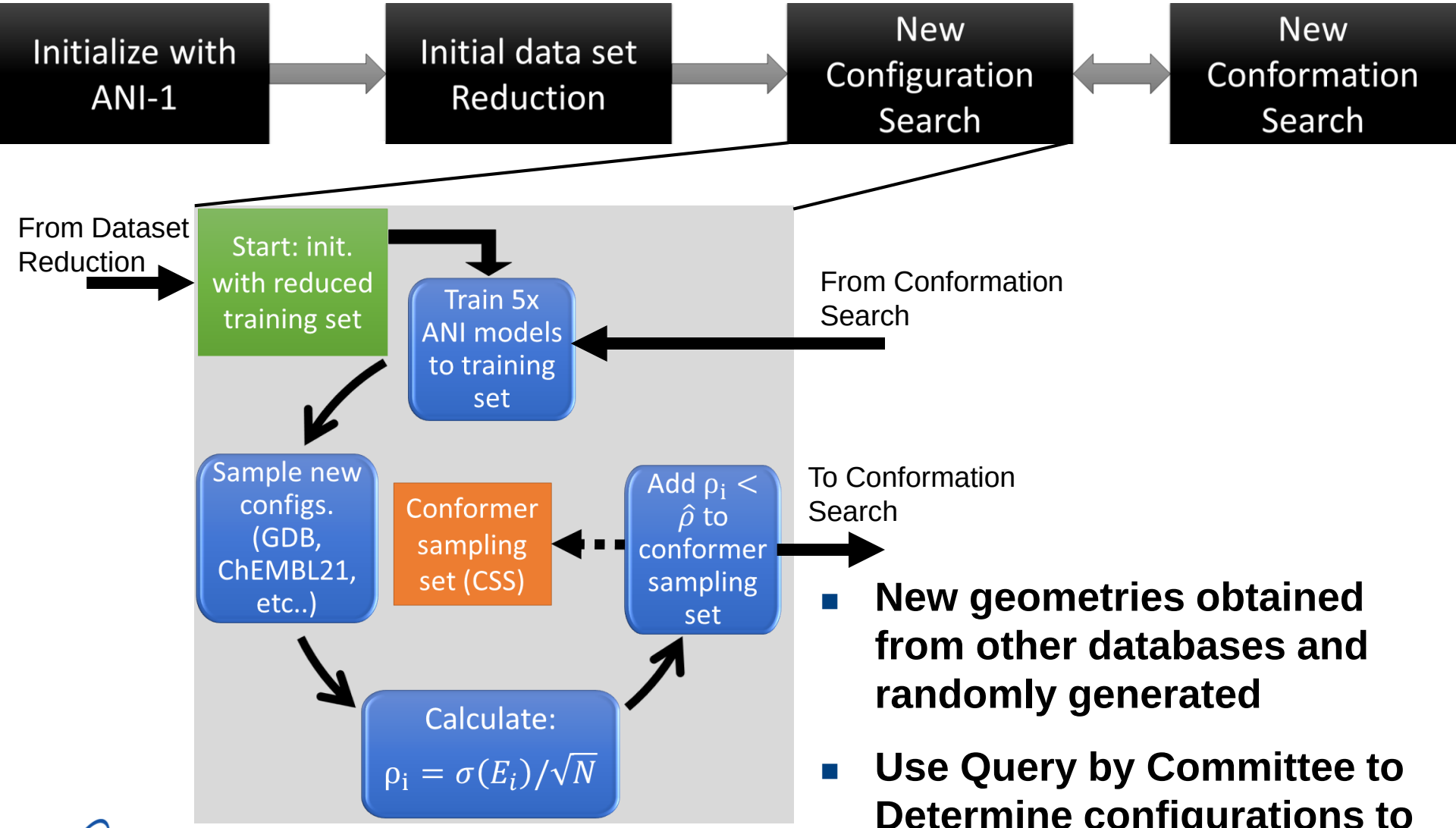
Active Learning and Database Construction



- **Dataset reduction cycles reduce the size of the initial dataset**
- **When only a low percentage (5%) of data is continuing to fail the accuracy check, it is all added to the reduced training set**
- **All data is either accurately predicted or part of the training set**
  - Being part of the training set does not guarantee accurate prediction

Smith, J. S., Isayev, O., Nebgen, B., Roitberg, A. E. "Active learning in bio-chemical space for the development of universal machine learned molecular potentials" In Preparation.

Active Learning and Database Construction

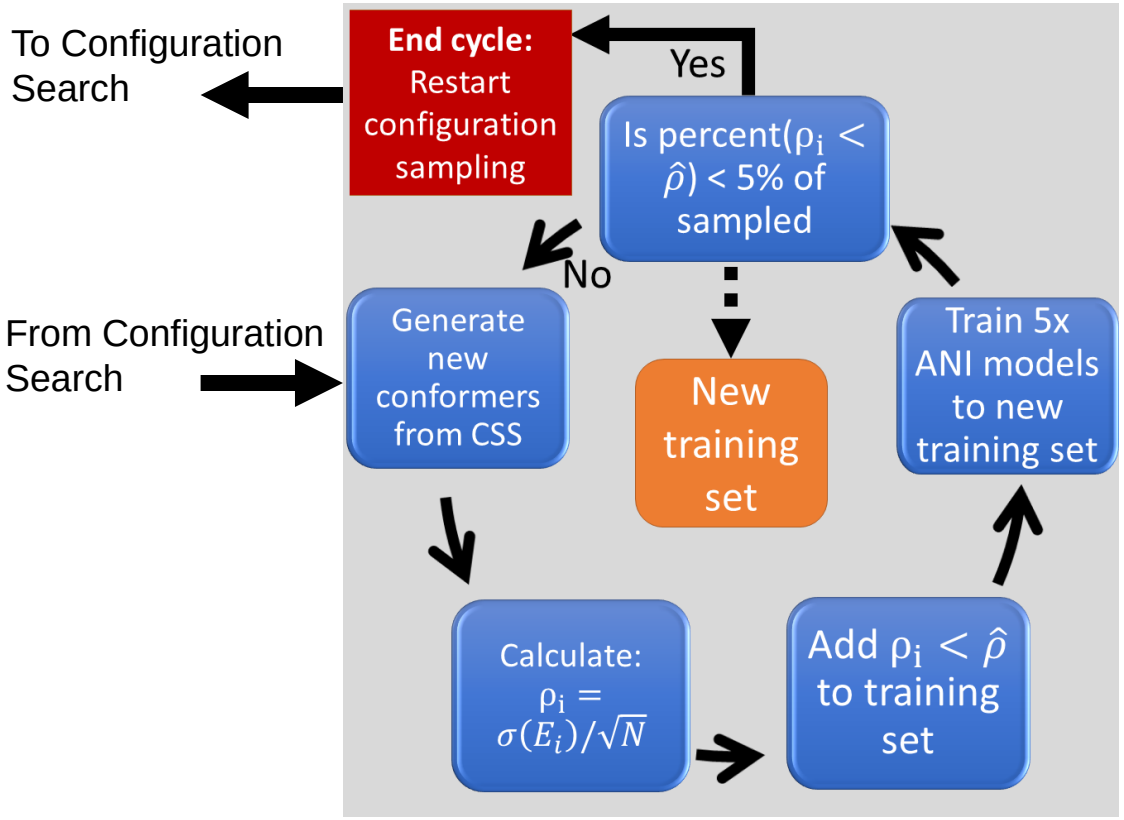


- **New geometries obtained from other databases and randomly generated**
- **Use Query by Committee to Determine configurations to add to database**

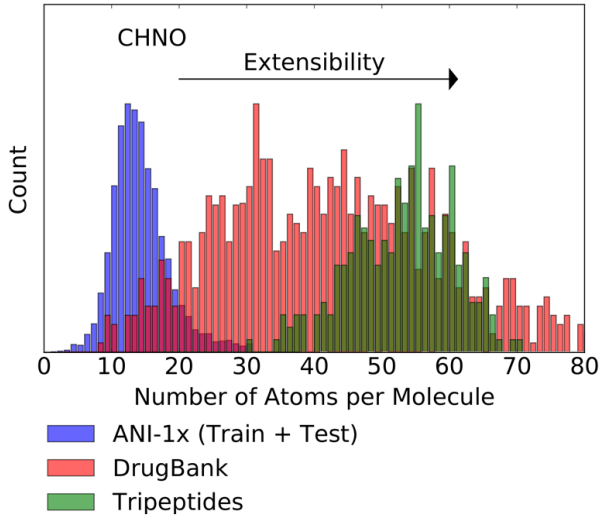
Active Learning and Database Construction

Initialize with ANI-1 → Initial data set Reduction → New Configuration Search ↔ New Conformation Search

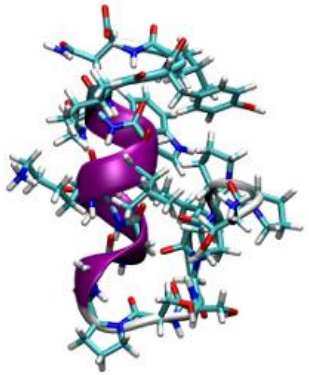
- **New Geometries generated by normal mode sampling**
- **Use Query by Committee**
  - $\sigma(E_i)$ is the ensemble standard deviation of energy prediction
  - N is the number of atoms
- **Add new conformations until less than 5% fail ($\rho_i < \hat{\rho}$)**

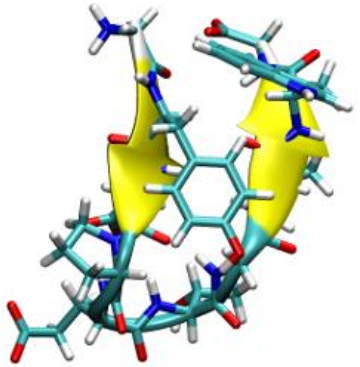


Accuracy Versus Active Learning Iteration

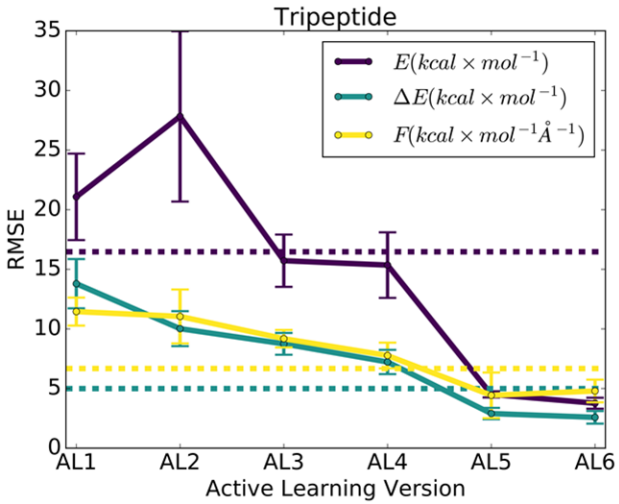


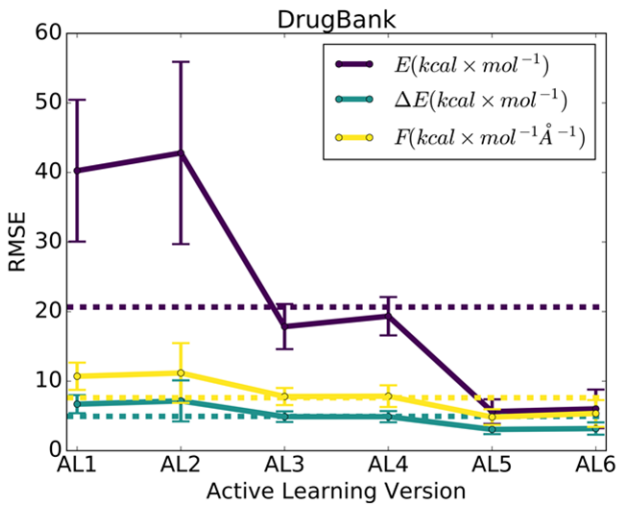
Chignolin

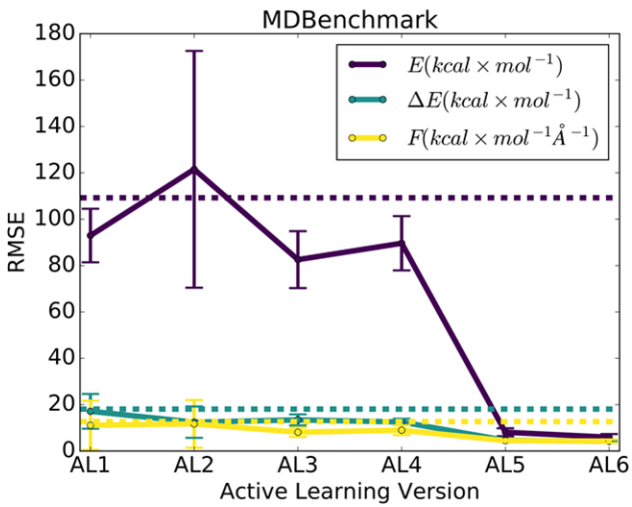




Trp-cage

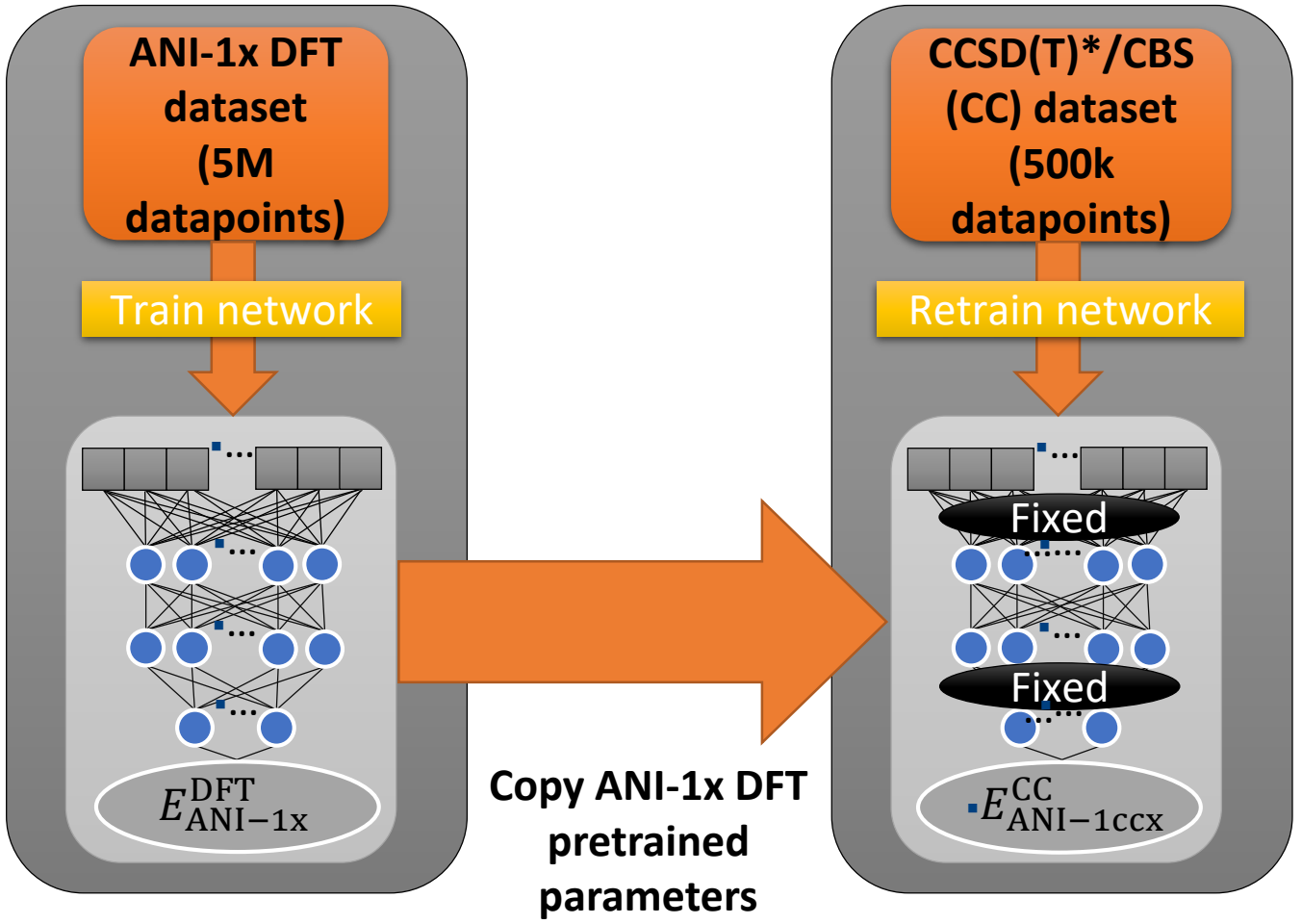






CCSD Transfer Learning

- Subsample 10% of ANI-1x training data (0.5M of 5M)
- Recompute CCSD(T)/CBS level
- 340k parameters fixed, re-train 60k
- 10^7 faster than DFT



Predictions on molecules with 10-13 heavy atoms

ANI-1ccx predicts accurate forces compared to MP2/TZ

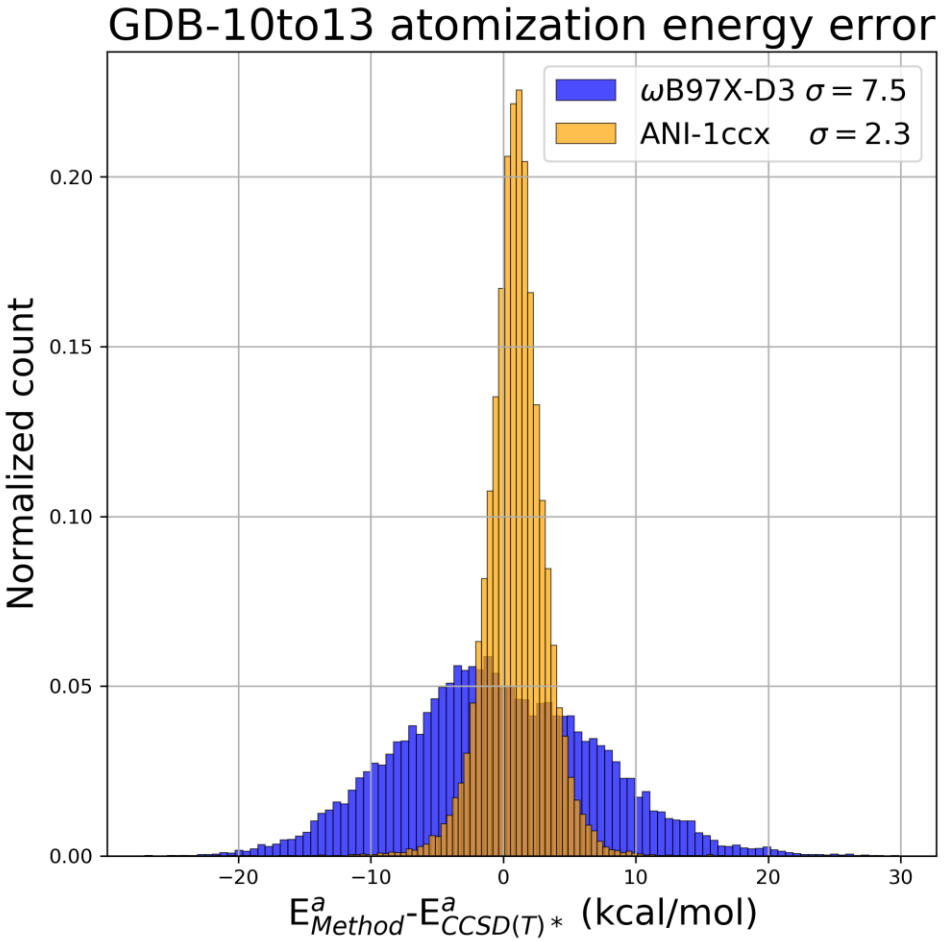
	ANI-1ccx	ANI-1x	ωB97X-D3	MP2/DZ
MP2/TZ	3.4	4.7	3.7	4.6

Units: kcal/mol/A (error MAD)

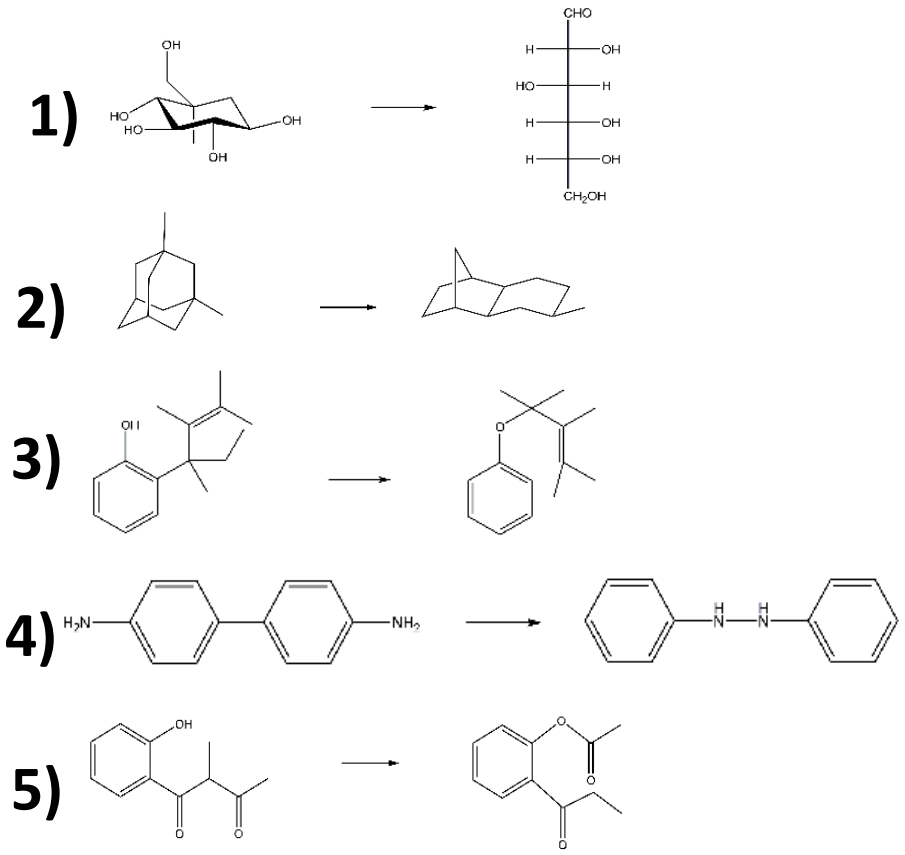
ANI-1ccx predicts accurate relative energies compared to CCSD(T)

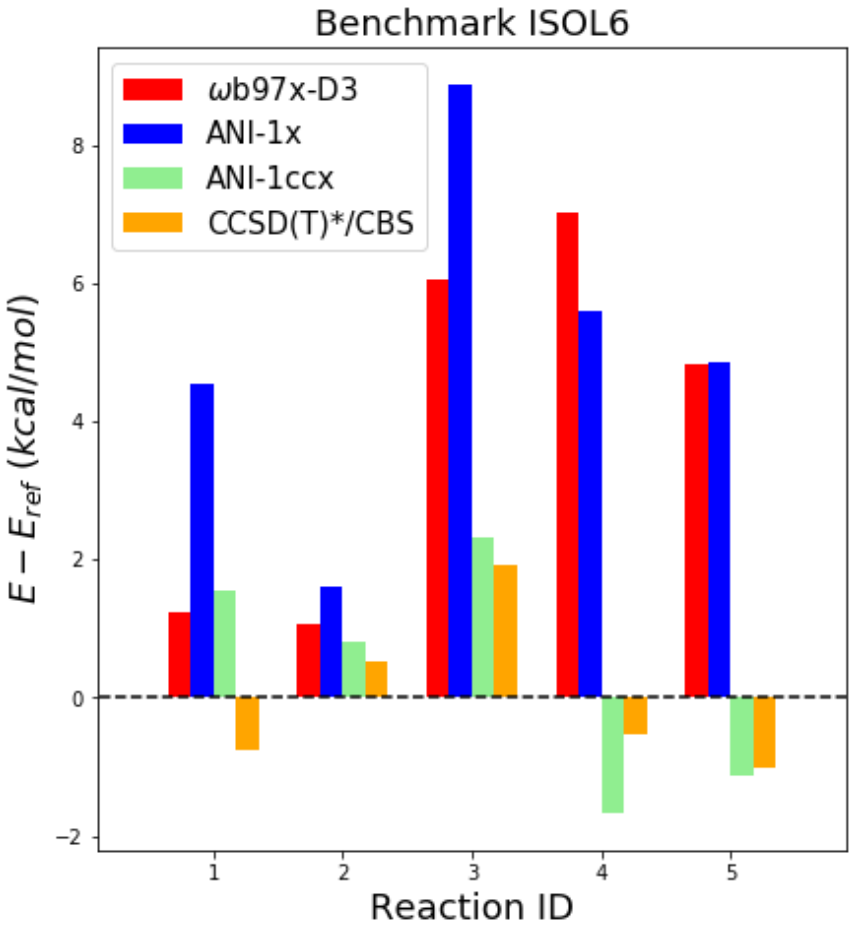
	ANI-1ccx	ANI-1ccx-R	ANI-1x	ωB97X-D3
MAD	1.5	1.8	2.0	1.4

Units: kcal/mol

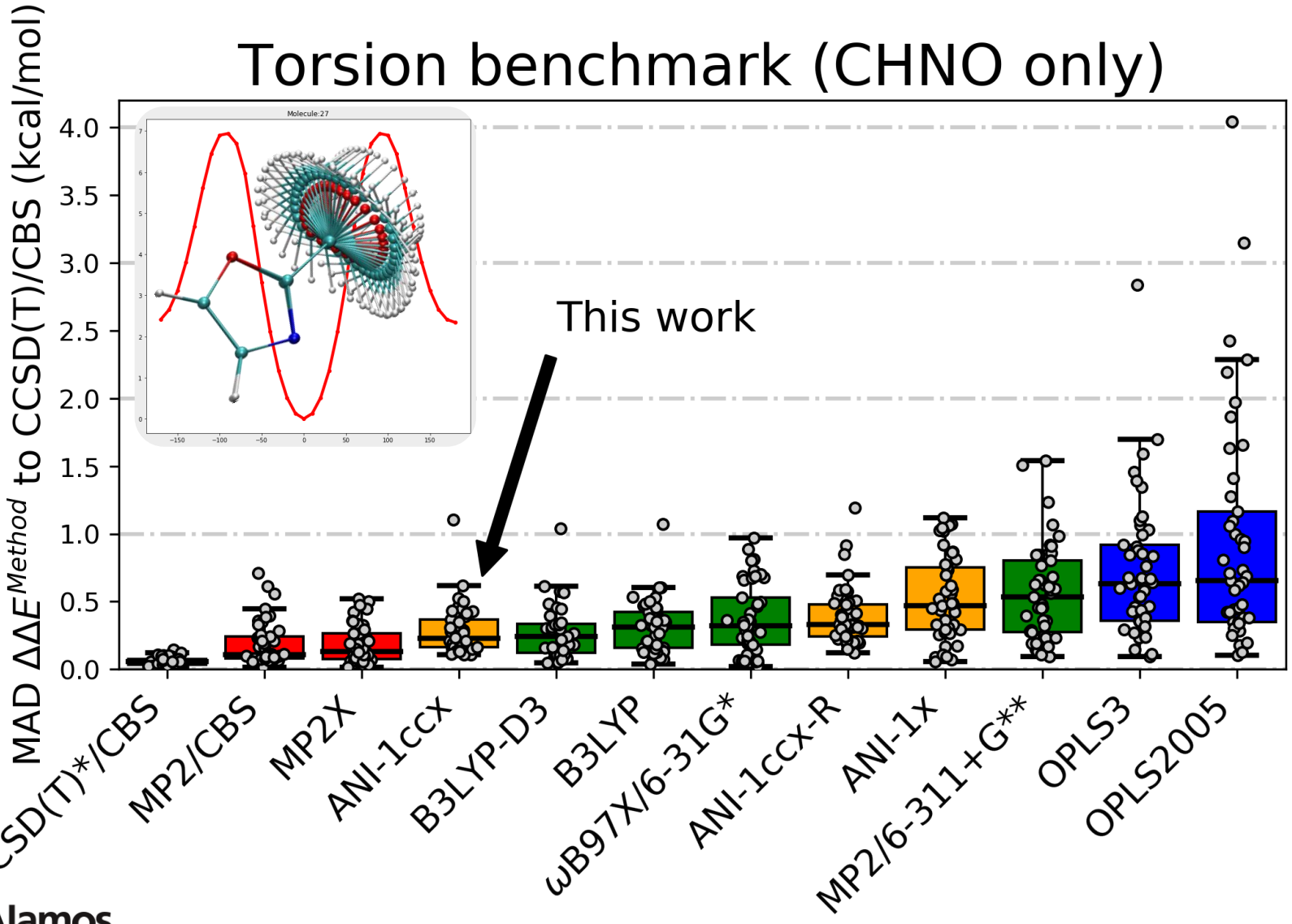


Isomerization Benchmark ISOL6

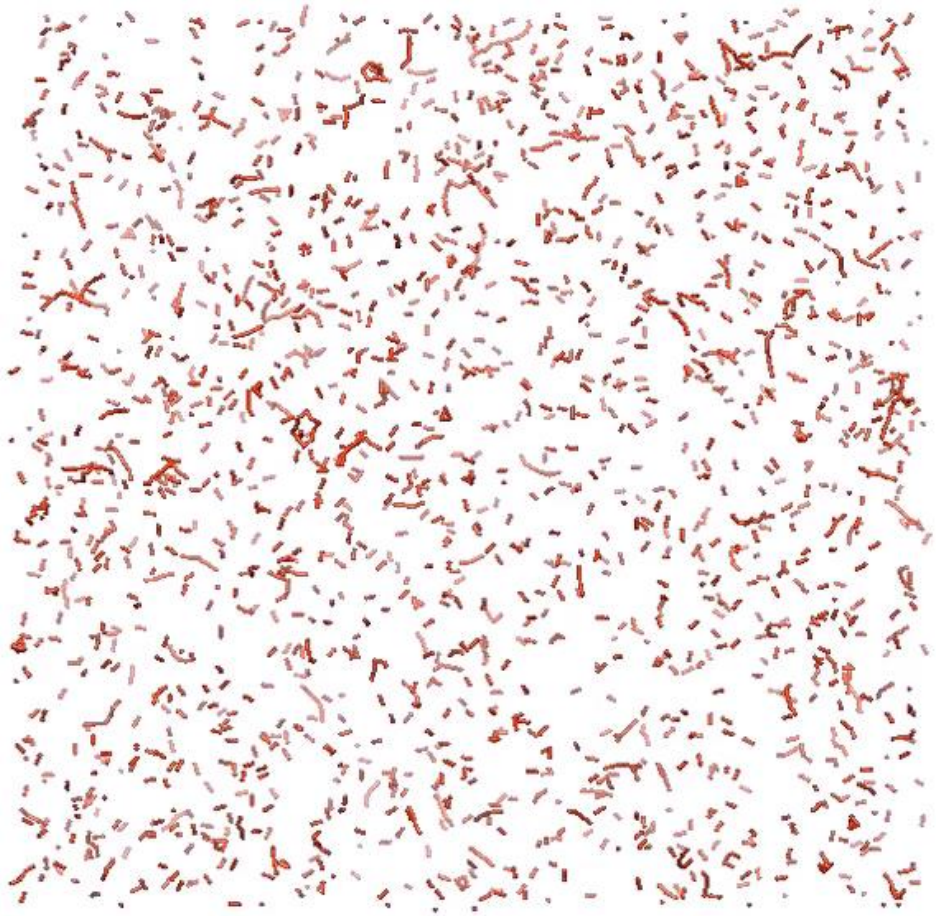




Torsional Benchmark



Fast and Accurate Molecular Dynamics



Carbon nanoparticles/sheets nucleation [4000 atoms in 60Å box at 2500K]
5ns MD simulation. 2 days on 1080-Ti. 800,000 atomic gradients / s

Machines Learning Properties

Different Networks for Different Applications

Network Name	Layer 1				Layer 2 (3)			
	Distances	Cutoff (Å)	Hard Cutoff (Å)	Number of Units	Distances	Cutoff (Å)	Hard Cutoff (Å)	Number of Units
R=4.0	8	2.5	4.0	20	8	2.5	4.0	20
R=6.0	10	3.0	6.0	20	8	2.5	4.0	20
R=8.0	20	5.5	8.0	40	20	5.5	8.0	40

Able to generate different networks with different hyperparameters (distance cut-off, atomic layers, distance bins, ect.)

Generally, the smaller the network the more generalizable but less accurate.

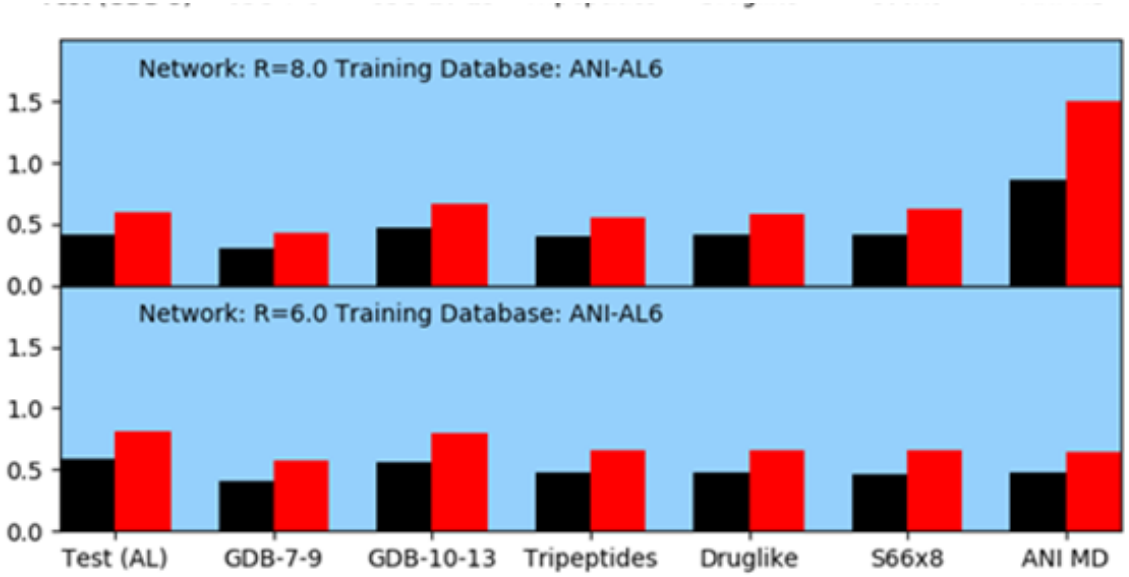
Difficulties associated with training large networks to databases of small molecules due to untrained parameters.

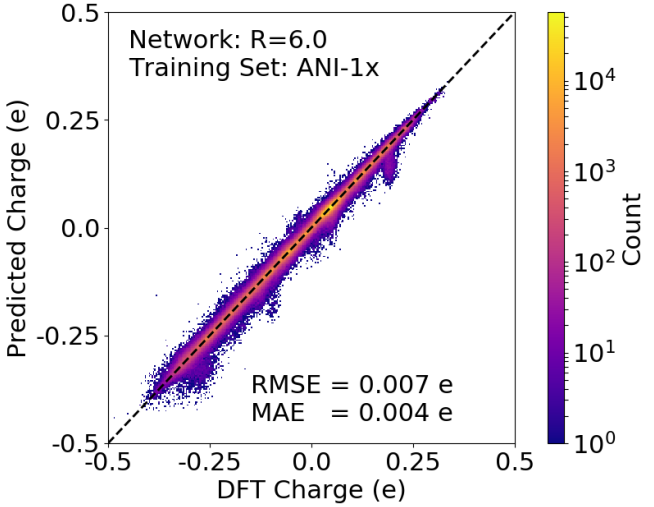
Nebgen, B., Lubbers, N., Smith, J. S., Sifain, A., Lokhov, A., Barros, K., Tretiak, T. “Transferable Charge Prediction for Molecular Systems Using Deep Neural Networks” In Preparation.

Larger Databases

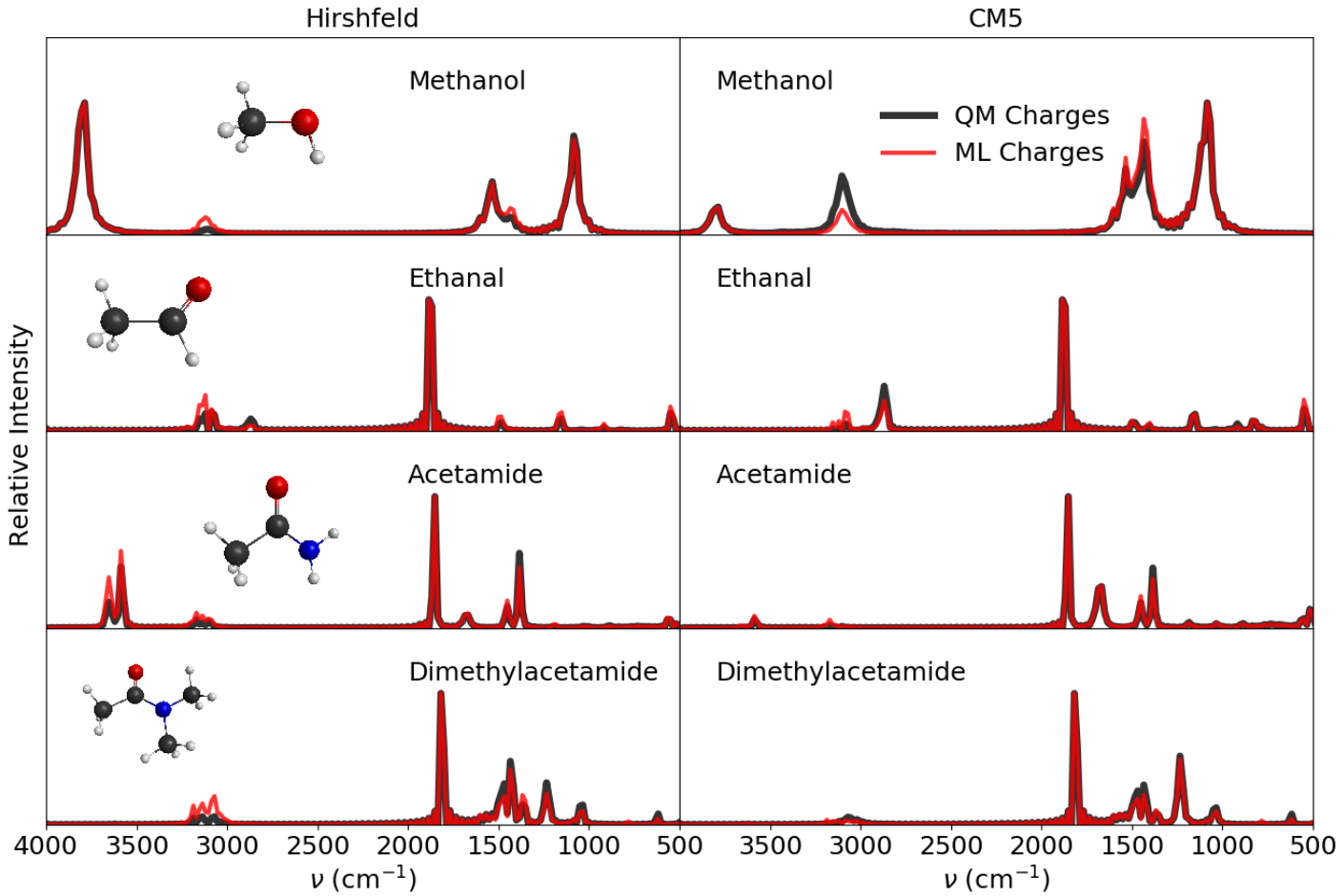
A network trained on a small database is still able to produce reasonable results on larger molecules.

Higher accuracy can be achieved if desired.



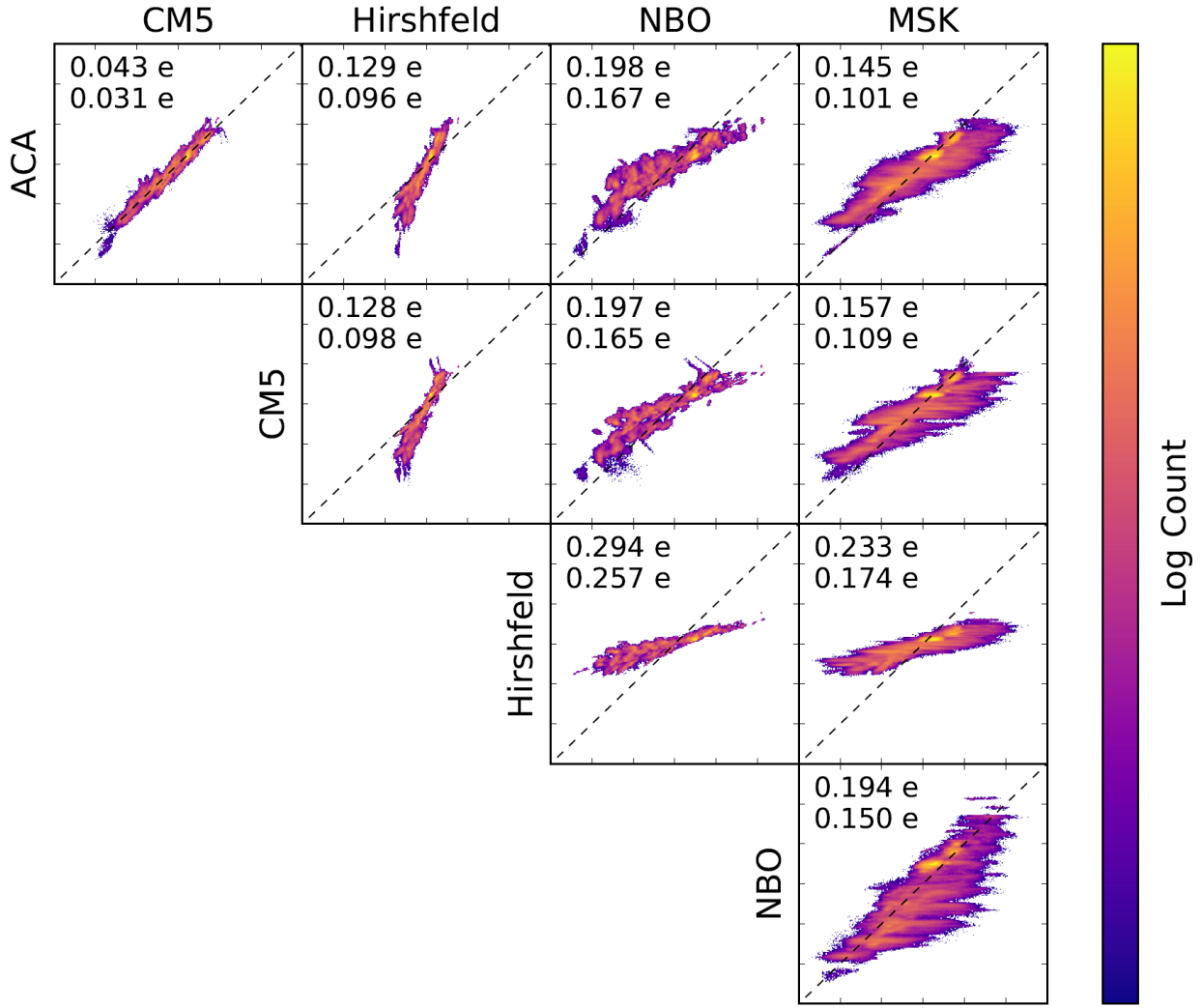


Building IR Spectra

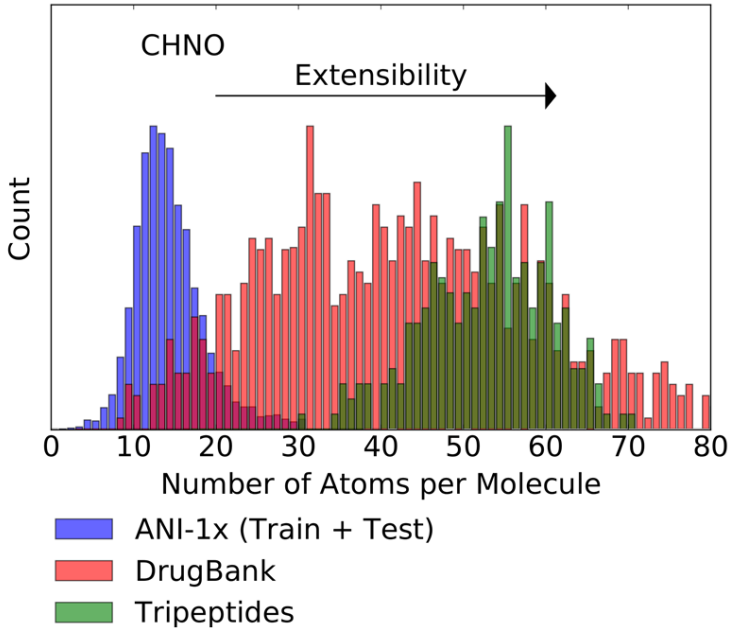


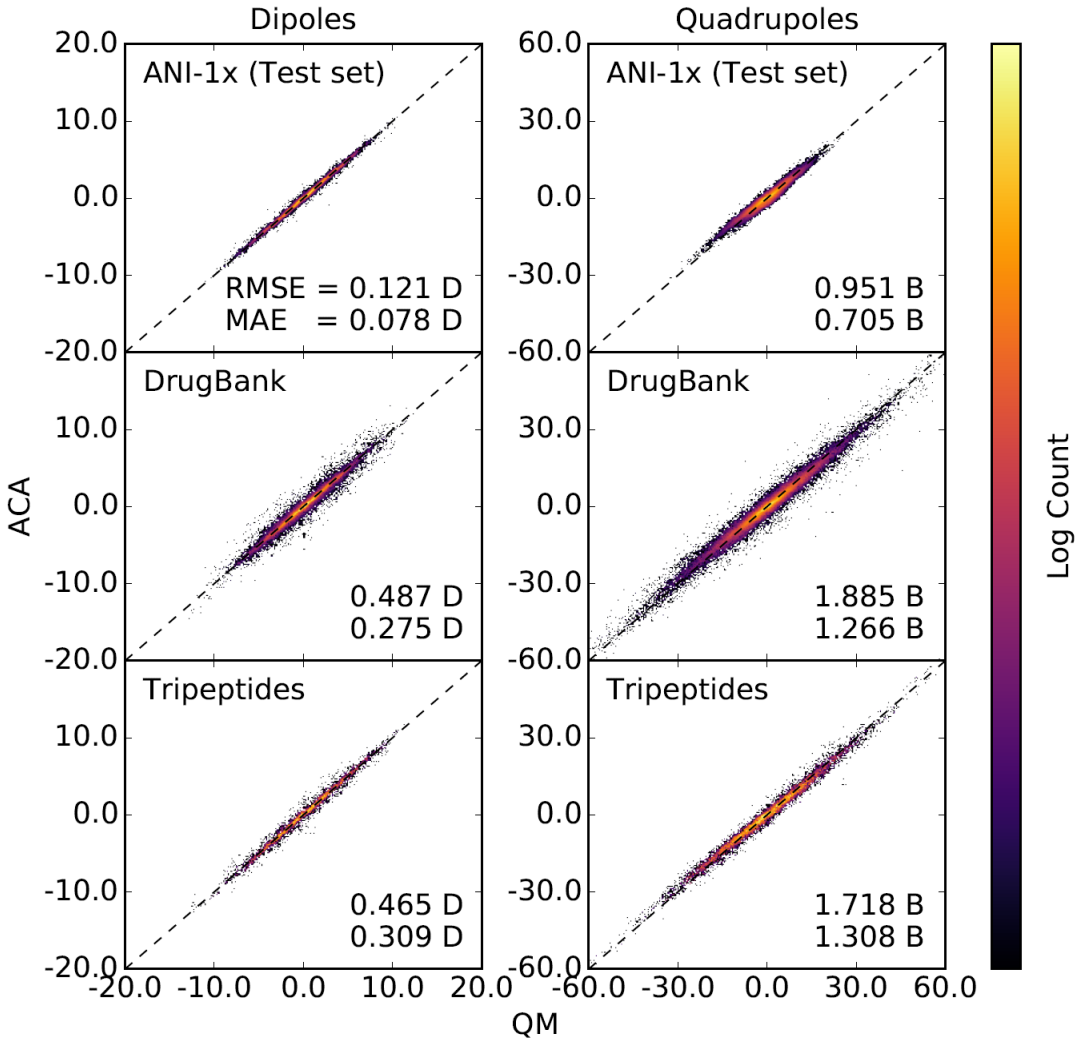
IR spectrum generated completely using ML. ANI generated a trajectory, HIP-NN generated charges at each point.

Charge Models Do Not Agree

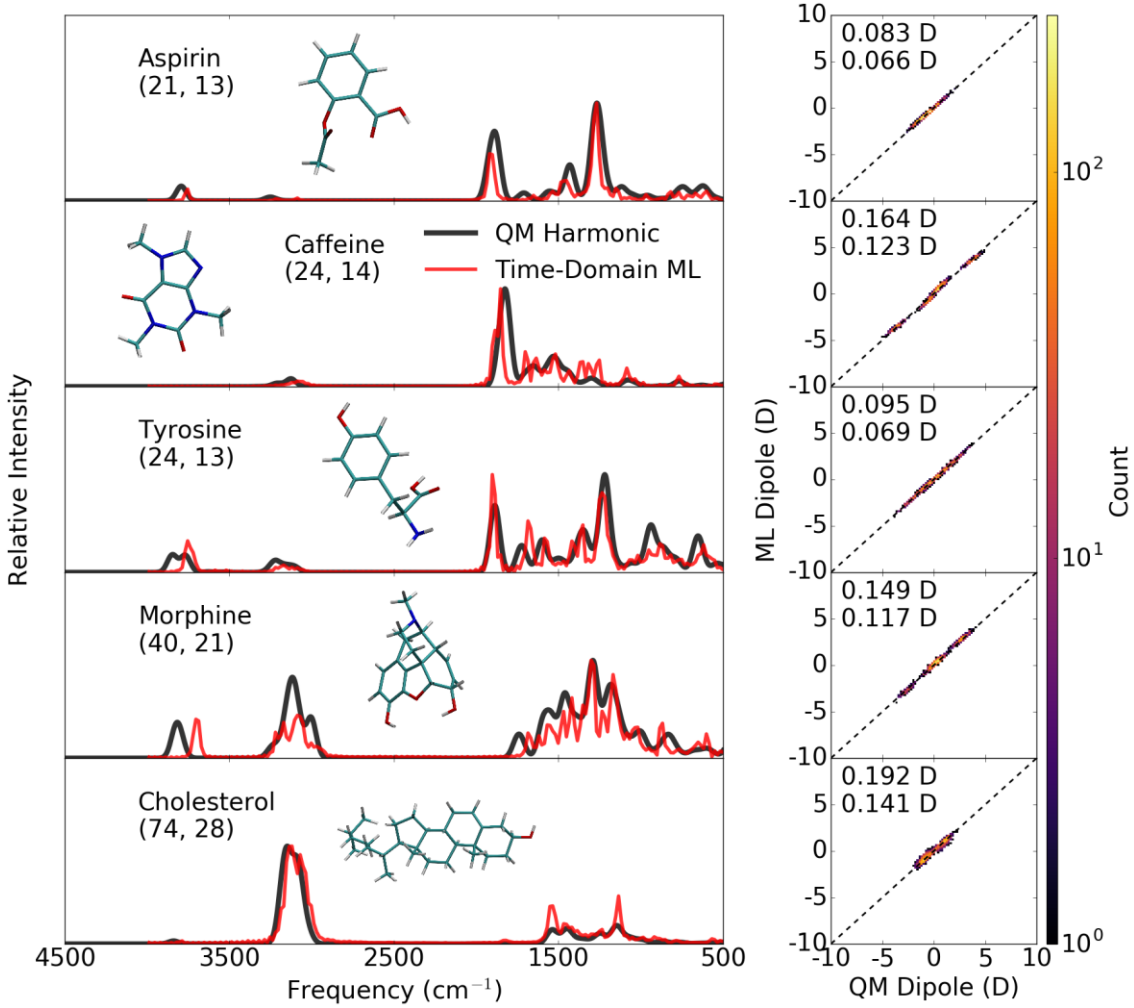


Training Directly to Dipoles





IR Spectra on Larger Molecules



Learning Hamiltonians

Hückel Theory and Other Model Hamiltonians

Extended Hückel theory is one of the simplest methods for simulating the electronic structure of Molecules.

$$H_{ij} = \frac{1.75}{2}\left(H_i + H_j\right)S_{ij}$$

$H_{carbon,\,S} = -21.4\ eV;\ H_{carbon,\,P} = -11.4\ eV;\ H_{hydrogen,\,S} = -13.6\ eV$

S is overlap matrix

Solve the following equation

$$H\psi = ES\psi$$

Can we optimize the atomic site parameters to match DFT?

Error Function

We need to quantify the error between Hückel and DFT in order to optimize the Hamiltonian.

$$Orbital\ Energy\ Error = \sum_{i}^{N_{orb}} \left(\varepsilon_i^{DFT} - \varepsilon_i^{Hückel}\right)^2$$

$$Regularization = \sum_{i}^{N_{parameter}} \left(H_i^{Hückel} - H_i^{Fit}\right)^2$$

$$Density\ Error = \sum_{i}^{N_{sites}} \left(\rho_i^{DFT} - \sum_{j}^{N_{occ}} \psi_{i,j}^2\right)^2$$

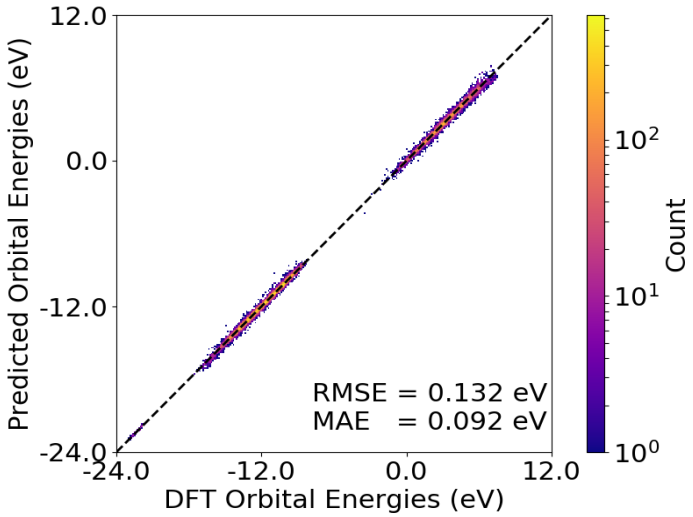
$$L = W_1 \times Orbital\ Energy\ Error + W_2 \times Regularization + W_3 \times Density\ Error$$

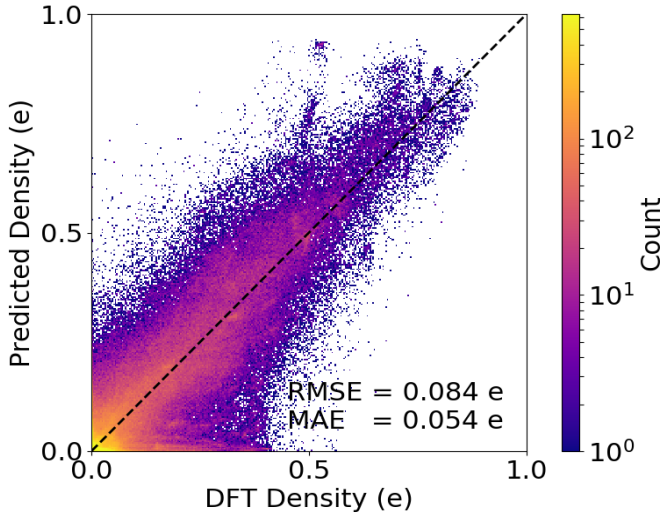
Minimize L over space of all H_i.

N_{orb} can be any number of valence and or virtual orbitals.

Training Dataset

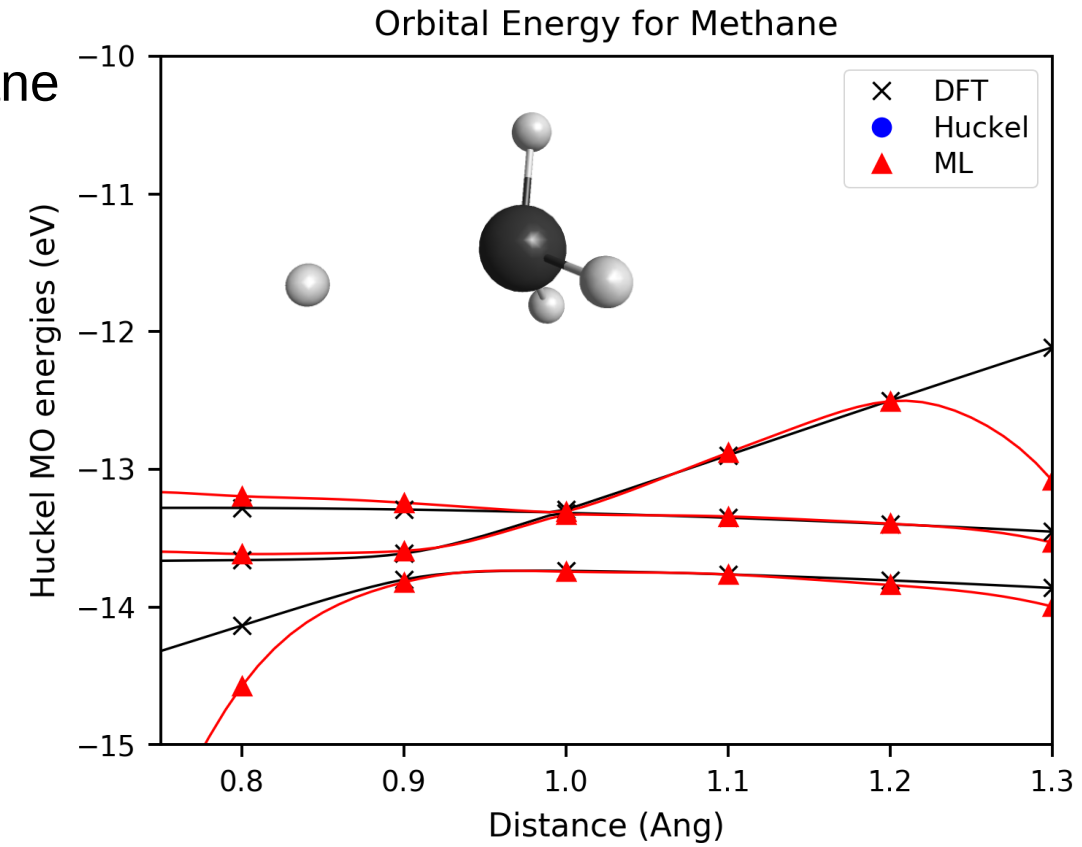
- 100,000 element database made of molecules with 14 atoms or less (5 heavy atoms or less)
- Small Database because diagonalization required at every computation



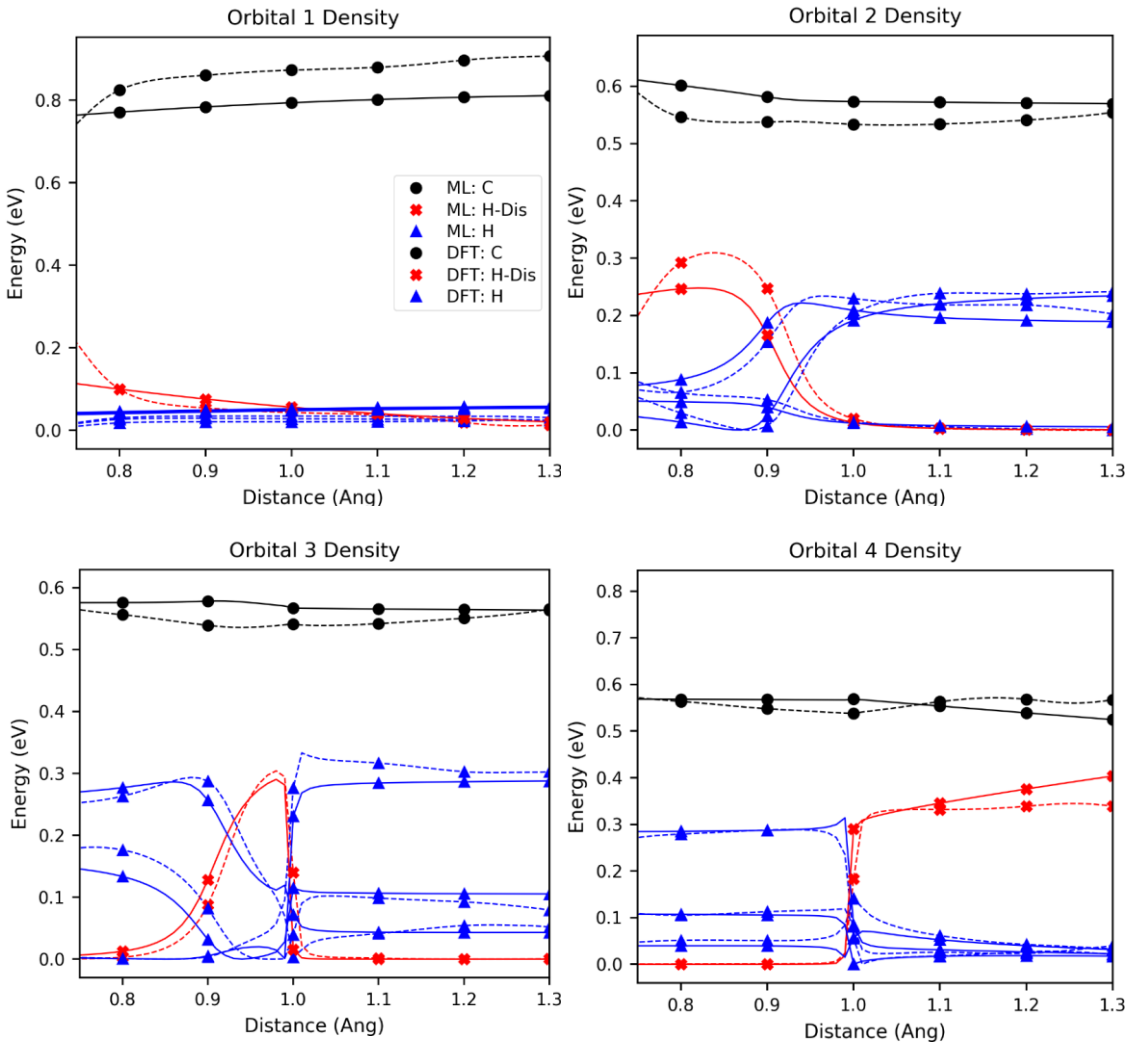


Interesting Orbital Properties are Still Captured

Eigenvalues across methane bond dissociation



Capture orbital character



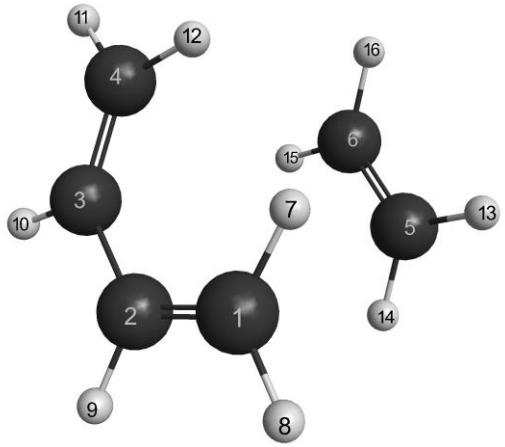
ML-dotted

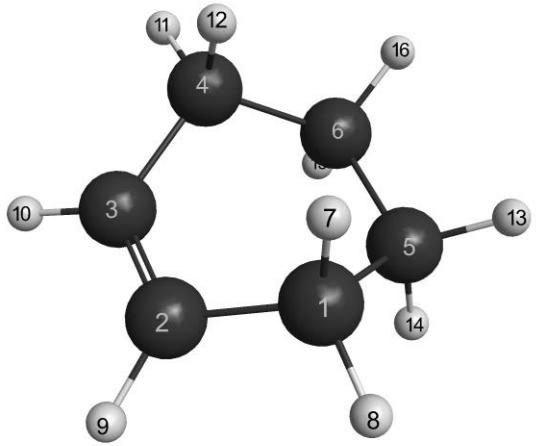
Solid-DFT

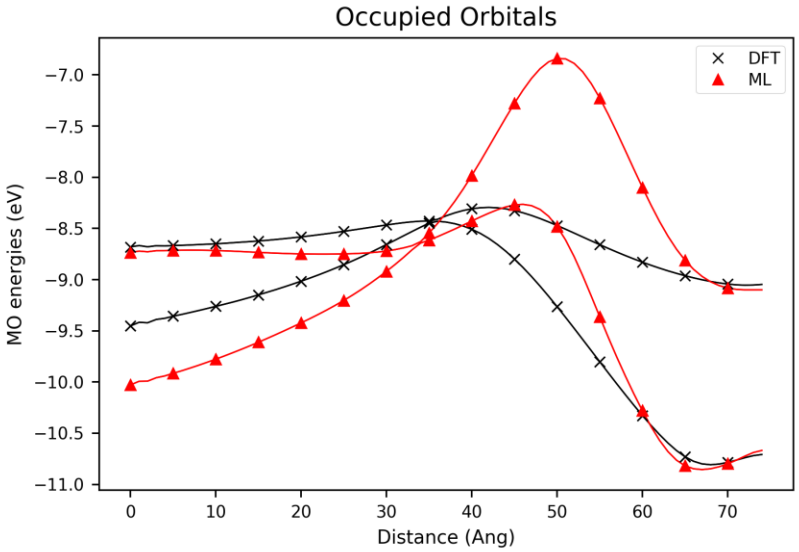
Mulliken charge contribution from each orbital plotted against the methane bond dissociation.

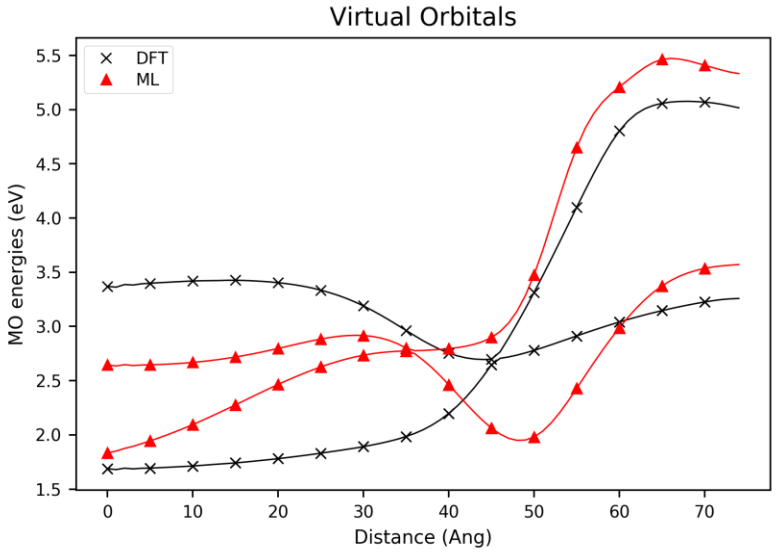
ML is able to attribute the correct state character to each MO.

Ring Cyclization

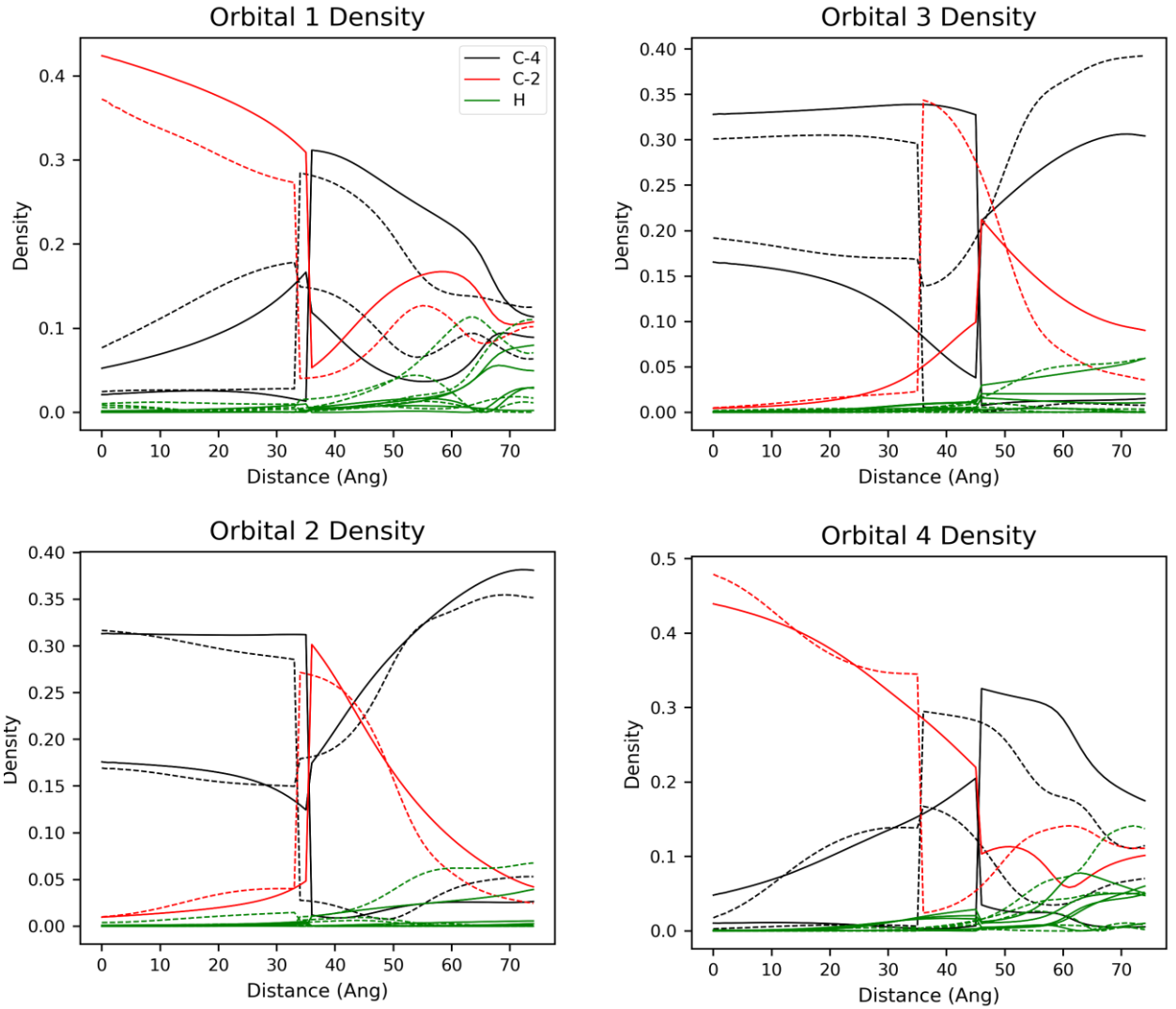








Qualitative orbital character is easier than energies



Thank you for listening