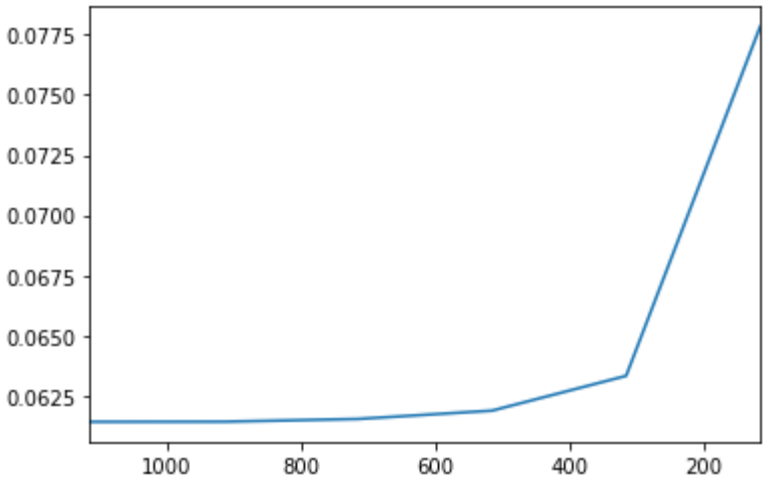
Report

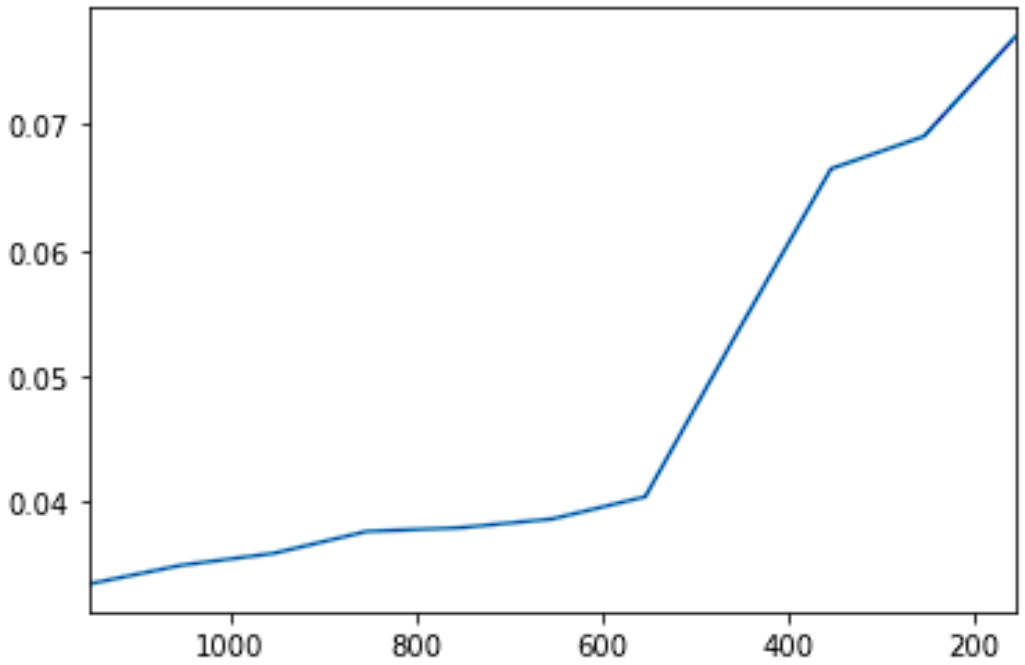
First Test on Glycine Molecules (4 degrees)



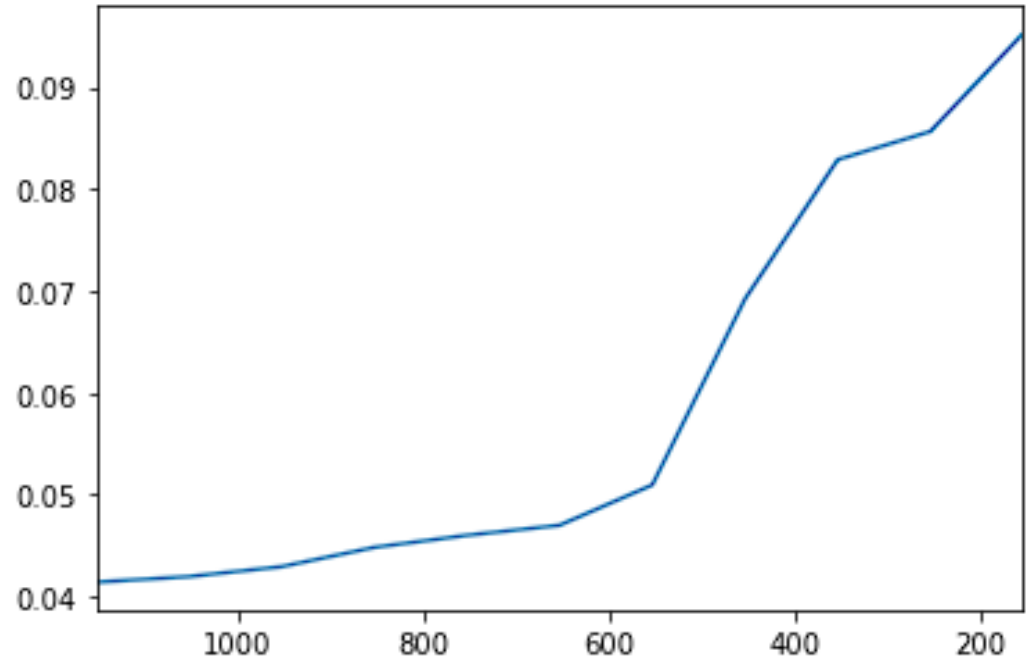
The RMSD did not increase much until the terms are eliminated to about 300 terms, compared to the original about 1100 terms.

Test on 2-body Water Molecules

training set errors



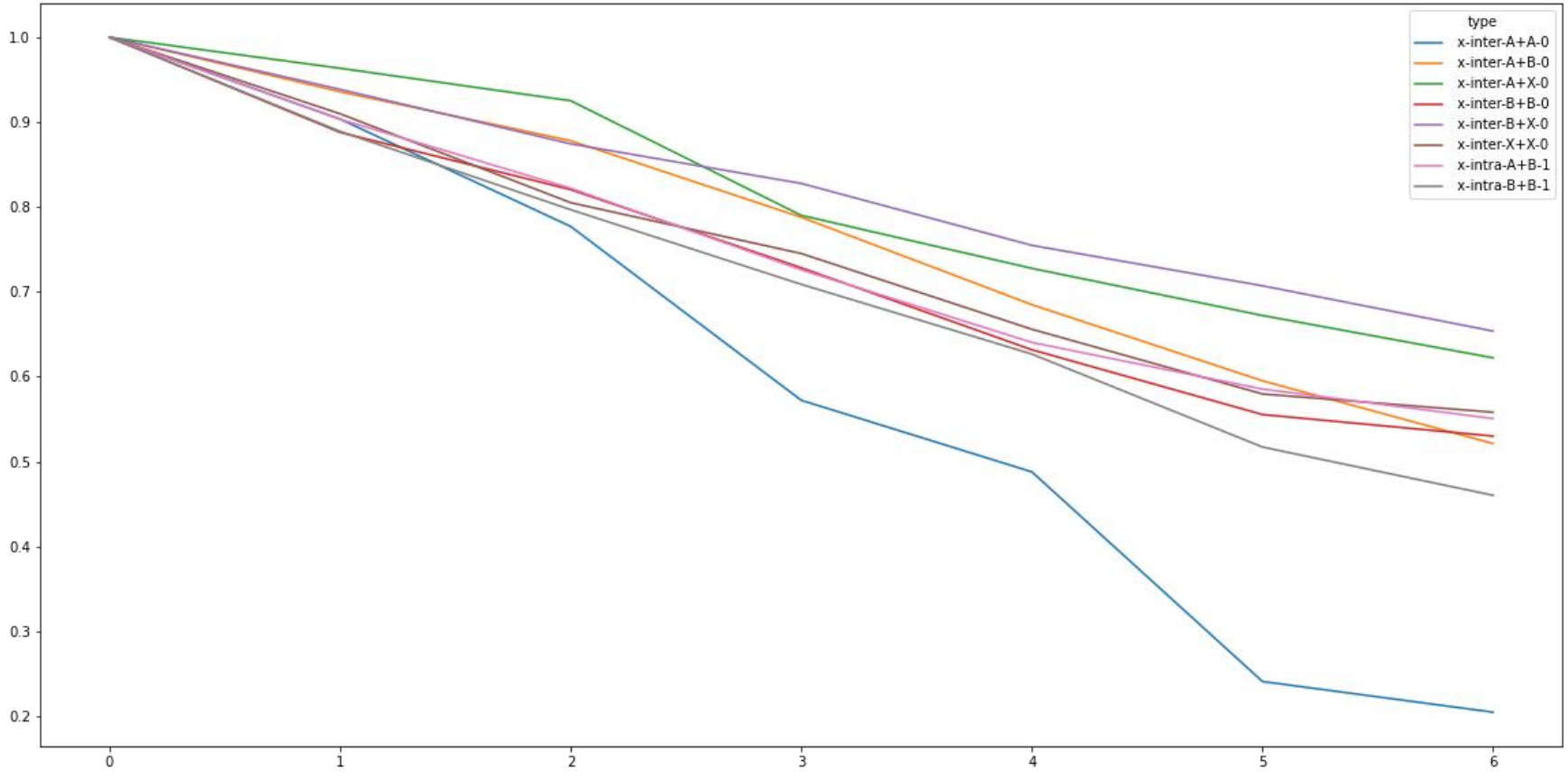
test set errors



Information of which terms are eliminated/remained

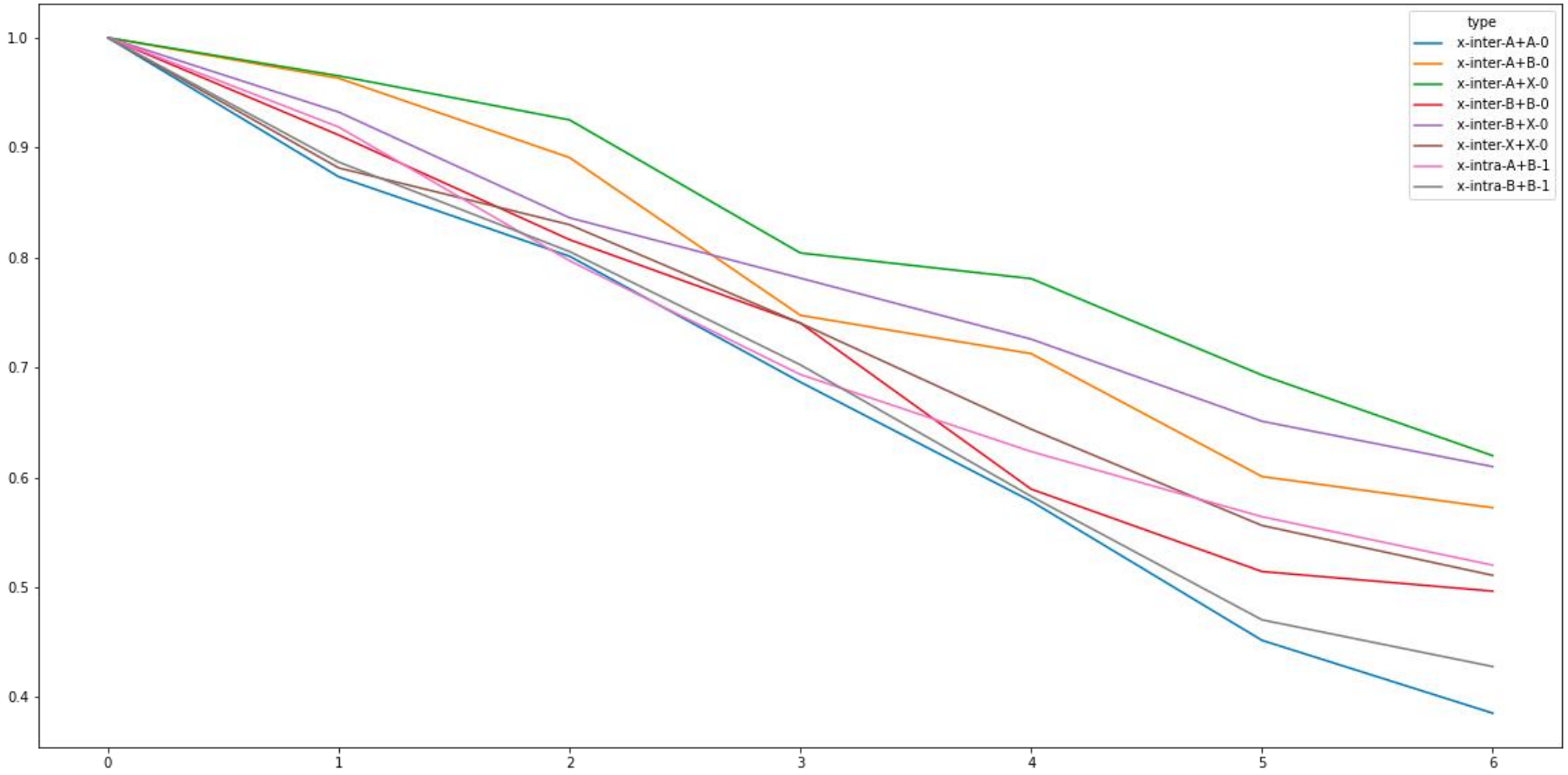
	A1, a, A2, b, x-inter-A+A-0	A1, a, B1, a, x-intra-A+B-1	A1, a, B2, a, x-intra-A+B-1	A1, a, B3, b, x-inter-A+B-0	A1, a, B4, b, x-inter-A+B-0	A1, a, X3, b, x-inter-A+X-0	A1, a, X4, b, x-inter-A+X-0	A2, b, B1, a, x-inter-A+B-0	A2, b, B2, a, x-inter-A+B-0	A2, b, B3, b, x-intra-A+B-1	...	B2, a, X4, b, x-inter-B+X-0	B3, b, B4, b, x-intra-B+B-1	B3, b, X1, a, x-inter-B+X-0	B3, b, X2, a, x-inter-B+X-0	B4, b, X1, a, x-inter-B+X-0	B4, b, X2, a, x-inter-B+X-0	X1, a, X3, b, x-inter-X+X-0	X1, a, X4, b, x-inter-X+X-0	X2, a, X3, b, x-inter-X+X-0	X2, a, X4, b, x-inter-X+X-0
0	0	0	0	0	0	1	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	0	0	0	1	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	1	0	0	0
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1148	0	1	0	0	0	0	0	0	0	1	...	0	0	0	0	0	0	0	0	0	0
1149	0	2	0	0	0	1	0	0	1	0	...	0	0	0	0	0	0	0	0	0	0
1150	0	1	1	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1151	0	1	0	1	0	0	0	0	0	0	...	0	1	0	0	0	0	0	0	0	0
1152	0	1	1	0	0	0	0	0	0	0	...	0	0	1	0	0	0	0	0	0	0

Information of which terms are eliminated/remained

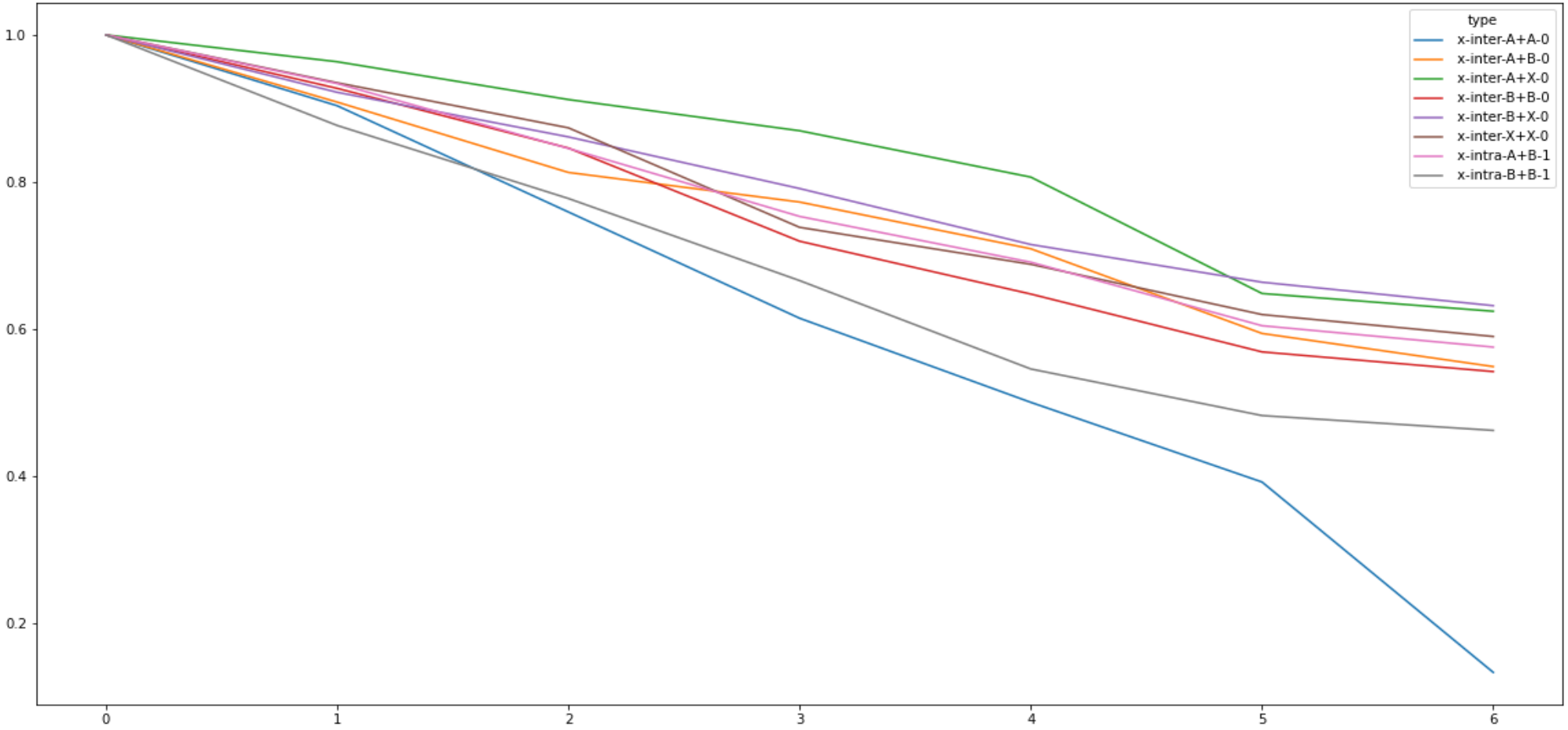


The normalized proportion of the remaining terms after each round of eliminations compared with the initial states without eliminations.

Information of which terms are eliminated/remained



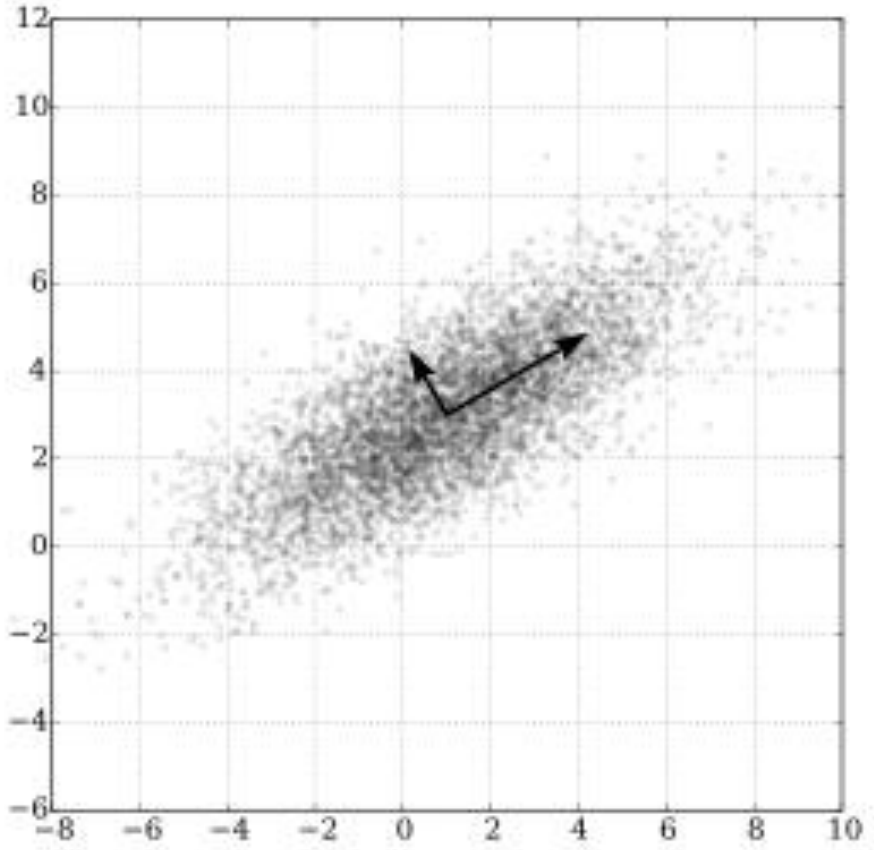
Similar patterns observed

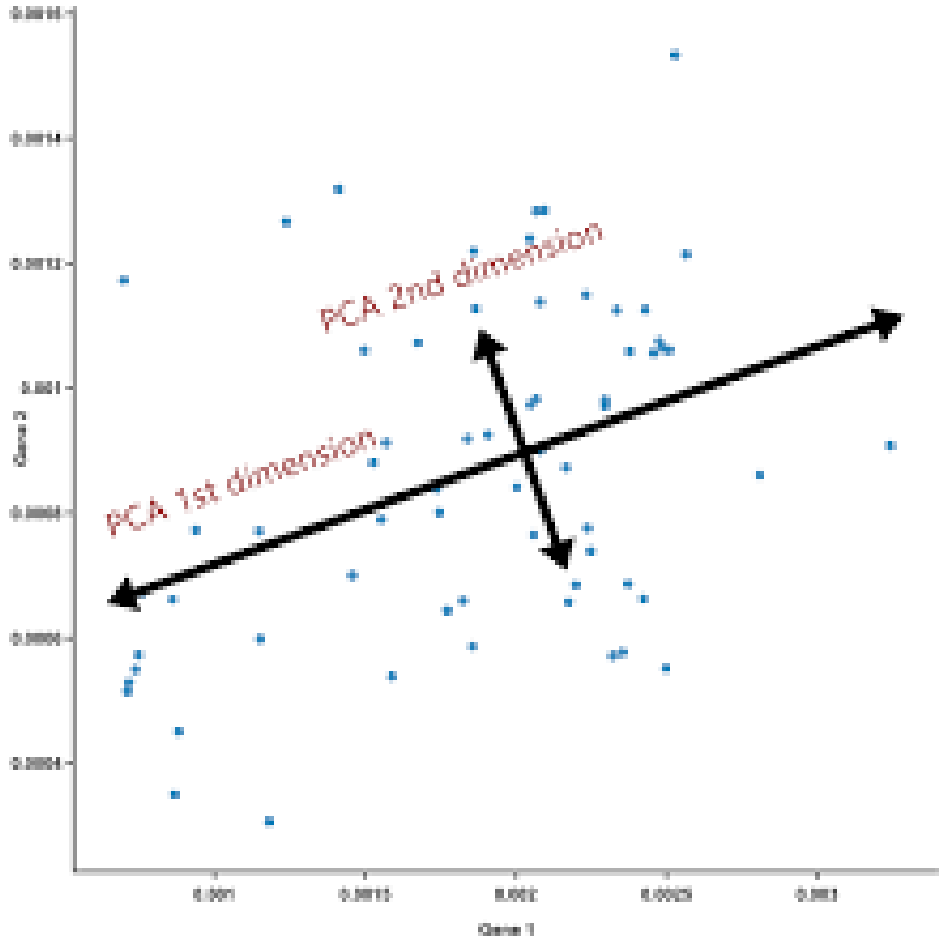


Limitation on RFE

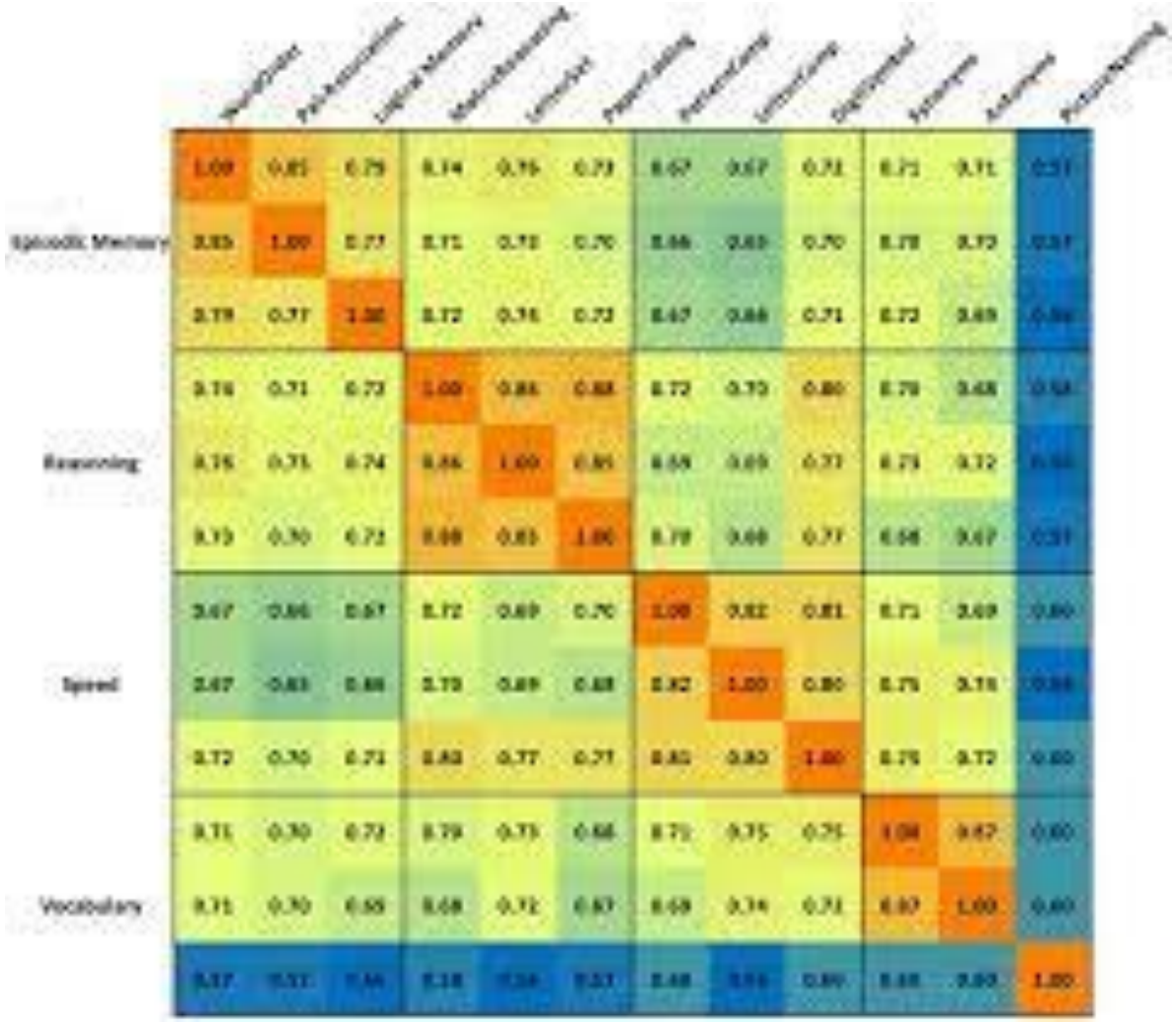
```
Atom names ['A1a', 'B1a', 'B2a', 'Z1a', 'Z2a', 'A2b', 'B3b', 'B4b', 'Z3b', 'Z4
b']
File ./poly.log already exists, moving existing ./poly.log to ./poly.log.backup
-1 to make way for new file.
Finding permutations...
Generating terms up to degree 5...
|=========================================================================
======================|
8 possible degree 1 terms, now filtering them...
There were 6 accepted degree 1 terms.
70 possible degree 2 terms, now filtering them...
There were 63 accepted degree 2 terms.
505 possible degree 3 terms, now filtering them...
There were 491 accepted degree 3 terms.
3260 possible degree 4 terms, now filtering them...
There were 593 accepted degree 4 terms.
18697 possible degree 5 terms, now filtering them...
There were 5384 accepted degree 5 terms.
There were 6537 accepted terms over all
```

Idea from PCA





Correlation Matrix




```
def eval_monomial(monomial, configuration):
    monomial_value = 1
    for variable_index, degree in enumerate(monomial):
        monomial_value *= variable[variable_index] ** degree
    return monomial_value
```

---

where variables are computed as:
variable[variable_index] = e^(-kd)

Results

Pearson correlation coefficient:

The threshold I used is the testing p-value for testing non-correlation.

smaller p-value → more statistically significant

higher p-value → the non-correlation can’t be rejected

Configurations I used to test from ruihan:

1 body-molecule with 12 atoms, C(O)(C(H)(H)(H))N(H)C(H)(H)(H)

Performance

Run MB-Fit on new selected terms

Time: over 100 hours → about 4 hours

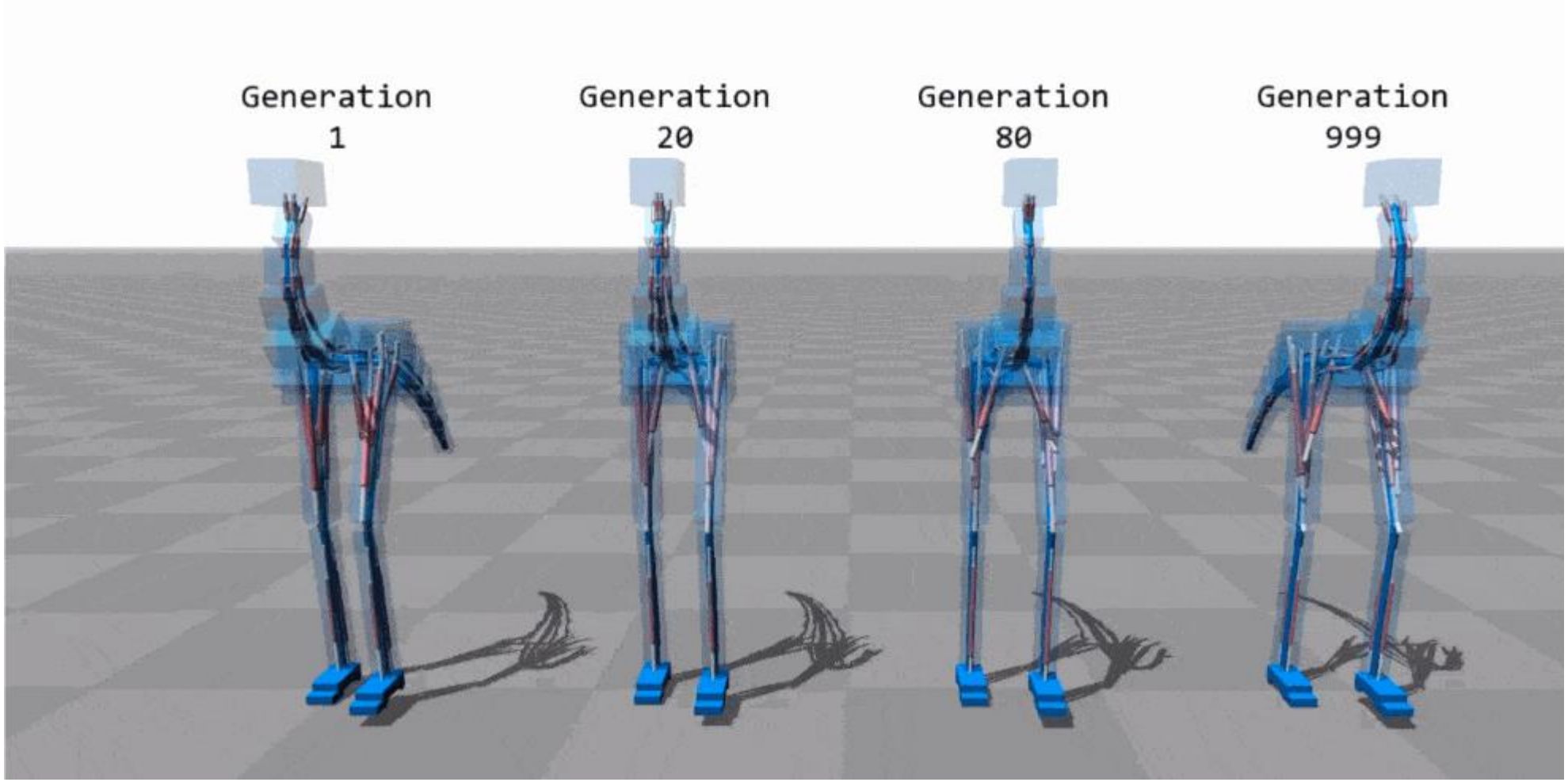
RMSE: 0.081 → 0.72

Issues and Further improvements:

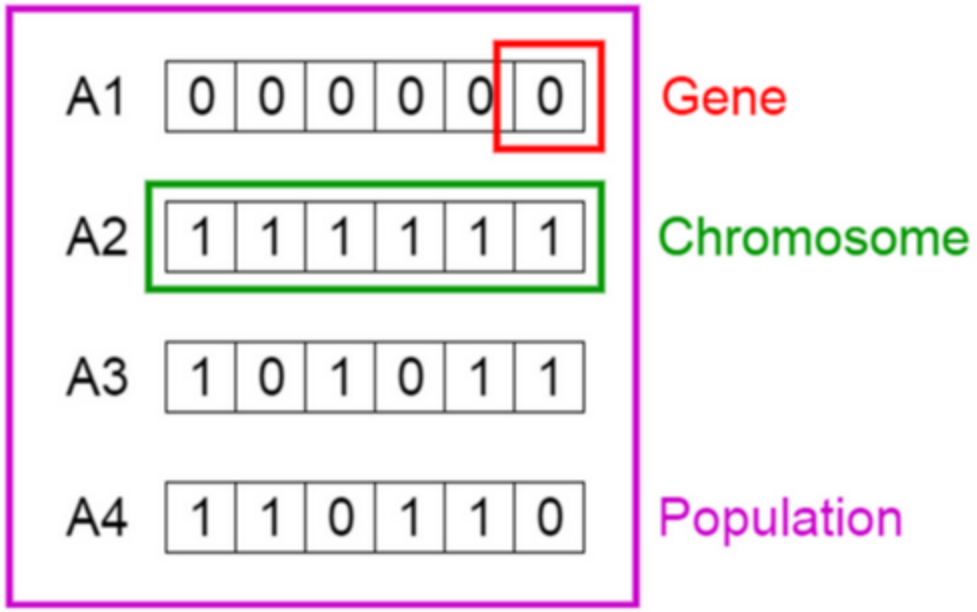
1. The approximation of non-linear parameters (more important)
2. Filter's choosing strategy, it might be too far (from 8000 to 1800)

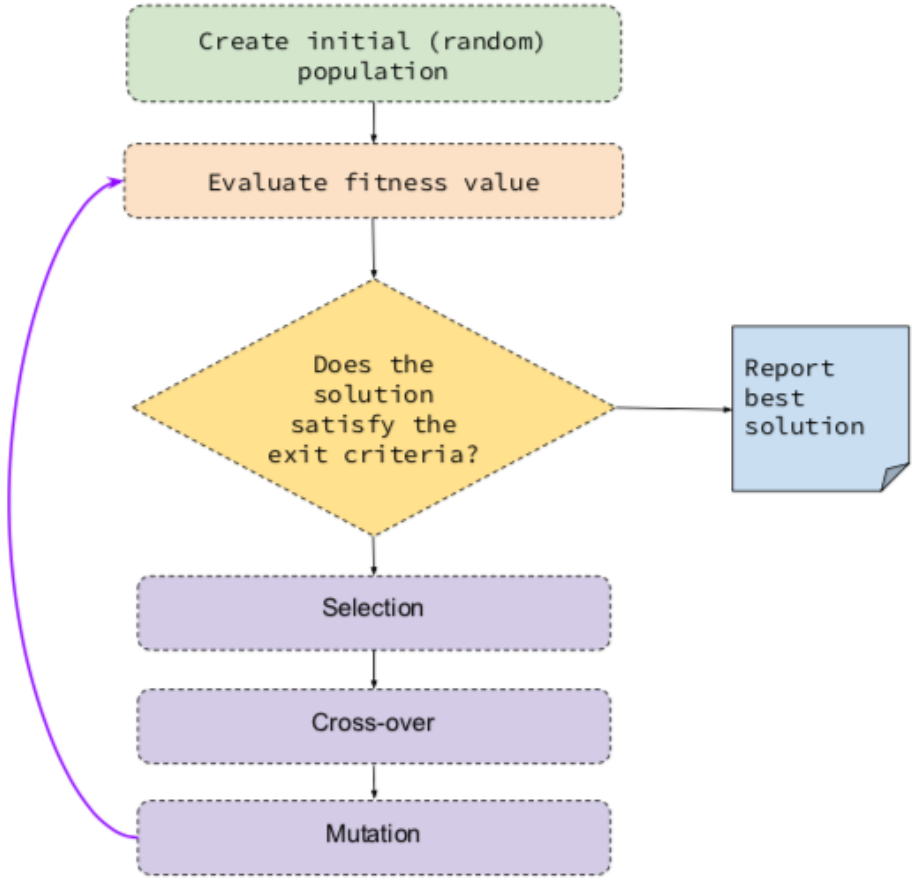
If the issues can't be solved, might move to GA.

GA (Genetic Algorithms)



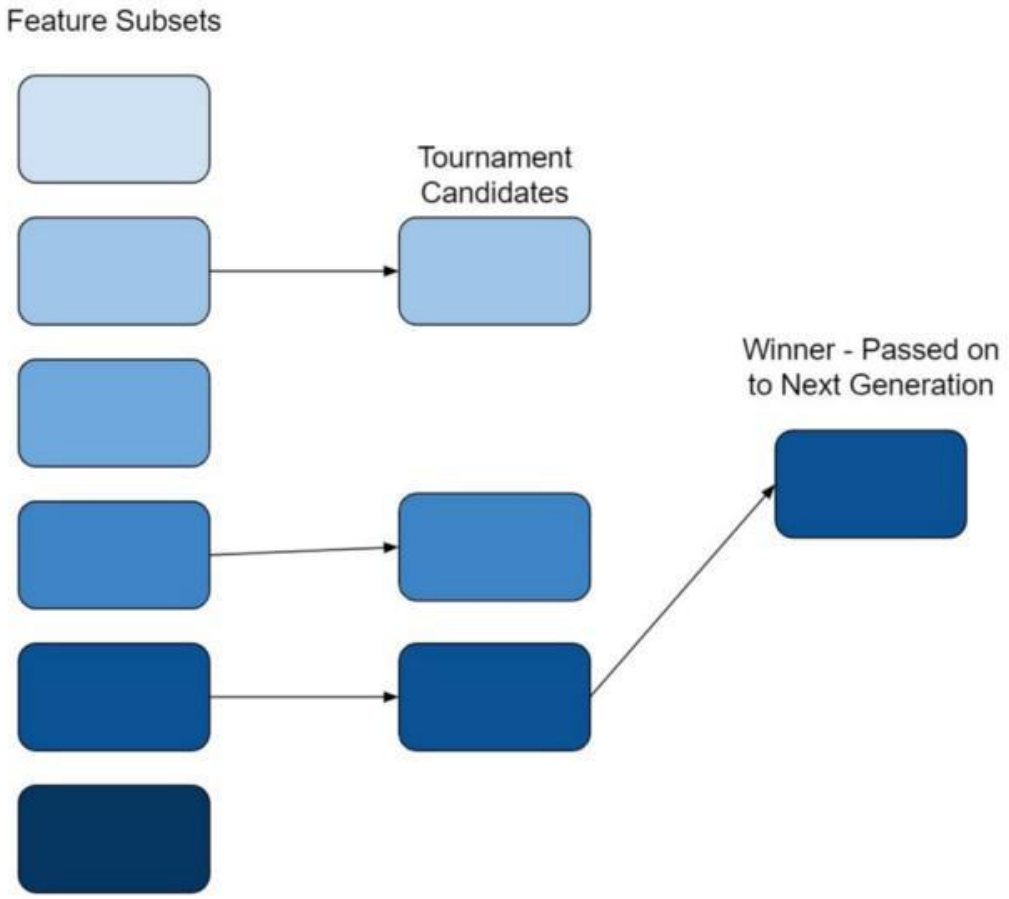
GA Process and its operators

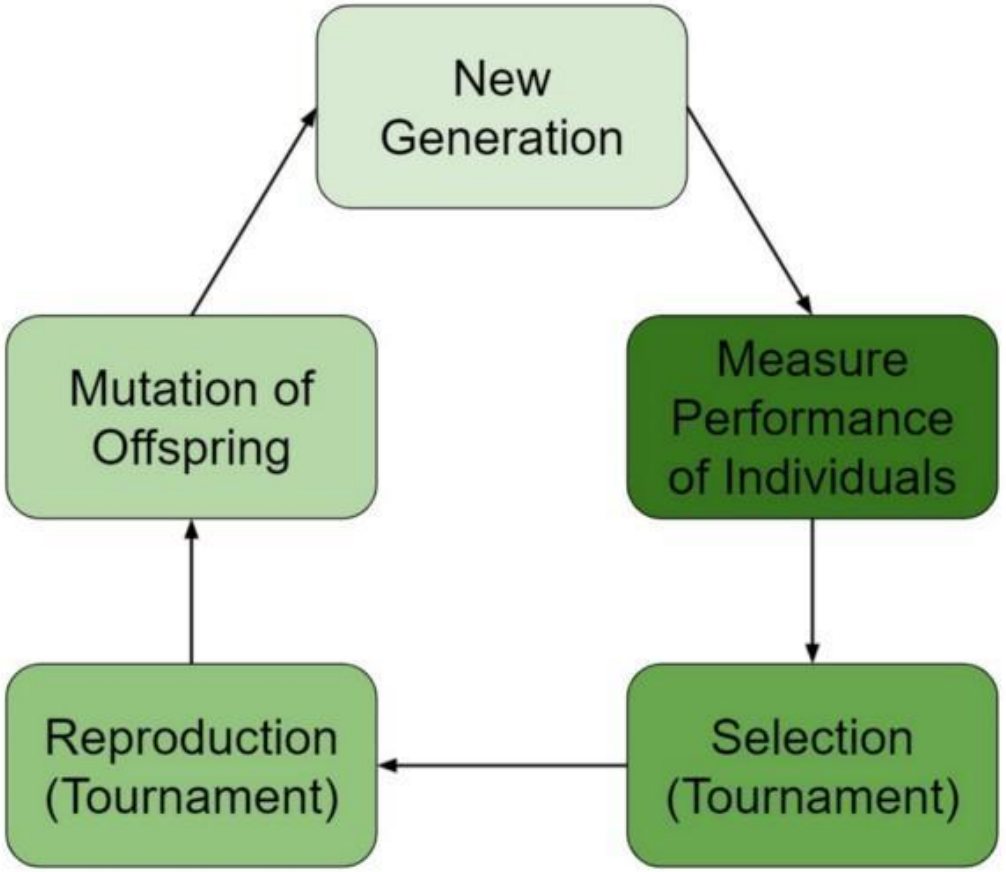




https://blog.datascienceheroes.com/

More about GA





Updates on Neural Network

Article | Open Access | Published: 04 May 2022

E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials

Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P. Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E. Smidt & Boris Kozinsky

Nature Communications **13**, Article number: 2453 (2022) | Cite this article

In this paper they said their neural network (called nequip), can use much less dataset to reach even better accuracy

Training Data

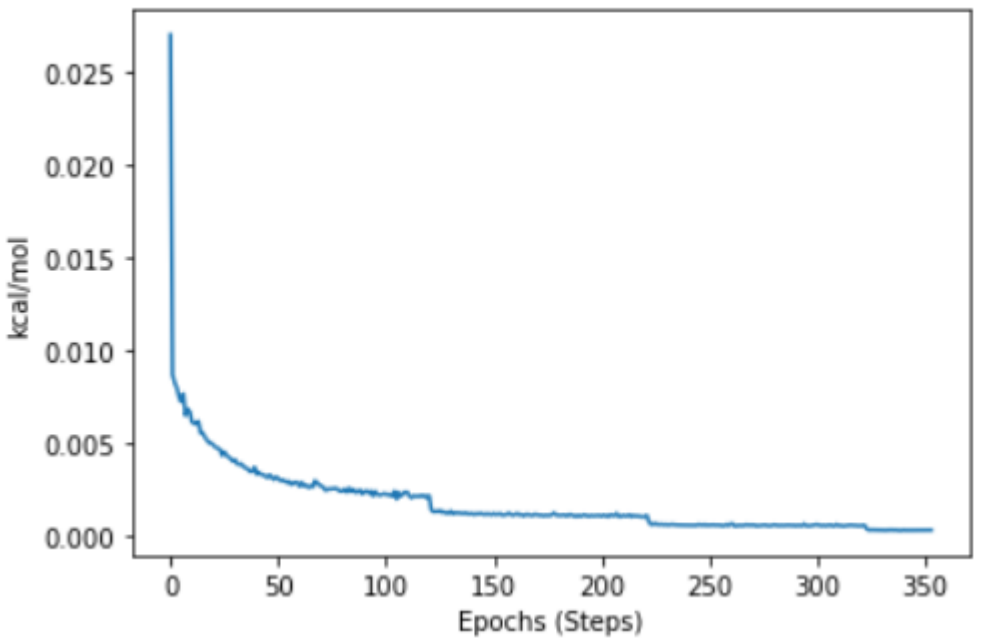
Training data came from Yaoguang, Iteration-3, combined training sets from 198K to 368K. There are 87177 samples in total, each sample is a box of water molecules containing 256 water molecules.

Because the original training set is quite large and I found the model running very slow even with gpu, I instead tested the models using two different smaller sizes training set sampled from the original training set (from each sub training sets from 198K to 368K):

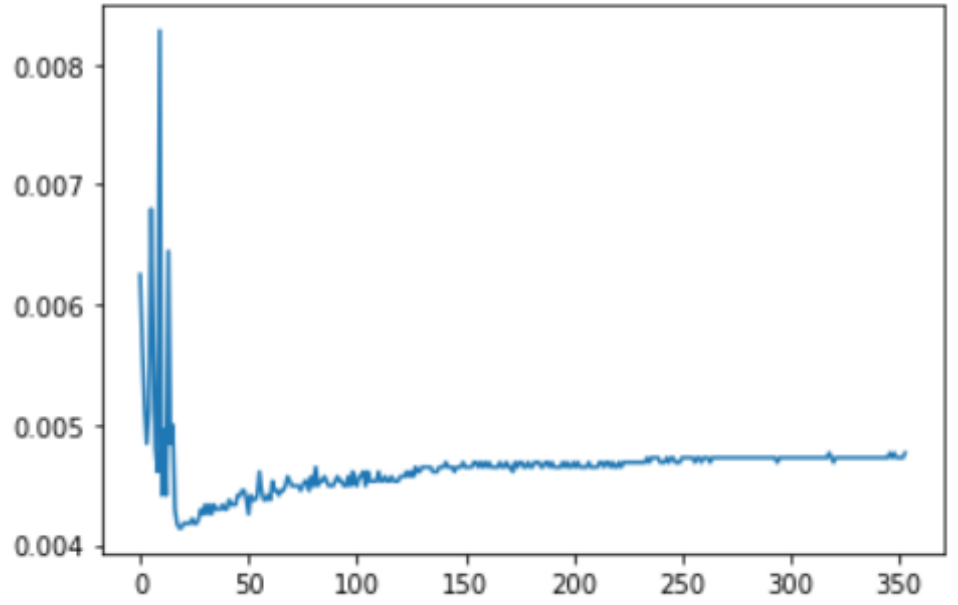
1. About 22000 samples
2. About 1300 samples

Performance: 22000 samples

Train



Test

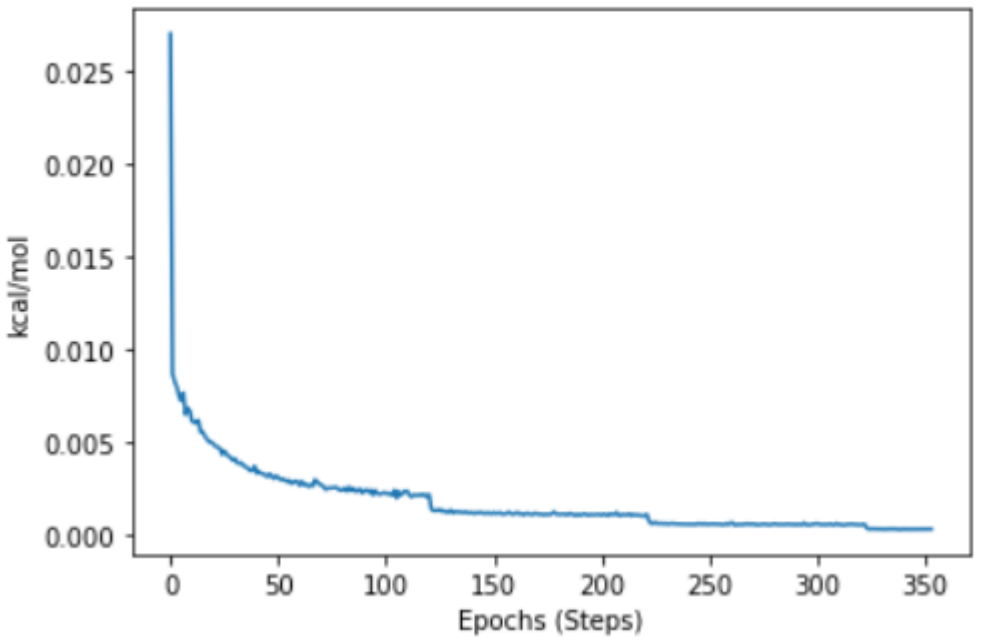


Overfit!!

Training	Testing
8.43e-05	0.0013

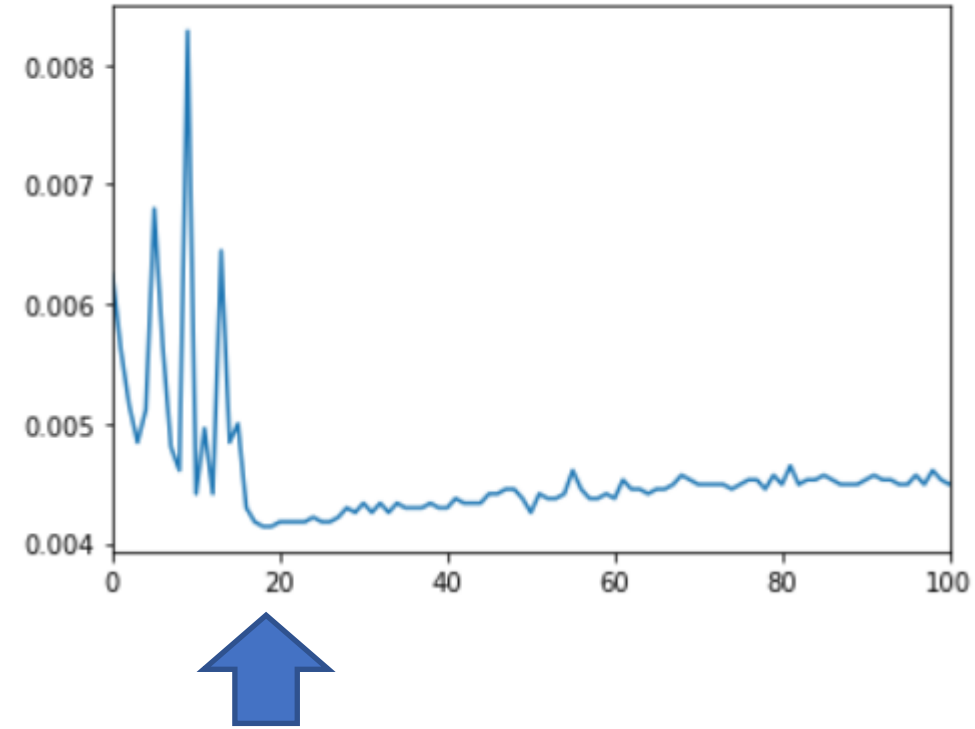
Performance: 22000 samples

Train



Test

Zoom-in version

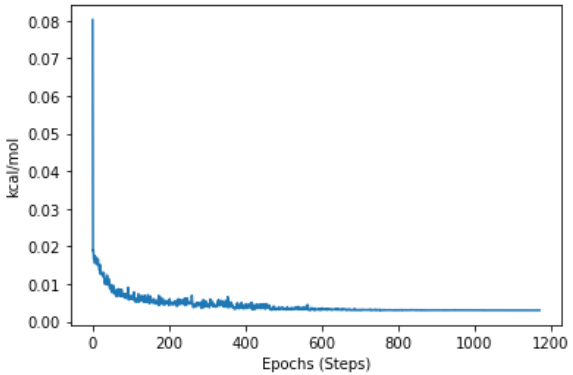


Reached best rmsd at around epoch 19

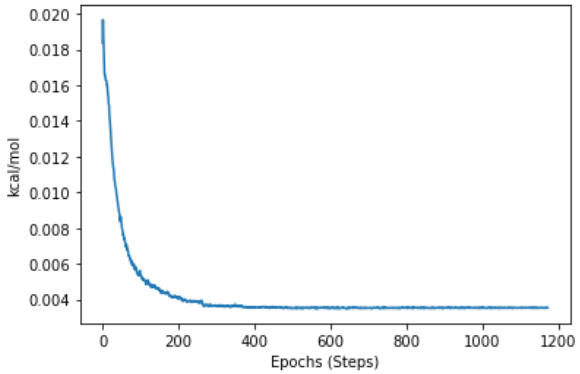
Training	Testing
8.43e-05	0.0013

Performance: 1300 samples

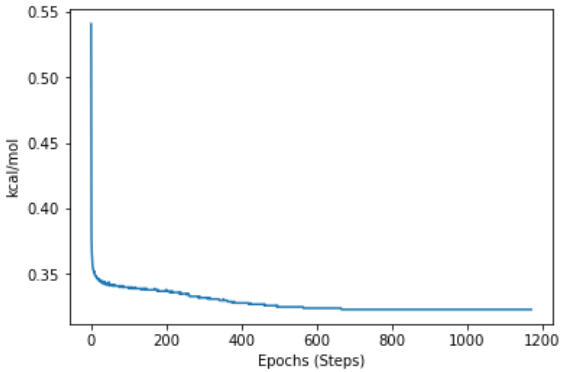
Train Energy



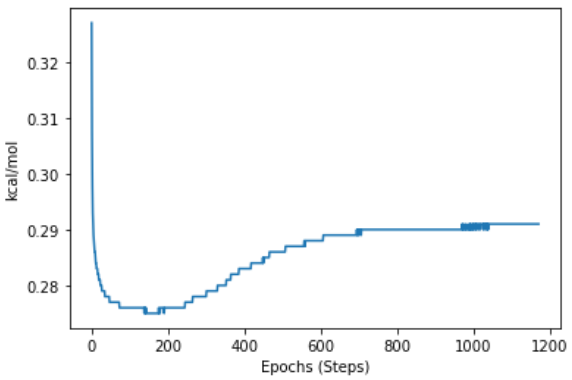
Test Energy



Train Forces



Test Forces



Energy Train	Energy Test	Forces Train	Forces Test
0.00301	0.00353	0.323	0.291

Evaluate Model on Smaller Clusters

	Monomers	Dimers	Trimers	Tetramers
f_rmse	0.299152	0.303673	0.304719	0.294754
e_rmse	0.278719	0.584196	0.821500	1.057976

Simulations

LAMMPS:

Problem:

LAMMPS using nequip pair-style running slow (or no output at all) for 15 hours even when using gpu. (no error and the job just keep running without any output). Possible reasons: MPI is not supported, NPT simulation is not supported (only NVT is supported now)

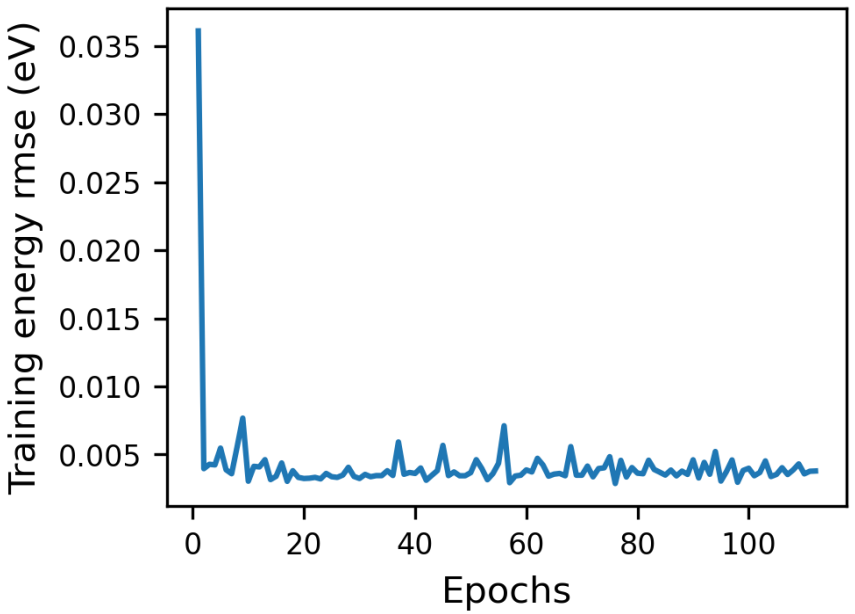
ASE:

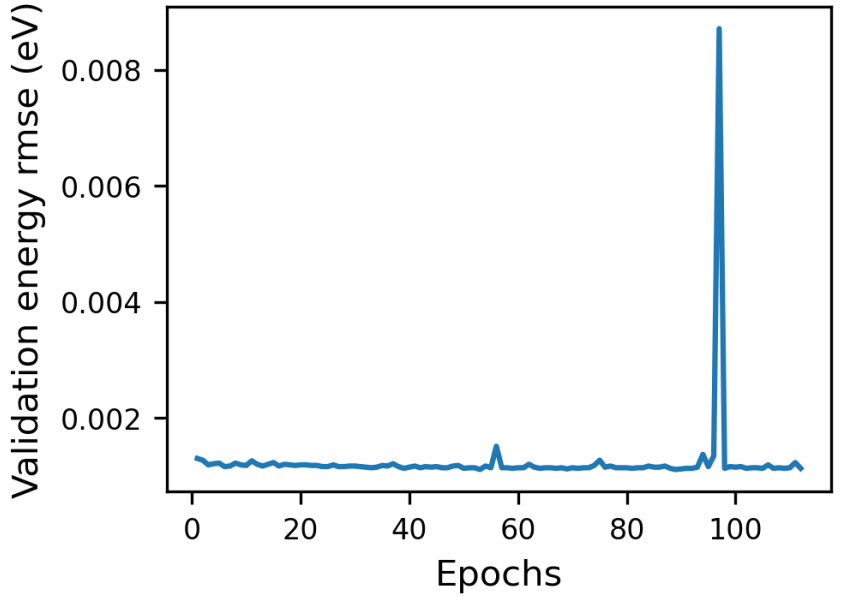
Problem:

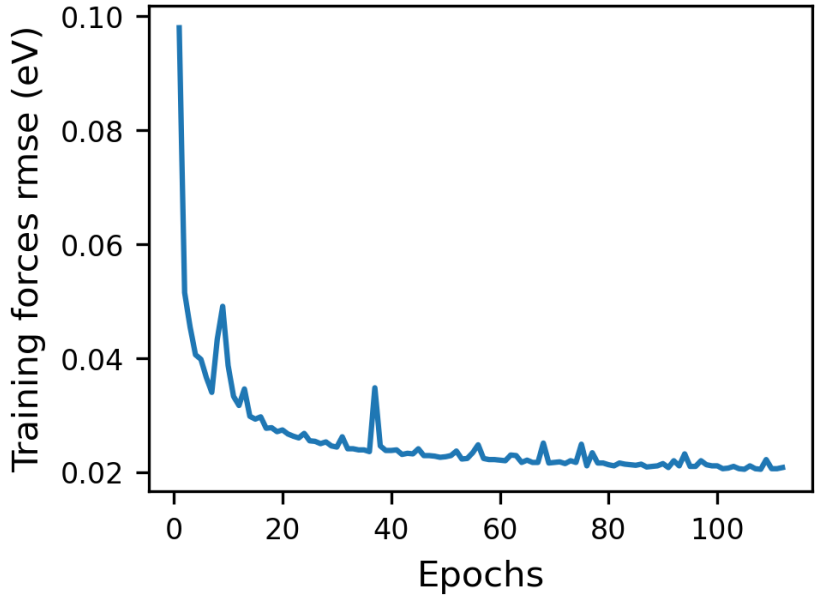
Not stable: can run simulation fast and getting outputs, but the simulation start becoming unstable after few hundreds of steps.

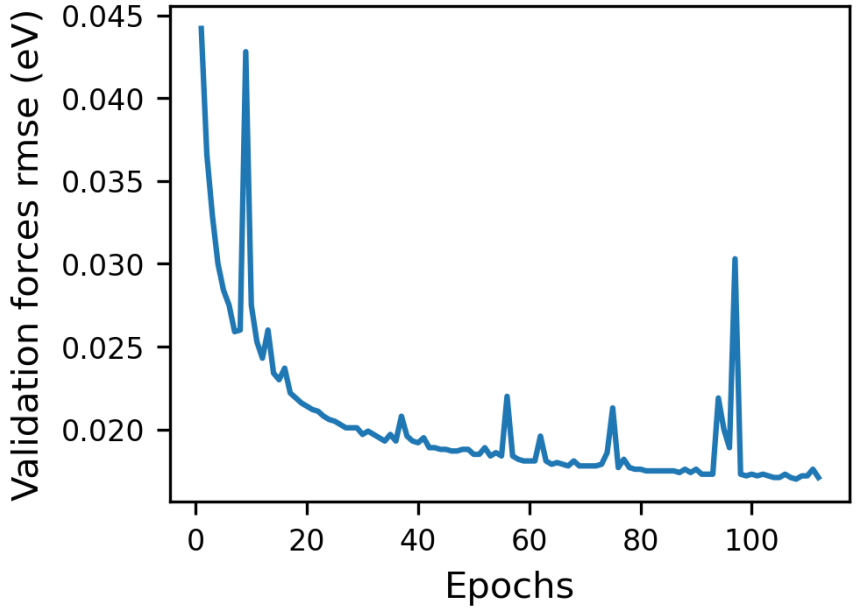
(need to spend a little more time to learn how to use ASE)

New Model: Performance

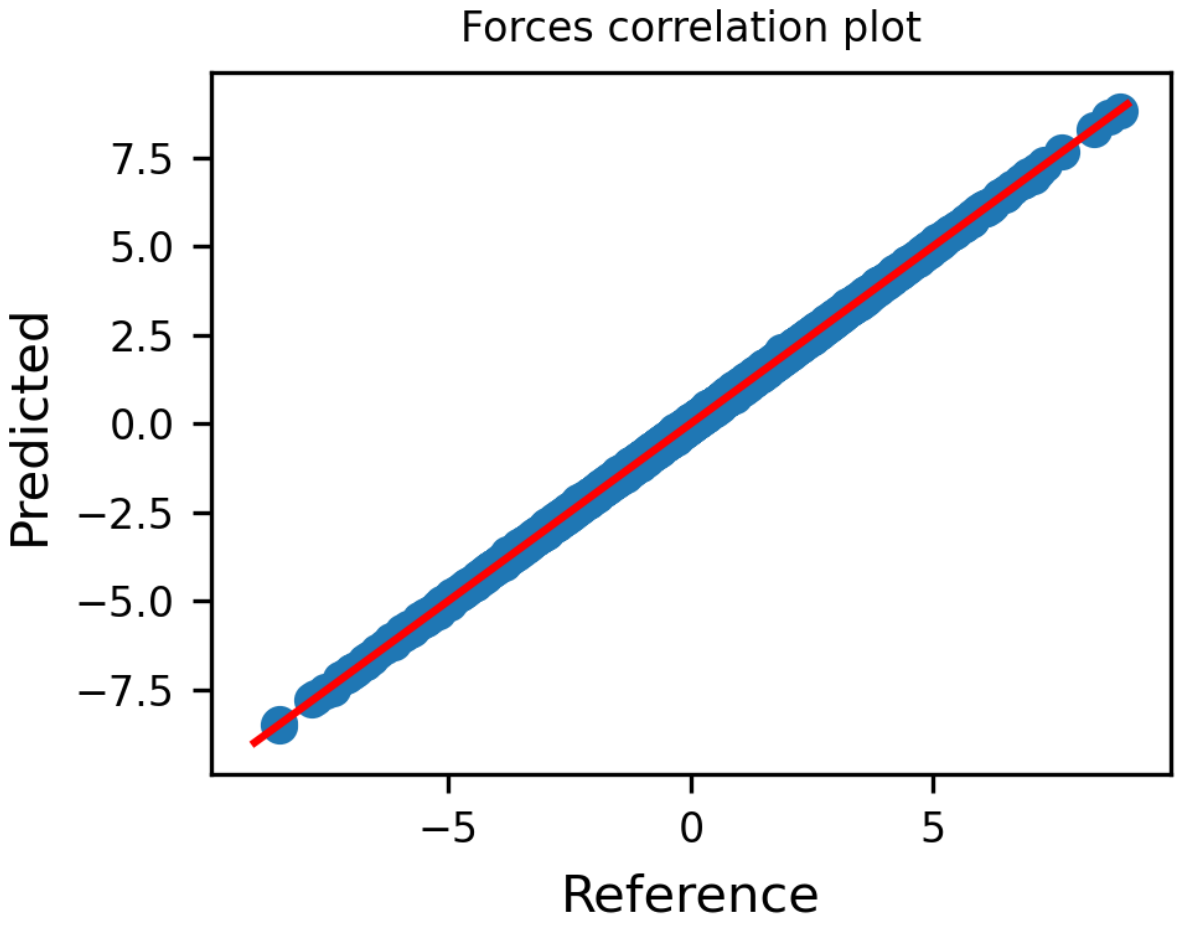


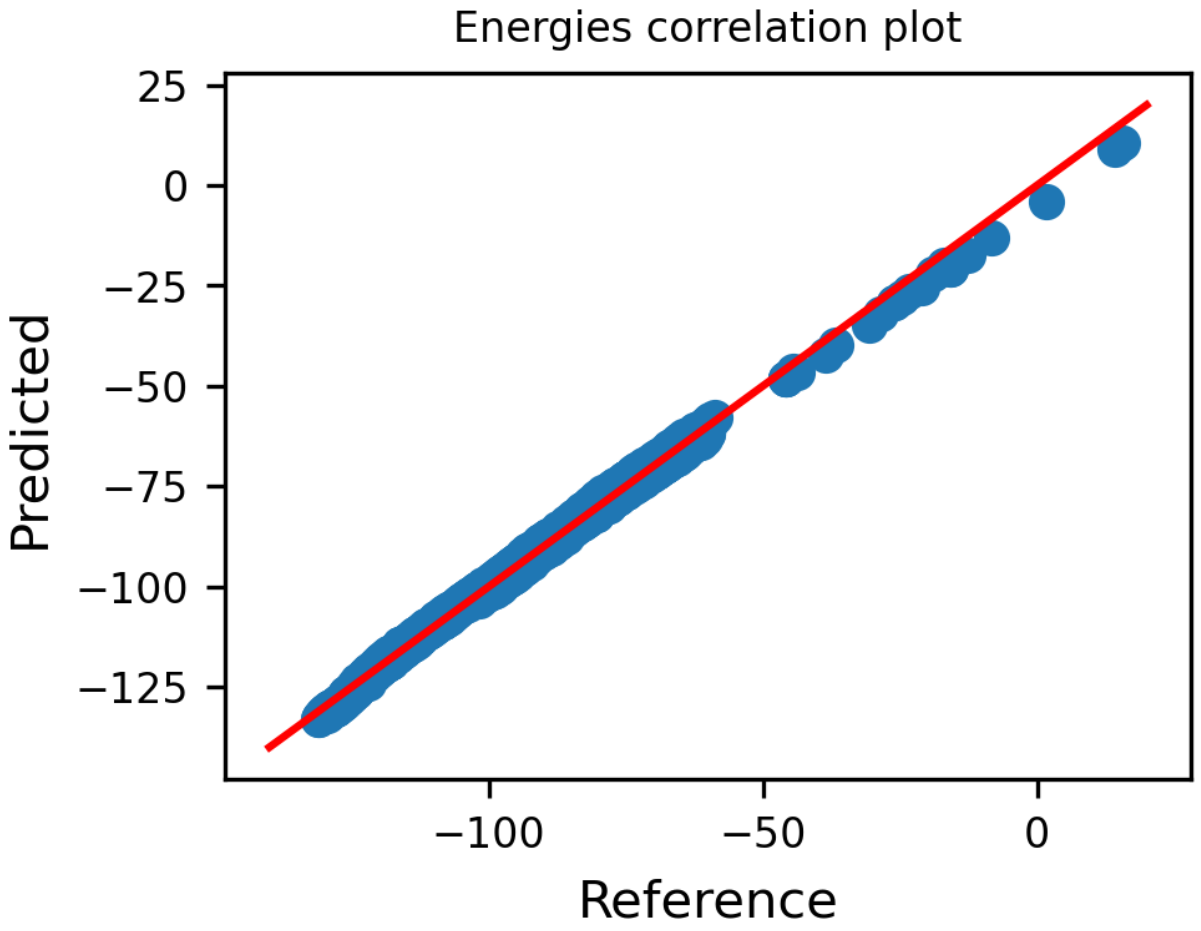




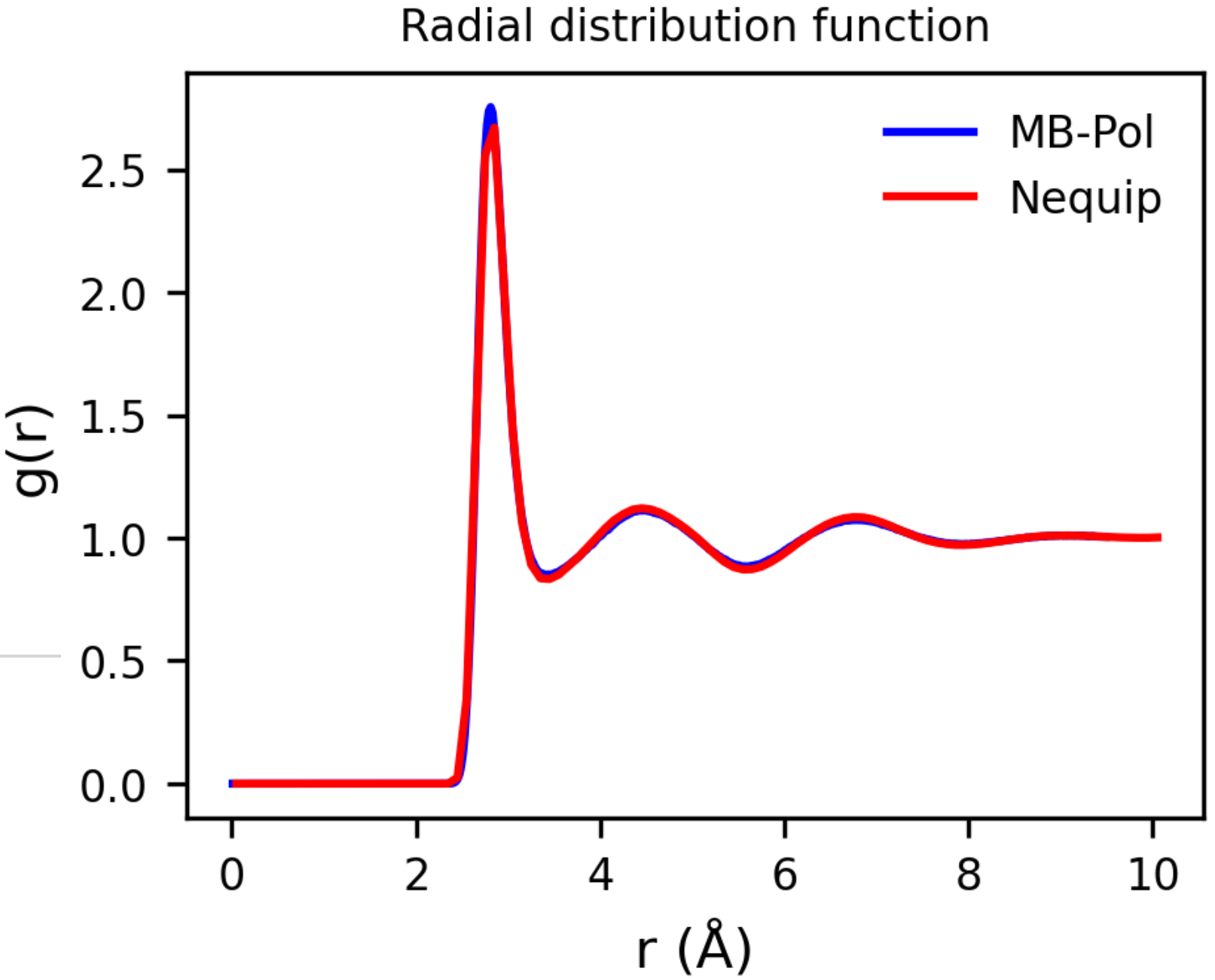


New model: Correlation plots

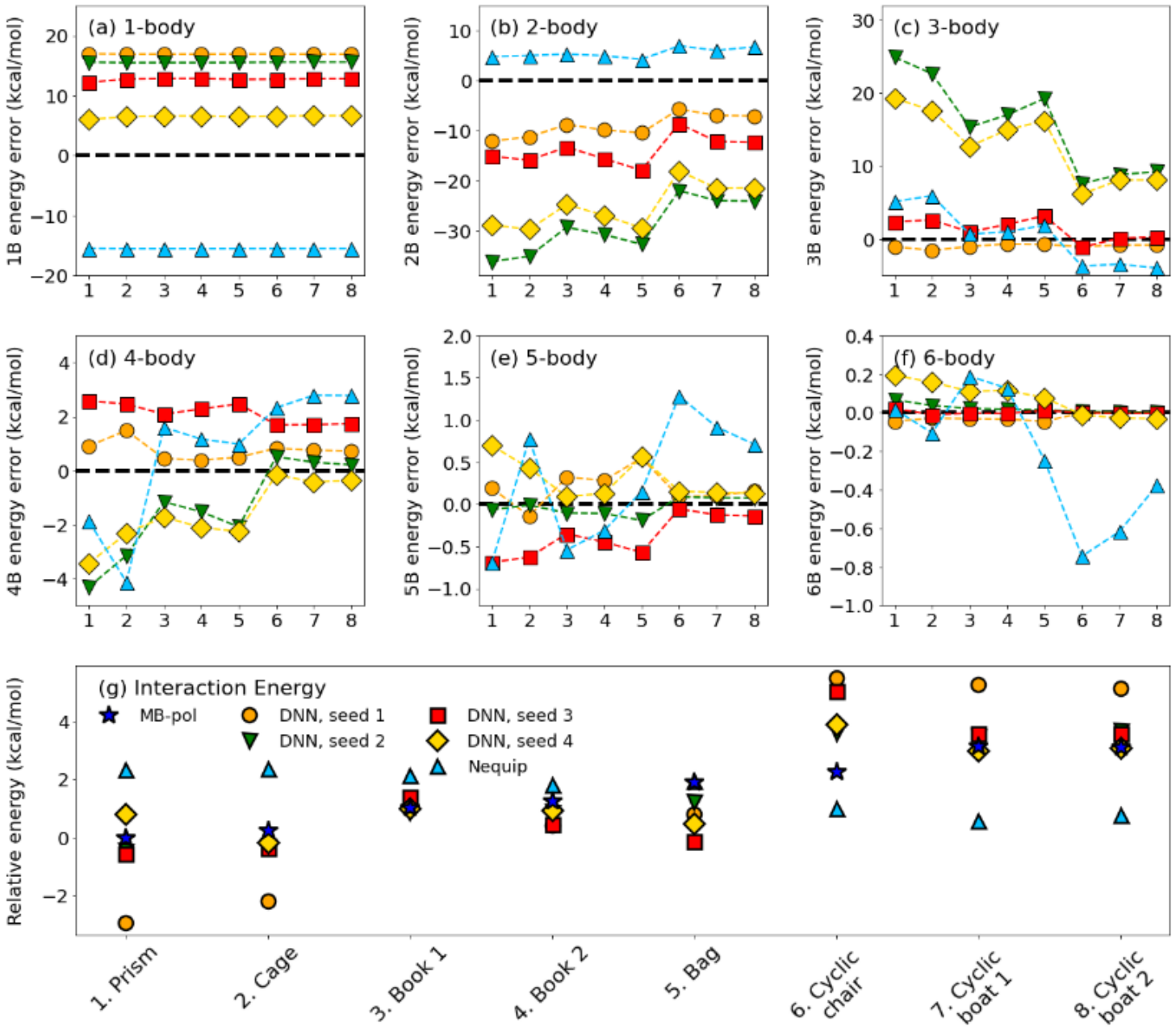


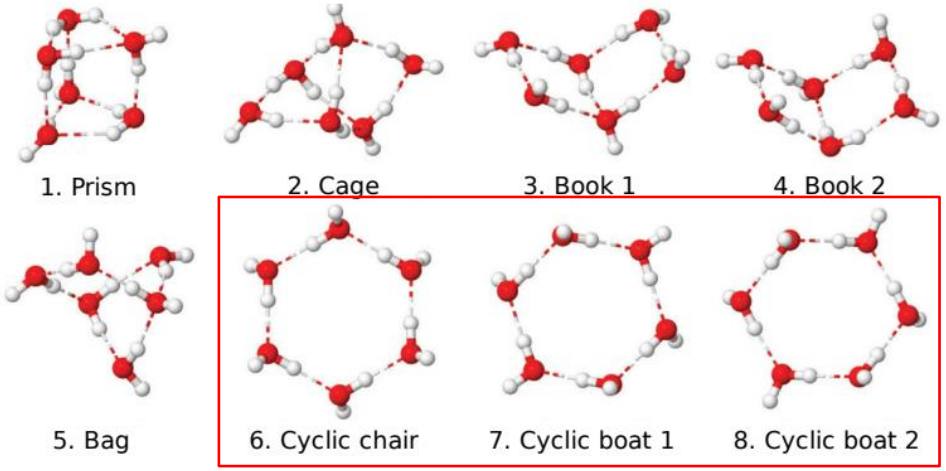


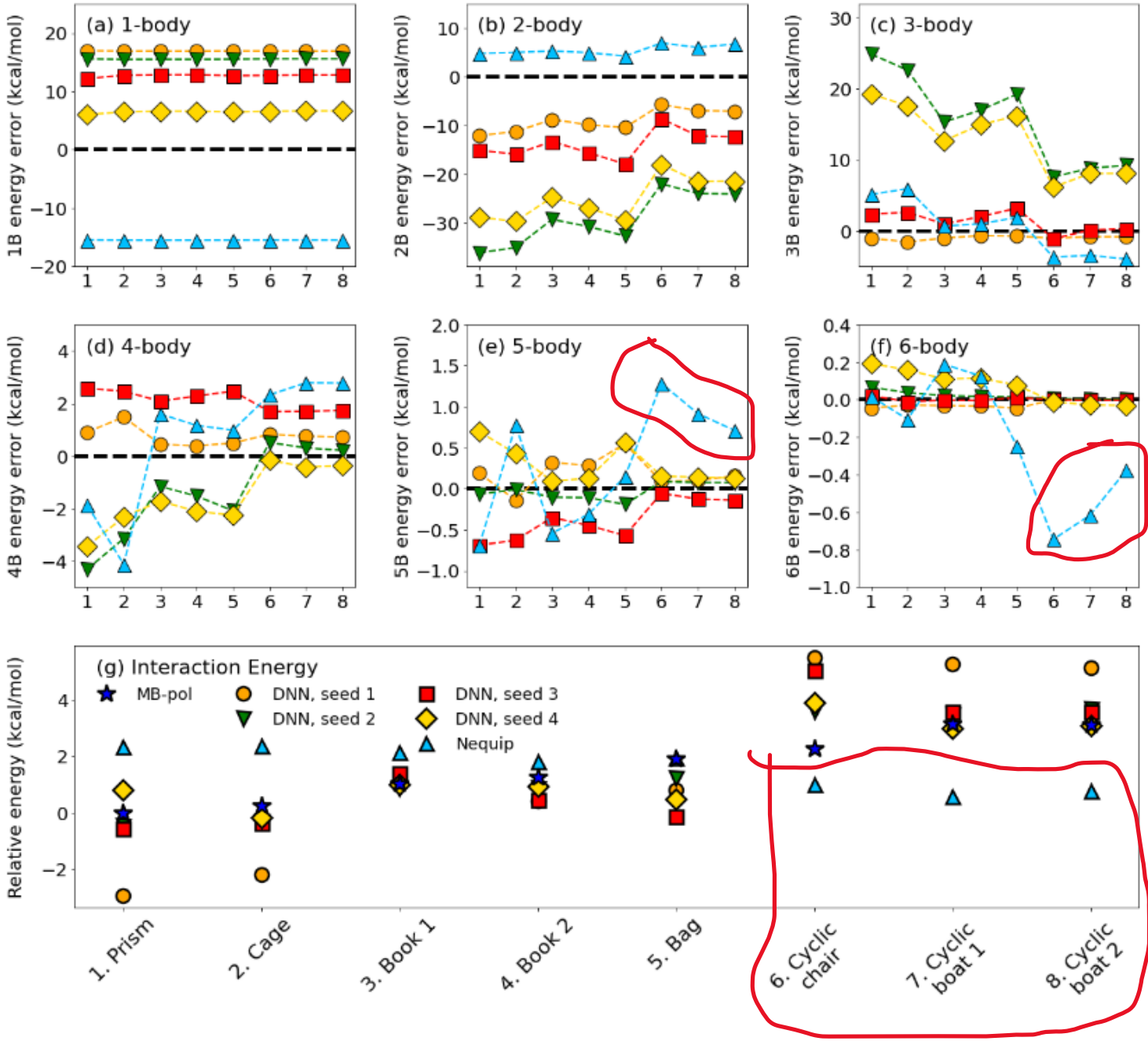
Simulation and RDF



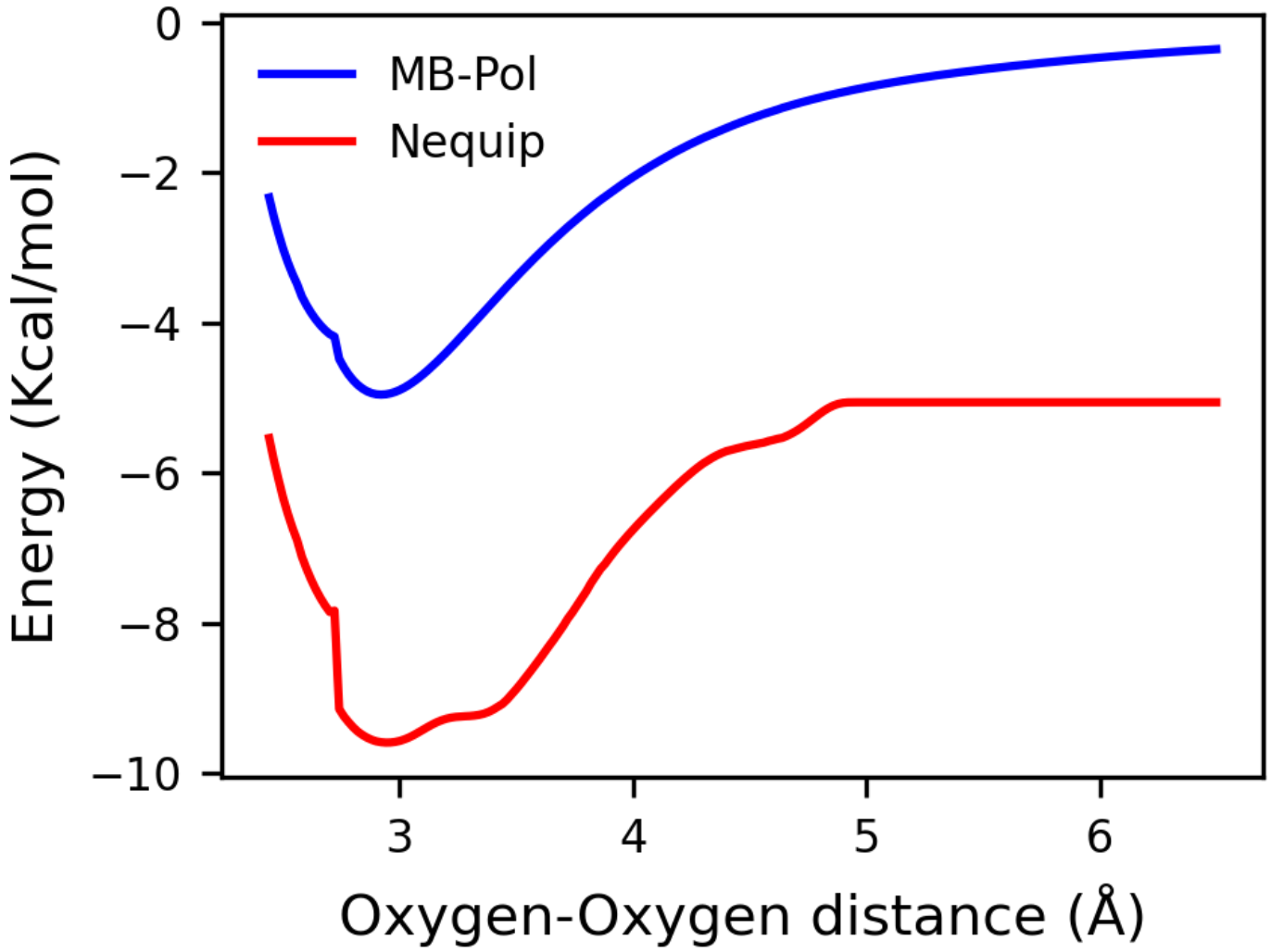
Many-body Decomposition



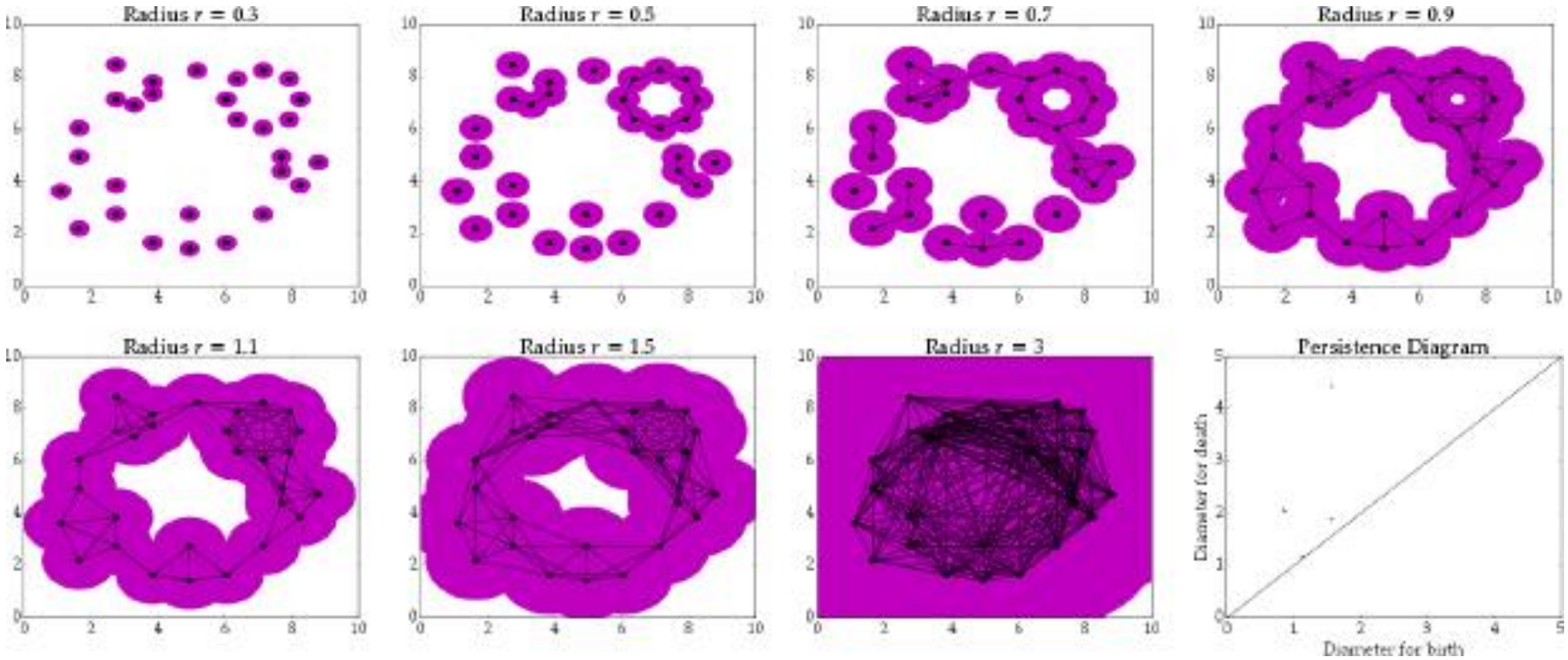

1. Prism
2. Cage
3. Book 1
4. Book 2
5. Bag
6. Cyclic chair
7. Cyclic boat 1
8. Cyclic boat 2


(a) 1-body
1B energy error (kcal/mol)
(b) 2-body
2B energy error (kcal/mol)
(c) 3-body
3B energy error (kcal/mol)
(d) 4-body
4B energy error (kcal/mol)
(e) 5-body
5B energy error (kcal/mol)
(f) 6-body
6B energy error (kcal/mol)
(g) Interaction Energy
MB-pol
DNN, seed 1
DNN, seed 2
DNN, seed 3
DNN, seed 4
Nequip
Relative energy (kcal/mol)
1. Prism
2. Cage
3. Book 1
4. Book 2
5. Bag
6. Cyclic chair
7. Cyclic boat 1
8. Cyclic boat 2

Dimer Scan


MB-Pol
Nequip
Energy (Kcal/mol)
0
−2
−4
−6
−8
−10
3
4
5
6
Oxygen-Oxygen distance (Å)

Topological Data Analysis: Permutation Equivariant



Topological Data Analysis: Ethane