Pacific Northwest
NATIONAL LABORATORY

Geometric Deep Learning for Molecule Generation

January 20, 2021

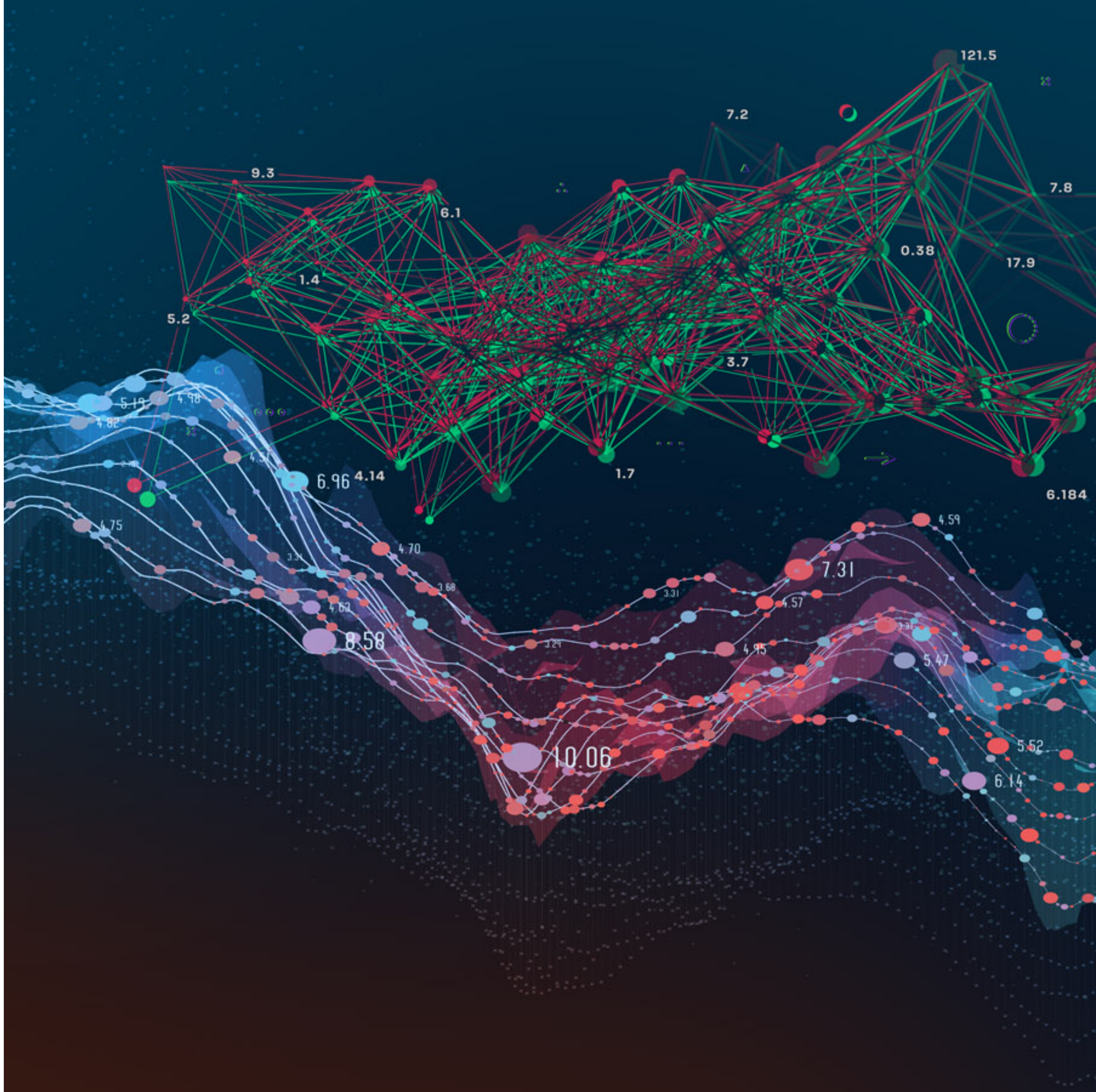
Sutanay Choudhury

Pacific Northwest National Laboratory

Joint work with Jenna Pope, Sotiris Xantheas,
Malachi Schram, Neeraj Kumar,
James Ang (Pacific Northwest National Laboratory),
Logan Ward, Ian Foster (Argonne National Laboratory)
Joseph Heindel (University of Washington)

U.S. DEPARTMENT OF ENERGY BATTELLE

PNNL is operated by Battelle for the U.S. Department of Energy

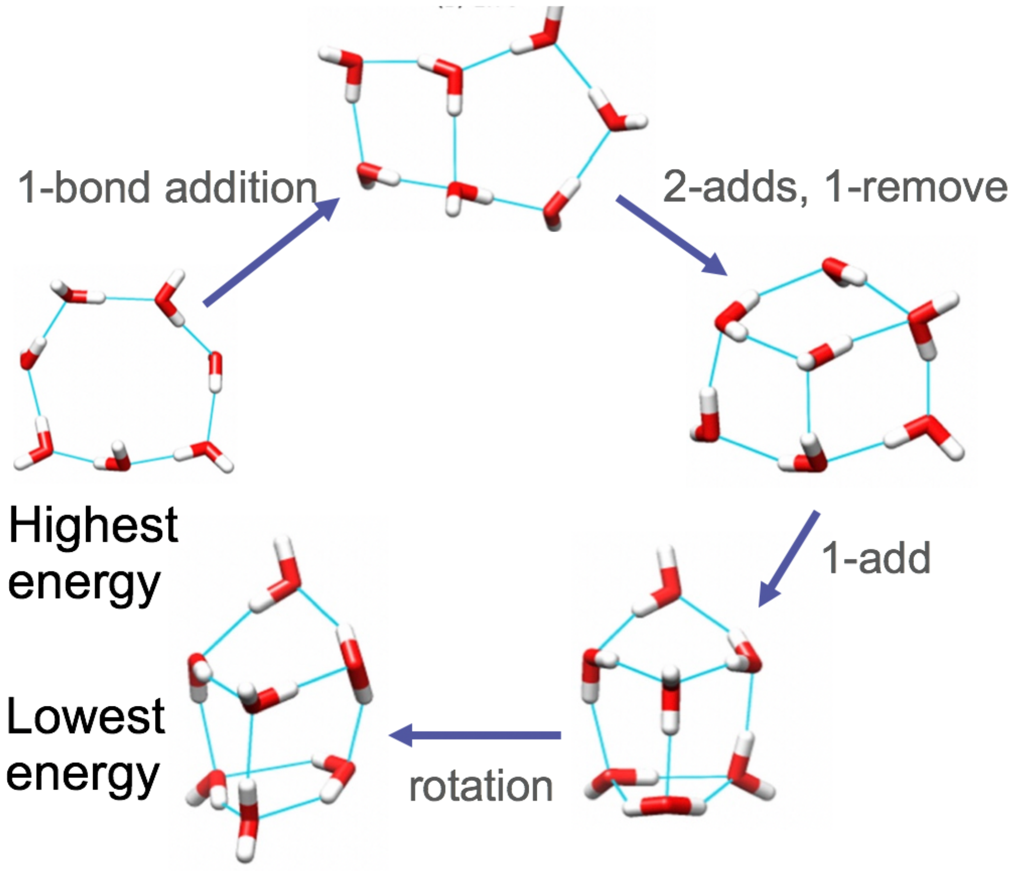


Key Takeaways

- Which ML method you want to use for designing a new molecule? What are the trade-offs between various methods?
    - More specifically, Variational Autoencoders vs. Deep Reinforcement Learning
- If the target molecule structures exhibit strong structural/geometric properties, how do we incorporate that knowledge into the ML methods?

Molecule Design problem

How do we automate the design of chemical structures that have interesting properties?



1-bond addition

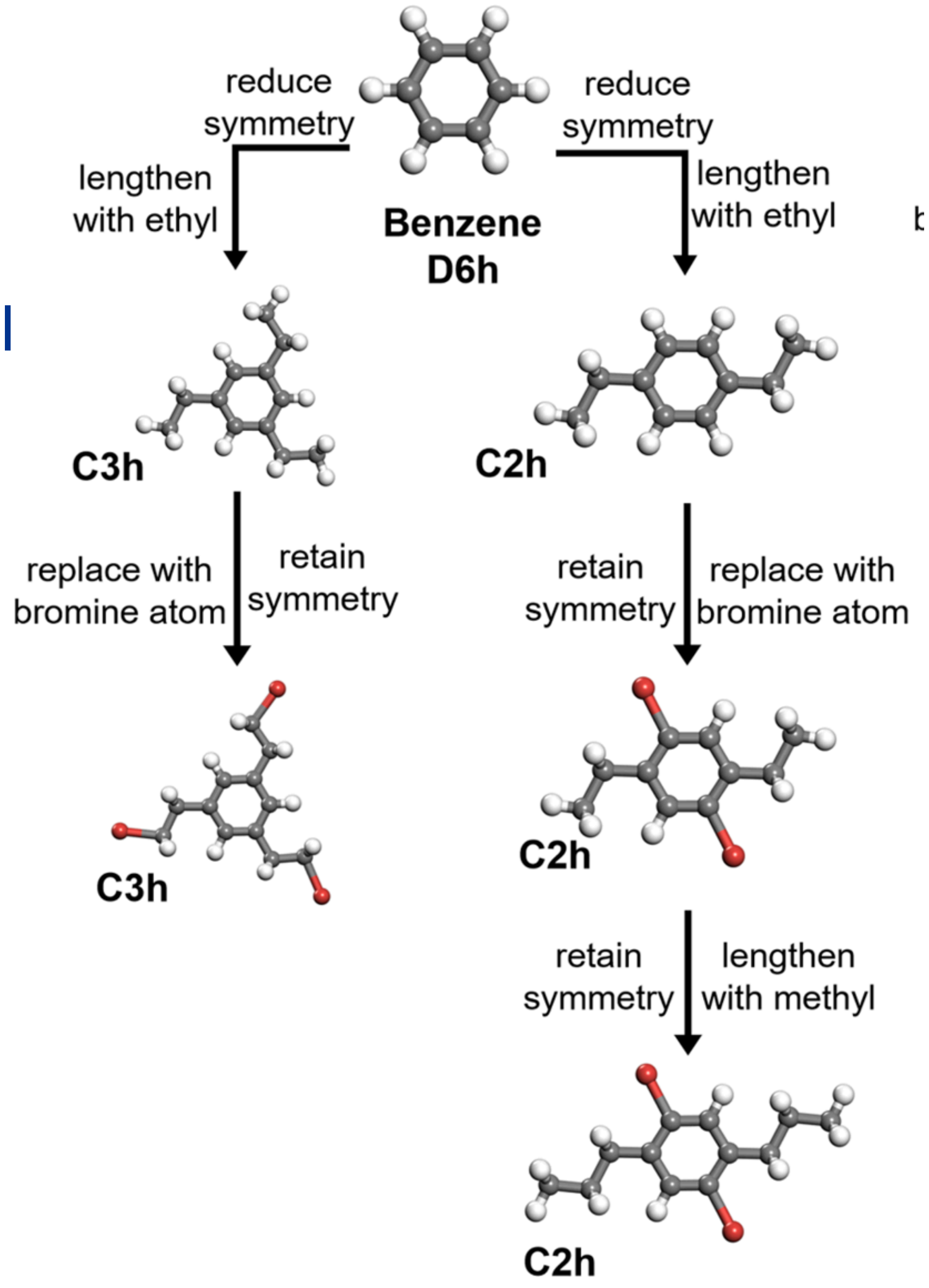
2-adds, 1-remove

1-add

rotation

Highest energy

Lowest energy



reduce symmetry

lengthen with ethyl

Benzene D6h

reduce symmetry

lengthen with ethyl

C3h

C2h

replace with bromine atom

retain symmetry

retain symmetry

replace with bromine atom

C3h

C2h

retain symmetry

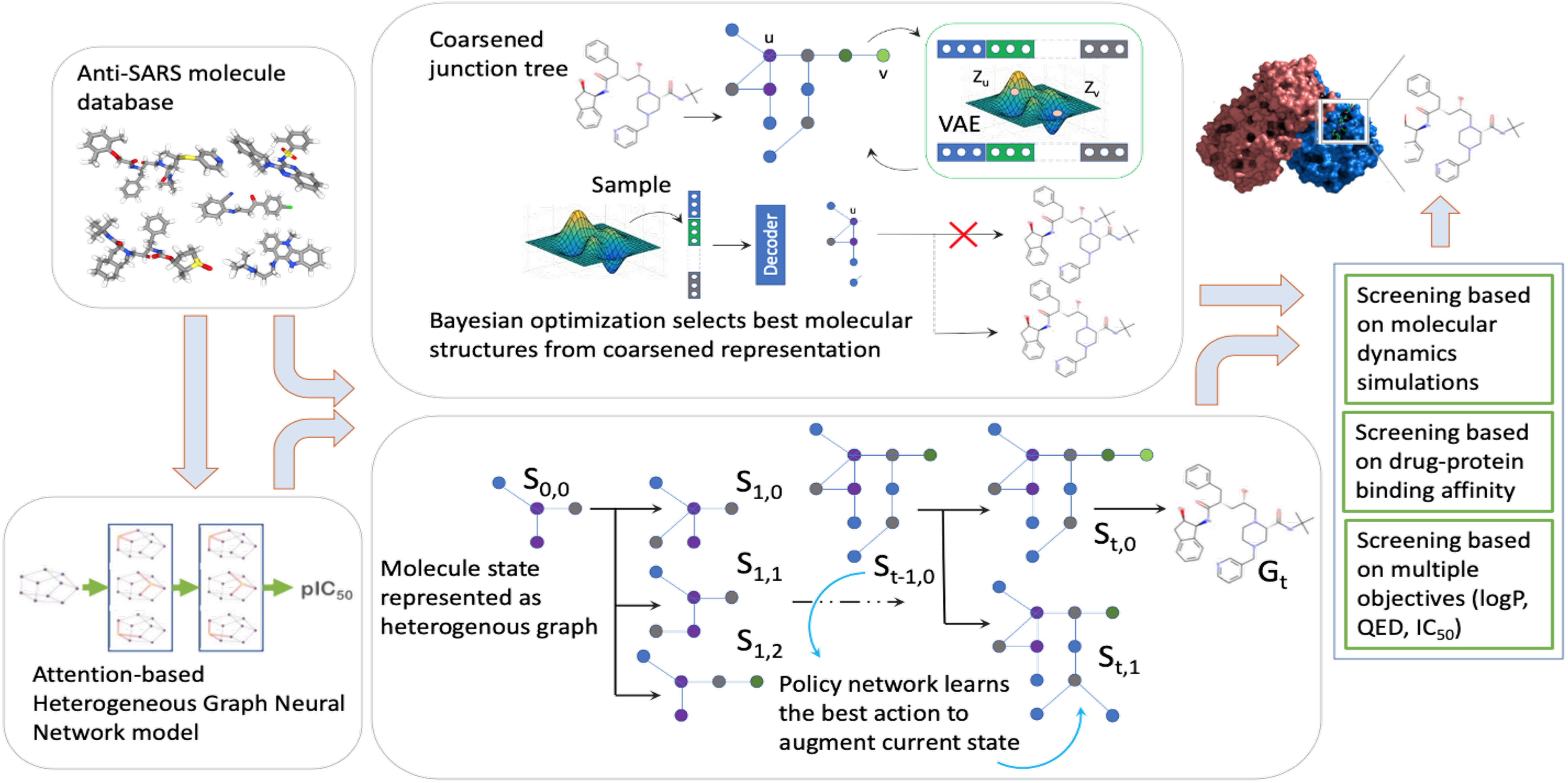
lengthen with methyl

C2h

Why is this hard?

- If we were to compose a molecule with n functional groups from a library of N functional groups, the size of the search space would be on the order of permutations (N, r):
  - N=100, r=10, search space: 6.28 * 10^19
  - N=100, r=20, search space: 1.31 * 10^39
  - N=200, r=10, search space: 8.14 * 10^22
  - N=200, r=20, search space: 3.92 * 10^45
- Our goal with machine learning is to avoid the exhaustive enumeration of the search space.

Variational Autoencoder vs Deep Reinforcement Learning



Anti-SARS molecule database

Attention-based Heterogeneous Graph Neural Network model

Coarsened junction tree

Bayesian optimization selects best molecular structures from coarsened representation

Molecule state represented as heterogenous graph

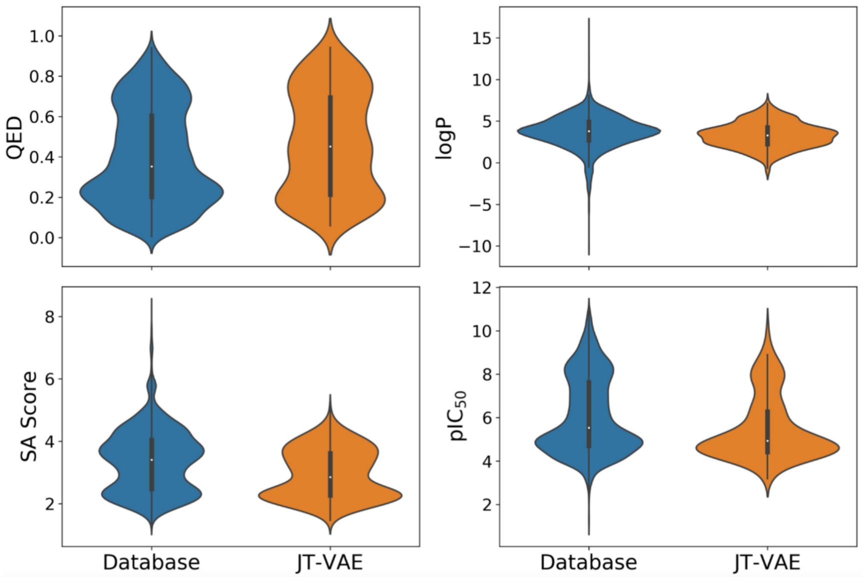
Policy network learns the best action to augment current state

Screening based on molecular dynamics simulations

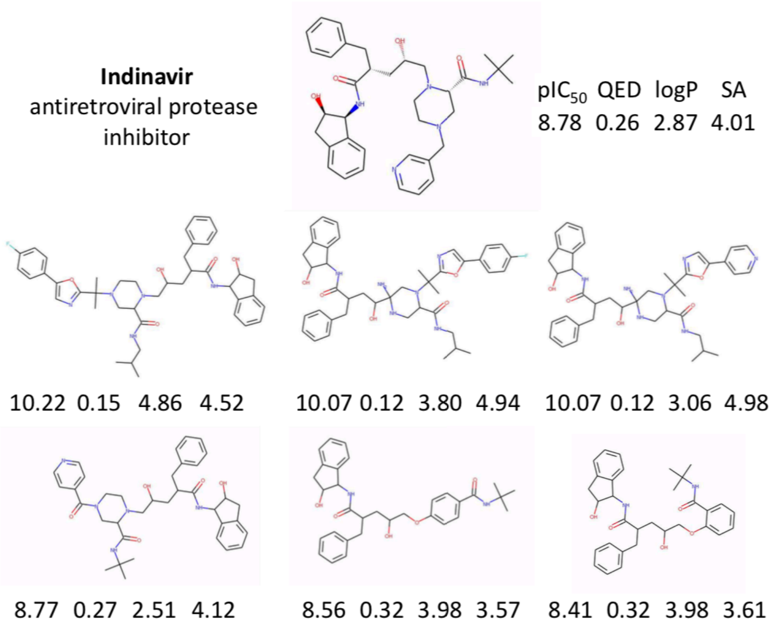
Screening based on drug-protein binding affinity

Screening based on multiple objectives (logP, QED, IC$_{50}$)

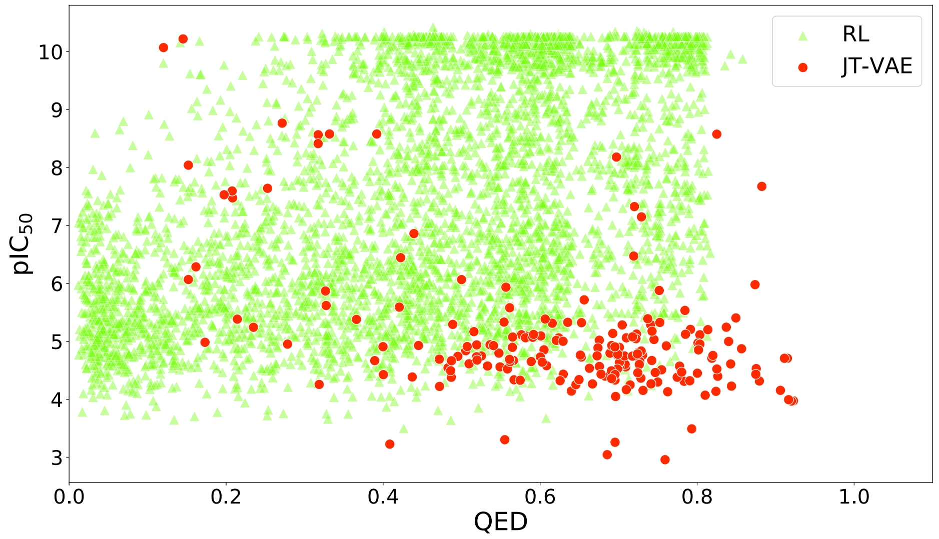
Pros and Cons



If you want a molecule that is close to ones existing in your database, use VAE.

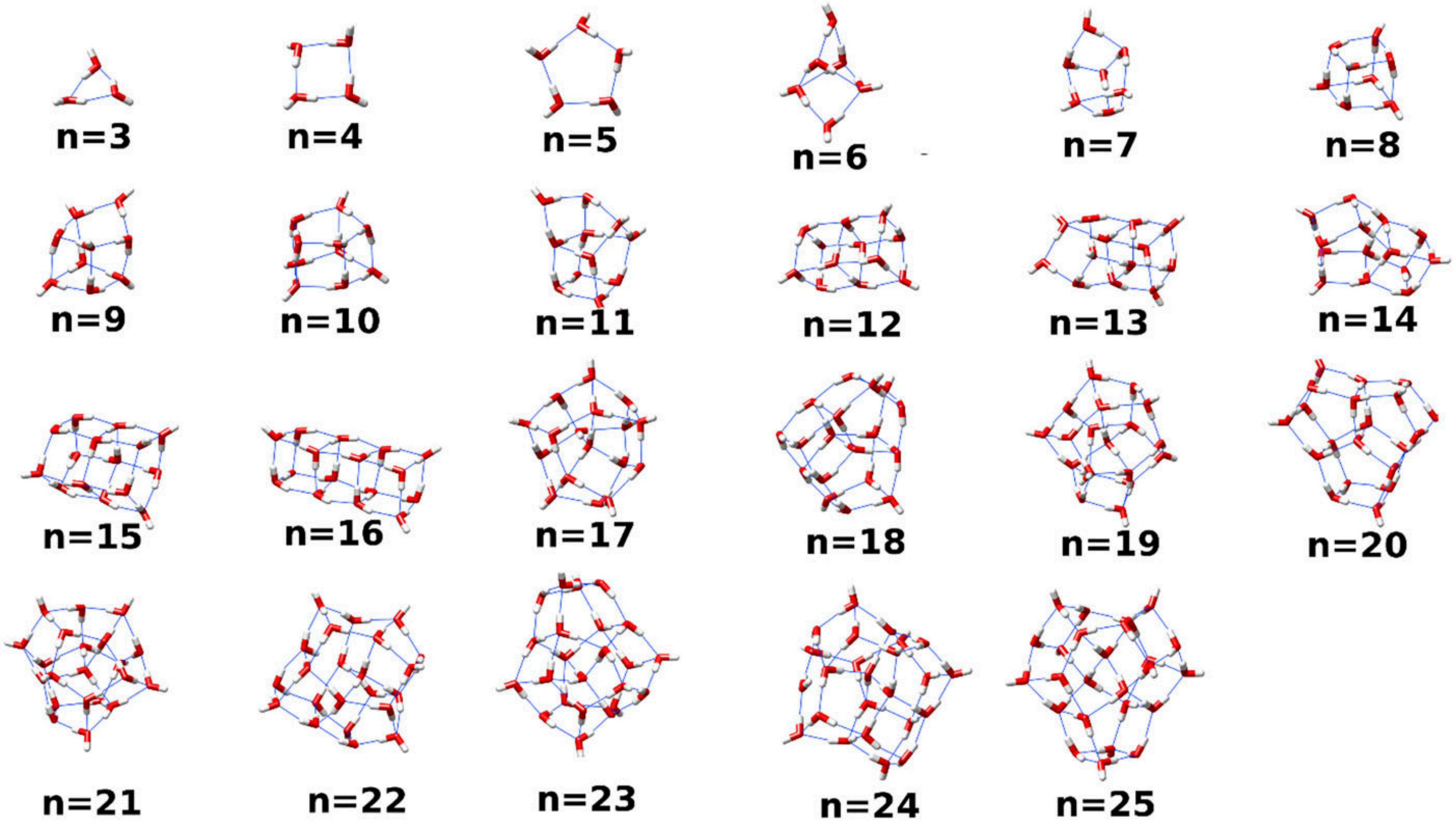


One of our top molecules (generated by JT-VAE) was a match to a widely researched COVID-19 therapeutic.



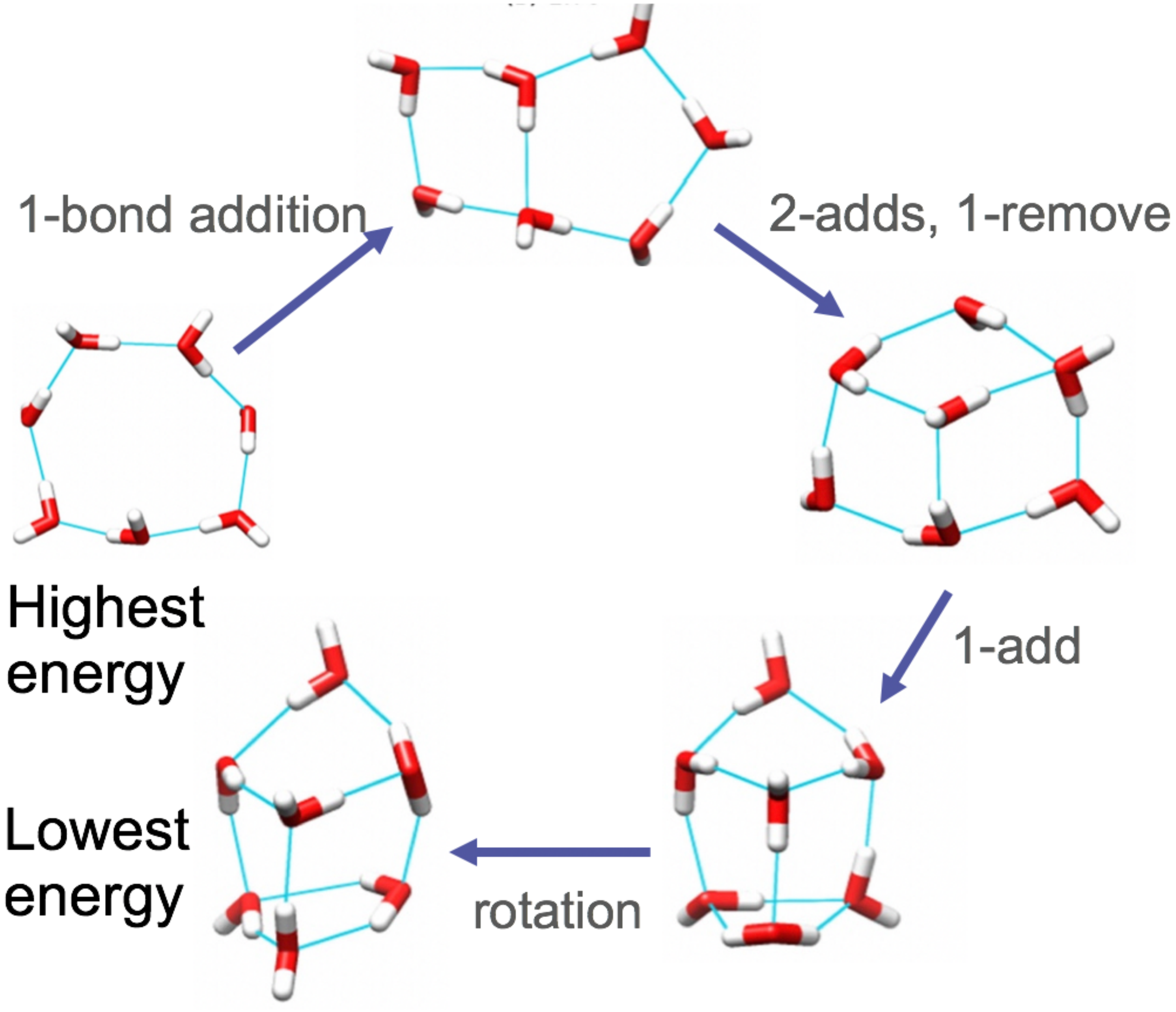
Deep RL is a promising approach to find novel candidates that we will miss if just ”searching where the light is.”

HydroNet: A ML Benchmark for Modeling Intermolecular Interactions

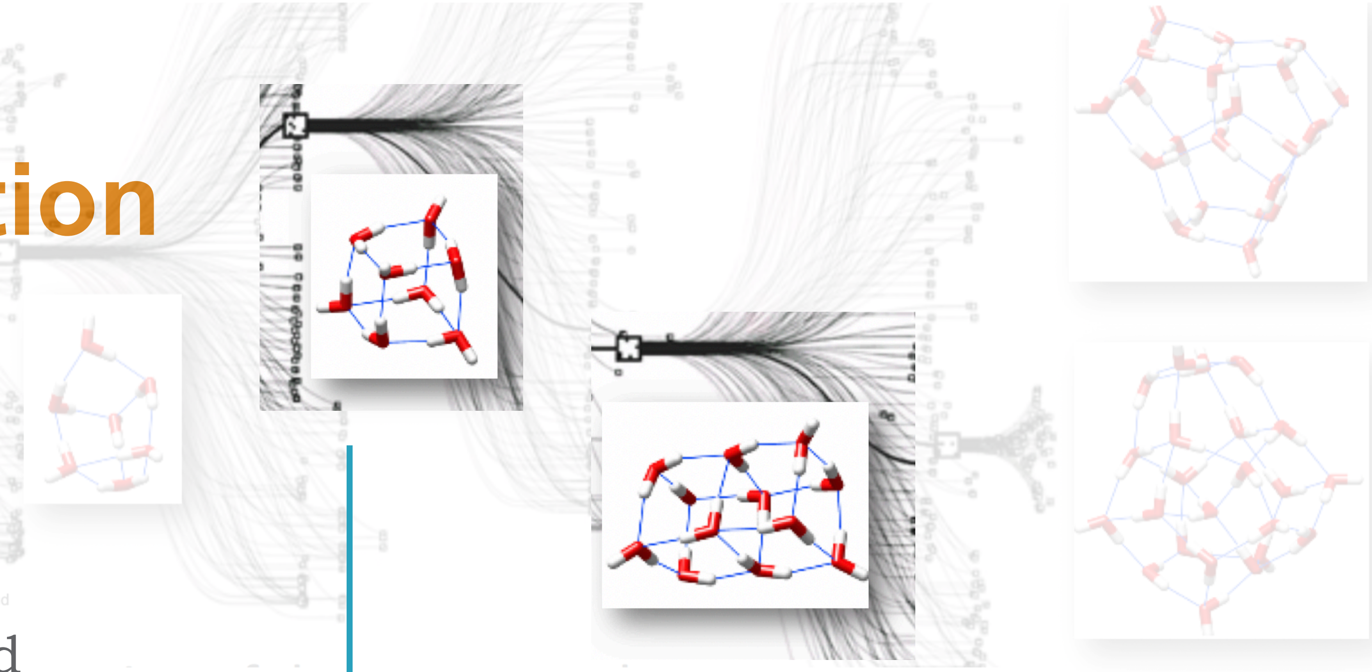


https://exalearn.github.io/hydronet [3]

Goal: Learning a model to generate low energy water clusters

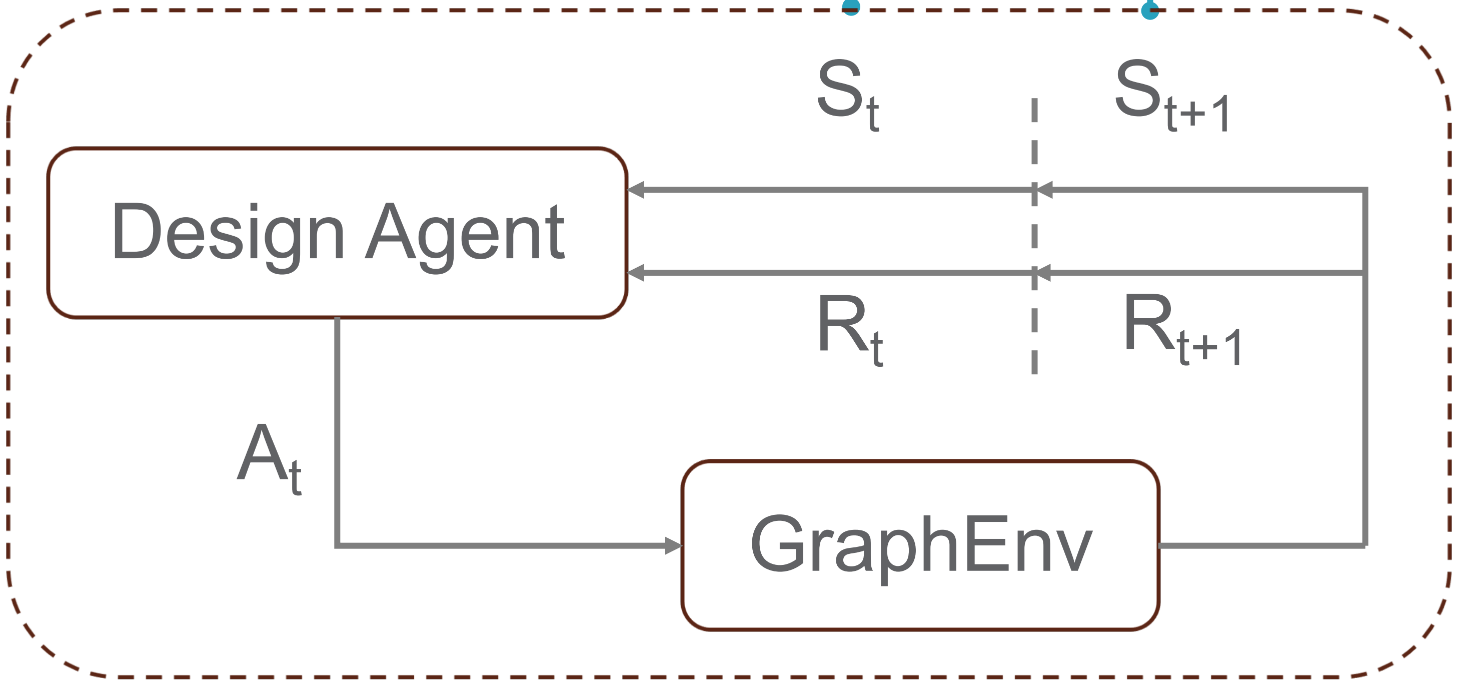


Deep RL Formulation



Combinatorial game tree picture courtesy: Google's DeepMind

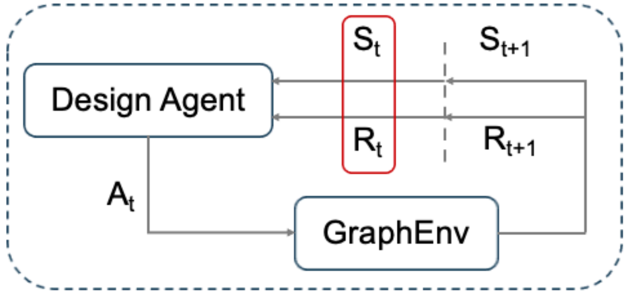
State represented via Attributed Graph and graph-theoretic chemical descriptors

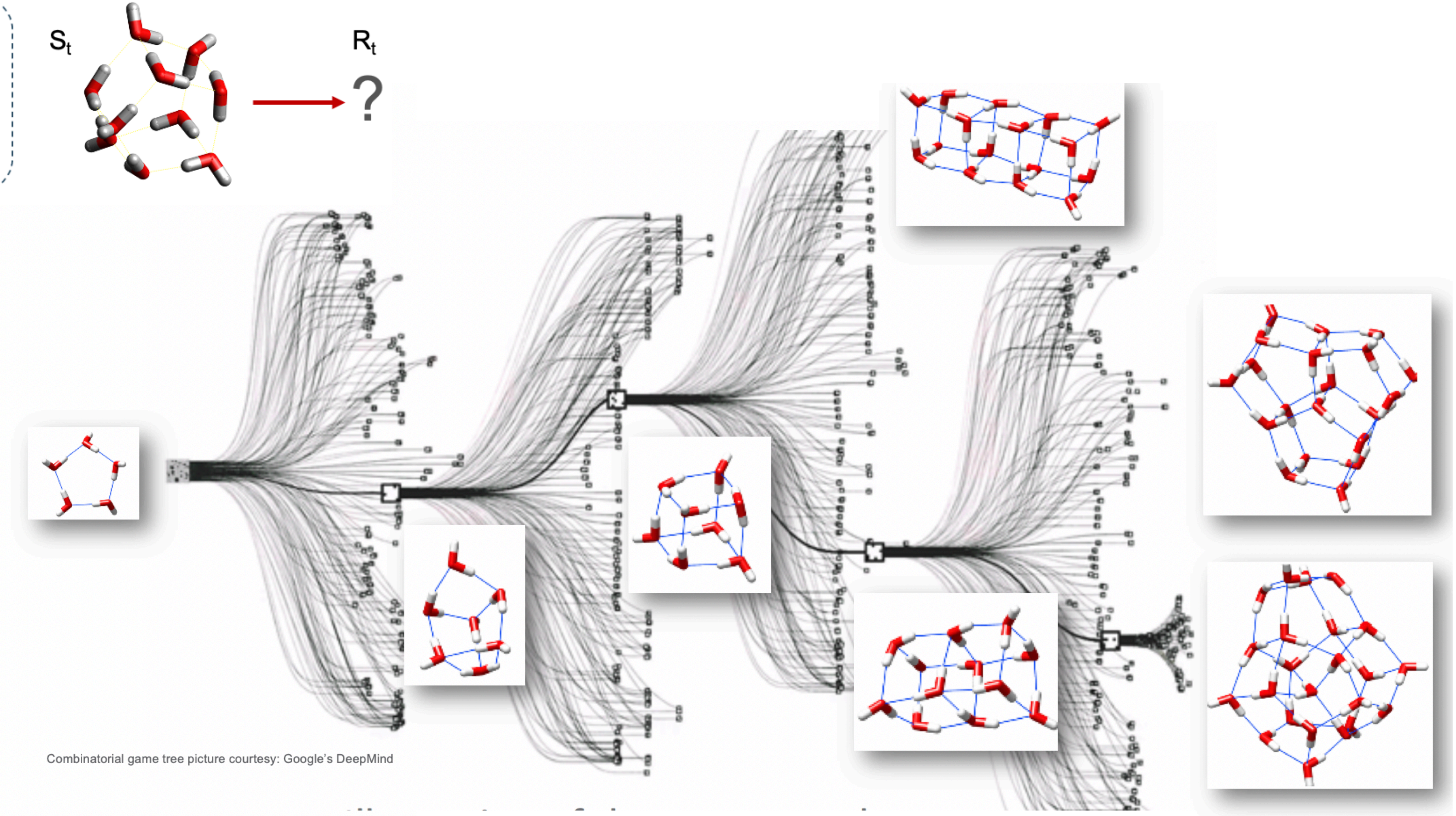


Reward estimated via chemical descriptors and/or graph neural network based surrogate models

Environment represented via Attributed Graphs

State space exploration



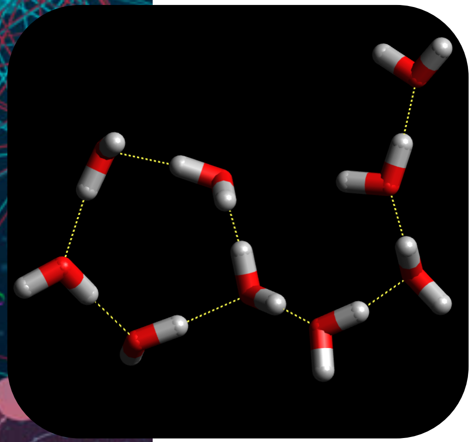


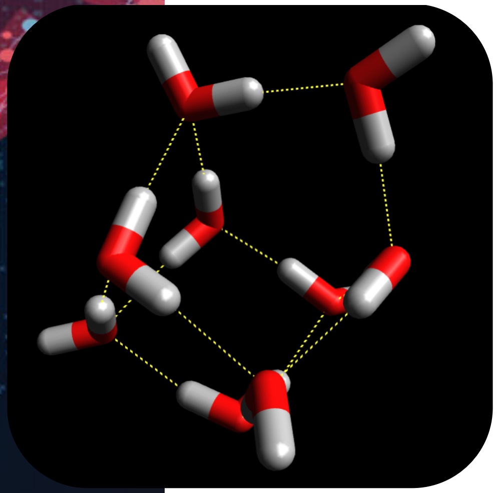
Combinatorial game tree picture courtesy: Google's DeepMind

Pacific Northwest
NATIONAL LABORATORY

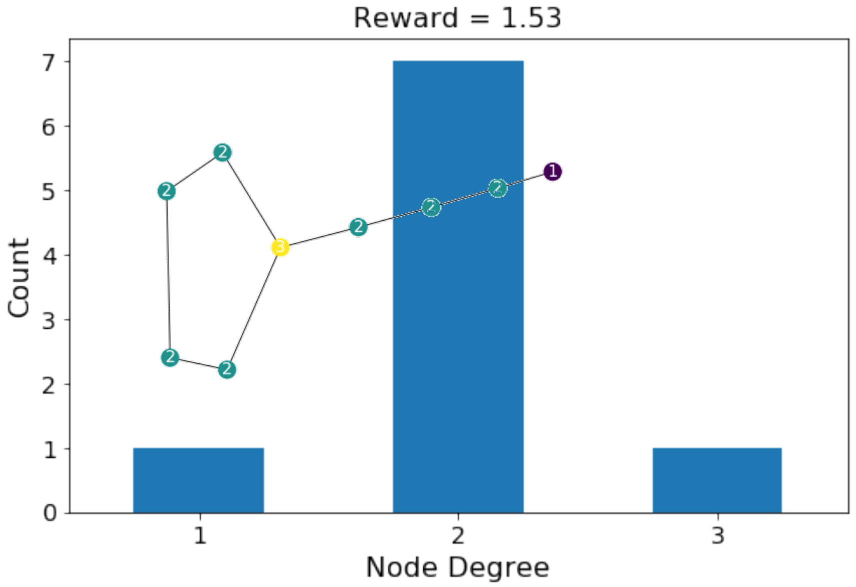
Graph-theoretic Reward Components

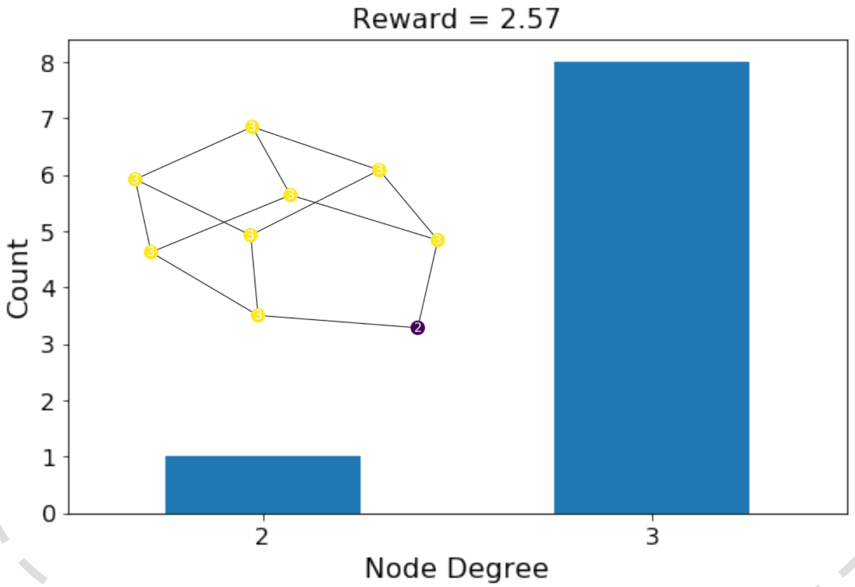
$(H_2O)_9$





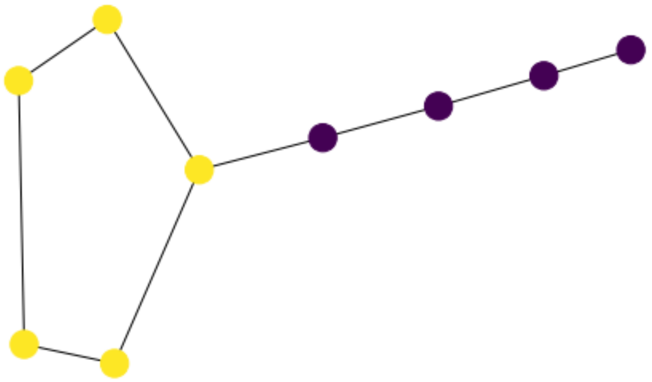
Degree



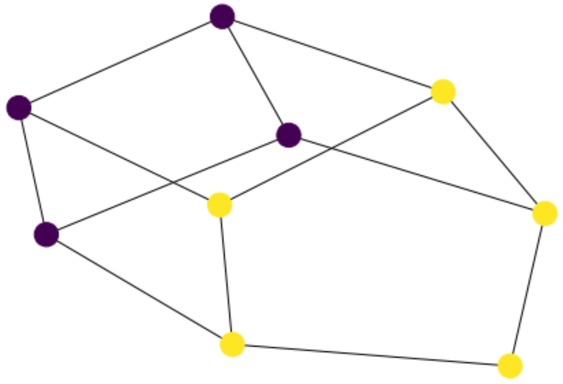


Geometric Cycles

0 Trimers 0 Tetramers
1 Pentamers 0 Hexamers

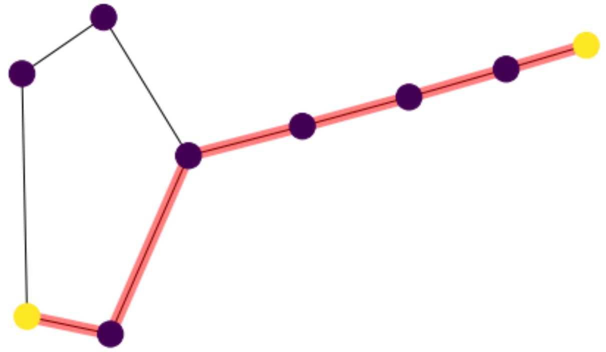


0 Trimers 4 Tetramers
2 Pentamers 3 Hexamers

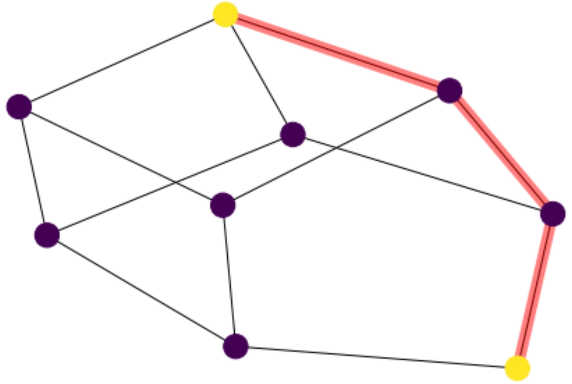


Path Measures

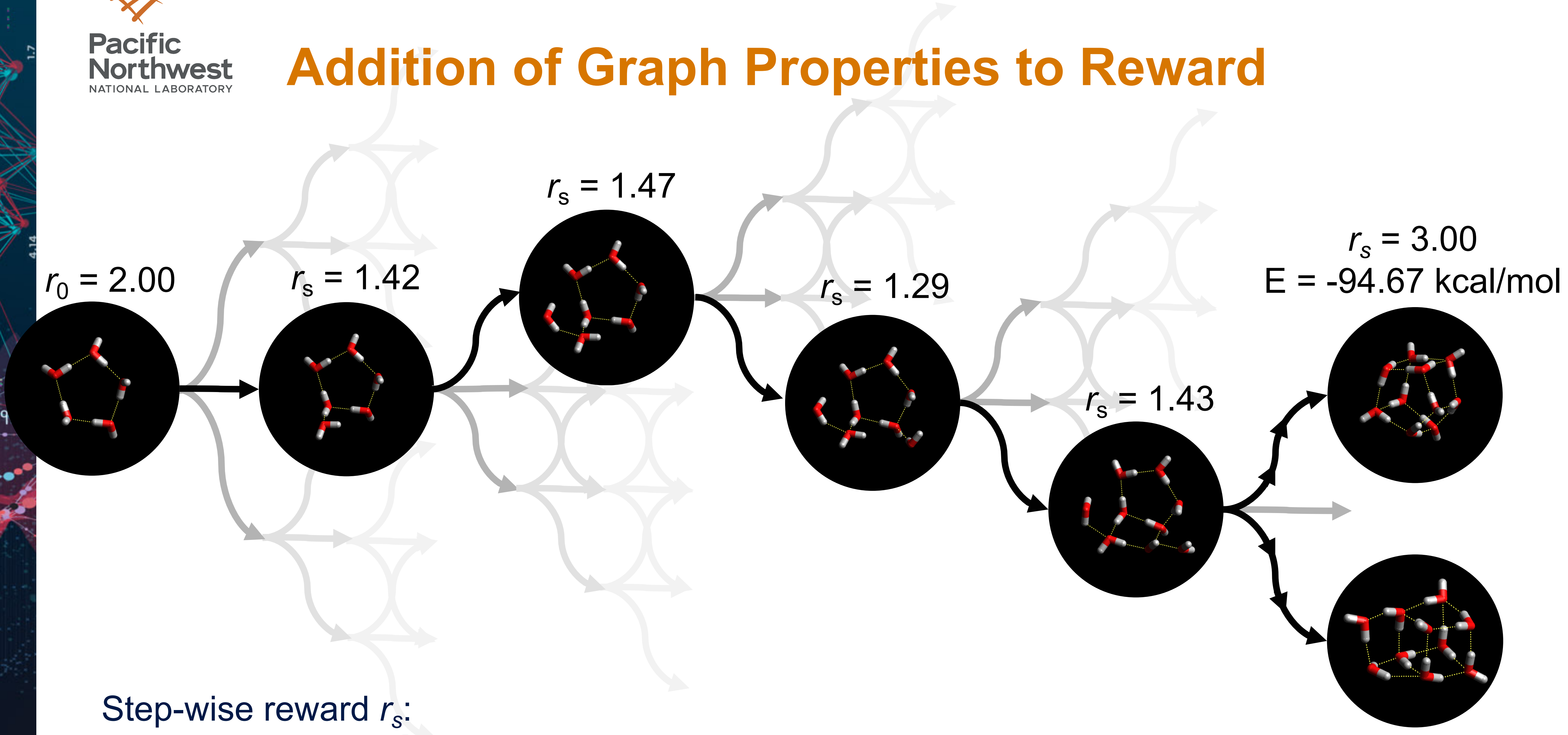
Diameter = 6
Avg Shortest Path Length = 2.75



Diameter = 3
Avg Shortest Path Length = 1.89



Addition of Graph Properties to Reward



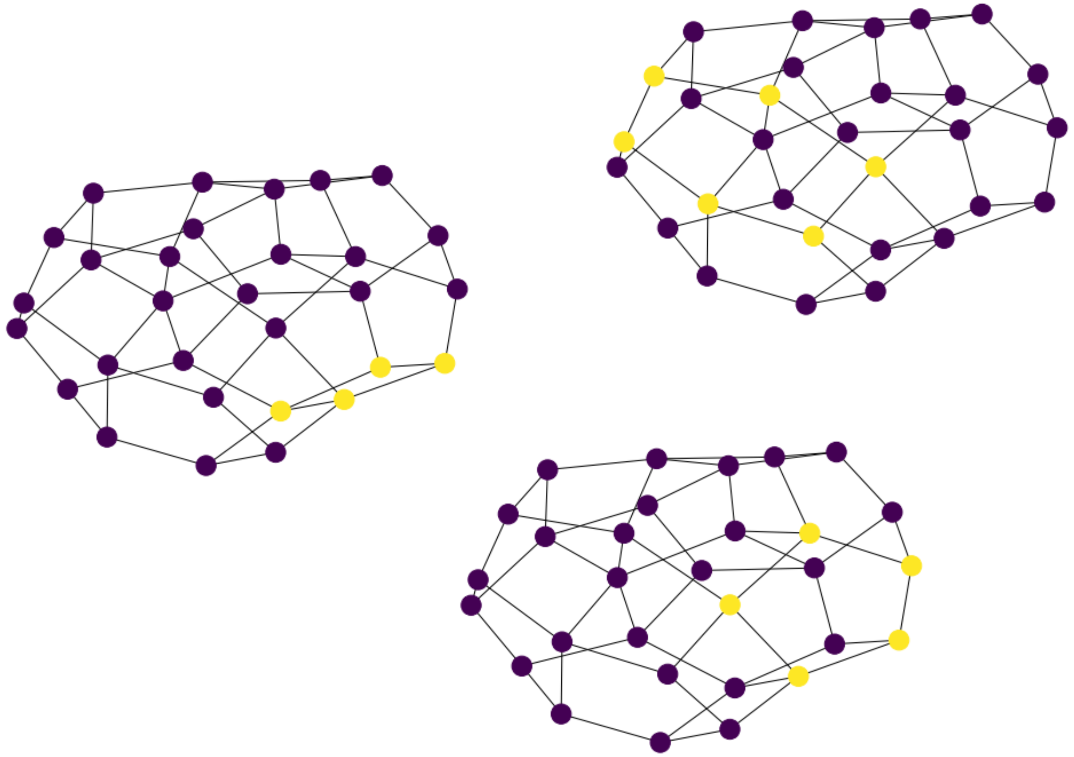
Step-wise reward r_s:

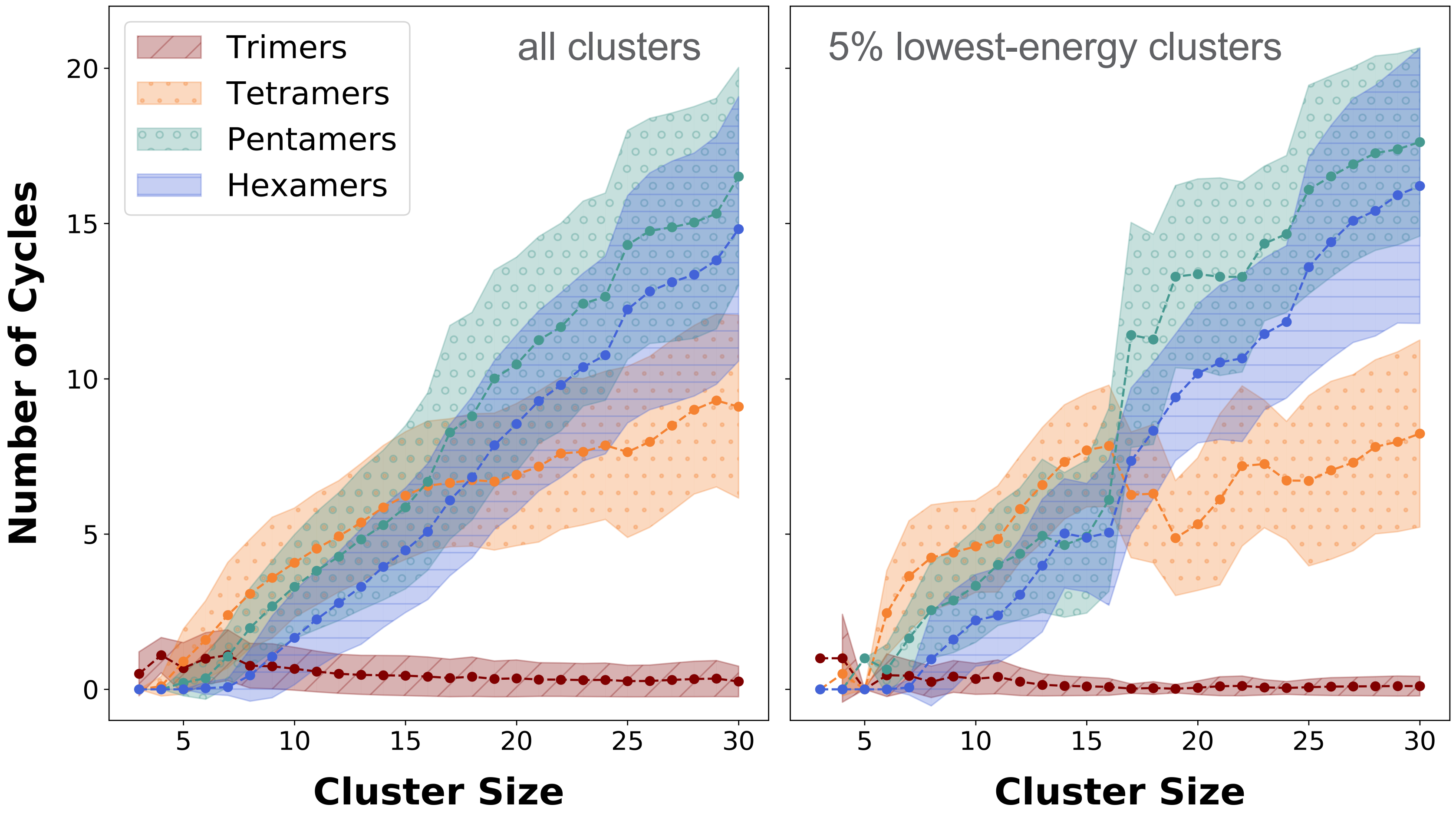
$$r_s = E(D_s) - \sqrt{var(D_s)}$$

where D_s is the degree distribution at step s

Structural measure preserving Molecule Generation

- The distribution of structural motifs evolve with scale

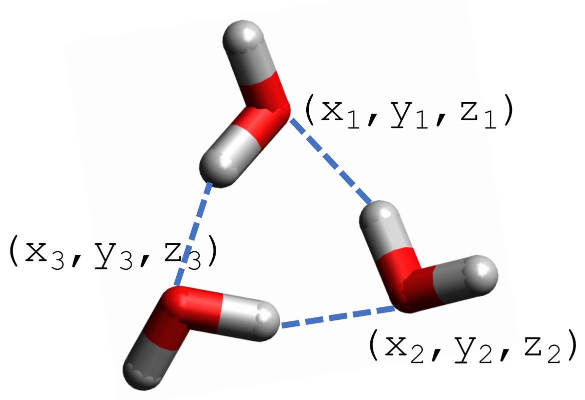


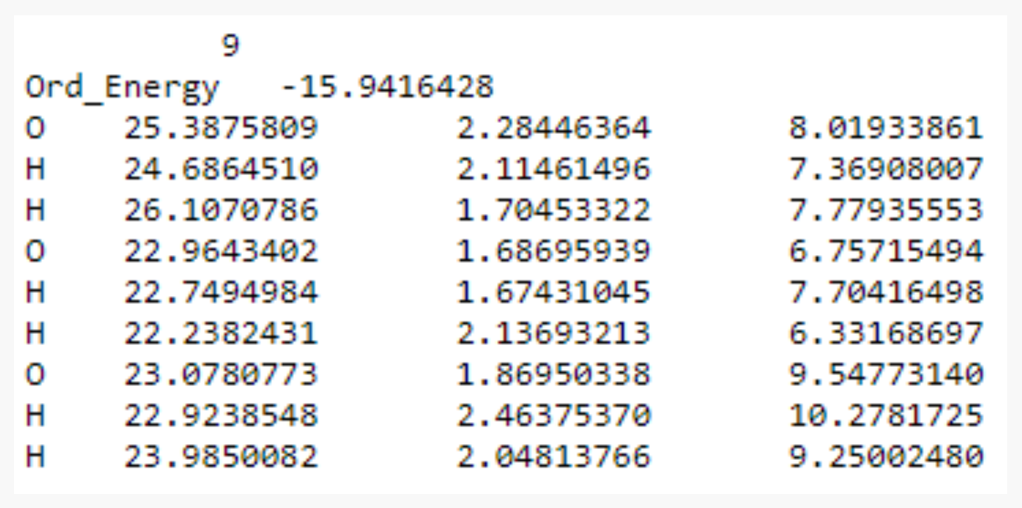


HydroNet: Multi-representation Benchmark

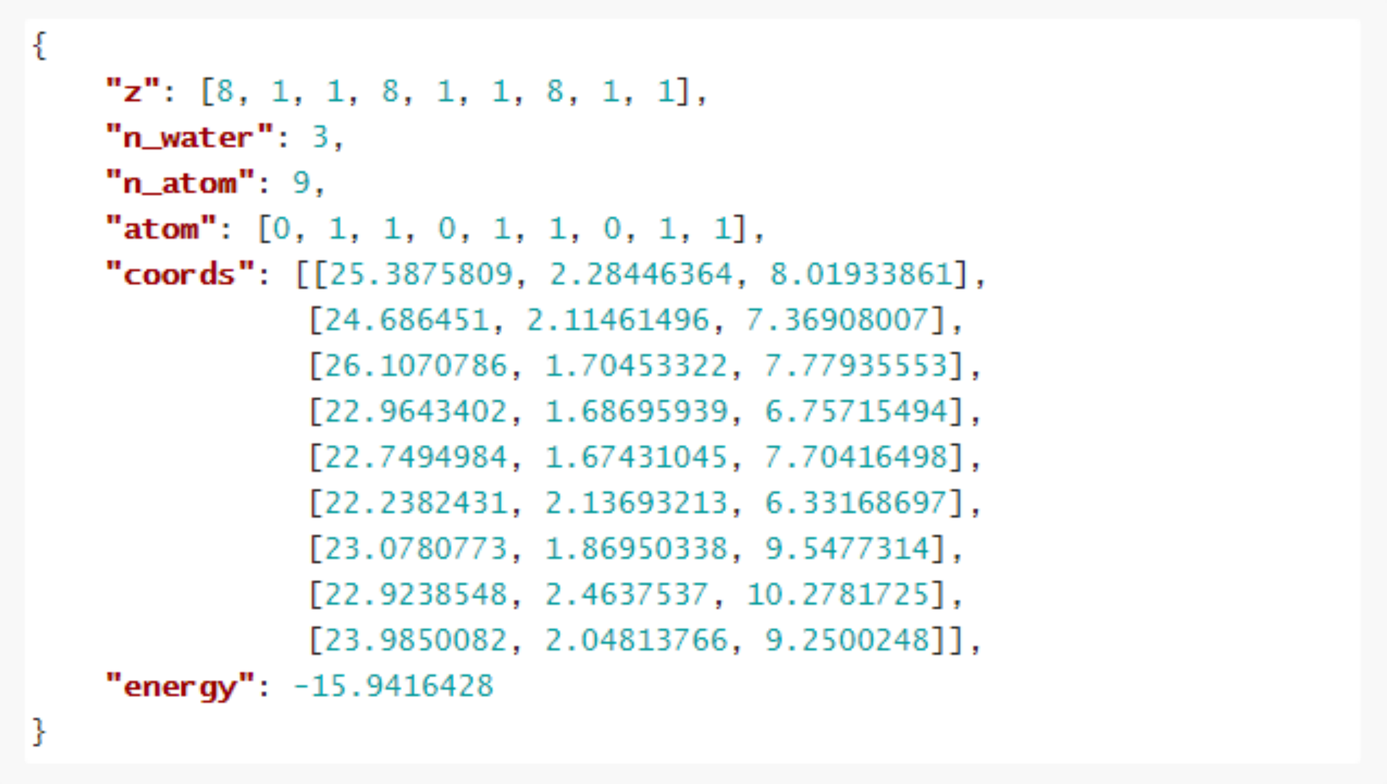
- Provide data with graph and coordinates information
- Provide pre-trained model for energy prediction using both graph neural networks and coordinate based convolutional neural network[1]
- Provide tools for validating quality of generated molecules

Geometry: coordinates





```
             9
Ord_Energy    -15.9416428
O     25.3875809         2.28446364         8.01933861
H     24.6864510         2.11461496         7.36908007
H     26.1070786         1.70453322         7.77935553
O     22.9643402         1.68695939         6.75715494
H     22.7494984         1.67431045         7.70416498
H     22.2382431         2.13693213         6.33168697
O     23.0780773         1.86950338         9.54773140
H     22.9238548         2.46375370        10.2781725
H     23.9850082         2.04813766         9.25002480
```



```
{
    "z": [8, 1, 1, 8, 1, 1, 8, 1, 1],
    "n_water": 3,
    "n_atom": 9,
    "atom": [0, 1, 1, 0, 1, 1, 0, 1, 1],
    "coords": [[25.3875809, 2.28446364, 8.01933861],
               [24.686451, 2.11461496, 7.36908007],
               [26.1070786, 1.70453322, 7.77935553],
               [22.9643402, 1.68695939, 6.75715494],
               [22.7494984, 1.67431045, 7.70416498],
               [22.2382431, 2.13693213, 6.33168697],
               [23.0780773, 1.86950338, 9.5477314],
               [22.9238548, 2.4637537, 10.2781725],
               [23.9850082, 2.04813766, 9.2500248]],
    "energy": -15.9416428
}
```

References

1. Bilbrey J.A., J. Heindel, M. Schram, P. Bandyopadhyay, S.S. Xantheas, and S. Choudhury. 2020. "A Look Inside the Black Box: Using graph-theoretical descriptors to interpret a Continuous-Filter Convolutional Neural Network (CF-CNN) trained on the global and local minimum energy structures of neutral water clusters." Journal of Chemical Physics 153, no. 2:024302.
2. Choudhury S., L. Ward, J.A. Bilbrey, M. Schram, S.S. Xantheas, J. Heindel, and M. Schwarting, et al. 02/05/2020. "ExaLearn-Design: RL-driven Computational Design at Exascale." Exascale Computing Project Annual Meeting, Houston, Texas.
3. Choudhury, S., Bilbrey, J.A., Ward, L., Xantheas, S.S., Foster, I., Heindel, J.P., Blaiszik, B. and Schwarting, M.E., 2020. HydroNet: Benchmark Tasks for Preserving Intermolecular Interactions and Structural Motifs in Predictive and Generative Models for Molecular Data. *NeurIPS Workshop on Physical Sciences*.

Pacific Northwest
NATIONAL LABORATORY

Thank you

ECP
EXASCALE COMPUTING PROJECT