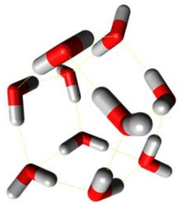
A Look Inside the Black Box: Using Graph-Theoretical Descriptors for the Post-Hoc Interpretation of Neural Networks

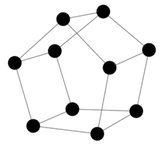
Jenna A. Bilbrey, Joseph P. Heindel, Malachi Schram, Pradipta Bandyopadhyay, Sotiris S. Xantheas, Sutanay Choudhury*

Database Analysis

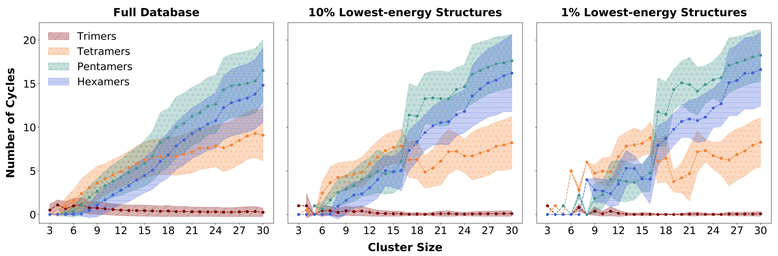
- Newly published database of over 5 million neutral water cluster $(H_2O)_N$ minima of sizes N=3–30 within 5 kcal/mol from the putative minimum for each cluster size (doi: 10.1063/1.5128378)
- Potential energies computed using the flexible, polarizable Thole-Type Model (TTM2.1-F, version 2.1) interaction potential for water
- Structural sampling performed using the Monte Carlo Temperature Basin Paving sampling method



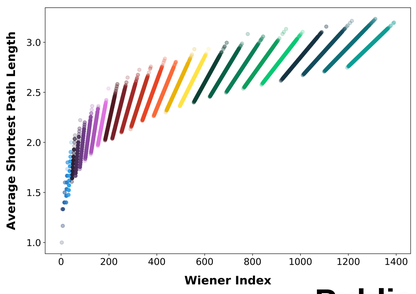
- Structures were transformed into graphs
  - ← All-atoms graph: each vertex is an atom and each edge is a covalent or hydrogen bond
  - → Projected graph: each vertex is a water molecule and each edge is a hydrogen bond



- Descriptors derived from graph theory were computed for the graphs



Cycle counts for all clusters in the database (left) and the 10% (middle) and 1% (right) lowest-energy clusters. The means are given as circles, and the shaded areas show the standard deviation.

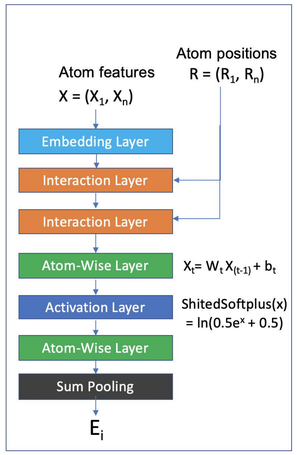


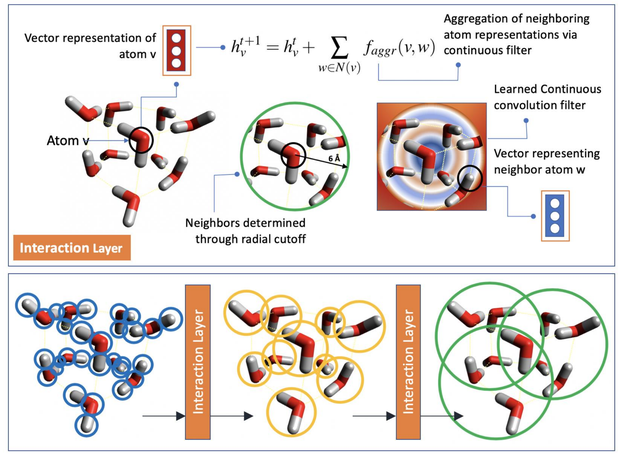
Wiener index vs the average shortest path length for all clusters in the database. Each point is colored by cluster size. The slope of each grouping is $^2/_{N(N-1)}$, i.e., the inverse of the number of pairs in the system.

Publication

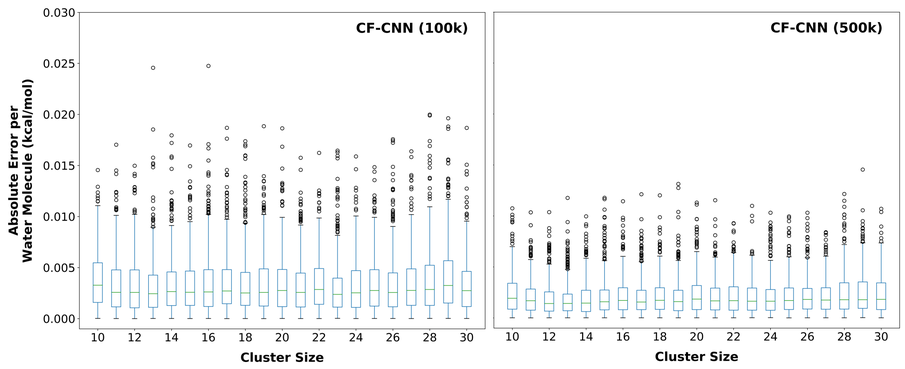
A look inside the black box: Using graph-theoretical descriptors to interpret a Continuous-Filter Convolutional Neural Network (CF-CNN) trained on the global and local minimum energy structures of neutral water clusters. *J. Chem. Phys*. **153**, 024302 (2020). https://doi.org/10.1063/5.0009933

Continuous-filter Convolutional Neural Network (CF-CNN)





- Filters learned during training provide a unique internal representation incorporating symmetry and ordering invariances
- Convolutional layers learn a representation of the pair-wise interactions between atoms to predict the contribution of each atom to the desired property
- Because each atom is represented by its neighborhood, the CF-CNN can accurately predict the potential energy of structures both smaller and larger than those in the training set



Box and whisker plots showing the absolute error per water molecule (kcal/mol) on the test set for CF-CNNs trained on 100,000 (left) and 500,000 (right) water clusters. Boxes extend from the lower to upper quartiles with a line at the median value, bars extend to 1.5 times the interquartile range, and outliers are plotted individually. Both plots share a *y*-axis for comparison.

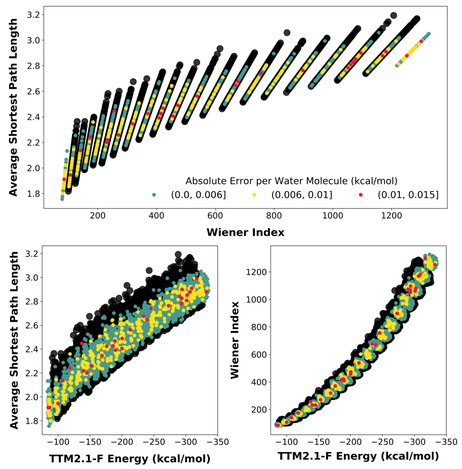
Data and Code Repositories

Data: Cartesian coordinates, all-atoms graphs, projected graphs, and energies available at:
https://sites.uw.edu/wdbase/

Code: Graph generation, graph-based analysis, and SchNetPack version available at:
https://github.com/exalearn/molecular-graph-descriptors

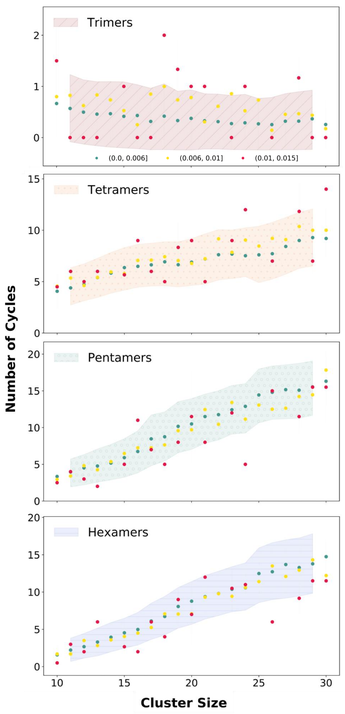
Post-hoc Analysis

- Graph descriptors developed for database analysis are used to provide a post-hoc interpretation of our trained CF-CNN



Wiener index vs average shortest path length (top), Wiener index vs energy (bottom left), and average shortest path length vs energy (bottom right) of projected graphs in the training and test sets. Clusters in the training set are shown in black, while those in the test set are colored according to the absolute error per water molecule (kcal/mol) in the CF-CNN prediction.

- The full range of configuration space is contained in the training set
- Clusters deviating from the training set mean were less well learned than structures with values close to the mean
- The structural space covered by the training set influences the CF-CNN predictions and should be considered alongside the chemical space when developing training sets



Mean cycle counts in the training and test sets. Shaded regions show one standard deviation from the mean for the training set. Test set clusters are shown as points colored according to the absolute error per water molecule (kcal/mol) in the CF-CNN prediction.

Test structures far from the training set mean were associated with larger errors, indicating clusters deviating from the training set mean – though within the configuration space learned by the CF-CNN – were less well learned than structures with values close to the mean.

U.S. DEPARTMENT OF ENERGY BATTELLE

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

For additional information, contact:
Sutanay Choudhury | (509) 375-3978 | sutanay.choudhury@pnnl.gov

9/10/2020 PNNL-SA-156026

Pacific Northwest NATIONAL LABORATORY

www.pnnl.gov