HydroNet: Benchmark Tasks for Preserving Intermolecular Interactions and Structural Motifs in Predictive and Generative Models for Molecular Data

Sutanay Choudhury, Jenna A. Bilbrey, Logan Ward, Sotiris S. Xantheas, Ian Foster, Joseph Heindel, Ben Blaiszik and Marcus Schwarting

I. INTRODUCTION

- Intermolecular and long-range interactions are central to phenomena as diverse as gene regulation, topological states of quantum materials, electrolyte transport in batteries, and the universal solvation properties of water.
- We present a set of challenge problems for preserving intermolecular interactions and structural motifs in machine-learning approaches to chemical problems, through the use of a recently published dataset of 4.95 million water clusters held together by hydrogen bonding interactions and resulting in longer range structural patterns.
- The dataset provides spatial coordinates as well as two types of graph representations, to accommodate a variety of machine-learning practices.

II. DATASET

- The dataset of 4.95 million water cluster minima is the largest collection of water cluster minima reported to date

https://exalearn.github.io/hydronet/

III. GRAPH-THEORETIC DESCRIPTORS FOR THE DATASET

- Stability of chemical systems is known to depend on the connectivity of water molecules, as well as other geometrical/atomic/coarse bonding properties of the water clusters. These properties are suitable graph-theoretical descriptors.

IV. TASK-I: POTENTIAL ENERGY PREDICTION

This task is to predict the potential energy of a water cluster without the use of expensive ab initio methods, in two settings. In the geometry-to-energy setting, we assume the geometry of the structure is known accurately. In the graph-to-energy setting, we require predictions to be made from the connectivity of the water molecules only.

\[ m_l^{+1} = \sum w(x) M_i(b_{l_i}a_{l_i}a_{r_i}) \]

\[ b_{l_i}^{+1} = b_{l_i} + m_l^{+1} \]

\[ \alpha_{r_i}^{+1} = \alpha_{r_i} + b_{l_i}^{+1} + M_i(b_{l_i}a_{l_i}a_{r_i}) \]

V. INTEGRATING GRAPH AND SPATIAL INFORMATION

Both tasks require awareness of geometric and structural information, especially in the context of minimum energy water clusters. The performance of graph neural networks is limited when they do not account for geometric information.

REFERENCES