An Unsupervised Deep Learning Approach for Feature Extraction from Molecular Dynamics Simulation of Lipid Membranes

Piyush Karande[1], Helgi I. Ingólfsson[2], Timothy S. Carpenter[2], Harsh Bhatia[3], Peer-Timo Bremer[3], Felice C. Lightstone[2], Brian C. Van Essen[3]

[1]Computational Engineering Division; [2]Physical and Life Sciences; [3]Center for Applied Scientific Computing. Lawrence Livermore National Laboratory

Introduction

- Mutations in RAS protein lead to cancer initiation and growth.
- RAS interacts with, and affects domain formation in lipid cell membranes.
- Domain formation detected in large-scale molecular dynamics (MD) simulations using hand-engineered features and clustering algorithms.
- We present an unsupervised deep learning approach for automatic feature extraction.

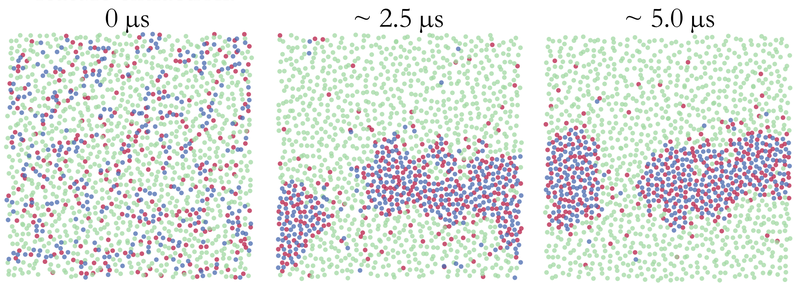


Figure 1. Progression of a 3-component coarse-grain (CG) simulation that results in formation of a domain. CHOL● DIPC● DPPC●

Molecular Structure

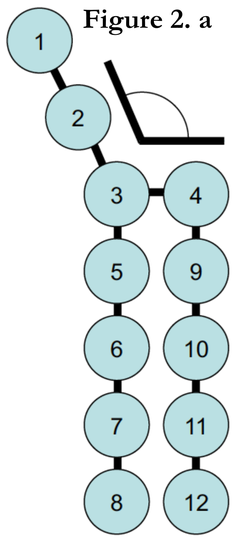
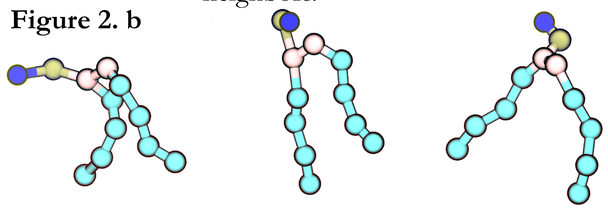


Figure 2. a

- Martini CG simulations model molecules with "beads"; 8-12 in these simulations.
- Typically, a bead represents four heavy atoms with associated hydrogens.
- Figure 2. a shows each bead bonded to 1-3 other beads.
- Structure defined by bond lengths and angles.
- Parameters have a certain range of motion allowing molecules to assume different configurations (2. b).
- Structure encoded using hand-engineered features: positions, height, lipid tilt, and order parameter of lipid tail.
- Interactions and neighborhoods encoded with lipid area, and number/characteristics of neighbors.

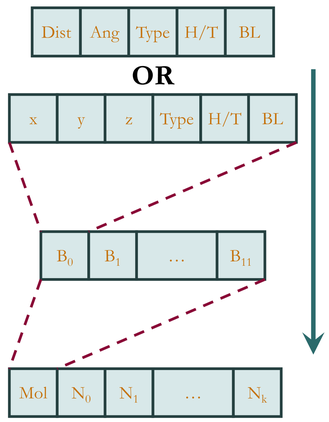
Figure 2. b



Input Data Format

- Data formatted to include both structural and neighborhood information.
- Structural data:
  - relative radial(2)/cartesian(3) coordinates
  - the type of molecule(3)
  - head or tail bead(2)
  - bond lengths with other beads(12)
- Radial coordinates make data rotationally invariant.
- Neighborhood: structural data of k closest neighbors appended to the molecule.

Figure 3.



Autoencoder

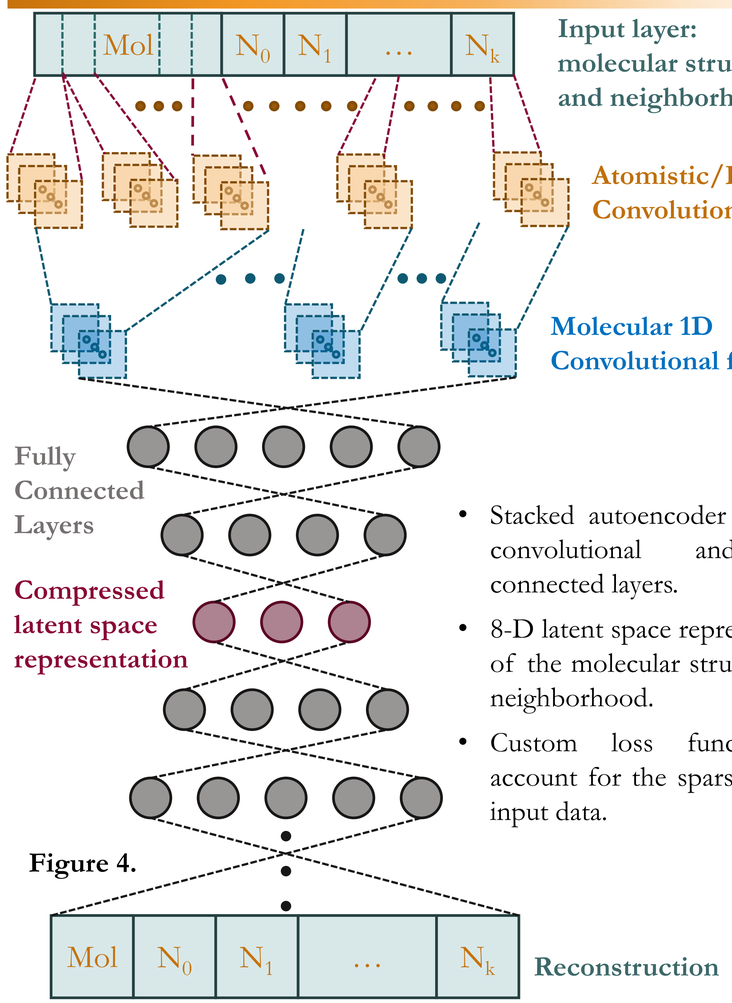


Figure 4.

- Stacked autoencoder with 1D convolutional and fully connected layers.
- 8-D latent space representation of the molecular structure and neighborhood.
- Custom loss function to account for the sparsity in the input data.

HDBSCAN Clustering

- Figure 5. a shows a typical frame with single domain (periodic boundaries).
- Learned features produced robust and physically-meaningful clustering stable to the changes in input parameters of HDBSCAN.
- Hand-engineered features required careful design and weighting to produce desirable results.

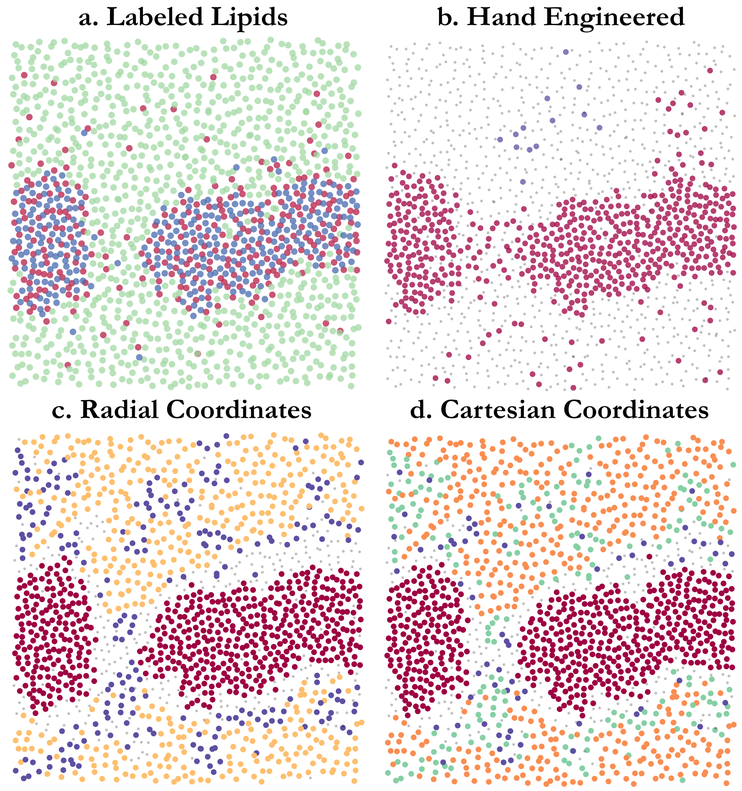


Figure 5. HDBSCAN clustering on simulation frame in a. using different feature vectors (b-d). Colors represent lipids in a. and clusters in b-d. Smaller black dots represent lipids labelled as noise.

Future Directions

- Use CANcer Distributed Learning Environment (CANDLE) for hyperparameter and network architecture optimization.
- Improve the model to train from multiple simulations and create a informative representation of molecules from unseen simulations.
- Train on complex simulations with several different molecule types.
- Use the learned compressed representation to quantify the state of a simulation and create a predictive model to reduce simulation time.

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. LLNL-POST-751701