Professorship of Multiscale Modeling of Fluid Materials
TUM School of Engineering and Design
Technical University of Munich

Encoding Chemistry: Molecular Embeddings for ML

Project Description

Machine learning models for molecular and materials systems rely heavily on how structures are represented. Different embedding strategies-from traditional fingerprints to graph-based message passing embeddings and continuous 3D representations-offer distinct advantages and limitations. Molecular representations are essential because they translate complex chemical structures into numerical forms that models can learn from, capturing key information about atoms, bonding, and geometry. They are widely used in property prediction, molecular design, reaction modeling, and materials screening, where the quality of the representation strongly influences accuracy, data efficiency, and generalization. This project aims to systematically investigate, compare, and develop molecular representation methods for downstream property prediction tasks.

Objectives

1. **Survey Existing Molecular Embedding Techniques**

   Identify and summarize the main ways molecules and materials can be represented numerically. This includes classical fingerprints, graph-based models, and 3D geometric descriptors.

2. **Benchmark Performance for Direct Property Prediction**

   Test different representations by using them as inputs to machine learning models and evaluating how well they predict molecular or material properties (e.g., energy, stability, reactivity).

3. **Develop or Improve a New Embedding Method**

   Design or refine a representation that addresses limitations found in existing methods, such as better capturing geometry, long-range interactions, or chemical environments.

Application Process

If interested, email m.sanocki@tum.de / linying.zhang@tum.de with:

1. A brief introduction (background, interests, and motivation).
2. Your transcript of records.