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

Generative Machine Learning for 3D Coarse-Grained Bio-Aggregate Structure Modeling

1 Project Description

The self-assembly of peptides into nanostructures (e.g., fibers, tubes, vesicles) plays a vital role in biomaterials and synthetic biology. Generating these 3D aggregate structures directly from peptide sequence information (e.g., FASTA format) is a key challenge that could accelerate peptide design. While machine learning has shown success in protein folding and molecular modeling, there is limited understanding of how best to represent aggregate morphologies in a way that enables generative models to learn and synthesize them effectively.

This project focuses on the output representation problem: **what is the most effective way to represent a self-assembled aggregate structure (e.g. composed of hundreds of peptides) in 3D so that ML models can learn to generate such configurations from sequences**? To solve this, the student will survey existing generative ML approaches applied to 3D structure generation in soft-matter systems and molecular modeling — especially focusing on how these models encode and generate the 3D output structure. Then, using available datasets (e.g., synthetic data, polymer or lipid assemblies, and $\sim$ 100 peptide aggregate trajectories), they will implement and benchmark multiple 3D representation strategies such as voxel grids, point clouds and implicit neural fields (e.g. occupancy or SDF). Each method will be compared in terms of learnability and model performance (reconstruction fidelity, generation plausibility), data requirements, and suitability for peptide self-assembly use cases. The final outcome will be a critical evaluation of how generative models could best represent and produce aggregate morphologies, setting the foundation for future conditional generation from peptide sequences.

2 Objectives

1. **Literature Review:** Review existing ML-based 3D generative modeling efforts in soft matter, materials science, and molecular biology. Focus on how the 3D output structure is represented and how successful these representations are for learning spatial configurations. The input will always be a structured sequence (e.g. FASTA), but the project centers on the output: the aggregate.

2. **Representation Benchmarking:** Implement baseline generative models (e.g. diffusion models) using multiple 3D representations (voxels, point clouds, implicit fields, graphs). Evaluate performance on synthetic data and soft-matter aggregate datasets.

3. **Dataset Strategy and Aggregation:** Define best practices for dataset construction from coarse-grained MD trajectories (e.g., time-averaged final configurations vs dynamic windows). Use synthetic examples and real-world soft-matter systems for training.

4. **Extrapolation to Peptide Self-Assembly:** Based on the benchmarking, evaluate which output representations are most promising for future conditional generation from peptide sequences. If time permits and existing data is enough, test a synthetic end-to-end task: from simple peptide descriptors or one-hot FASTA inputs to generated aggregates.

3 Prerequisites

If you do not fill all requirements but are still very interested in the topic, please feel free to apply — we value motivation very highly!

- Proficiency in Python and experience with machine learning with PyTorch
- Familiarity or basic understanding of 3D data types (e.g. voxel grids, point clouds, meshes)

- Basic understanding of generative modeling (e.g. autoencoders, GANs, or diffusion models)
- Interest in biomolecular simulation (beneficial but not required)

4 Contact

If interested, please email Nuno Costa (Chair of Multiscale Modeling of Fluid Materials) at nuno.costa@tum.de with:

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