genPI: Generative Polymer Informatics for Structured Polymer Discovery

Halil Ibrahim Erdogan[1], Prof. Dr. Christopher Künneth[1]

[1]Computational Materials Informatics, Faculty of Engineering Science, University of Bayreuth, 95447 Bayreuth, Germany

Abstract Text :

The design of novel polymer materials remains a challenging and time-intensive task in chemistry, often relying on expert knowledge and costly experimental procedures. Recent advances in machine learning provide new opportunities to accelerate this process by enabling the automated generation of candidate structures. However, existing approaches often struggle to ensure chemical validity while also accounting for practical synthesizability, and are often limited by the quality and structure of available training data.

In this project, we investigate machine learning methods for the generation of polymer structures under chemically informed constraints. Our approach focuses on sequence-based representations of polymers, allowing the model to learn patterns from existing data and generate new candidates within a structured chemical space. Essential chemical rules, such as valency and bonding consistency, are incorporated to maintain validity, while selected constraints may be relaxed to enable exploration of novel structures.

A central aspect of this work is the development of improved datasets and benchmarks that better reflect realistic chemical design tasks while remaining suitable for machine learning models. By carefully balancing complexity and learnability, we aim to provide data that supports both robust model training and meaningful evaluation. In addition, we examine how dataset design influences model behavior and generalization.

Another key challenge addressed in this work is the estimation of synthesizability without performing laboratory experiments. To tackle this, we explore computational proxies and learned evaluation strategies that approximate whether a generated polymer can realistically be synthesized.

Overall, this work aims to establish a coherent framework that combines data design, polymer generation, constraint-aware modeling, and evaluation into a unified pipeline. By aligning machine learning methods more closely with chemical knowledge and practical constraints, we aim to reduce the gap between computer science and chemistry, ultimately supporting the discovery of novel and synthesizable polymer materials.