Generative inverse design of polymers in extremely low-data regimes

Sid Kumar[1,†]

[1]Delft Technical University, the Netherlands

† corresponding author's email: Sid.Kumar@tudelft.nl

Machine learning (ML) for molecule design typically requires either datasets comprising millions of samples or smaller datasets supplemented by large-scale pretraining followed by fine-tuning or transfer learning. In practice, acquiring such volumes of data through experiments or simulations is infeasible. We address a fundamental challenge: how can polymer molecules be reliably designed with target physical properties in extreme low-data regimes, e.g., fewer than 100 data points?

We introduce the Matrix-Grammar Chemistry-Informed Generation (MCG) framework, a zero-shot, non-parametric approach for inverse molecular design that requires neither pretraining nor fine-tuning. MCG represents molecular graphs as matrices and identifies a sequence (grammar) of matrix operations (production rules) to generate structurally related molecules. The central premise is that matrix-based grammars provide an inherently data- and compute-efficient representation for generative design of molecules. MCG explicitly incorporates domain-specific chemical rules to ensure chemical validity, class specificity, and improved synthesizability by construction. MCG enables conditional inverse design of polymers with targeted properties, including acrylates, self-healing vitrimers, high-temperature polyimides, and covalent organic frameworks, using as few as 100 data points and without reliance on pretraining or large external datasets. Selected MCG-designed polymers are currently being pursued for synthesis, with preliminary assessments by chemists indicating experimental feasibility.