Efficient exploration of the chemical space through the composition of property-specific SMILES language models

S. Nakata, Kobe University

Today, we have access to billions of virtual compounds that can be synthesized on demand. However, exhaustive exploration of this chemical space is not feasible. To efficiently explore the vast chemical space, we have developed a novel approach to generating diverse drug-like compounds while optimizing for desired target activities. Our approach composes prior and property-specific SMILES language models to control the compound generation process. The prior model is trained on a large on-demand compound library to cover a wide range of synthetically accessible chemical space. The property-specific models are fine-tuned on small datasets with experimentally determined activity values to capture the structure-activity relationships. Each token in the SMILES string is sampled using the weighted sum of the log probabilities.

In this way, we can control the generation process to efficiently explore the chemical space. We investigate the effectiveness of our approach by generating compounds with selective activity profiles against a predefined set of kinases. Evaluation of the generated molecules using activity prediction models, molecular docking, and ligand similarity analysis shows that our approach is beneficial for biasing the generation process toward desired target activities while preserving the synthesizability and structural diversity of the prior.