Abstract

論 文 の 内 容 の 要 旨

論文題目　Molecular Structure Generation
Using Small Chemical Dataset
(小規模の化学データセットを利用した分子構造生成)

氏　名　井上　貴央

The structure generator, which generates molecular structures by computer, can propose structures satisfying the desired properties by using statistical models and datasets of the properties of interest. Therefore, structure generators are expected to streamline the process of compound design. However, due to experimental costs and other reasons, the available datasets are often small, and in such situations, the statistical model may overfit and fail to generate structures with good performance.

In this thesis, we develop a method that can generate structures with good performance when a statistical model-based structure generator is trained on a small chemical dataset consisting of about 1000 compounds with the desired properties. We proposed two methods to avoid overfitting: one is to use DAECS, a structure generator based on a statistical model without deep learning, and the other is to use data augmentation in a graph-based deep learning model to enhance the effect of transfer learning.

The structure generator DAECS tended to have low diversity of the generated structures. For this reason, we added structural modification rules that reduce the need for multiple applications of structural modification rules to a single application, and designed a seed structure selection algorithm so that the seed structures that undergo structural changes become diverse.

Through the case study, it was confirmed that these methods could actually diversify the generated structures and generate new structures that are not included in the training dataset.

We also designed a method for JT-VAE, a deep structure generator, to enhance the effect of transfer learning, which uses a large molecular structure dataset along with a small chemical dataset to prevent overfitting. Specifically, we designed a data augmentation that perturbates the feature vectors for some vertices with standard normal random numbers during message passing in GNN. It was confirmed that this data augmentation enhanced the prediction performance of the QSPR model trained on a small chemical dataset and the structure generation performance of JT-VAE trained by transfer learning.

Although there remain some challenges related to the design of the structure generator and the scalability of the proposed method to training with a much smaller number of samples, it is expected to contribute to the efficiency of compound design in situations where only small-scale chemical datasets are available.