Data Representations for Deep Learning and Chemistry-Related Applications

Sheng Hu
Specially Appointed Assistant Professor
Affiliation: ICReDD, Hokkaido University
e-mail: hu.sheng@icredd.hokudai.ac.jp
Specialty: Deep Learning on Graphs, Data Wrangling, Information Retrieval

ABSTRACT

The ultimate goal of modern drug discovery is to find the target molecules with desired chemical properties, while the potential chemical space of drug-like compounds is 10^{23}–10^{60}.

Visual graph query composition can assist modern drug discovery. Instead of exploratory searching given a subgraph query in graph databases and showing matched graphs, we prefer to use generative models to grow novel molecules on the given subgraph. In the applications of drug discovery, such a subgraph query is usually called a scaffold (i.e., privileged or bioactive scaffold), and performs as a core structure in the molecule to preserve the preferable bioactivity properties. The generated novel molecular graphs are supergraphs of the scaffold thus being guaranteed to contain the scaffold to reveal the chemical properties. Fixing the scaffold usually dramatically reduces the search space of the desired drug thus saving experts' time and cost.

Due to surprising success of deep neural network (DNN) models these days, two categories of representations used in DNN-based models emerge in the drug discovery domain. (1) simplified molecular input line entry system (SMILES) strings representation. Several early works were proposed to learn the SMILES grammar using RNN architectures and then generate corresponding SMILES strings from the trained models. These methods have limitations to learning the unrelated grammar and thus have low chemical validity from generated SMILES. A recent trend is (2) undirected labeled graph representation. It is more natural to learn the original graph structure by using graph neural networks (GNNs). This representation can easily achieve higher chemical validity. In our work, we adopt the graph-based representation along with GNN models as our generative models. GNN models are employed during the visual graph query composition process to generate completed candidates.

In this talk, we present our recent study[1] on a GNN-based molecular graph generation system, to allow users to edit the intermediate graph candidates during the molecule design process in multiple steps, utilizing the edit operation to predict the user's real intention to improve effectiveness. We design an interactive substructure-by-substructure adopt process to verify this idea. This process guarantees the involvement of user decisions to interact with a generative user-centered AI system, which differentiates our work from previous studies that generate graphs in a single run. We test our system using a real-world molecular dataset containing nearly one million graphs to show the performance.

REF.

1. S. Hu, I. Takigawa, C. Xiao, Edit-Aware Generative Molecular Graph Autocompletion for Scaffold Input, **The 36th AAAI Conference on Artificial Intelligence (AAAI) Workshop on Deep Learning on Graphs (DLG)** (2022)