Scaffold-Based Molecular Design Using a Graph Generative Model

Jaechang Lim,[a] Sang-Yeon Hwang,[a] Seungsu Kim,[b] Seokhyun Moon[a] and Woo Youn Kim[a,c]

[a] *Department of Chemistry, KAIST, Daejeon 34141, South Korea*
[b] *School of Computing, KAIST, Daejeon 34141, South Korea*
[c] *KI for Artificial Intelligence, KAIST, Daejeon 34141, South Korea*
s.hwang@kaist.ac.kr

The vast expanse of chemical space and accumulation of data have been more and more stirring the utilization of machine learning in molecular design. Going beyond allowing fast screening of molecules through efficient prediction of properties, recent advances in generative models enable one to design molecules with desired properties *de novo*. The structural characteristics of generated molecules vary depending on various factors, such as the molecules used for learning and the desired properties. On the other hand, certain applications, particularly in drug design, commonly require new molecules to possess a particular scaffold showing promising functionality. One way of generating molecules having a desired scaffold is defining a generative model whose underlying distribution over molecules is conditioned on the kinds of contained scaffolds. In such a case, however, the probabilistic nature also allows molecules with irrelevant scaffolds, and there can be limitations in categorically representing arbitrary kinds of scaffolds. To tackle these problems, we developed a scaffold-based graph generative model, realizing the architecture as a variational autoencoder [1]. Our model represents a molecular scaffold as a graph and extends it to a supergraph by sequentially adding vertices and edges. By learning the strategy of adding atoms and bonds to scaffolds, our model guarantees with certainty the existence of a desired scaffold in the generated molecules and is free from the representability of scaffold kinds. Our evaluation using unobserved as well as observed scaffolds shows that nearly all of the generated graphs conform to valid valency and that nearly all of the resulting molecules are distinct. When singly or jointly conditioned on molecular properties such as the molecular mass, topological polar surface area, and octanol–water partition coefficient, our model showed good control over the properties in that the property values of the generated molecules had mostly tolerable deviations from the target values. Upon these results, we hope our present work provides a promising approach to designing novel molecules with retained core structures and controlled properties.

References

1. J. Lim, S.-Y. Hwang, S. Kim, S. Moon and W. Y. Kim, *arXiv:1905.13639* (2019).