2019年6月5日　統計数理研究所　オープンハウス

Inverse molecular design with machine translation model

Zhang Qi　　総合研究大学院大学 統計科学専攻 博士後期課程1年

1 INTRODUCTION

1.1 Background

The goal of drug and material design is to identify novel molecules that have certain desirable properties. The main challenge for the chemist is to select and examine molecules from a large search space which has been estimated that involves 10^{60} drug-like molecules. A molecular generator is a desirable tool to narrow down the enormous search space. Hopefully, the generator can identify the promising hypothetical molecules with a predefined set of desired properties.

1.2 Related work

Segler et al.[1] presented an alternative sequential generation algorithm based on Recurrent neural networks. Rafael et al.[2] convert the discrete representations of molecules to and from a multidimensional continuous representation by variational autoencoder. Ikehata et al.[3] realize inverse design by incorporating expert knowledge into the optimization procedure, via improved Bayesian sampling with sequential Monte Carlo. Jin et al.[4] present a junction tree variational autoencoder for generating molecular graphs.

1.3 Problem

Most of the current works generate the molecule sequentially, by which the generation variety will decrease due to the cumulative error. On the other hand, the reactant information of each generated molecule does not include in the generation model, in another word, even if some hypothetical molecules are generated by the computer, how to generate them by a chemical reaction is still a big issue for chemists.

2 METHODS

We propose a machine translation model based algorithm for molecular generation and inverse molecular design. The proposed method is expected to have the following advantages:

・A massive modification of molecule can be achieved by fewer steps which offer generated molecule in diversity.

・Molecules are generated by a chemical reaction prediction model, which supplies the reactant information for real reaction guidance.

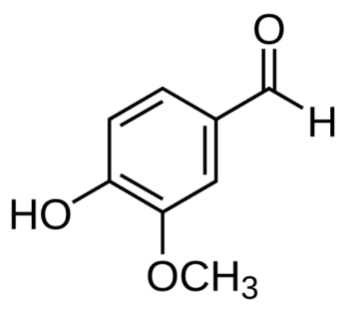
2.1 Molecular representation

The first step in our work is to build a descriptor of each molecule such that they can be processed into the machine learning algorithm.

Most generative modeling so far has worked with molecular graphs. There are several ways to represent graphs for machine learning. The most popular way is the SMILES string representation.

Molecule: Vanillin

Structure:

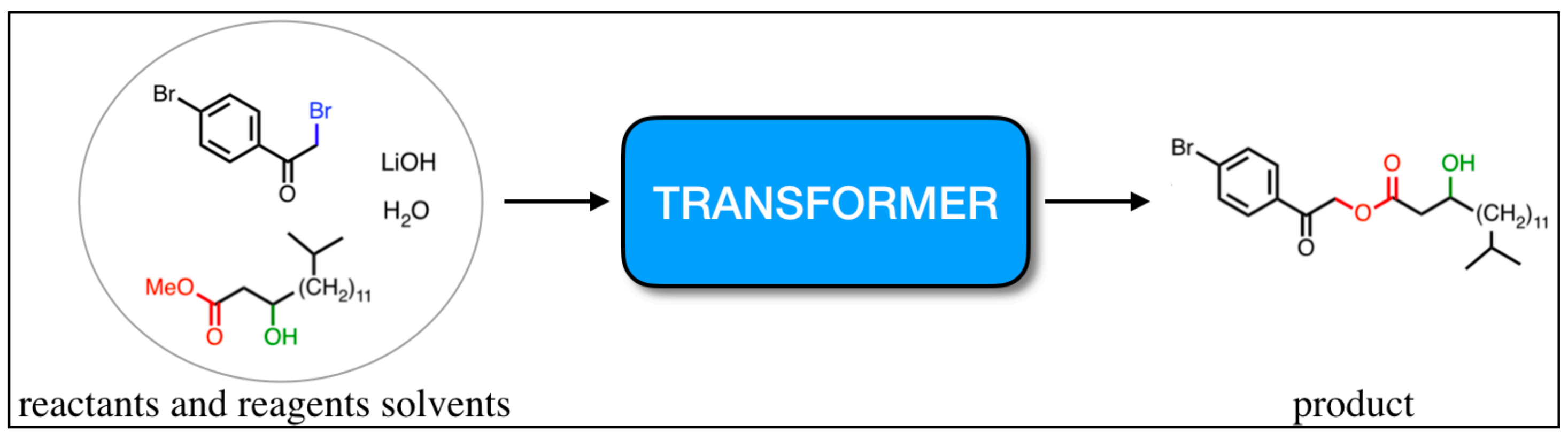


SMILES Formula: O=Cc1ccc(O)c(OC)c1OCc1cc(C=O)ccc1O

Representation of SMILES representation

2.2 Molecular generator

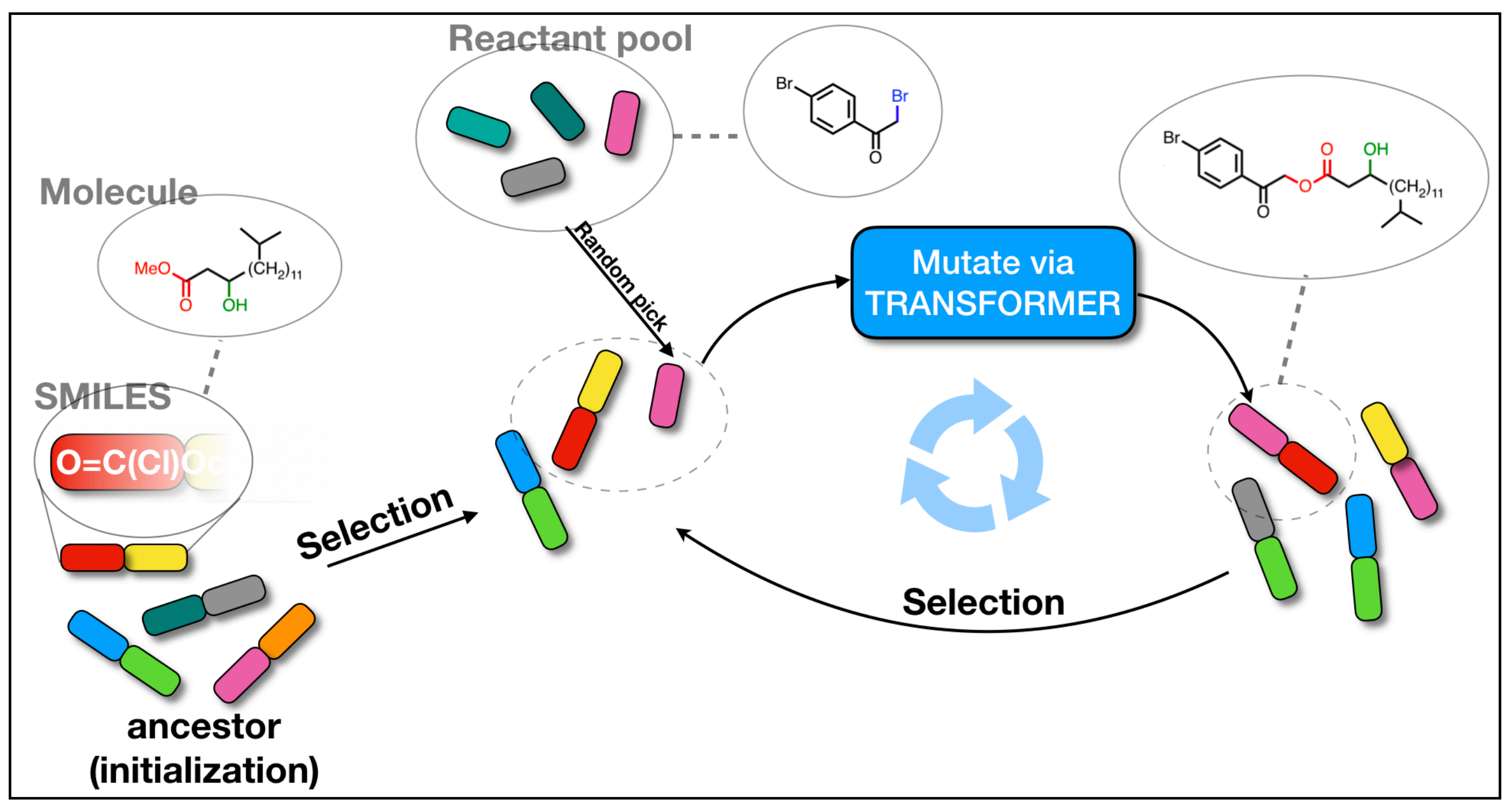
We use the transformer for molecule mutation. Transformer is a machine translator that uses attention concept which helped improve the performance of neural machine translation applications.



The above figure illustrates the transformer used for the molecular generation, where the input is a concatenated form of reactants, reagents, and solvents, the output is the predicted reaction production. Both input and output are represented as SMILES strings.

2.3 Inverse molecular design

We propose a genetic algorithm based architecture for the inverse molecular design where the task is to generate molecules with a predefined set of desired properties. Unlike the original genetic algorithm, our current work only involves selection and mutation process. The following two stages are alternately operated over generations: 1, perform artificial chemical reaction on molecules in the current generation with reactants randomly picked from a pre-designed reactant pool by the transformer. 2, the products of the transformer is selected based on the desired properties:

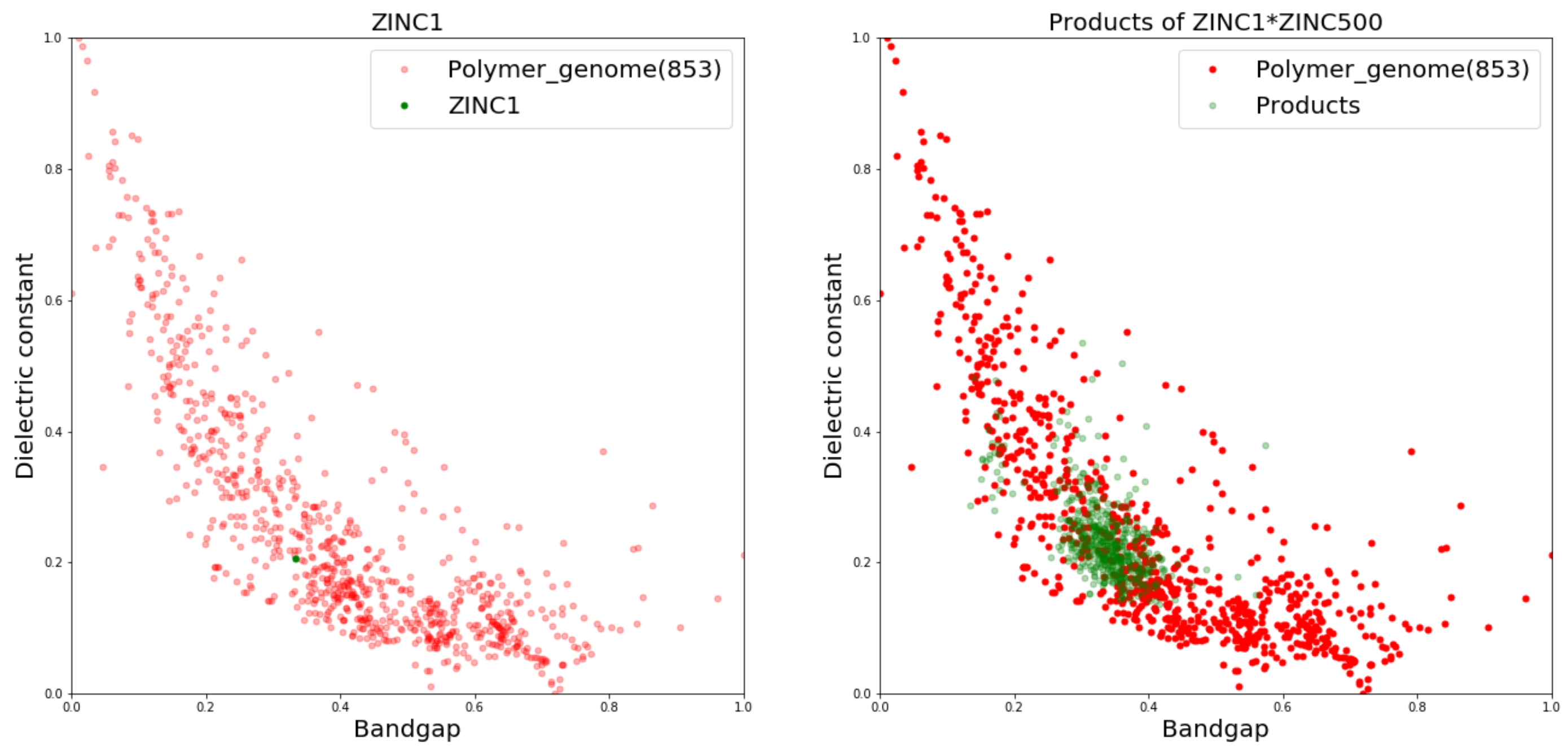


Representation of molecule generation process

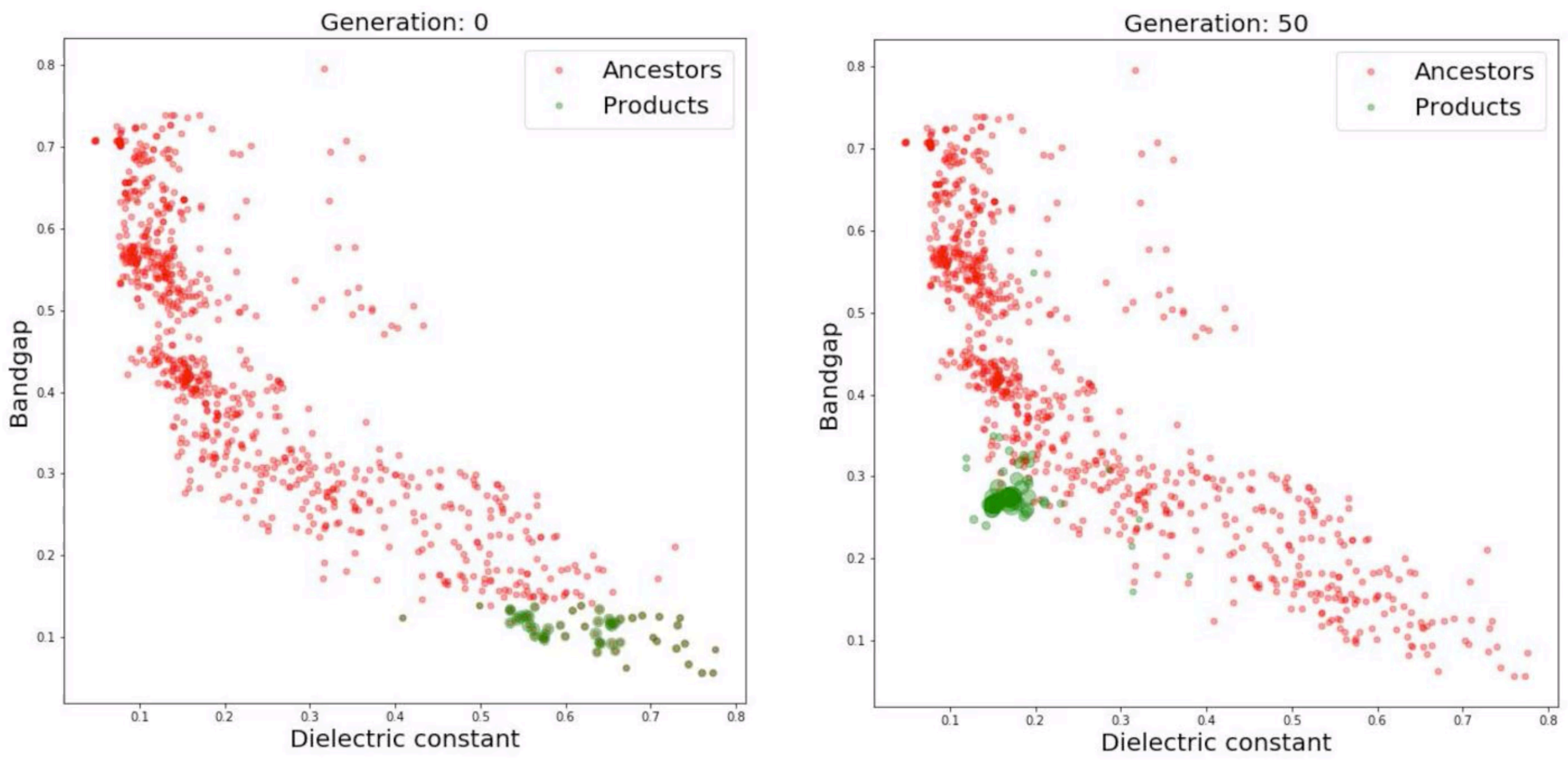
3 RESULT and DISCUSSION

We currently focus on generating a molecule library with low dielectric constant and low bandgap.

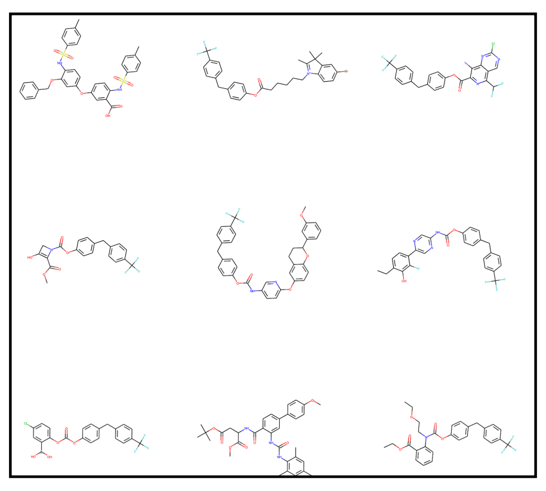
The following figure shows an example of mutation process by the transformer, one molecule is selected from the current generation (left), we perform artificial chemical reaction 500 times on this molecule with 500 different reactants from the reactant pool, the property of the productions are as follow (right).



The following figure is the illustration of the proposed algorithm, we initialize the first generation by choosing the molecules with low bandgap and high dielectric constant (left), after 50 iterations, both of the generations move towards the desired area (right).



Some of the product structures are shown below:



4 FUTURE WORK

The following works are considered as the future direction:

1, A quantitative measurement of the proposed method.

2, Generation process of desired molecules within fewer steps (reactions).

REFERENCE

[1] Segler, Marwin HS, et al. "Generating focused molecule libraries for drug discovery with recurrent neural networks." ACS central science 4.1 (2017): 120-131.

[2] Gómez-Bombarelli, Rafael, et al. "Automatic chemical design using a data-driven continuous representation of molecules." ACS central science 4.2 (2018): 268-276.

[3] Ikebata, Hisaki, et al. "Bayesian molecular design with a chemical language model." Journal of computer-aided molecular design 31.4 (2017): 379-391.

[4] Jin, Wengong, Regina Barzilay, and Tommi Jaakkola. "Junction tree variational autoencoder for molecular graph generation." arXiv preprint arXiv:1802.04364 (2018).