第八屆台灣工業與應用數學會年會

National Cheng Kung University, Tainan, Taiwan, Jul 24, 2020

Assisted Design of Chemical Structures and Properties Prediction via Deep Generative Model

Tung-Ching Hsieh (謝東景)(1)(2), Chia-Yung Jui (芮嘉勇)(2), Yao-Chun Wang (王耀群)(3), Jia-Han Li (李佳翰)(1)*, Wen-Jay Lee (李玟頡)(2)*

(1) Department of Engineering Science and Ocean Engineering, National Taiwan University, No. 1, Sec. 4, Roosevelt Rd., Taipei, Taiwan, 10617

(2) National Center for High-Performance Computing, No. 22, Keyuan Rd., Central Taiwan Science Park, Taichung, Taiwan, 40763

(3) Daxin Materials Corp., No. 15, Keyuan 1st Rd., Central Taiwan Science Park, Taichung, Taiwan, 40763

Email: tc.hsiehhsieh@gmail.com

NARLabs
National Applied Research Laboratories
National Center for High-performance Computing

Abstract

To accelerate the exploration of novel materials, the deep-learning-based inverse design for the intelligent discovery of organic molecules was introduced by experts in computational materials. The novel molecules with desired properties can be generated from the trained continuous latent space, which describes the high-level feature for chemical structure. In addition to continuous representation which relates to actual chemical structures, we reschedule each loss in our model and search optimal molecules with precise chemical structures efficiently. This approach helps us not only to generate the valid chemical structures by our deep generative model precisely but also to correspond molecules to the perspective of physical significance in each chemical property. We implement organic molecules on our model with electrical properties.

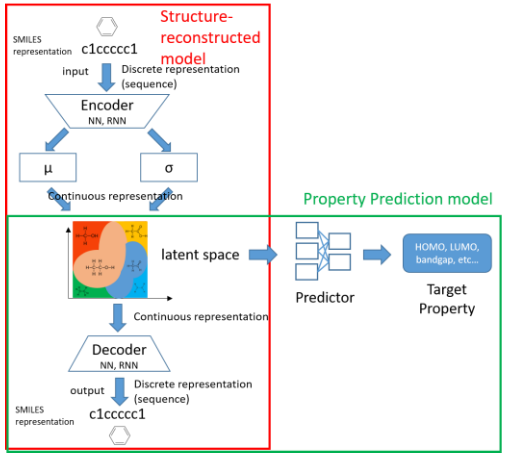
Problem Description

Over the past decades, researchers dedicated to searching new materials with desired properties by conducting enormous experiments and simulations. However, discovering novel materials with desired properties remains an expensive task and takes a long period of time. Deep learning (DL) gives a potential solution to reduce the cycle of development and explore new chemical structures. Large amounts of novel candidate molecules with predicted properties can be obtained by DL model instead of through trial and error with simulations and experiments. Therefore, we can efficiently reduce the potential candidates of materials by machine-learning-based inverse design [1]. Due to the difficulty of searching large areas of chemical space, we need to establish a general model to search optimal molecules described by continuous representation [2]. However, how to construct a general model that can be implemented by the DL algorithm with domain knowledge in physic and chemistry is still a problem needed to be solved.

Methodology

The proposed variational autoencoder (VAE) model for inverse design is combined as follows. Both the encoder and decoder consist of the word embedding for SMILES string with 48 dimension sizes and 2-layers gated recurrent units (GRU) with 512 units of hidden states. The property predictor consists of 2-layers neural networks with 28 and 14 nodes. Since the VAE model underestimates the reconstruction loss, it breaks the balance with KL loss [3]. Since the posterior collapse phenomenon of the RNN autoencoder, the architecture tends to ignore the subset of latent variables and causes the latent space less meaningful. We can adjust the reconstruction loss to recover the balance of loss [4]. The VAE was trained by 106,598 molecules and validated by 26,649 molecules in the QM9 dataset [5].

Figure 1. VAE architecture of deep generative inverse design model. It combines with GRU-based sequence-to-sequence model for chemical structure reconstruction and property prediction model with multilayer perceptron (MLP) structure. As the model learn the dataset with SMILES string structure and numerical property values, we can find the connection between the physical significance of chemical structure and chemical properties on the latent space.


Structure-reconstructed model
SMILES representation
c1ccccc1
input
Discrete representation (sequence)
Encoder
NN, RNN
μ
σ
Continuous representation
latent space
Continuous representation
Decoder
NN, RNN
output
Discrete representation (sequence)
SMILES representation
c1ccccc1
Property Prediction model
Predictor
HOMO, LUMO, bandgap, etc...
Target Property

Results

Analysis of latent space

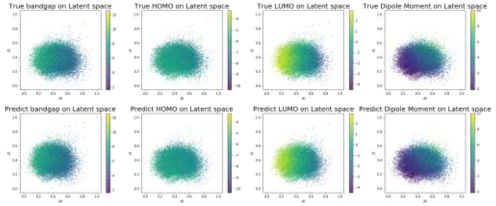


Figure 2. Two-dimensional PCA analysis of latent space for VAE model. The selected properties are shown in the color bar.

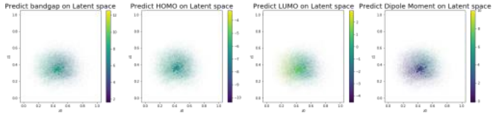


Figure 3. Random sampling in two-dimensional PCA analysis of latent space with desired properties. The selected properties are shown in darker colors.

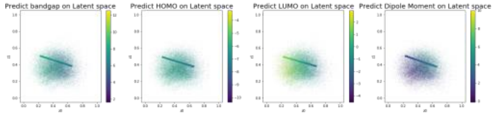


Figure 4. Interpolation between 2 molecules in two-dimensional PCA analysis of latent space. The selected properties are shown in darker colors

Representation of sampling results

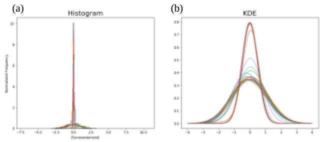


Figure 5. Representation of the (a) Histogram (b) Kernel Density Estimation (KDE) of each latent dimension of VAE

Model	Model 1	Model 2	Model 3	Re-balancing VAE [4]
Validity	0.9063	0.9100	0.9129	0.9009
Accuracy	0.9912	0.9903	0.9841	0.9266

Table 1. Results of structure reconstruction

	Pearson Correlation	MAE	R^2
bandgap	0.9733	0.2423	0.9378
HOMO	0.9318	0.1730	0.8604
LUMO	0.9815	0.2400	0.9434
Dipole Moment	0.8975	0.5176	0.7828

Table 2. Results of property prediction

Analysis of predictor

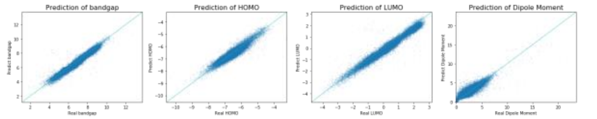


Figure 6. Relation of chemical properties between the real value of properties from the QM9 validation dataset and predicted properties from the VAE model.

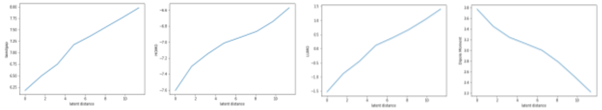


Figure 7. Relation of chemical properties for interpolation between 2 molecules.

Discussion and Conclusions

VAE served as a generative model that can be more precise to reconstruct the SMILES representation by adjusting the loss weights and the reduced latent dimension. Besides, the predictor also correlates chemical properties with the high-level feature on latent space. That is, the similar molecules cluster after latent space is reorganized by jointing the property prediction. When we choose two molecules with different properties, we can find a continuous relationship on not only the latent space but also the chemical properties. The results are shown in figure 4, 7 and 9. Electronegativity of different atoms, ring structure, and molecular geometry affect the electric dipole. Besides, the relation among different properties discovered by the model also matches the theory of molecular geometry. The electric dipole moment shows a negative correlation with the bandgap of the materials. That is, the deep generative model learns more physical significance on the continuous representation.

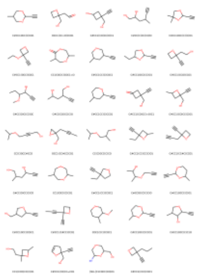


Figure 8. Random sampling of valid chemical structure with desired properties.

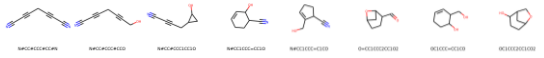


Figure 9. Interpolation between 2 molecules of valid chemical structure with desired properties. The numerical progress results are shown in figure 6.

References

[1] B. Sanchez-Lengeling, A. Aspuru-Guzik, *Science*. 361 (2018) 360–365.

[2] A. Aspuru-Guzik, et. al, *ACS Cent. Sci*. (2018) 4, 268−276

[3] J. Lucas, et. al, ICLR 2019 Workshop

[4] J. Huang, et. al, *arXiv*. (2019) 1910.00698

[5] Z. Wu, et. al, *arXiv*. 7 (2017) 1703.00564.