Bayesian Optimization with Gaussian Process Assisted by Deep Learning for Material Designs

Shin Kiyohara[1,*]

[1] *Institute for Materials Research, Tohoku University, Sendai, Japan*

[Introduction]

Bayesian optimization (BO) is a powerful method that enables efficient exploration of vast material spaces and is now widely used for material discovery. However, the efficiency of conventional Bayesian optimization based on Gaussian processes (GP) heavily depends on the choice of descriptors. In this study, we constructed a deep kernel learning (DKL) model that combines deep learning with a Gaussian process and applied it to BO. Our method was evaluated under different conditions by applying it to (1) first-principles calculation data, where crystal structure information is available, and (2) experimental data, where only compositional information is available. We assessed the efficiency of BO using DKL and developed a new approach for material discovery.

[Methods]

The DKL architecture used in this study consists of two components: CGCNN (Crystal Graph Convolutional Neural Network) [1] and GP. The crystal information is first vectorized by the CGCNN component and then used as input to the kernel of the GP.

The search flow of our study is shown in Figure 1. The search begins with a small number of randomly selected data points, followed by Bayesian optimization. For comparison, we also conducted searches using a conventional GP. In the conventional GP approach, only descriptors evaluated as important by random forest and permutation importance were used [2]. Since the efficiency of Bayesian optimization significantly depends on the initial dataset, optimization was performed 50 times, and the efficiency was evaluated based on the average number of cycles required to reach convergence.

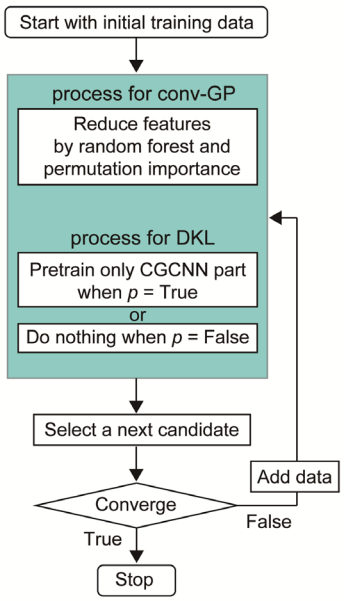

Figure 1. The BO scheme adopted in this study.

[Results]

Using the above method, we explored materials with the maximum bandgap among 922 oxides obtained from first-principles calculations[3]. When the initial dataset contained 10 points, the conventional GP-based search found the maximum value in an average of 75 cycles (excluding the initial data points). In contrast, DKL found the maximum value in an average of 21 cycles.

References

[1] Xie, T.; Grossman, J. C. *Phys. Rev. Lett.*, **2018**, 120, 145301.

[2] Takahashi, A.; Kumagai, Y.; Aoki, H.; Tamura, R.; Oba, F. *STAM:Methods*, **2022**, 2, 55-66.

[3] Kumagai, Y.; Tsunoda, N.; Takahashi, A.; Oba, F. *Phys. Rev. Mater.*, **2021**, 5, 123803.