Data assimilation model combined with symbolic regression for predicting band gaps of perovskite photocatalytic compounds.

Shuan Shinkawa
Nara Institute of Science and Technology

To design water-splitting photocatalysts with visible-light response, predicting band gaps using machine learning is useful and is expected to improve the efficiency of materials discovery. While machine learning requires large training datasets, experimental data is often limited. First-principles calculations can be used instead, although they inherently include systematic errors due to physical approximations. Previous studies have suggested that assimilating computational data with experimental data can correct predictions and improve accuracy. However, such methods often struggle to handle complex, nonlinear systematic errors.

In this presentation, we propose a new data assimilation framework that determines a correction function to address the problem of nonlinear systematic errors. We initially focus on reducing the complexity of high-dimensional features by using symbolic regression for dimensionality reduction, aiming to capture band gap narrowing in photocatalysts. In the next step, symbolic regression is applied to the prediction residuals to select the optimal algebraic formula as a correction model. We then demonstrate the effectiveness of this approach by showing an improvement in prediction accuracy, using a specific photocatalyst candidate system as a case study.