# Abstract

**Aya Fujiwara (2220027)**

Materials Informatics (MI) aims to accelerate the discovery and understanding of materials through data science, especially in cases where complexity exceeds human intuition. However, there are two main challenges to implementing MI in practical applications: one is the lack of data that meets the quality, scale, and consistency required for effective machine learning, and the other is the need for advanced domain knowledge to design descriptors (numerical representations of catalysts). To address the data challenge, Taniike and his research group have developed their own datasets using high-throughput experimentation (HTE). Additionally, they introduced an automatic feature engineering (AFE) method that enables the design of descriptors without domain-specific knowledge, thereby addressing the descriptor challenge. This technique generates numerous features (descriptor candidates) from the physicochemical properties of catalyst components and extracts those that are relevant for describing catalyst performance.

The oxidative coupling of methane (OCM) is a catalytic reaction that directly synthesizes ethylene ($C_2H_4$) from methane ($CH_4$). This process is significantly more energy-efficient than conventional synthesis routes via syngas. However, due to the higher stability of $CH_4$ compared to $C_2H_4$, achieving high yields remains challenging, necessitating the development of high-performance catalysts. Taniike and his group have accumulated a dataset containing performance data for over 600 catalysts in OCM using their HTE system. Using this data as the training dataset, they have expanded the scope of catalyst exploration with machine learning techniques, such as decision tree analysis and support vector regression (SVR). These approaches, however, adopted one-hot encoding to represent catalyst compositions, where a catalyst component is represented as either 0 (element x is absent) or 1 (element x is present). The limitation of this method is that it treats catalyst compositions as symbols without providing insight. While researchers may recognize that catalysts A and B share common physical properties, the machine learning model perceives only common symbols. For efficient catalyst exploration, it would be ideal to define some physical features as a descriptor, however, achieving this in such a complex catalytic reaction is highly challenging. This illustrates a core challenge in MI: the need for a machine learning framework capable of generating hypotheses without requiring prior knowledge of the target system, particularly for the vast exploration of candidate materials.

From the perspective that “researchers derive insights from multiple observations,” when observational data is limited, it becomes difficult to reject competing hypotheses. Thus, researchers conduct repeated control experiments to refine their hypotheses and arrive at accurate conclusions. This hypothesis refinement process is heavily reliant on a researcher’s intuition, impacting both the speed and quality of research outcomes. If this process could be integrated into a machine learning framework and

streamlined with high-throughput experimentation, it could establish a versatile research framework applicable to a wide range of experimental research.

In this thesis, a refinement process is simulated using an adaptive catalyst design cycle that combines AFE, farthest point sampling (FPS), and HTE, targeting OCM catalysts. The study aims to develop and validate a robust machine learning model, demonstrating its utility in understanding catalyst design principles and supporting efficient catalyst development.

In **Chapter 2**, an efficient catalyst discovery process combining data-driven AFE, active learning, and HTE was demonstrated targeting BaO-supported catalysts for OCM. Through the refinement of feature space across diverse catalysts via AFE and HTE, predictive accuracy was enhanced, leading to the discovery of high-performance catalysts with $C_2$ yields exceeding 15%.

Building on this methodology, **Chapter 3** expands the scope to multiple catalyst supports, including BaO, CaO, $La_2O_3$, $TiO_2$, and $ZrO_2$. The analysis revealed two main patterns: one where a single element, such as La on BaO, dominated performance, achieving high yields without additional elements; and another, exemplified by CaO, where a combination of multiple elements, particularly alkaline earth metals with Cs, was essential. These findings highlighted distinctive design rules for each support.

In **Chapter 4**, active learning efficiency was explored to achieve high-accuracy learning with smaller datasets. Additionally, I explored the possibility of applying design hypotheses derived from known supports to new supports. We tested whether design hypotheses from the five supports in Chapter 3 could guide predictions for $Y_2O_3$ supports. This approach confirmed the transferability of design insights across supports.

The catalysts newly developed in Chapters 2, 3, and 4 are synthesized in **Chapter 5**, which highlights the high-performance catalysts created throughout the thesis.

In summary, this thesis presents an original adaptive catalyst design cycle with recursive feature engineering, proving the utility of the resulting machine learning models in understanding catalyst design heuristics and enhancing catalyst development. By exploring a broad range of compositions for OCM catalysts, this approach successfully identified high-performance catalysts and valuable design heuristics.

**Keywords:** Catalyst informatics, Machine learning, High-throughput experimentation, Oxidative coupling of methane, Descriptor design technology