



Title	Development of a Method to Reduce the Number of Synthetic Experiments by Combining Bayesian Optimization and Reaction Barrier Calculations
Author(s)	岡田, 拓明
Degree Grantor	北海道大学
Degree Name	博士(理学)
Dissertation Number	甲第15860号
Issue Date	2024-03-25
DOI	https://doi.org/10.14943/doctoral.k15860
Doc URL	https://hdl.handle.net/2115/94462
Type	doctoral thesis
File Information	OKADA_Hiroaki.pdf



Development of a method to reduce the number of
synthetic experiments by combining Bayesian optimization and
reaction barrier calculations

ベイズ最適化と反応障壁計算を組み合わせた

合成実験数削減法の開発

Hiroaki Okada

岡田 拓明

Hokkaido University

March 25, 2024

Contents

1	General Introduction	1
1.1	Organic Chemistry	1
1.1.1	Organic Reaction	1
1.1.2	Reaction Mechanism	2
1.1.3	What's the Challenge?	4
1.2	Quantum Chemical Calculation	5
1.2.1	Schrödinger Equation	5
1.2.2	Density Function Theory	7
1.2.3	Potential Energy Surface	8
1.2.4	Automatic Reaction Path Search	10
1.2.5	Prediction of Reaction Outcome	12
1.3	Informatics	12
1.3.1	Machine Learning Algorithm	12
1.3.2	Machine Learning Algorithm and Chemistry	14
1.3.3	Optimization Algorithm	16
1.3.4	Optimization Algorithm and Chemistry	17
1.4	Purpose of the Study	18
1.5	Overview of This Thesis	19
	References	21

2	Methods	29
2.1	Artificial Force Induced Reaction Method	29
2.1.1	AFIR Function	29
2.1.2	AFIR Path	31
2.1.3	Systematic Reaction Path Search	32
2.2	Geometry Generation Algorithm	32
2.2.1	Input	33
2.2.2	Algorithm	34
2.3	Genetic Algorithm	37
2.3.1	Algorithm	38
2.4	Bayesian Optimization	40
2.4.1	Algorithm	42
2.4.1.1	Surrogate Model	42
2.4.1.2	Acquisiton Function	48
2.4.1.3	Parallelization	51
2.5	Chemspeed ISYNTH Robotic Platform	53
2.5.1	Synthesizer	54
2.5.2	Scriptfile	54
2.5.2.1	Zone	55
2.5.2.2	Task	56
	References	56
3	A Dataset of Computational Reaction Barriers for the Claisen Rearrangement: Chemical and Numerical Analysis	58
3.1	Introduction	58

3.2	Methods	60
3.2.1	Barrier Calculation	60
3.2.2	Genetic Algorithm	62
3.3	Results and Discussion	64
3.3.1	Barrier Distribution	64
3.3.2	Chemical Analysis	65
3.3.3	Numerical Analysis	69
3.4	Conclusion	70
3.5	Data Availability	71
	References	71
4	On Accelerating Substrate Optimization Using Computational Gibbs	
	Energy Barriers: A Numerical Consideration Utilizing a Computational	
	Dataset	75
4.1	Introduction	75
4.2	Method	81
4.2.1	Barrier Dataset Utilized	81
4.2.2	Input and Output	83
4.2.3	Implementation of Bayesian Optimization	84
4.3	Results and Discussion	86
4.3.1	Performance Comparison	86
4.3.2	Heatmap Analysis	88
4.3.3	Limitation	90
4.4	Conclusion	91
4.5	Data Availability	92

4.6	Code Availability	92
	References	92
5	Regioselectivity Maximization of the Thermal Azide-alkyne Cycloadditions Based on Bayesian Substrate Optimization	100
5.1	Introduction	100
5.2	Method	102
5.2.1	Reaction Space	102
5.2.2	Substrate Optimization Scheme	103
5.2.3	Implementation of Bayesian Optimization	105
5.2.4	Experimental Details	108
5.2.5	Computational Details	109
5.3	Results and Discussion	111
5.3.1	Optimization Results	111
5.3.2	Substituent Effects Analysis	113
5.3.3	Performance Comparison	117
5.4	Conclusion	119
	References	120
6	General Conclusion	128
	Acknowledgements	132

Chapter 1

General Introduction

1.1 Organic Chemistry

Organic chemistry is the field of artificial synthesis of compounds. Today, organic compounds are used in various products, becoming indispensable tools in human life. This chapter introduces the history, overview, and challenges of organic chemistry.

1.1.1 Organic Reaction

Until the early 19th century, it was believed that organic compounds could only be synthesized by the mysterious power of life [1]. Substances that could be synthesized by humans were classified as inorganic, whereas those believed to originate only from living organisms were classified as organic. However, in 1828, Friedrich Wöhler discovered that urea could be synthesized by heating ammonium cyanate (Figure 1.1) [2]. This discovery demonstrated that organic compounds could be synthesized from inorganic substances. This historic discovery, known as the Wöhler synthesis, is considered to be the beginning of modern organic chemistry. Since then, organic chemistry has developed as a new field of chemistry and achieved remarkable success. In 1973, Robert Woodward's group achieved the total synthesis of vitamin B₁₂ [3]. This legendary achievement is still considered a

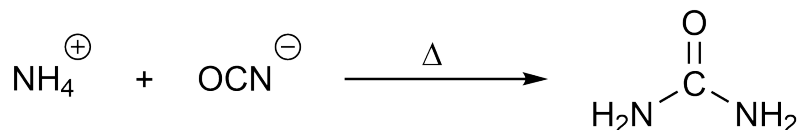


Figure 1.1. The Wöhler synthesis.

monumental achievement in organic synthesis. Today, compounds synthesized by organic chemistry are used in various products, including fibers, plastics, liquid crystals, semiconductors, pharmaceuticals, and agrochemicals, making organic chemistry inseparable from human life.

Now, what is the definition of organic chemistry? In essence, organic chemistry is the field of chemistry that deals with the study of carbon. Carbon exhibits a unique ability to form covalent bonds with up to four other atoms, allowing for the creation of molecules ranging from simple hydrocarbons like methane to complex biomolecules. This unique property of carbon allows the synthesis of compounds with various functionalities through organic reactions.

1.1.2 Reaction Mechanism

There are various types of organic reactions, ranging from simple to complex reactions. For simplicity, a model one-step reaction where two molecules, A and B, combine to produce the product AB is considered as an example (Figure 1.2). In a one-step reaction, there is a reactant, product, and transition state. The reactant passes through the transition state and reaches the product while undergoing a structural change. The Gibbs free energy difference between the reactant and the transition state is denoted as $\Delta\Delta G$, representing the energy barrier that the reaction must overcome. The Gibbs free energy difference between the reactant and the product is expressed as ΔG , indicating whether the chemical reaction proceeds spontaneously. Generally, organic reactions characterized

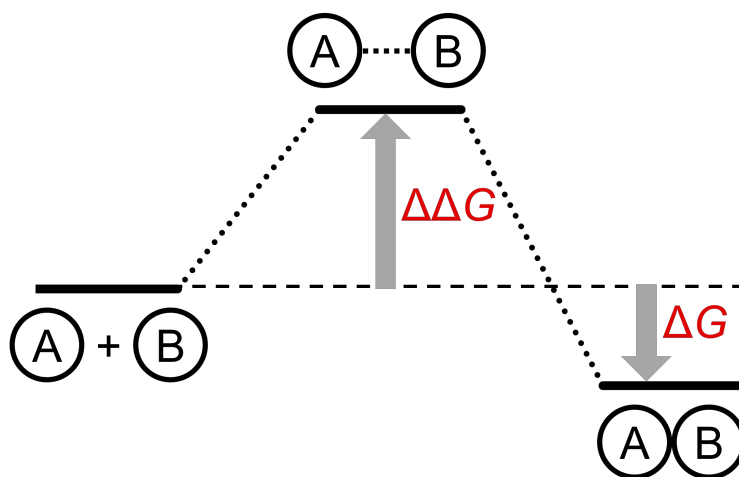


Figure 1.2. An energy diagram for a one-step reaction. $\Delta\Delta G$ represents the reaction barrier, while ΔG represents the reaction energy.

by high activation barriers exhibit reduced reactivity and yield. Furthermore, organic reactions associated with negative reaction energies proceed spontaneously, yielding products. Even in multi-step reactions involving multiple transition states and intermediates between reactants and products, analyzing the reaction barriers of rate-determining steps and the reaction energies associated with each elementary step provides insights into reaction mechanisms and kinetics.

Analysis of structural changes is also crucial for understanding reaction mechanisms. While the mechanisms of organic reactions are diverse, the movement of electrons fundamentally governs the formation and dissociation of chemical bonds. For example, a covalent bond forms when two atoms share a pair of electrons. Resonance structures in a molecule arise due to the delocalization of electron pairs. Even in complex catalytic cycles, the reaction mechanism is often interpreted in terms of the transfer of electrons. Thus, understanding the movement of electrons is essential for understanding the mechanisms of chemical reactions.

1.1.3 What’s the Challenge?

Understanding the Gibbs free energy changes and the movement of electrons facilitates the analysis of reactivity and mechanisms of organic reactions. However, the most important aspect of organic synthesis is efficiently discovering the organic reactions that yield the desired products through a desired process. In other words, designing a synthetic pathway for a new compound by using knowledge of established organic reactions is critical. Therefore, the objective in reaction development is to identify the optimal combination of substrates and reaction conditions, which is often uncertain due to the complex, high-dimensional nature of chemical reactions dependent on various parameters. Reactivity and selectivity of chemical reactions are influenced by a variety of factors, including substrates, transition metals, ligands, acids, bases, and solvents. These species exhibit different electronic and steric properties, stability, and solubility, all of which significantly impact the outcome of chemical reactions. In addition, numerical parameters such as stoichiometry, reaction temperature, concentration, and time, as well as addition rates, mixing conditions, and workup procedures are additional crucial factors influencing reactivity and selectivity. Since unsuitable substrates, reaction conditions, and procedures reduce reactivity, selectivity, and generality, optimization of the combination of chemical parameters is an unavoidable task in reaction screening.

Traditionally, chemists have explored the chemical space using empirical or heuristic approaches based on their experience and intuition. For instance, they narrowed down the search space through knowledge of similar reactions or routine-dependent methods to identify regions where optimal solutions were thought to exist. These approaches have the disadvantage that applying prior knowledge to highly novel reactions is challenging, and the search process is subject to chemist bias. Hence, the challenge in reaction development lies in efficiently screening reactions at minimal cost. In recent years, quantum chemical

calculations and informatics have been used as promising approaches to address this issue. In the following sections, an overview of quantum chemical calculations and informatics, and their applications to organic synthesis will be provided.

1.2 Quantum Chemical Calculation

Quantum chemistry, which began with the discovery of the Schrödinger equation [4], has undergone a remarkable development in just 100 years. In this section, an overview of the Schrödinger equation and density functional theory will be introduced. The systematic search for potential energy surfaces by quantum chemical calculations is also described.

1.2.1 Schrödinger Equation

The time-independent Schrödinger equation can be expressed as

$$\hat{H}\Psi = E\Psi \quad (1.1)$$

where, $\hat{H}$ is the Hamiltonian operator, Ψ is the wave function, and E is the energy eigenvalue. Using atomic units ($m = 1, e = 1, \hbar = 1$), $\hat{H}$ can be written as

$$\hat{H} = \sum_{I=1}^M \left(-\frac{1}{2M_I} \nabla_I^2 \right) + \sum_{i=1}^N \left(-\frac{1}{2} \nabla_i^2 \right) - \sum_{i=1}^N \sum_{I=1}^M \frac{Z_I}{r_{iI}} + \sum_{\substack{i,j=1 \\ (i>j)}}^N \frac{1}{r_{ij}} + \sum_{\substack{I,J=1 \\ (I>J)}}^M \frac{Z_I Z_J}{R_{IJ}} \quad (1.2)$$

where the first term represents the non-relativistic kinetic energy of the nuclei, the second term is the kinetic energy of the electrons, the third term accounts for the Coulomb interaction between nuclei and electrons, the fourth term represents the Coulomb interaction between electrons, and the fifth term is the Coulomb interaction between nuclei. M_I and Z_I denote the mass and charge of the I -th nucleus, r_{iI} is the distance between the i -th electron and the I -th nucleus, r_{ij} is the distance between the i -th and j -th electrons, and R_{IJ} is the distance between the I -th and J -th nuclei.

Here, an approximation known as the Born-Oppenheimer approximation is introduced [5]. Atomic nuclei are significantly heavier than electrons, with a mass difference of over three orders of magnitude. Consequently, electrons move much faster than nuclei. Therefore, when considering the motion of electrons, it is assumed that the motion of nuclei can be neglected. Under the Born-Oppenheimer approximation, the kinetic energy of the nuclei is considered zero, and the Coulomb interaction between nuclei is treated as a constant. Thus, $\hat{H}$ can be reformulated as the electronic Hamiltonian operator $\hat{H}^{\text{el}}$ given by

$$\hat{H}^{\text{el}} = \sum_{i=1}^N \left(-\frac{1}{2} \nabla_i^2 \right) - \sum_{i=1}^N \sum_{I=1}^M \frac{Z_I}{r_{iI}} + \sum_{\substack{i,j=1 \\ (i>j)}}^N \frac{1}{r_{ij}} \quad (1.3)$$

Under this assumption, the Schrödinger equation for the electronic states is given by

$$\hat{H}^{\text{el}} \Psi^{\text{el}} = E^{\text{el}} \Psi^{\text{el}} \quad (1.4)$$

where Ψ^{el} and E^{el} are the electronic wave function and the electronic energy eigenvalue, respectively. Solving the Schrödinger equation yields Ψ^{el} and E^{el} . However, the Schrödinger equations cannot be solved for systems other than the hydrogen atom. Therefore, some form of approximation is necessary.

One of the promising methods for approximately solving the Schrödinger equation is molecular orbital theory and density functional theory. In molecular orbital theory, an N -electron wave function is approximated by single-electron wave functions based on the independent electron approximation. A representative method in molecular orbital theory is the Hartree-Fock method [6]. In the Hartree-Fock method, an N -electron wave function

is expressed by the following the Slater determinant [7]:

$$\Psi(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N) = \frac{1}{\sqrt{N!}} \begin{vmatrix} \chi_1(\mathbf{x}_1) & \chi_2(\mathbf{x}_1) & \cdots & \chi_N(\mathbf{x}_1) \\ \chi_1(\mathbf{x}_2) & \chi_2(\mathbf{x}_2) & \cdots & \chi_N(\mathbf{x}_2) \\ \vdots & \vdots & & \vdots \\ \chi_1(\mathbf{x}_N) & \chi_2(\mathbf{x}_N) & \cdots & \chi_N(\mathbf{x}_N) \end{vmatrix} \quad (1.5)$$

where, χ denotes spin orbitals, and $\mathbf{x}$ denotes the position $\mathbf{r}$ and spin σ of an electron. Constraints are imposed to ensure that χ satisfies orthonormality, and using variational methods to minimize the $\hat{H}^{\text{el}}$, the Schrödinger equation is approximated by the Hartree-Fock equation:

$$\left[-\frac{1}{2}\nabla^2 + V(\mathbf{r}) + \sum_{j=1}^{N/2} (2\hat{J}_j - \hat{K}_j) \right] \chi_i(\mathbf{x}) = \epsilon_i \chi_i(\mathbf{x}) \quad (1.6)$$

where $-\frac{1}{2}\nabla^2$ is the kinetic energy of an electron, $V(\mathbf{r})$ is the Coulomb interaction between nuclei and an electron, $\hat{J}_j$ is the Coulomb operator, $\hat{K}_j$ is the exchange operator, and ϵ_i is the orbital energy of the spin orbital $\chi_i(\mathbf{x})$. In practice, the spatial orbitals $\phi_i(\mathbf{r})$ constituting the spin orbitals $\chi_i(\mathbf{x})$ are approximated as linear combinations of basis functions, and the expansion coefficients of the basis functions are determined self-consistently.

1.2.2 Density Function Theory

Density functional theory (DFT) is a type of first-principles calculation method and a crucial theory widely used in theoretical studies using quantum chemical calculations. The main feature of DFT is the use of the electron density instead of the wave function, to determine the energy of the ground state [8]. For this purpose, it is required that the energy of the ground state is a functional of the electron density, and it must satisfy the variational principle with the electron density as a trial function. These conditions are established by the Hohenberg-Kohn theorems [9].

Practically employed DFT is known as Kohn-Sham DFT [10]. In Kohn-Sham DFT, the following Kohn-Sham equation is considered:

$$\left[-\frac{1}{2}\nabla^2 + V(\mathbf{r}) + \int \frac{\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r}' + V_{\text{xc}}(\mathbf{r}) \right] \varphi_i(\mathbf{r}) = \varepsilon_i \varphi_i(\mathbf{r}) \quad (1.7)$$

where the first term is the kinetic energy of the electrons, the second term is the Coulomb interaction between nuclei and electrons, the third term is the Coulomb interaction between electrons, and the fourth term is the exchange-correlation potential. $\varphi_i(\mathbf{r})$ represents the Kohn-Sham orbitals, and ε_i is the Kohn-Sham orbital energy corresponding to $\varphi_i(\mathbf{r})$. In addition, $V_{\text{xc}}(\mathbf{r})$ is the first-order derivative of the exchange-correlation energy $E_{\text{xc}}(\mathbf{r})$ with respect to the electron density:

$$V_{\text{xc}}(\mathbf{r}) = \frac{\delta E_{\text{xc}}(\mathbf{r})}{\delta \rho} \quad (1.8)$$

$E_{\text{xc}}(\mathbf{r})$ includes the electron exchange energy, the electron correlation energy, and the difference between the true and approximated kinetic energies of the electrons. While the exact form of $E_{\text{xc}}(\mathbf{r})$ is unknown, practical Kohn-Sham DFT uses approximated exchange-correlation functionals. In solving the Kohn-Sham equations, $\varphi_i(\mathbf{r})$ is approximated as a linear combination of basis functions, and the coefficients of the basis functions are determined self-consistently, similar to the Hartree-Fock method.

1.2.3 Potential Energy Surface

The degree of freedom of an N -atom system is $3N$ since each atom has 3 degrees of freedom along the x -, y -, and z -axis directions. The translational degree of freedom is always 3, and the rotational degree of freedom is also 3 (2 in the case of linear molecules). Since translational and rotational motions don't contribute to structural changes, the degree of freedom of structural changes is $3N-6$. The potential energy as a function

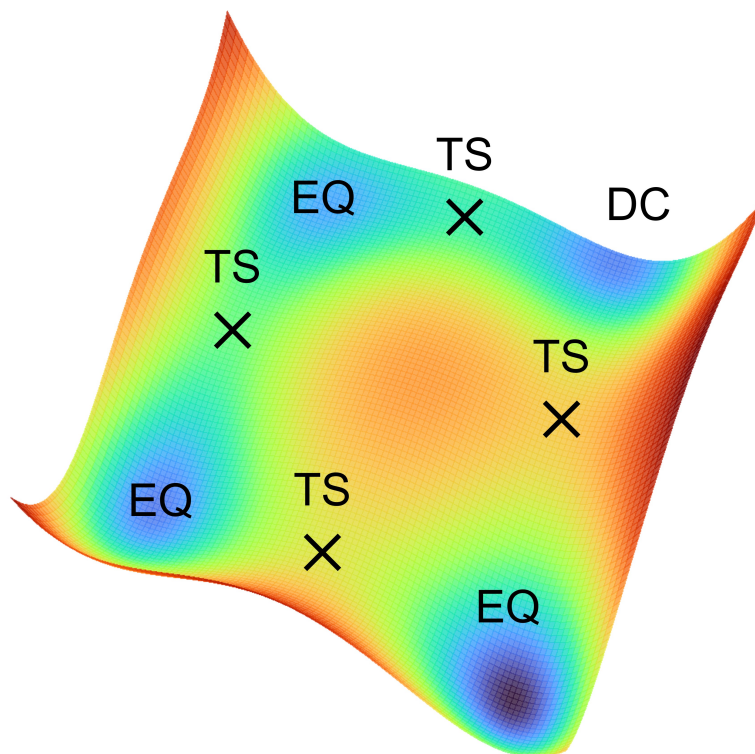


Figure 1.3. Model potential energy surface. Each equilibrium structure is connected by the transition states along the intrinsic reaction coordinate.

of $3N-6$ structural variables is called the potential energy surface (PES). Local minima on the PES correspond to the equilibrium (EQ) structures, and first-order saddle points are the transition state (TS) structures (Figure 1.3). In EQ, the gradient is 0 in all directions, and all eigenvalues of the Hessian matrix are positive. In TS, the gradient value is also 0, but there is only one negative eigenvalue, and the corresponding eigenvector represents the direction of the reaction path. The path from a TS down the steepest descent path in the direction of the eigenvector in the mass-weighted PES is called the intrinsic reaction coordinate (IRC) path [11]. Each EQ is connected by an IRC path via the TS. Furthermore, pathways from EQ to dissociated channels (DC) via TS also exist.

Reactants and products of a chemical reaction correspond to EQs on the PES, typically connected via multiple TS and meta-stable EQs. Hence, by exploring reaction

pathways with low reaction barriers from reactants to products, the reactivity and mechanisms of the reaction can be predicted by calculations. Traditionally, the systematic exploration of reaction pathways has been considered challenging. However, recently developed reaction path exploration methods enable the comprehensive exploration of chemically important PES regions.

1.2.4 Automatic Reaction Path Search

Numerous reaction path exploration methods have been proposed. Among them, the anharmonic downward distortion following (ADDF) method [12, 13] and the artificial force induced reaction (AFIR) method [14, 15], along with their utilization in the global reaction route mapping (GRRM) strategy [16, 17], are groundbreaking approaches that enable systematic and automatic exploration of reaction paths.

The ADDF method utilizes the anharmonicity of the PES near EQ. Figure 1.4 shows the harmonic potential centered at each EQ. However, the actual potential surface looks like a red line smoothly connecting the EQs and TS and is distorted below the harmonic potential (Figure 1.4). This effect is called anharmonic downward distortion (ADD). The presence of ADD implies the existence of a reaction path in that direction. Therefore, by tracing the reaction path from the EQ in the direction of the ADD maximum, it is possible to reach an approximate TS. In the actual algorithm, all ADDs can be found by considering the hypersphere in the scaled normal coordinates, which are the normal coordinates scaled by the corresponding vibrational eigenvalues, and examining all locations where the energy is minimum, i.e., where the ADD is maximum, on the hypersphere. The method to trace the reaction paths along with ADD on the hypersphere is called the scaled hypersphere search (SHS) method, and the method to find all ADDs on the hypersphere is called the iterative optimization elimination (IOE) method.

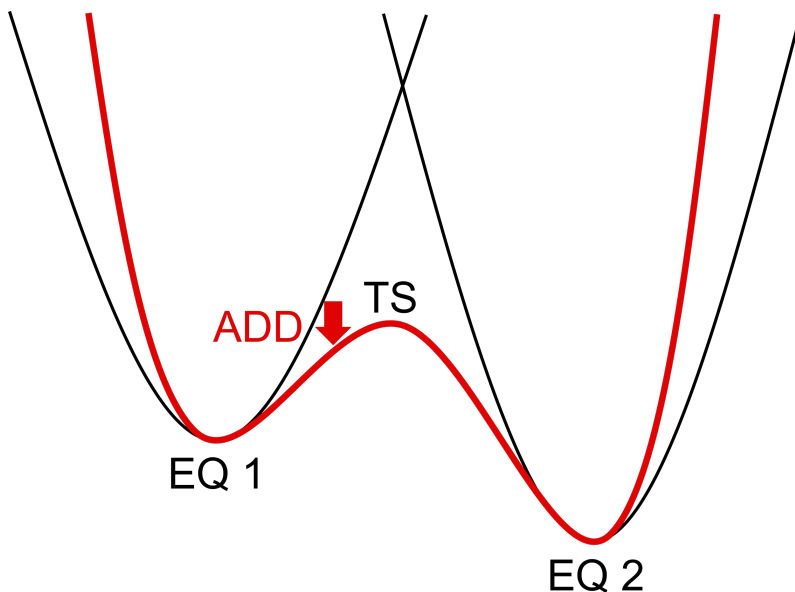


Figure 1.4. Typical potential curve (red line) and the harmonic potentials (black line). The ADD is shown by a down arrow.

In the AFIR method, part of a chemical structure is considered a fragment and a virtual force is applied between the fragments to induce a structural change. The AFIR method can find a reaction path from one EQ to another EQ by a simple geometry optimization calculation. Moreover, since multiple patterns of fragment can be defined for an EQ, it is possible to explore the reaction path from one EQ to multiple EQs by systematically changing the fragment that artificial force is applied. The AFIR method can also be applied to EQs newly found by the AFIR method. Hence, the automatic reaction path search is enabled by systematically adding the artificial force to all EQs. Furthermore, in geometry optimization using the AFIR method, the structure with the maximum potential energy without artificial forces is considered to be the approximate TS structure. Details of the AFIR method are presented in Chapter 2.1.

In the ADDF and the AFIR methods, the actual TS is obtained from the approximate TS using the quasi-Newton method. Hence, the systematic application of the ADDF

and AFIR methods enables the systematic exploration of chemical reaction path networks and the prediction of reaction barriers.

1.2.5 Prediction of Reaction Outcome

Quantum chemical calculations allow the analysis of various physical parameters such as charges, dipole moments, bond distances, and frontier molecular orbital energies. There have also been many studies that rationalize experimental reactivity and selectivity of organic reactions by exploring intermediate and TS structures [18–29]. In recent years, advancements in computer performance and the development of automatic reaction path search methods have enabled reaction screening by quantum chemical calculations prior to experimental validation [30–38]. Some studies even succeeded in the prediction of reaction yield without experiments [39, 40]. Such computation-driven reaction screening has the potential to significantly reduce experimental costs required for reaction screening by narrowing down desirable substrates and experimental conditions through quantum chemical calculations.

1.3 Informatics

Informatics is a technique for extracting rules from available data to construct a model that predicts outcomes of unknown inputs and explores candidate solutions that give better results. This section introduces the concept of machine learning algorithms and optimization algorithms and their application in chemistry.

1.3.1 Machine Learning Algorithm

Machine learning algorithms are methods that enable computer programs to learn from data and recognize rules and patterns behind the data to solve problems [41, 42].

Machine learning algorithms can be broadly classified into three main categories.

1. **Supervised learning:** Supervised learning is a method that uses input and output data pairs as training data to learn the underlying knowledge or rules within the dataset. The primary objective of supervised learning is to construct models that can predict outputs for unknown inputs with high accuracy. Such models are termed regression models when the objective variable is continuous and classification models when it is discrete. Linear regression, random forest regression, and Gaussian process regression are some of the well-known regression algorithms.
2. **Unsupervised learning:** Unsupervised learning involves extracting patterns, structures, and other insights from training data, which is comprised solely of input data, without corresponding output data. In contrast to supervised learning, there is no objective variable in unsupervised learning. Hence, there is no feedback from the results. k -means clustering and principal component analysis are typical examples of unsupervised learning algorithms.
3. **Reinforcement learning:** Reinforcement learning does not use a dataset and begins learning with no prior knowledge. An agent chooses an action and subsequently receives feedback, often in the form of rewards, from the environment based on its actions. Through this iterative process, the agent learns the relationship between its actions and the received feedback and finds a policy to choose the optimal action that maximizes the cumulative reward. The multi-armed bandit algorithm is one of the popular reinforcement learning algorithms.

This section focuses on the application of supervised learning to the regression problem. In the regression problem, the training data consists of the input $\mathbf{x}$ and corresponding

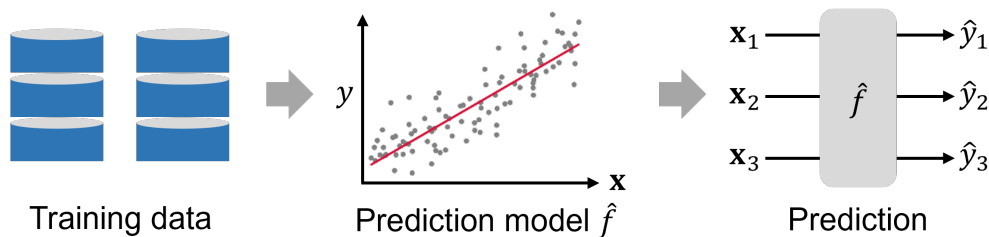


Figure 1.5. A conceptual diagram of regression using machine learning. The computer learns the relationship between input and output in the training data and utilizes it to predict output for new inputs.

output y . The $\mathbf{x}$ is a column vector consisting of several descriptors x_i .

$$\mathbf{x}^\top = (x_1, x_2, \dots, x_n) \quad (1.9)$$

It is commonly assumed that y represents the function value $f(\mathbf{x})$, obtained by transforming $\mathbf{x}$ through the objective function f , with the addition of observation noise ϵ following a normal distribution $\mathcal{N}(0, \sigma^2)$:

$$\begin{cases} y = f(\mathbf{x}) + \epsilon \\ \epsilon \sim \mathcal{N}(0, \sigma^2) \end{cases} \quad (1.10)$$

Commonly, the objective function f is a black-box function. Therefore, regression algorithms construct a prediction model $\hat{f}$ by learning the relationship between $\mathbf{x}$ and y . Using $\hat{f}$, the predicted output $\hat{y}_*$ for a new input $\mathbf{x}_*$ can be expressed as follows

$$\hat{y}_* = \hat{f}(\mathbf{x}_*) \quad (1.11)$$

The outline of the regression algorithm is shown in Figure 1.5.

1.3.2 Machine Learning Algorithm and Chemistry

The use of machine learning in chemistry has its roots in the Bell-Evans-Polanyi (BEP) principle (1936) [43, 44] and the Hammett equation (1937) [45]. The BEP principle

is a linear relationship in the difference in reaction barrier for different reactions of the same type is proportional to the difference in the reaction enthalpy. The Hammett equation is also a linear relationship in which the logarithmic ratio of the equilibrium constants or reaction rate constants of two aromatic reactants with substituents in *meta* or *para* position is proportional. These two principles are one of the linear free energy relationships (LFERs) that predict the specific chemical property from the available data by linear regression.

Subsequently, quantitative structure-activity relationship (QSAR) [46] and quantitative structure-property relationship (QSPR) [47] have subsequently been developed to model qualitative relationships between compound structures and pharmacological activity or physical parameters. QSAR/QSPR constructs prediction models for activity and physical property using descriptors that represent various features of chemical structure. Subsequently, through an inverse analysis of the prediction model, it becomes possible to design compounds that exhibit desired activity and physical properties.

With the development of machine learning algorithms, reliable methods that are capable of regression and classification of non-linear and complex data have been used in chemistry. Early research used clustering [48, 49], neural networks [50, 51], support vector machines [51], decision forests [52], and random forests [53] to analyze and predict the objective valuable. In recent years, machine learning has been more widely used in organic chemistry. For instance, it has been used in predicting reaction yield [54–56], selectivity [57–60], designing optimal reaction conditions [61], proposing reaction pathways [62, 63], and performing retrosynthesis [64]. These successes prove that machine learning can significantly accelerate the discovery of new reactions that previously relied on the intuition and experiment of chemists.

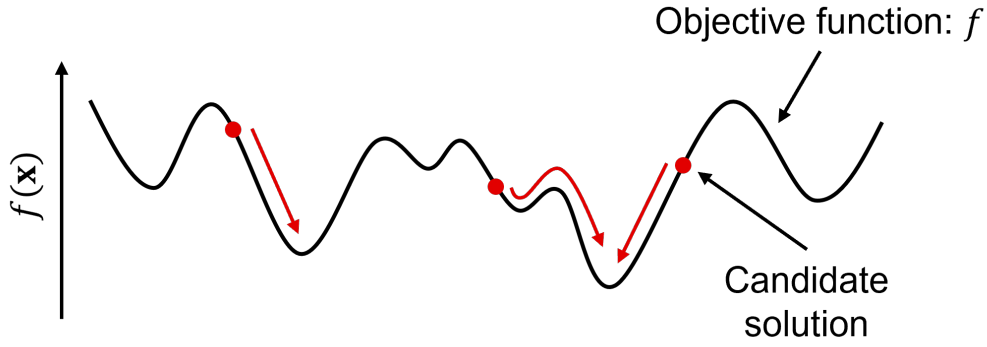


Figure 1.6. Conceptual diagram of the optimization algorithm. The search space is efficiently explored and the location of local and global minima are searched.

1.3.3 Optimization Algorithm

Optimization algorithms are techniques used to find the best solution or set of parameters for a given problem. These problems often involve the minimization or maximization of a black-box function (Figure 1.6).

Optimization algorithms can be broadly classified into continuous optimization algorithms and discrete optimization algorithms. Continuous optimization algorithms are methods that find the optimal solution from a continuous objective function and can be divided into methods that use differential information, e.g., gradient and Hessian, of the function and those that do not. The gradient descent method and Newton's method are representative examples of the former, and the simplex method and the golden section search method are representative examples of the latter. Discrete optimization algorithms, on the other hand, are methods for solving optimization problems whose solutions have a combinatorial structure, and there is a wide range of them, including dynamic programming, shortest path algorithms, and many metaheuristics.

1.3.4 Optimization Algorithm and Chemistry

In chemistry, there are various optimization problems, and various optimization algorithms are used to solve them. Geometry optimizations searching local minima (EQs) and saddle points (TSs) on the PES are continuous optimization problems. EQs can be found by using the steepest descent algorithm to change the molecular structure in the opposite direction of the gradient of the PES. In addition, the correct TSs can be found by using Newton’s method to search for a saddle point from a molecular structure close to the TS. Since the actual PES is a complex function in a $3N-6$ dimensional space, more advanced methods such as the conjugate gradient algorithm, the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm, and the quasi-Newton method are employed as optimization algorithms for practical use.

Discrete optimization algorithms also contribute to solving optimization problems in chemistry. For example, by considering the reaction path network as an undirected graph where EQs are vertices and TSs are edges, graph algorithms can be applied. For example, using Kruskal’s algorithm, one can find the minimum spanning tree of the reaction path network [65]. The Floyd-Warshall algorithm can also be used to find the shortest path between all EQs by constructing a weighted undirected graph considering the height of the reaction barrier for each elementary step as the weight of an edge [66].

Metaheuristics have been employed since the 20th century to accelerate *in silico* optimization. For example, the partial swarm algorithm is used for crystal structure prediction [67], molecular screening [68], and searching the minimum energy structure of carbon clusters [69]. The basing-hopping algorithm and the genetic algorithm are also employed to systematically explore molecular conformations and discover low energy structures [70–74].

Bayesian optimization has also emerged as a promising method to solve optimiza-

tion problems. Bayesian optimization can find optimal solutions with a small number of trials by balancing exploration and exploitation to explore the domain efficiently [75, 76]. Bayesian optimization has been used in many studies to design materials [77–79], catalysts [80], batteries [81], thin-film [82], optimize reaction conditions [83–86], and maximize yields [87–89] and selectivities [90]. Some studies combine robotic synthesizers and Bayesian optimization to discover the best reactions in a closed loop without the intervention of a chemist [79–82, 86, 88–90]. A unique aspect of the Bayesian optimization algorithm is its use of machine learning models as surrogates. In other words, Bayesian optimization is an algorithm that combines machine learning and optimization strategies. Details of Bayesian optimization are presented in Chapter 2.4.

As described above, optimization algorithms and chemistry are closely related. In particular, Bayesian optimization is considered a powerful method.

1.4 Purpose of the Study

Bayesian optimization can reduce the number of experiments required for reaction screening. Reaction barrier calculations can directly predict experimental reactivity and selectivity of chemical reactions by evaluating the Gibbs free energy change. These facts raise a specific question: "Can the combination of Bayesian optimization and barrier calculations further reduce the number of synthetic experiments required for reaction screening?"

In reaction screening, it is not difficult to utilize Bayesian optimization and barrier calculations simultaneously. For example, since the reaction barrier predicted by barrier calculations directly reflects the reactivity or selectivity of the reaction, scores of candidate reactions proposed by Bayesian optimization can be evaluated by barrier calculations instead of experiments. In other words, Bayesian optimization can be performed by utilizing

theoretical reactivity and selectivity data obtained from barrier calculations, rather than relying solely on experimental results. Therefore, there is no inherent issue in employing both experimental and computational data in Bayesian optimization. With this approach, it is intuitively expected that if the deviation between experimental and computational results is sufficiently small, the convergence of Bayesian optimization can be accelerated and the number of experiments required for reaction screening can be reduced by utilizing the computational results.

However, it is not clear whether such an approach can effectively reduce the number of experiments. Therefore, in this study, I developed a method to reduce the number of synthetic experiments by combining Bayesian optimization and barrier calculations. The performance of the developed method was benchmarked using a reaction barrier dataset to explore the feasibility of using barrier calculations to reduce the number of experiments required for reaction screening. Furthermore, the developed method was applied to actual organic reactions to investigate its applicability.

1.5 Overview of This Thesis

This thesis is composed of six chapters. Chapter 1 is a general introduction and reviewed the challenges in reaction screening. Quantum chemical calculations and informatics approach for organic reactions and the objectives of this study were also explained.

In Chapter 2, an overview of the methods employed in this study was provided. First, the concept of the AFIR method was introduced. Then, the algorithm for substituting atoms, adding substituents, and placing molecules for a model structure was explained. In addition, the concept and the algorithm of a genetic algorithm and Bayesian optimization were discussed. Finally, an overview of the Chemspeed ISYNTH robotic platform employed in organic synthesis was presented.

In Chapter 3, reaction barrier datasets that can benchmark the performance of optimization algorithms were created. Focusing on the Claisen rearrangement, 100,000 allyl vinyl ether molecules with different substituent combinations were generated. The reaction barriers for each molecule in the Claisen rearrangement were calculated at the semi-empirical PM7 level using the AFIR method. The substituent effects that lower the reaction barrier in the datasets were systematically analyzed, and it was shown that chemically and numerically appropriate reaction barriers were obtained. In addition, the performance of the genetic algorithm was evaluated using created datasets and compared to the performance of the random search. The results showed that created reaction barrier datasets of the Claisen rearrangement can be used to evaluate the performance of optimization algorithms.

In Chapter 4, a method to reduce the number of experiments required for substrate optimization was developed by combining Bayesian optimization and reaction barrier calculations. I hypothesized that the number of required experiments could be reduced by performing Bayesian optimization using theoretical reactivity data as well as experimental reactivity data. However, the relationship between the magnitude of computational deviation, computational batch size, and the number of experiments that can be reduced is not clear. Therefore, I benchmarked the optimization performance of the developed method through virtual experiments and virtual computations using the reaction barrier dataset of the Claisen rearrangement. The results indicated that even with a computational error of approximately 20 kJ/mol, the developed method outperformed the conventional Bayesian optimization.

In Chapter 5, the developed method was applied to maximize the regioselectivity of the thermal azide-alkyne cycloaddition. The thermal azide-alkyne cycloaddition is a powerful reaction to form triazoles from azides and alkynes, generally producing a mixture

of two isomers, i.e., 1,5-disubstituted isomer and 1,4-disubstituted isomer. To maximize the regioselectivity of the 1,5-disubstituted isomer, I optimized the combination of azide and alkyne substrates. Three substrate optimization algorithms that employ Bayesian optimization and quantum chemical calculation were used to optimize the combination of substrates. As a result, the combinations that show 74.7% 1,5-regioselectivity and 86.4% 1,4-regioselectivity were found.

Chapter 6 is a general conclusion, and I summarize this thesis and outline future directions.

References

- [1] C. Wentrup, *European Journal of Organic Chemistry* **2022**, 2022, e202101492.
- [2] F. Wöhler, *Annalen der Physik* **1828**, 88, 253–256.
- [3] R. B. Woodward, *Pure and Applied Chemistry* **1973**, 33, 145–178.
- [4] E. Schrödinger, *Annalen der Physik* **1926**, 384, 361–376.
- [5] M. Born, R. Oppenheimer, *Annalen der Physik* **1927**, 389, 457–484.
- [6] V. Fock, *Zeitschrift für Physik* **1930**, 61, 126–148.
- [7] J. C. Slater, *Physical Review* **1929**, 34, 1293–1322.
- [8] L. H. Thomas, *Mathematical Proceedings of the Cambridge Philosophical Society* **1927**, 23, 542–548.
- [9] P. Hohenberg, W. Kohn, *Physical Review* **1964**, 136, B864–B871.
- [10] W. Kohn, L. J. Sham, *Physical Review* **1965**, 140, A1133–A1138.
- [11] K. Fukui, *Accounts of Chemical Research* **1981**, 14, 363–368.

- [12] K. Ohno, S. Maeda, *Chemical Physics Letters* **2004**, *384*, 277–282.
- [13] S. Maeda, T. Taketsugu, K. Morokuma, K. Ohno, *Bulletin of the Chemical Society of Japan* **2014**, *87*, 1315–1334.
- [14] S. Maeda, K. Morokuma, *The Journal of Chemical Physics* **2010**, *132*, 241102.
- [15] S. Maeda, Y. Harabuchi, *WIREs Computational Molecular Science* **2021**, *11*, e1538.
- [16] K. Ohno, S. Maeda, *The Journal of Physical Chemistry A* **2006**, *110*, 8933–8941.
- [17] S. Maeda, K. Ohno, K. Morokuma, *Physical Chemistry Chemical Physics* **2013**, *15*, 3683.
- [18] Nobuaki. Koga, Keiji. Morokuma, *Chemical Reviews* **1991**, *91*, 823–842.
- [19] K. N. Houk, P. H.-Y. Cheong, *Nature* **2008**, *455*, 309–313.
- [20] W. Thiel, *Angewandte Chemie International Edition* **2014**, *53*, 8605–8613.
- [21] W. M. C. Sameera, S. Maeda, K. Morokuma, *Accounts of Chemical Research* **2016**, *49*, 763–773.
- [22] S. Maeda, K. Ohno, *Journal of the American Chemical Society* **2008**, *130*, 17228–17229.
- [23] P. J. Donoghue, P. Helquist, P.-O. Norrby, O. Wiest, *Journal of the American Chemical Society* **2009**, *131*, 410–411.
- [24] M. Hatanaka, S. Maeda, K. Morokuma, *Journal of Chemical Theory and Computation* **2013**, *9*, 2882–2886.
- [25] E. Hansen, A. R. Rosales, B. Tutkowski, P.-O. Norrby, O. Wiest, *Accounts of Chemical Research* **2016**, *49*, 996–1005.

- [26] Y. Guan, V. M. Ingman, B. J. Rooks, S. E. Wheeler, *Journal of Chemical Theory and Computation* **2018**, *14*, 5249–5261.
- [27] W. Matsuoka, Y. Harabuchi, S. Maeda, *ACS Catalysis* **2022**, *12*, 3752–3766.
- [28] W. Matsuoka, Y. Harabuchi, S. Maeda, *ACS Catalysis* **2023**, *13*, 5697–5711.
- [29] K. N. Houk, F. Liu, *Accounts of Chemical Research* **2017**, *50*, 539–543.
- [30] P. Friederich, G. d. P. Gomes, R. D. Bin, A. Aspuru-Guzik, D. Balcells, *Chemical Science* **2020**, *11*, 4584–4601.
- [31] S. Ahn, M. Hong, M. Sundararajan, D. H. Ess, M.-H. Baik, *Chemical Reviews* **2019**, *119*, 6509–6560.
- [32] A. R. Rosales, J. Wahlers, E. Limé, R. E. Meadows, K. W. Leslie, R. Savin, F. Bell, E. Hansen, P. Helquist, R. H. Munday, O. Wiest, P.-O. Norrby, *Nature Catalysis* **2019**, *2*, 41–45.
- [33] M. Foscato, V. R. Jensen, *ACS Catalysis* **2020**, *10*, 2354–2377.
- [34] T. Mita, Y. Harabuchi, S. Maeda, *Chemical Science* **2020**, *11*, 7569–7577.
- [35] M. Mukai, K. Nagao, S. Yamaguchi, H. Ohmiya, *Bulletin of the Chemical Society of Japan* **2022**, *95*, 271–277.
- [36] S. Yamaguchi, *Organic & Biomolecular Chemistry* **2022**, *20*, 6057–6071.
- [37] S. Gallarati, R. Fabregat, R. Laplaza, S. Bhattacharjee, M. D. Wodrich, C. Corminboeuf, *Chemical Science* **2021**, *12*, 6879–6889.
- [38] S. M. Maley, D.-H. Kwon, N. Rollins, J. C. Stanley, O. L. Sydora, S. M. Bischof, D. H. Ess, *Chemical Science* **2020**, *11*, 9665–9674.

- [39] Y. Sumiya, Y. Harabuchi, Y. Nagata, S. Maeda, *JACS Au* **2022**, *2*, 1181–1188.
- [40] H. Hayashi, H. Katsuyama, H. Takano, Y. Harabuchi, S. Maeda, T. Mita, *Nature Synthesis* **2022**, *1*, 804–814.
- [41] A. L. Samuel, *IBM Journal of Research and Development* **1959**, *3*, 210–229.
- [42] P. Langley, *Machine Learning* **2011**, *82*, 275–279.
- [43] R. P. Bell, C. N. Hinshelwood, *Proceedings of the Royal Society of London. Series A - Mathematical and Physical Sciences* **1997**, *154*, 414–429.
- [44] M. G. Evans, M. Polanyi, *Transactions of the Faraday Society* **1936**, *32*, 1333–1360.
- [45] L. P. Hammett, *Journal of the American Chemical Society* **1937**, *59*, 96–103.
- [46] R. Perkins, H. Fang, W. Tong, W. J. Welsh, *Environmental Toxicology and Chemistry* **2003**, *22*, 1666–1679.
- [47] T. Le, V. C. Epa, F. R. Burden, D. A. Winkler, *Chemical Reviews* **2012**, *112*, 2889–2919.
- [48] C. Hansch, S. H. Unger, A. B. Forsythe, *Journal of Medicinal Chemistry* **1973**, *16*, 1217–1222.
- [49] B. R. Kowalski, C. F. Bender, *Journal of the American Chemical Society* **1972**, *94*, 5632–5639.
- [50] S. A. Hiller, V. E. Golender, A. B. Rosenblit, L. A. Rastrigin, A. B. Glaz, *Computers and Biomedical Research* **1973**, *6*, 411–421.
- [51] S. Doniger, T. Hofmann, J. Yeh, *Journal of Computational Biology* **2002**, *9*, 849–864.

- [52] W. Tong, H. Hong, H. Fang, Q. Xie, R. Perkins, *Journal of Chemical Information and Computer Sciences* **2003**, *43*, 525–531.
- [53] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, B. P. Feuston, *Journal of Chemical Information and Computer Sciences* **2003**, *43*, 1947–1958.
- [54] A. M. Żurański, J. I. Martinez Alvarado, B. J. Shields, A. G. Doyle, *Accounts of Chemical Research* **2021**, *54*, 1856–1865.
- [55] K. Wu, A. G. Doyle, *Nature Chemistry* **2017**, *9*, 779–784.
- [56] J. M. Granda, L. Donina, V. Dragone, D.-L. Long, L. Cronin, *Nature* **2018**, *559*, 377–381.
- [57] J. P. Reid, M. S. Sigman, *Nature* **2019**, *571*, 343–348.
- [58] A. Milo, E. N. Bess, M. S. Sigman, *Nature* **2014**, *507*, 210–214.
- [59] Z. L. Niemeyer, A. Milo, D. P. Hickey, M. S. Sigman, *Nature Chemistry* **2016**, *8*, 610–617.
- [60] A. F. Zahrt, J. J. Henle, B. T. Rose, Y. Wang, W. T. Darrow, S. E. Denmark, *Science* **2019**, *363*, eaau5631.
- [61] H. Gao, T. J. Struble, C. W. Coley, Y. Wang, W. H. Green, K. F. Jensen, *ACS Central Science* **2018**, *4*, 1465–1476.
- [62] J. N. Wei, D. Duvenaud, A. Aspuru-Guzik, *ACS Central Science* **2016**, *2*, 725–732.
- [63] C. W. Coley, R. Barzilay, T. S. Jaakkola, W. H. Green, K. F. Jensen, *ACS Central Science* **2017**, *3*, 434–443.
- [64] M. H. S. Segler, M. Preuss, M. P. Waller, *Nature* **2018**, *555*, 604–610.

- [65] T. Yoshimura, S. Maeda, T. Taketsugu, M. Sawamura, K. Morokuma, S. Mori, *Chemical Science* **2017**, *8*, 4475–4488.
- [66] Y. Sumiya, Y. Tabata, S. Maeda, *ChemSystemsChem* **2020**, *2*, e1900022.
- [67] Y. Wang, J. Lv, L. Zhu, Y. Ma, *Physical Review B* **2010**, *82*, 094116.
- [68] R. Winter, F. Montanari, A. Steffen, H. Briem, F. Noé, D.-A. Clevert, *Chemical Science* **2019**, *10*, 8016–8024.
- [69] G. Jana, A. Mitra, S. Pan, S. Sural, P. K. Chattaraj, *Frontiers in Chemistry* **2019**, *7*.
- [70] A. Nayeem, J. Vila, H. A. Scheraga, *Journal of Computational Chemistry* **1991**, *12*, 594–605.
- [71] D. J. Wales, J. P. K. Doye, *The Journal of Physical Chemistry A* **1997**, *101*, 5111–5116.
- [72] N. Nair, J. M. Goodman, *Journal of Chemical Information and Computer Sciences* **1998**, *38*, 317–320.
- [73] M. J. Vainio, M. S. Johnson, *Journal of Chemical Information and Modeling* **2007**, *47*, 2462–2474.
- [74] A. Supady, V. Blum, C. Baldauf, *Journal of Chemical Information and Modeling* **2015**, *55*, 2338–2348.
- [75] E. Brochu, V. M. Cora, N. de Freitas, *A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning*, **2010**.

- [76] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, N. De Freitas, *Proceedings of the IEEE* **2016**, *104*, 148–175.
- [77] H. C. Herbol, M. Poloczek, P. Clancy, *Materials Horizons* **2020**, *7*, 2113–2123.
- [78] T. Ueno, T. D. Rhone, Z. Hou, T. Mizoguchi, K. Tsuda, *Materials Discovery* **2016**, *4*, 18–21.
- [79] S. Langner, F. Häse, J. D. Perea, T. Stubhan, J. Hauch, L. M. Roch, T. Heumueller, A. Aspuru-Guzik, C. J. Brabec, *Advanced Materials* **2020**, *32*, 1907801.
- [80] B. Burger, P. M. Maffettone, V. V. Gusev, C. M. Aitchison, Y. Bai, X. Wang, X. Li, B. M. Alston, B. Li, R. Clowes, N. Rankin, B. Harris, R. S. Sprick, A. I. Cooper, *Nature* **2020**, *583*, 237–241.
- [81] P. M. Attia, A. Grover, N. Jin, K. A. Severson, T. M. Markov, Y.-H. Liao, M. H. Chen, B. Cheong, N. Perkins, Z. Yang, P. K. Herring, M. Aykol, S. J. Harris, R. D. Braatz, S. Ermon, W. C. Chueh, *Nature* **2020**, *578*, 397–402.
- [82] B. P. MacLeod, F. G. L. Parlane, T. D. Morrissey, F. Häse, L. M. Roch, K. E. Dettelbach, R. Moreira, L. P. E. Yunker, M. B. Rooney, J. R. Deeth, V. Lai, G. J. Ng, H. Situ, R. H. Zhang, M. S. Elliott, T. H. Haley, D. J. Dvorak, A. Aspuru-Guzik, J. E. Hein, C. P. Berlinguette, *Science Advances* **2020**, *6*, eaaz8867.
- [83] F. Häse, L. M. Roch, C. Kreisbeck, A. Aspuru-Guzik, *ACS Central Science* **2018**, *4*, 1134–1145.
- [84] B. J. Shields, J. Stevens, J. Li, M. Parasram, F. Damani, J. I. M. Alvarado, J. M. Janey, R. P. Adams, A. G. Doyle, *Nature* **2021**, *590*, 89–96.
- [85] A. M. Schweidtmann, A. D. Clayton, N. Holmes, E. Bradford, R. A. Bourne, A. A. Lapkin, *Chemical Engineering Journal* **2018**, *352*, 277–282.

- [86] N. Sugisawa, H. Sugisawa, Y. Otake, R. V. Krems, H. Nakamura, S. Fuse, *Chemistry-Methods* **2021**, *1*, 484–490.
- [87] N. V. Faurschou, R. H. Taaning, C. M. Pedersen, *Chemical Science* **2023**, *14*, 6319–6329.
- [88] M. Christensen, L. P. E. Yunker, F. Adediji, F. Häse, L. M. Roch, T. Gensch, G. dos Passos Gomes, T. Zepel, M. S. Sigman, A. Aspuru-Guzik, J. E. Hein, *Communications Chemistry* **2021**, *4*, 1–12.
- [89] N. H. Angello, V. Rathore, W. Beker, A. Wołos, E. R. Jira, R. Roszak, T. C. Wu, C. M. Schroeder, A. Aspuru-Guzik, B. A. Grzybowski, M. D. Burke, *Science* **2022**, *378*, 399–405.
- [90] A. M. K. Nambiar, C. P. Breen, T. Hart, T. Kulesza, T. F. Jamison, K. F. Jensen, *ACS Central Science* **2022**, *8*, 825–836.

Chapter 2

Methods

2.1 Artificial Force Induced Reaction Method

The artificial force induced reaction (AFIR) method is a technique to induce conformational changes of chemical structures in a simulation by applying virtual external forces between molecular substructures called fragments [1, 2]. A systematic search for structural changes in a chemical reaction using the AFIR method reveals the entire reaction path network, including many equilibrium structures (EQs), transition states (TSs), and dissociation channels (DCs). This enables us to predict the reactivity and selectivity of chemical reactions by quantum chemical calculations without experiments. In this study, the AFIR method was used to systematically search for chemical reaction pathways and calculate reaction barriers to predict theoretical reactivity and selectivity.

2.1.1 AFIR Function

The idea of the AFIR method is simple. It is to induce chemical reactions by pushing and pulling fragments apart. As an example, consider a model chemical reaction where two fragments A and B react to form a product AB. Normally, there is a reaction barrier between $A + B$ and AB. Therefore, even if a geometry optimization is performed from

A + B, the reaction barrier cannot be overcome and AB cannot be reached (Figure 2.1). Therefore, in the AFIR method, an artificial force is applied to push A and B together. This increases the energy of A + B on the potential energy surface (PES) and eliminates the reaction barrier. By performing the geometry optimization on the PES with such an artificial force applied, a reaction pathway from A + B to AB is obtained (Figure 2.1). The strength of the artificial force is a function directly proportional to the distance r_{AB} between A and B. In the more general case, the artificial force term is given by the following equation:

$$F^{\text{AFIR}}(\mathbf{Q}) = E(\mathbf{Q}) + \rho\alpha \frac{\sum_{i \in A} \sum_{j \in B} \omega_{ij} r_{ij}}{\sum_{i \in A} \sum_{j \in B} \omega_{ij}} \quad (2.1)$$

where $E(\mathbf{Q})$ is the potential energy of the atomic coordinates $\mathbf{Q}$. ρ is set to 1.0 or -1.0. If $\rho = 1.0$, an attractive (pushing) force is applied, and if $\rho = -1.0$, a repulsive (pulling) force is applied. α is the strength of the artificial force, which is described in detail below. The second term is given as a weighted sum of the distance r_{ij} between i th and j th atoms, and the weight ω_{ij} is given as

$$\omega_{ij} = \left[\frac{R_i + R_j}{r_{ij}} \right]^6 \quad (2.2)$$

where R_i and R_j are the covalent radii of i th and j th atoms, respectively. The force strength parameter α is obtained using the following equation:

$$\alpha = \frac{\gamma}{\left[2^{-\frac{1}{6}} - \left(1 + \sqrt{1 + \frac{\gamma}{\epsilon}} \right)^{-\frac{1}{6}} \right] R_0} \quad (2.3)$$

where R_0 and ϵ are the Ar-Ar Lennard-Jones parameters, which are set to 3.8164 Å and 1.0061 kJ/mol, respectively. α corresponds to the mean force that an Ar-Ar pair feels as it moves from the minimum point to the turning point in its direct collision with the collision energy γ . Thus, γ is called the model collision energy parameter and corresponds to an

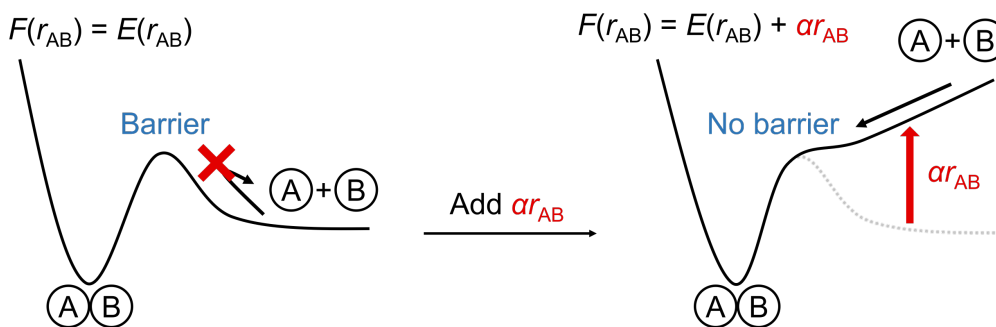


Figure 2.1. Potential energy surface and AFIR path

approximate upper limit of the barrier that the AFIR method can overcome. Thus, by adding artificial forces to the PES and virtually deleting TS, the products can be obtained efficiently. In practice, however, the information on TSs between reactants and products is most important to investigate reaction rates, barriers, mechanisms, and so on. Removing TS from the PES may seem to be contradictory to wanting to obtain TS. However, since $F(\mathbf{Q})$ contains the potential energy $E(\mathbf{Q})$ in its expression, it is possible to reconstruct $E(\mathbf{Q})$ by removing the artificial force term from $F(\mathbf{Q})$, and the energy maxima and minima can be obtained as approximate TS and approximate product structures, respectively.

2.1.2 AFIR Path

The reaction path obtained by the AFIR method is called the AFIR path. Since the AFIR path is located close to the intrinsic reaction coordinate (IRC) path [3], a local maximum and a local minimum on the AFIR path can be regarded as approximate structures of the product and TS. By using the quasi-Newton method, the actual TS structure can be obtained from the approximate TS. One can also find the reactant and product connected by TS by searching the IRC path from the TS. However, if the value of γ is inappropriate, the approximate TS structure will be far from the actual TS, and the search will not be successful. In such cases, the locally update plane (LUP) method

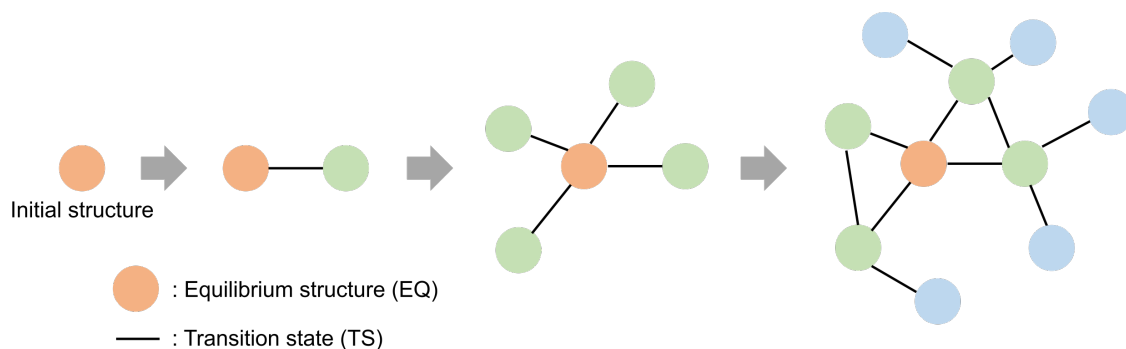


Figure 2.2. Schematic diagram of systematic reaction path search by the AFIR method

is used to relax the AFIR path and more accurately reproduce the IRC path [4].

2.1.3 Systematic Reaction Path Search

The AFIR method enables an automatic search of reaction paths from one EQ to another. By executing this process sequentially, a systematic reaction path search is possible. Specifically, since there are multiple possible fragment definitions for an EQ, reaction paths from one EQ to multiple EQs can be obtained by automatically defining different fragments for an EQ and sequentially applying the AFIR method. Furthermore, it is also possible to define fragments for newly obtained EQs and apply the AFIR method to them. Therefore, by repeating this process one after another, an entire reaction path network can be obtained (Figure 2.2). Recently, the rate constant matrix contraction (RCMC) method has been used to search only for reaction paths that proceed within a certain time scale, but the details of the RCMC method are omitted here [5].

2.2 Geometry Generation Algorithm

In computational chemistry, the generation of molecular structures to be computed is sometimes a time-consuming task. This is especially critical in cases where systematic reaction screening is required. For example, in the case of screening combinations of metal

ions, ligands, substituents, bases, etc., the number of possible combinations is enormous, and it is impractical to manually construct model reaction systems for all of them. In such cases, it would be very useful to have an algorithm that can automatically generate molecular geometries in model reaction systems. Although there is already some geometry generation software, such as molSimplify [6], these methods do not fully implement the functions needed in my research. Therefore, I implemented a simple original geometry generation algorithm. This program provides three functions: substitution of atoms, addition of substituents, and placement of molecules to a model reaction system. The program is implemented to generate geometries with minimal steric hindrance. Therefore, it does not consider changes in interatomic distances due to substitutions of atoms. It also does not consider changes in non-covalent interactions, e.g., formation of hydrogen bonds, C-H bonds, and π - π stackings, due to the addition of substituents or molecules. Therefore, the program is set up to automatically perform geometry optimization from the structures generated by my algorithm.

2.2.1 Input

The user creates an input file specifying the atomic coordinates of a model system with X atoms at the site of manipulation. For each X atom, the user specifies one of the following modes: substitution of atoms (A), addition of substituents (S), or placement of molecules (M). For mode S, the atom to which the substituent is attached must also be specified. A more advanced option is also implemented to consider multiple X atoms as a group and avoid duplicate combinations within the group. A list of candidate atoms, substituents, and molecules to be placed on each X atom is then specified. Substituents must always be considered in the form neutralized by the X atom. The coordinates of the X atom used for neutralization and the coordinates of the substituent must be written in

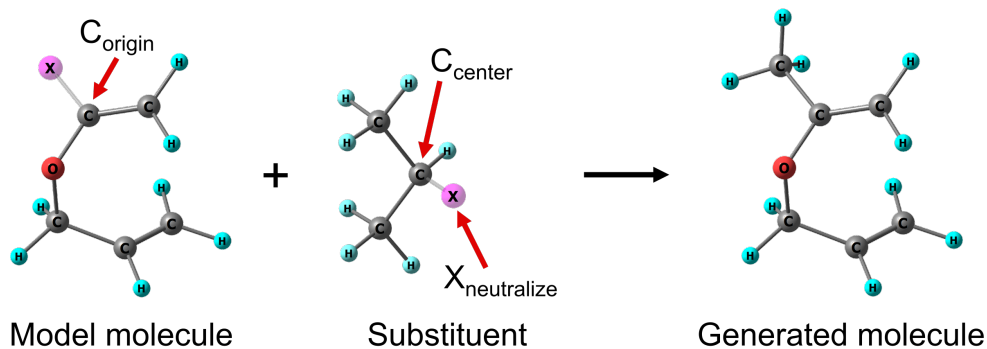


Figure 2.3. Addition of a substituent to a model molecule.

this order in the input.

The user can specify whether the geometry optimization for the generated structure should be performed automatically. Whether or not atoms unrelated to the substitution operation are treated as fixed atoms in the calculation depends on the options. The user can also specify the machine and computation nodes on which the calculations are executed. When performing geometry optimization, input and intermediate files with the level of theory and options to be used must be prepared in advance.

2.2.2 Algorithm

The program reads an input file created by the user. For each X atom, a list of atomic coordinates of all possible atoms, substituents, and molecules that could be placed at that X atom is defined. The number of possible combinations is the product of the number of all candidates for each X atom. In my implementation, all combinations are searched using depth-first search. If a particular combination is specified in the program arguments, only the corresponding combination is considered for geometry generation.

Substitution of atoms can be easily performed by simply replacing the string of atom names in the atomic coordinates of the X atom with the name of the new atom. Adding substituents, on the other hand, requires a slightly more complex procedure. As a simple

example, consider the case of introducing an isopropyl group ($\text{CH}(\text{CH}_3)_2$) at the position of an X atom in a model molecule (Figure 2.3)

The carbon atom to which the X atom of the model molecule is bonded is called C_{origin} . The X atom moves to a position where the distance between the X atom and the C_{origin} atom is twice the covalent bond radius of the carbon atom. This is done by considering the vector equation of the point P on the line passing through the two points of the X atom and the C_{origin} atom:

$$P = (1 - t)\text{C}_{\text{origin}} + tX \quad (2.4)$$

and adjusting the mediating variable t to an appropriate value so that the distance between point P and C_{origin} is twice the covalent bond radius of the carbon atom. The entire substituent geometry is then translated so that the C_{center} and X atoms overlap.

After that, the direction of the $\text{CH}(\text{CH}_3)_2$ is adjusted. This manipulation is done in two steps: 1. optimization of the substituent bond direction and 2. maximization of the sum of the distance between the substituent and the other atoms. In the optimization of the bonding direction of substituent, first, all atoms are moved so that C_{center} is at the origin, i.e., $(0, 0, 0)$. Next, a combination of rotation matrices around the x -, y -, and z -axes are acted on the substituent's coordinates to determine the rotation angle that minimizes the angle between the $\text{C}_{\text{center}}\text{--X}_{\text{neutralize}}$ bond and the $\text{C}_{\text{center}}\text{--C}_{\text{origin}}$ bond. The rotation

matrices around each axis are expressed by the following equations, respectively.

$$R_x(\theta) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos \theta & -\sin \theta \\ 0 & \sin \theta & \cos \theta \end{bmatrix} \quad (2.5)$$

$$R_y(\theta) = \begin{bmatrix} \cos \theta & 0 & \sin \theta \\ 0 & 1 & 0 \\ -\sin \theta & 0 & \cos \theta \end{bmatrix} \quad (2.6)$$

$$R_z(\theta) = \begin{bmatrix} \cos \theta & -\sin \theta & 0 \\ \sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (2.7)$$

This operation optimizes the bonding direction of the $\text{CH}(\text{CH}_3)_2$ group. $\text{X}_{\text{neutralize}}$ used to neutralize $\text{CH}(\text{CH}_3)_2$ is then removed. Next, substituents are rotated around the $\text{C}_{\text{center}}-\text{C}_{\text{origin}}$ axis to find the angle that maximizes the sum of the distance between the substituent and the other atoms. The rotation matrix around any axis $\mathbf{n} = (n_x, n_y, n_z)$, which satisfies $n_x^2 + n_y^2 + n_z^2 = 1$ in unit vector, is represented by the following equation.

$$R_{\mathbf{n}}(\theta) = \begin{bmatrix} \cos \theta + n_x^2(1 - \cos \theta) & n_x n_y(1 - \cos \theta) - n_z \sin \theta & n_z n_x(1 - \cos \theta) + n_y \sin \theta \\ n_x n_y(1 - \cos \theta) + n_z \sin \theta & \cos \theta + n_y^2(1 - \cos \theta) & n_y n_z(1 - \cos \theta) - n_x \sin \theta \\ n_z n_x(1 - \cos \theta) - n_y \sin \theta & n_y n_z(1 - \cos \theta) + n_x \sin \theta & \cos \theta + n_z^2(1 - \cos \theta) \end{bmatrix} \quad (2.8)$$

The above three operations, translation, bond direction optimization, and distance matrix maximization, enable the addition of substituents at appropriate positions. The above operations are summarized in Figure 2.4. The placement of molecules can likewise be performed by searching for the angle that maximizes the sum of the distance to the surrounding atoms.

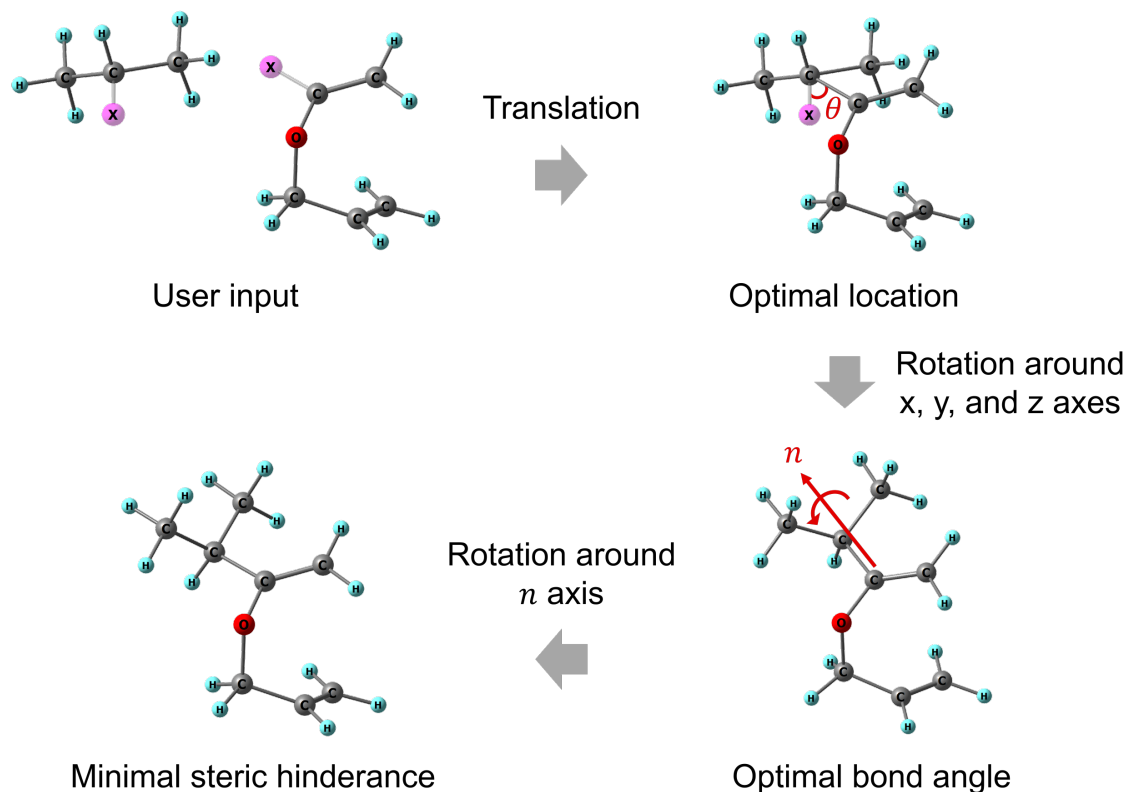


Figure 2.4. Scheme to add a substituent to a model molecule.

2.3 Genetic Algorithm

A genetic algorithm is an optimization algorithm that searches for the optimal solution to a complex optimization problem by mimicking the mechanism of natural selection through heredity and survival of the fittest in biological evolution [7]. In genetic algorithms, multiple candidate solutions are maintained as a population. Each candidate solution is called an individual and the component of each individual is called a gene. Genetic operations called selection, crossover, and mutation are repeatedly applied to this population to search for the solution that maximizes or minimizes the objective variable. This process is a multivariate optimization process that generates multiple candidate solutions in a multivariate space one after another, gradually narrowing the search space to a neighborhood of the optimal solution. The genetic algorithm is a metaheuristic that

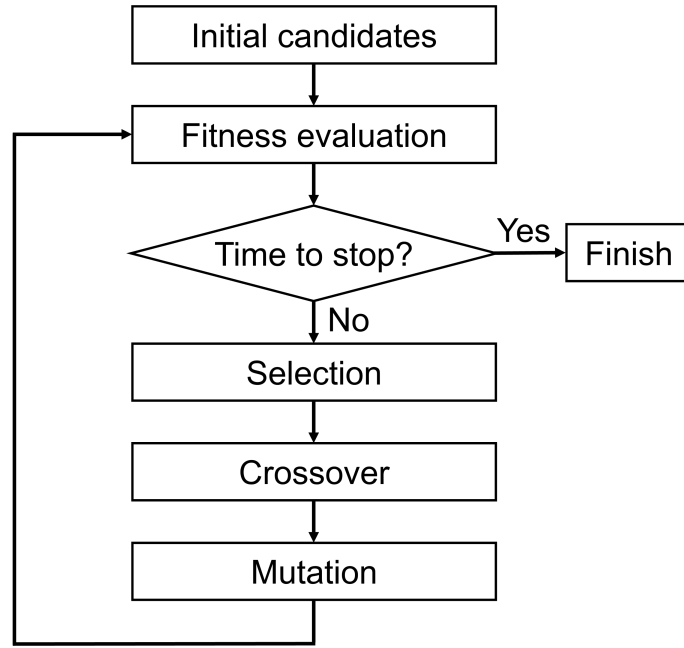


Figure 2.5. Flow of the genetic algorithm.

can be applied to a variety of problems, such as combinatorial optimization and NP-hard problems.

2.3.1 Algorithm

The algorithm of the genetic algorithm is shown in Figure 2.5. The initial population is generated using either randomly generated individuals or individuals generated by some algorithm to maximize diversity. In the evaluation of the fitness, the score of the objective variable for each individual is investigated.

Selection is an operation that models natural selection in living organisms. In nature, individuals that are adapted to their environment survive and multiply. On the other hand, individuals that are poorly adapted to the environment are eliminated and disappear from the population. Thus, the selection operation increases the number of individuals with high fitness for the objective function and decreases the number of individuals with low fitness. Typical selection methods include roulette selection, ranking selection, tournament

selection, and elite selection. In selection, it is important to maintain the diversity of individuals and genes in the population without removing too many individuals with low fitness. If only a small number of individuals are selected early in the optimization process, the diversity of candidate solutions within the population is lost. In this case, the subsequent crossover operation will not work effectively, leading to the premature convergence of the genetic algorithm.

Crossover is an operation that models the mating of organisms. In nature, offspring are born inheriting half of their parents' genes. Crossover mimics this mechanism by swapping parts of two individuals selected from a population with each other to create a new individual that did not exist in the population. The idea behind crossover is the simple assumption that "a solution that combines a good solution with a good solution will be a better solution". Typical crossover methods include one-point crossover, two-point crossover, uniform crossover, average crossover, arithmetic crossover, blended crossover, unimodal normal distribution crossover, and unitary crossover. In genetic algorithms, crossover may be the most crucial genetic operation. The optimization performance of the genetic algorithm depends largely on whether the search can maintain the genetic diversity of the population and perform crossover effectively.

In general, selection and crossover alone are insufficient to find the optimal solution. This is because selection and crossover alone produce only a limited range of individuals that are dependent on the parent individuals. Therefore, by mutation, a gene of an individual is randomly changed to a new gene with a low probability. Mutation prevents the genetic algorithm from getting stuck in a local minimum by maintaining genetic diversity. If the probability of mutation is too low, it leads to local convergence, and if it is too high, it increases the number of iterations required for convergence.

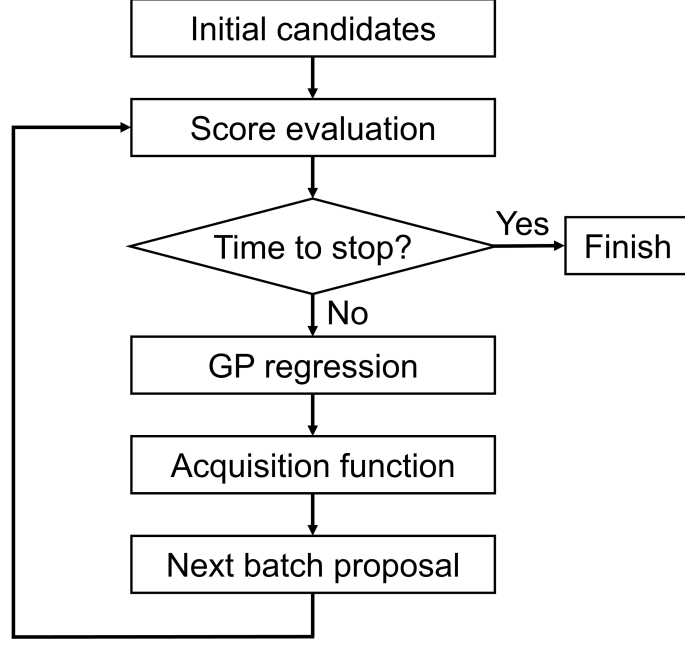


Figure 2.6. Flow of Bayesian optimization.

2.4 Bayesian Optimization

Bayesian optimization [8] is an optimization algorithm that search for an input $\mathbf{x}^*$ that maximizes or minimizes the black-box objective function f :

$$\mathbf{x}^* = \operatorname{argmax}_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) \quad (2.9)$$

where, $\mathbf{x}^*$ is the input that optimizes the objective function f and $\mathcal{X}$ denotes the set of possible inputs. Bayesian optimization is designed to examine the search space by balancing exploitation, which focuses on areas where data observations are sufficient, and exploration, which seeks areas where data observations are insufficient. This strategy enables Bayesian optimization to find optimal solutions with a small number of data and is particularly effective for solving optimization problems with high evaluation costs.

The algorithm of Bayesian optimization is illustrated in Figure 2.6. In Gaussian process (GP) regression, a Gaussian process is used as a surrogate model to predict the objective variable and evaluate its uncertainty. The acquisition function is utilized to

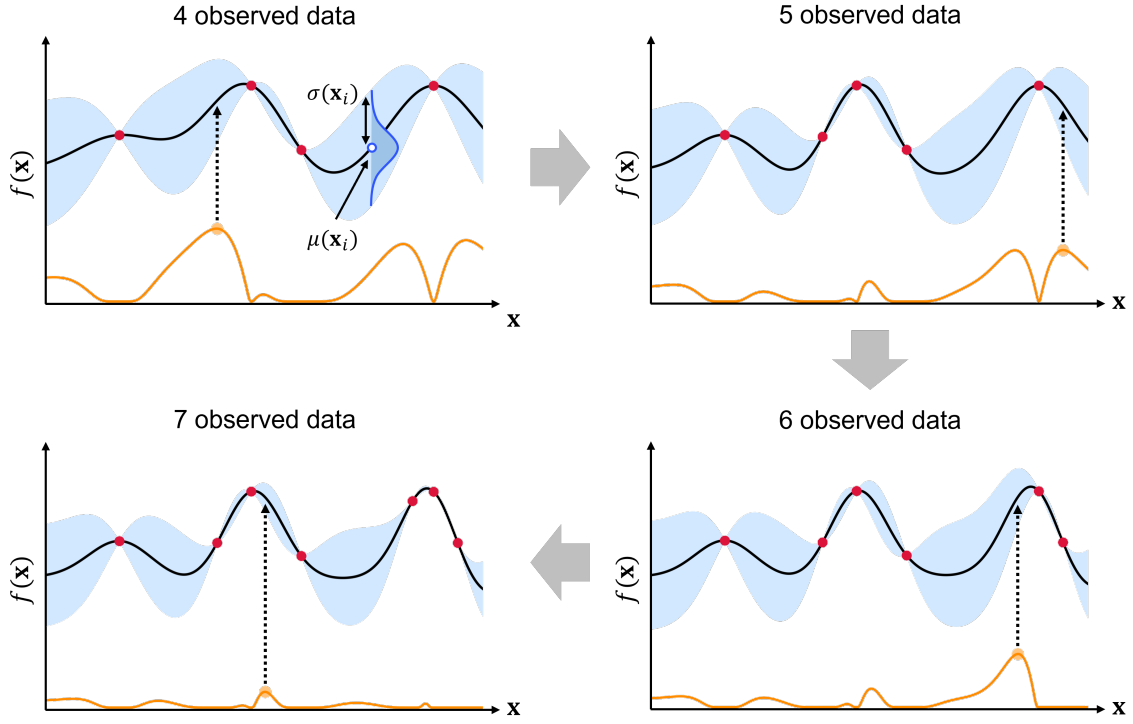


Figure 2.7. Example of using Bayesian optimization on a 1D search space. Red circles denote observed data points. The predictions of the GP regression are shown as black lines, and the uncertainty of the predictions (confidence interval) is shown as blue areas. The orange line shows the value of the acquisition function for each input. The input that maximizes the acquisition function is treated as the next search point.

assess the acquisition function value of each candidate solution and the candidate solution with the highest acquisition function value is typically selected as the next search point. An example of Bayesian optimization behavior for a one-dimensional input is shown in Figure 2.7. In the following sections, a detailed explanation of the surrogate model and the acquisition function is provided. Furthermore, parallelization methods for simultaneously proposing multiple candidate solutions in Bayesian optimization are explained.

2.4.1 Algorithm

2.4.1.1 Surrogate Model

A surrogate model approximates the behavior of a complex and computationally expensive black-box function. In Bayesian optimization, the surrogate model estimates the objective function f based on available observational data and predicts the output $f(\mathbf{x}_*)$ for new input $\mathbf{x}_*$. The most commonly used model for the surrogate model is GP regression. GP regression is a machine learning regression algorithm that assumes a Gaussian process for the prior distribution of the objective function f .

In GP regression, the posterior and predictive distributions of the Gaussian process are computed using observed data to estimate the expected value $\mu(\mathbf{x}_*)$ and the uncertainty of the prediction $\sigma(\mathbf{x}_*)$ for the new input $\mathbf{x}_*$. A Gaussian process is an infinite-dimensional Gaussian distribution. For any n , if the output vector for the input $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$:

$$\mathbf{f} = (f(\mathbf{x}_1), f(\mathbf{x}_2), \dots, f(\mathbf{x}_n)) \quad (2.10)$$

follows a Gaussian distribution $\mathcal{N}(\mu, \mathbf{K})$ with mean $\mu = (\mu(\mathbf{x}_1), \mu(\mathbf{x}_2), \dots, \mu(\mathbf{x}_n))$ and a covariance matrix $\mathbf{K}$ with elements $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$, then f is said to follow a Gaussian process denoted as:

$$f \sim \text{GP}(\mu(\mathbf{x}), k(\mathbf{x}_i, \mathbf{x}_j)) \quad (2.11)$$

In many cases, the mean function $\mu(\mathbf{x})$ is assumed to be a zero vector:

$$f \sim \text{GP}(0, k(\mathbf{x}_i, \mathbf{x}_j)) \quad (2.12)$$

In the following discussion, I present the equations under this assumption.

In general, the observed data y_i is obtained as the output $f(\mathbf{x}_i)$ of the function f plus the observation noise. Specifically, the observed data y_i is the true value $f(\mathbf{x}_i)$ derived

from f plus observation noise ϵ_i . It is commonly assumed that the observation error ϵ_i follows a normal distribution with mean 0 and variance σ^2 .

$$\begin{cases} y_i = f(\mathbf{x}_i) + \epsilon_i \\ \epsilon_i \sim \mathcal{N}(0, \sigma^2) \end{cases} \quad (2.13)$$

Here, given a set of observations:

$$\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\} \quad (2.14)$$

where

$$\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\} \quad (2.15)$$

is the input and

$$\mathbf{y} = \{y_1, y_2, \dots, y_n\} \quad (2.16)$$

is the corresponding output, the probability of observing $\mathbf{y}$ from the function f can be expressed as:

$$p(\mathbf{y} \mid \mathbf{f}) = \mathcal{N}(\mathbf{f}, \sigma^2 \mathbf{I}) \quad (2.17)$$

where $\mathbf{I}$ is the identity matrix. Thus, the probability of observing the output $\mathbf{y}$ for the input $\mathbf{X}$ can be expressed as:

$$p(\mathbf{y} \mid \mathbf{X}) = \int p(\mathbf{y}, \mathbf{f} \mid \mathbf{X}) d\mathbf{f} \quad (2.18)$$

$$= \int p(\mathbf{y} \mid \mathbf{f}) p(\mathbf{f} \mid \mathbf{X}) d\mathbf{f} \quad (2.19)$$

$$= \int \mathcal{N}(\mathbf{y} \mid \mathbf{f}, \sigma^2 \mathbf{I}) \mathcal{N}(\mathbf{f} \mid \mu, \mathbf{K}) d\mathbf{f} \quad (2.20)$$

$$= \mathcal{N}(\mu, \mathbf{K} + \sigma^2 \mathbf{I}) \quad (2.21)$$

This equation shows that when function f follows a Gaussian process, output $\mathbf{y}$ also follows

a Gaussian process. In addition, adding σ^2 to the diagonal elements of the covariance matrix $\mathbf{K}$ allows for the incorporation of the influence of observation noise.

The equation for predicting the output y_* at an unobserved $\mathbf{x}_*$ using GP regression is then described. In GP regression, the predictive distribution for y_* is expressed as a conditional probability:

$$p(y_* \mid \mathbf{X}, \mathbf{y}, x_*) = \mathcal{N}(\mu_*, \sigma_*^2) \quad (2.22)$$

Given a set of the training input $\mathbf{X}$ and the corresponding output $\mathbf{y}$, the mean function at point $\mathbf{x}_*$ is given by

$$\mu(\mathbf{x}_*) = \mathbf{k}_*^\top (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{y} \quad (2.23)$$

where $\mathbf{k}_*^\top$ is the vector of covariance function evaluations between $\mathbf{x}_*$ and each training input,

$$\mathbf{k}_*^\top = (k(\mathbf{x}_*, \mathbf{x}_1), k(\mathbf{x}_*, \mathbf{x}_2), \dots, k(\mathbf{x}_*, \mathbf{x}_n)) \quad (2.24)$$

and $\mathbf{K}$ is the $n \times n$ covariance matrix of the training inputs $\mathbf{X}$ obtained using a kernel function $k(\mathbf{x}_i, \mathbf{x}_j)$, and σ^2 is the noise variance. The variance function at point $\mathbf{x}_*$ is given by

$$\sigma^2(\mathbf{x}_*) = k_{**} - \mathbf{k}_*^\top (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{k}_* \quad (2.25)$$

where k_{**} is the covariance function evaluation at $\mathbf{x}_*$ itself,

$$k_{**} = k(\mathbf{x}_*, \mathbf{x}_*) \quad (2.26)$$

The kernel function represents the covariance between input $\mathbf{x}_i$ and $\mathbf{x}_j$

$$k(\mathbf{x}_i, \mathbf{x}_j) = \text{cov}(f(\mathbf{x}_i), f(\mathbf{x}_j)) \quad (2.27)$$

where $f(\mathbf{x}_i)$ and $f(\mathbf{x}_j)$ are the function value at $\mathbf{x}_i$ and $\mathbf{x}_j$. Commonly used kernel functions are the radial basis function (RBF) kernel function,

$$k_{\text{RBF}}(\mathbf{x}_i, \mathbf{x}_j) = \theta_1 \exp\left(-\frac{1}{2}r^2\right) \quad (2.28)$$

the Matérn1 kernel function,

$$k_{\text{matérn1}}(\mathbf{x}_i, \mathbf{x}_j) = \theta_1 \exp(-r) \quad (2.29)$$

the Matérn3 kernel function,

$$k_{\text{matérn3}}(\mathbf{x}_i, \mathbf{x}_j) = \theta_1 \left(1 + \sqrt{3}r\right) \exp\left(-\sqrt{3}r\right) \quad (2.30)$$

and the Matérn5 kernel function,

$$k_{\text{matérn5}}(\mathbf{x}_i, \mathbf{x}_j) = \theta_1 \left(1 + \sqrt{5}r + \frac{5}{3}r^2\right) \exp\left(-\sqrt{5}r\right) \quad (2.31)$$

where θ_1 represents the output scale parameter, which governs the magnitude of the change in prediction assessed by GP regression. r^2 is the square of Euclidean distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ scaled by the length scale parameter:

$$r^2 = (\mathbf{x}_i - \mathbf{x}_j)^\top \Theta^{-2} (\mathbf{x}_i - \mathbf{x}_j) \quad (2.32)$$

where Θ is a $d \times d$ diagonal matrix (d is the number of dimensions of $\mathbf{x}$) of the length scale parameters:

$$\Theta = \begin{pmatrix} \theta_{21} & 0 & \cdots & 0 \\ 0 & \theta_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \theta_{2d} \end{pmatrix} \quad (2.33)$$

When the automatic relevance determination (ARD) kernel is used, Θ becomes a non-

constant diagonal matrix and each dimension of the input $\mathbf{x}$ has a unique length scale parameter denoted as θ_{2j} for the j th dimension. Since the length scale parameter controls the smoothness and scale of variations in the function being modeled in GP regression, using the ARD kernel and learning a length scale parameter for each dimension allows us to consider the feature (descriptor) importance. Note that the importance of the features in the ARD kernel is not directly related to the weight parameters used in linear regression. The length scale parameter does not assume a linear relationship between inputs and outputs but simply considers the extent of the contribution of each feature by varying the amplitude of the function with respect to the direction of each input dimension. If the ARD kernel is not used, Θ is a constant diagonal matrix where θ_{2j} is uniform across all dimensions.

The hyperparameters of the kernel function such as the output scale parameter and length scale parameter determine the shape of the covariance function and need to be optimized to fit the observations $\mathcal{D}$. This operation is done via the type-II maximum likelihood estimation or the type-II maximum a posteriori (MAP) estimation. In the maximum likelihood estimation, the hyperparameters θ that maximizes the likelihood function $p(\mathbf{y} \mid \mathbf{X}, \theta)$ is evaluated. The likelihood function for a multivariate Gaussian distribution, obtained by marginalizing the dimensions other than the dimensions of the available data $\mathbf{y}$ from the Gaussian process, can be expressed by the following equation:

$$p(\mathbf{y} \mid \mathbf{X}, \theta) = \mathcal{N}(\mathbf{y} \mid \mathbf{0}, \mathbf{K} + \sigma^2 \mathbf{I}) \quad (2.34)$$

$$= \frac{1}{\sqrt{2\pi}^d \sqrt{|\mathbf{K} + \sigma^2 \mathbf{I}|}} \exp \left(-\frac{1}{2} \mathbf{y}^\top (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{y} \right) \quad (2.35)$$

Here, since the logarithmic function is a monotonically increasing function, maximizing the likelihood function $p(\mathbf{y} \mid \mathbf{X}, \theta)$ is equivalent to maximizing the log-likelihood function

$\log p(\mathbf{y} \mid \mathbf{X}, \theta)$. Thus, taking the logarithm on both sides of the equation:

$$\log p(\mathbf{y} \mid \mathbf{X}, \theta) = \log \left(\frac{1}{\sqrt{2\pi}^d \sqrt{|\mathbf{K} + \sigma^2 \mathbf{I}|}} \exp \left(-\frac{1}{2} \mathbf{y}^\top (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{y} \right) \right) \quad (2.36)$$

$$= -\frac{d}{2} \log(2\pi) - \frac{1}{2} \log |\mathbf{K} + \sigma^2 \mathbf{I}| - \frac{1}{2} \mathbf{y}^\top (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{y} \quad (2.37)$$

By maximizing this log-likelihood function, the hyperparameters of the kernel function can be optimized. This is achievable by finding the minimum of the negative log marginal likelihood function, typically using optimization techniques such as gradient descent. In my implementation, I utilized the Adam optimizer to find the minimum [9].

On the other hand, in the MAP estimation, the object is to determine the maximum value of the posterior distribution of θ given the observed data $\mathcal{D}$. The posterior distribution of θ is derived from Bayes' theorem:

$$p(\theta \mid \mathbf{y}, \mathbf{X}) = \frac{p(\mathbf{y} \mid \mathbf{X}, \theta) p(\theta)}{p(\mathbf{y} \mid \mathbf{X})} \quad (2.38)$$

Substituting the likelihood function $p(\mathbf{y} \mid \mathbf{X}, \theta)$ with a multivariate Gaussian distribution and taking the logarithm of both sides, the equation can be transformed:

$$\log p(\theta \mid \mathbf{y}, \mathbf{X}) = -\frac{d}{2} \log(2\pi) - \frac{1}{2} \log |\mathbf{K} + \sigma^2 \mathbf{I}| - \frac{1}{2} \mathbf{y}^\top (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{y} + \log P(\theta) + \text{const.} \quad (2.39)$$

Thus, the MAP estimation can be performed by maximizing $\log p(\theta \mid \mathbf{y}, \mathbf{X})$, which is the log-likelihood plus the log probability of the hyperparameters specified by the prior. In my implementation, a gamma prior is assigned to each hyperparameter. The gamma distribution is a type of continuous probability distribution and has two parameters called a shape parameter α and a rate parameter β . The choice of a gamma prior reflects prior knowledge or assumptions about the hyperparameters.

Finally, a surrogate model other than Gaussian process regression is introduced.

Random forest regression is sometimes used as an alternative surrogate model in Bayesian optimization [10]. Random forest regression is an ensemble learning algorithm that uses decision trees as weak learners and predicts the expected value of the objective variable by averaging the predictions $\hat{f}_i(\mathbf{x}_*)$ of each decision tree. In addition, by considering the standard deviation of the prediction of each decision tree as the uncertainty of the prediction, the empirical posterior distribution and posterior variance can be evaluated as in GP regression [11].

$$\mu(\mathbf{x}_*) = \frac{1}{n} \sum_{i=1}^n \hat{f}_i(\mathbf{x}_*) \quad (2.40)$$

$$\sigma^2(\mathbf{x}_*) = \frac{1}{n} \sum_{i=1}^n \left(\mu(\mathbf{x}_*) - \hat{f}_i(\mathbf{x}_*) \right)^2 \quad (2.41)$$

While GP regression approximates a continuous function, random forest regression approximates a discrete function. In cases where the search space is discrete, random forest regression may be more suitable as a surrogate model. However, random forest regression tends to exhibit poorer extrapolation performance in regions where no data exists compared to GP regression [8].

The surrogate model is sequentially updated by training a GP regression model each time observational data is added. When the kernel function and hyperparameters are appropriately chosen, the predictive accuracy of the surrogate model improves with increasing data. The sample paths drawn from a prior and posterior distribution obtained by GP regression are shown in Figure 2.8.

2.4.1.2 Acquisition Function

An acquisition function evaluates candidate solutions based on the posterior mean and posterior standard deviation obtained from GP regression. The probability of improvement (PI) function [12], expected improvement (EI) function [13], and upper confi-

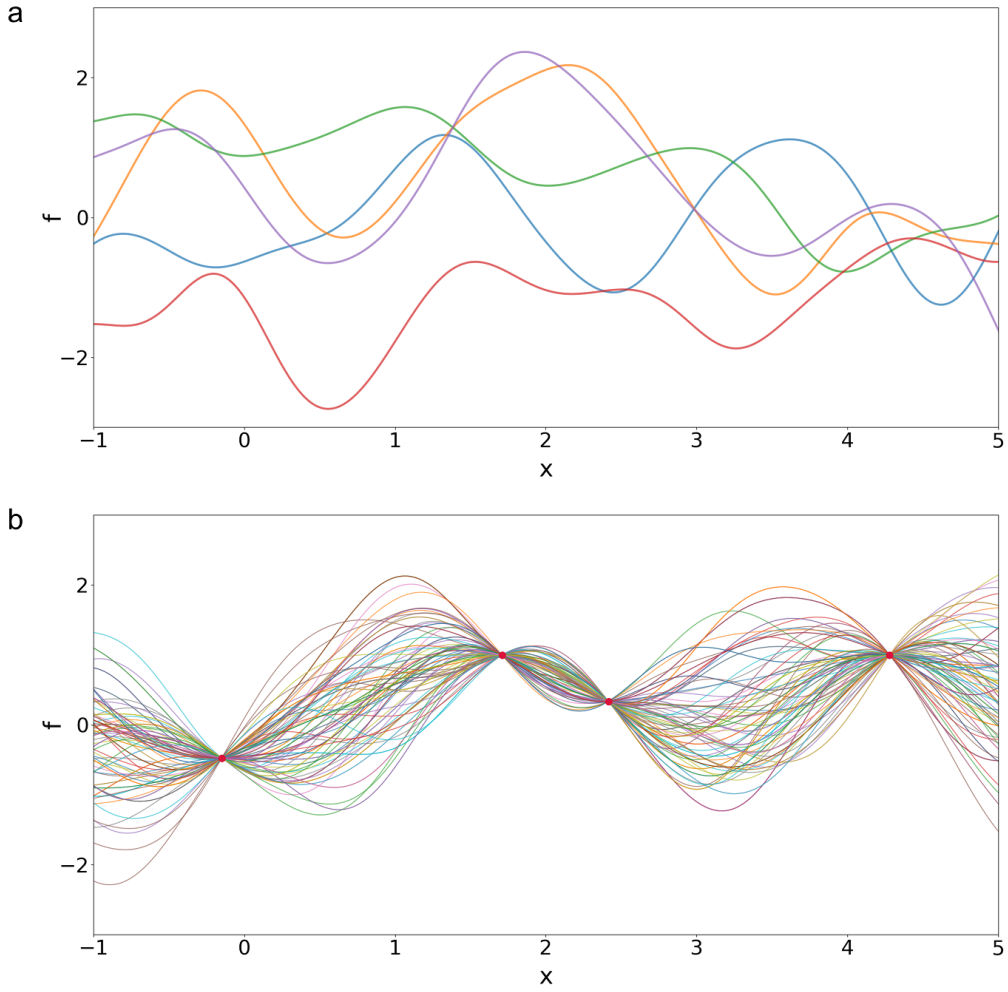


Figure 2.8. (a) Five sample paths drawn from a prior of a GP. (b) 100 sample paths drawn from a posterior distribution after observing 4 data.

dent bound (UCB) function [14] are the most widely used acquisition functions.

The PI function is defined as

$$\text{PI}(\mathbf{x}) = \Phi(Z) \quad (2.42)$$

$$Z = \frac{\mu(\mathbf{x}) - f(\mathbf{x}_{\text{best}}) - \xi}{\sigma(\mathbf{x})} \quad (2.43)$$

where $\Phi(Z)$ denote the cumulative density function of the standard normal distribution, $\mu(\mathbf{x})$ is a posterior mean predicted by GP regression, $f(\mathbf{x}_{\text{best}})$ is a function value of the current optimal solution, i.e., a maximum value of the current posterior mean, ξ is a

parameter that controls the trade-off between exploitation and exploration, $\sigma(\mathbf{x})$ is a posterior standard deviation obtained from GP regression. In my implementation, ξ is fixed to 0.01.

The EI function is given by

$$\text{EI}(\mathbf{x}) = (\mu(\mathbf{x}) - f(\mathbf{x}_{\text{best}}) - \xi) \Phi(Z) + \sigma(\mathbf{x}) \phi(Z) \quad (2.44)$$

$$Z = \frac{\mu(\mathbf{x}) - f(\mathbf{x}_{\text{best}}) - \xi}{\sigma(\mathbf{x})} \quad (2.45)$$

where $\phi(Z)$ denotes the probability density function of the standard normal distribution.

The UCB function is a technique that has been widely used in the multi-armed bandit problem. The most general expression of the UCB function is given:

$$\text{UCB}(\mathbf{x}) = \mu(\mathbf{x}) + \beta^{1/2} \sigma(\mathbf{x}) \quad (2.46)$$

where β is a parameter that controls the trade-off between exploitation and exploration. A smaller β emphasizes exploitation, while a larger β favors exploration. Various methods exist to determine β . In recent years, the GP-UCB function has been a reliable approach to the automatic selection of β [8]. In the GP-UCB function, β is given

$$\beta = 2 \log(|D| t^2 \pi^2 / 6\delta) \quad (2.47)$$

where $|D|$ is the size of the search space, and t is the total number of current observations, respectively. In my implementation, δ is set to 0.5, as employed in the previous study [15].

In Bayesian optimization, besides the PI function, EI function, and UCB function, various other acquisition functions have been proposed. For instance, Thompson sampling, as described in Chapter 2.4.1.3, is a powerful method for determining the next search point. Additionally, entropy search and predictive entropy search are representative types of acquisition functions. However, compared to the previously mentioned methods, these

techniques rely on somewhat complex assumptions and approximations.

2.4.1.3 Parallelization

In general, Bayesian optimization proposes only one candidate solution that maximizes the acquisition function per iteration. However, in practice, it is often more efficient to propose multiple candidate solutions at a time and evaluate their objective variables. Such an approach is called parallelization of Bayesian optimization, and various methods have been proposed. Here I focus on the Kriging believer algorithm and Thompson sampling. In the following description, the batch size is denoted as N .

Algorithm 1 Kriging Believer Algorithm

```

1:  $batch \leftarrow \{\}, beliefs \leftarrow \{x, y\}$ 
2: for  $i = 1$  to  $N$  do
3:   Fit GP model to beliefs
4:    $x_i \leftarrow \operatorname{argmax}_x f_i(x)$ 
5:    $batch \leftarrow batch \cup \{x_i\}$ 
6:    $beliefs \leftarrow beliefs \cup \{x_i, y_i\}$ 
7: end for
8: return  $batch$ 

```

In the Kriging believer algorithm (Algorithm 1), the posterior mean of GP regression of the candidate solution with the largest acquisition function value is temporarily regarded as observed data and added to a database [8]. Using this database, GP regression and the evaluation of the acquisition function are performed again to propose a new candidate solution. By repeating this process N times, N candidate solutions can be proposed simultaneously (Figure 2.9). As Doyle and co-workers note, the process of the Kriging believer algorithm closely resembles the process in shogi or chess, where a player predicts, "If the opponent plays this move, then it's optimal for me to play this move," to choose the best move [15].

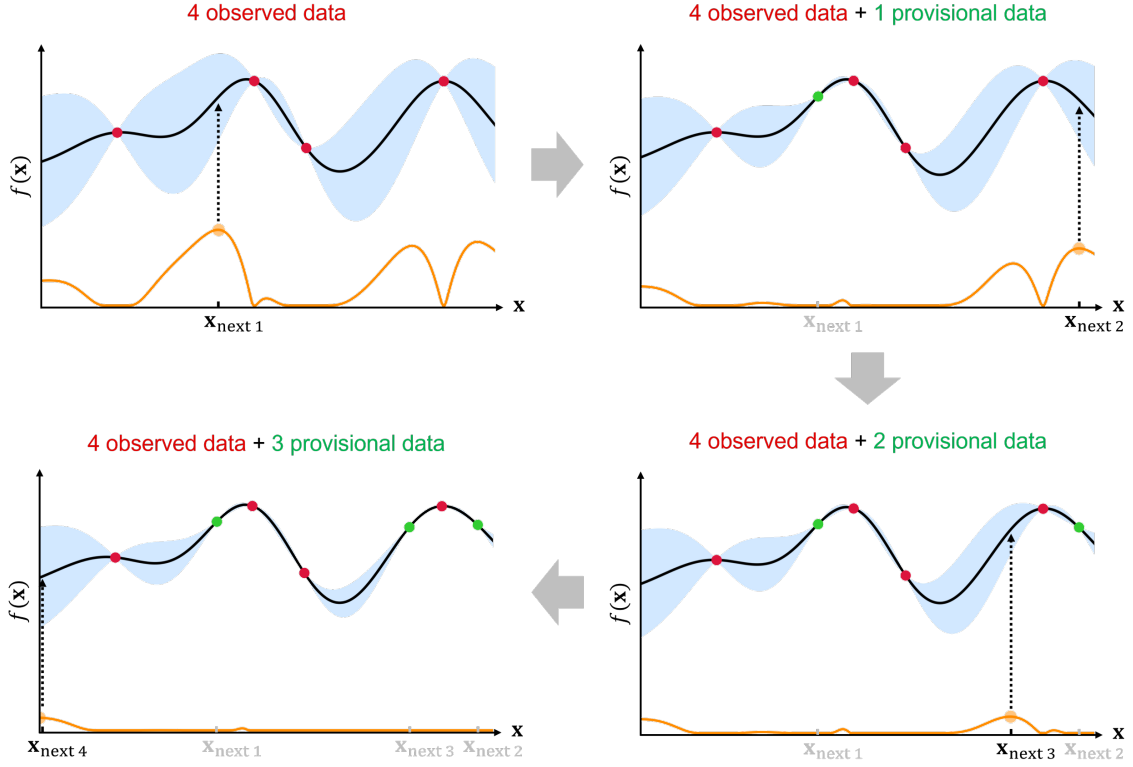


Figure 2.9. Flow of the Kriging believer algorithm. The posterior mean of the GP regression on the input that maximizes the acquisition function is temporarily added to the dataset as a provisional observational data. Bayesian optimization is then repeated a batch size number of times.

Algorithm 2 Thompson Sampling

```

1:  $batch \leftarrow \{\}$ 
2: for  $i = 1$  to  $N$  do
3:    $Sample\ g_i(x) \sim GP_{posterior}$ 
4:    $x_i \leftarrow \operatorname{argmax}_{x \in X} g_i(x)$ 
5:    $batch \leftarrow batch \cup \{x_i\}$ 
6: end for
7: return  $batch$ 

```

Thompson sampling (Algorithm 2) is a widely used decision-making technique in the multi-armed bandit problem. In the context of Bayesian optimization, Thompson sampling is used to obtain the input that maximizes the sample path drawn from a multivariate normal distribution. Specifically, in Thompson sampling, N sample paths are drawn from

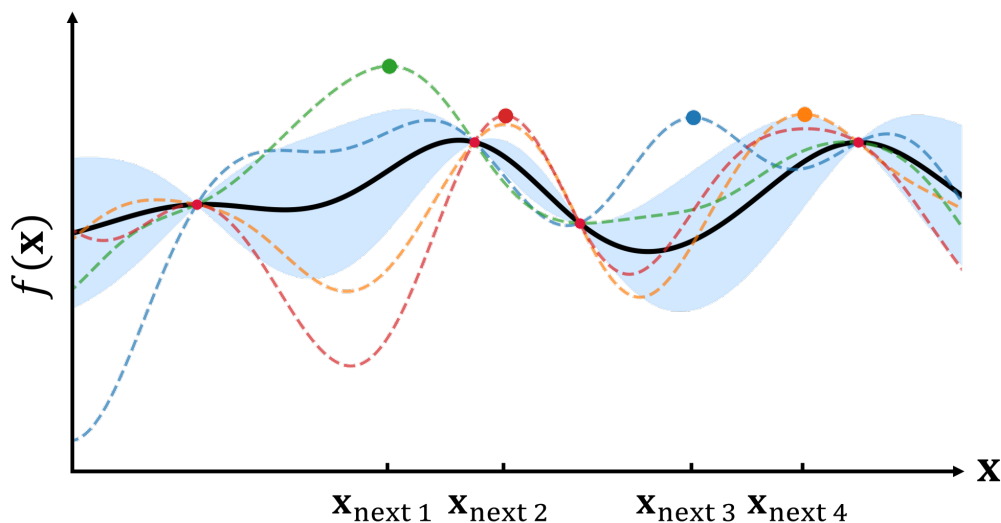


Figure 2.10. Flow of Thompson sampling. Sample paths are drawn from the posterior distribution of the GP a number of times equal to the batch size. The inputs that maximize each sample path are then treated as the next search batch.

the posterior distribution of the objective function obtained by GP regression [16]. Each sample path represents a possible function and the input that maximizes each sample path is selected as a candidate solution (Figure 2.10).

2.5 Chemspeed ISYNTH Robotic Platform

The Chemspeed robotic platform is a laboratory automation system developed to automate various processes in the field of chemistry. In particular, ISYNTH is a system specialized for the automation of organic synthesis. Organic synthesis using the Chemspeed ISYNTH robot can automatically perform common experimental operations such as dispensing liquid or solid reagents, stirring, heating, cooling, nitrogen flushing, and work-up of reaction mixtures. Robot synthesis has three major advantages: 1. it can achieve highly reproducible experiments through mechanical manipulation; 2. even experiments with toxic or explosive reagents can be performed safely; and 3. it can run multiple exper-

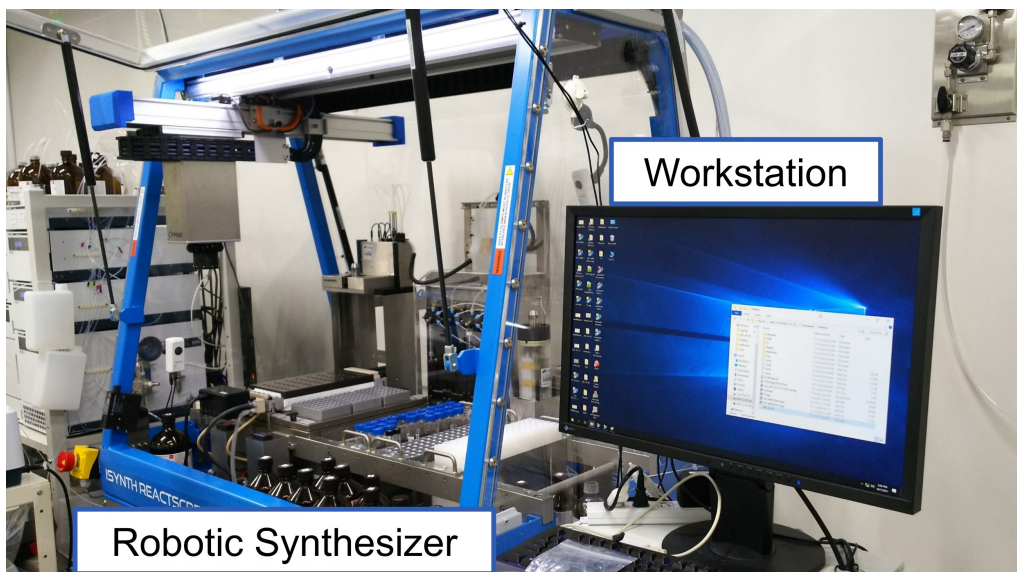


Figure 2.11. Chemspeed ISYNTH robotic platform.

iments simultaneously, thus significantly reducing human costs in processing experimental batches. Because of these advantages, the Chemspeed ISYNTH robotic platform was used for organic synthesis in this study.

2.5.1 Synthesizer

The Chemspeed ISYNTH robotic platform consists of a workstation and a robotic synthesizer (Figure 2.11). The equipment used depends on the type of the organic experiment, so different devices must be used for different purposes (Figure 2.12). It is also possible to connect third-party equipment to the Chemspeed ISYNTH robotic platform to automate the execution of experiments and the analysis of experimental results.

2.5.2 Scriptfile

In robot synthesis, all experimental operations are executed according to a script file created by the user. To create a script file, the user must define "zone" and "task".

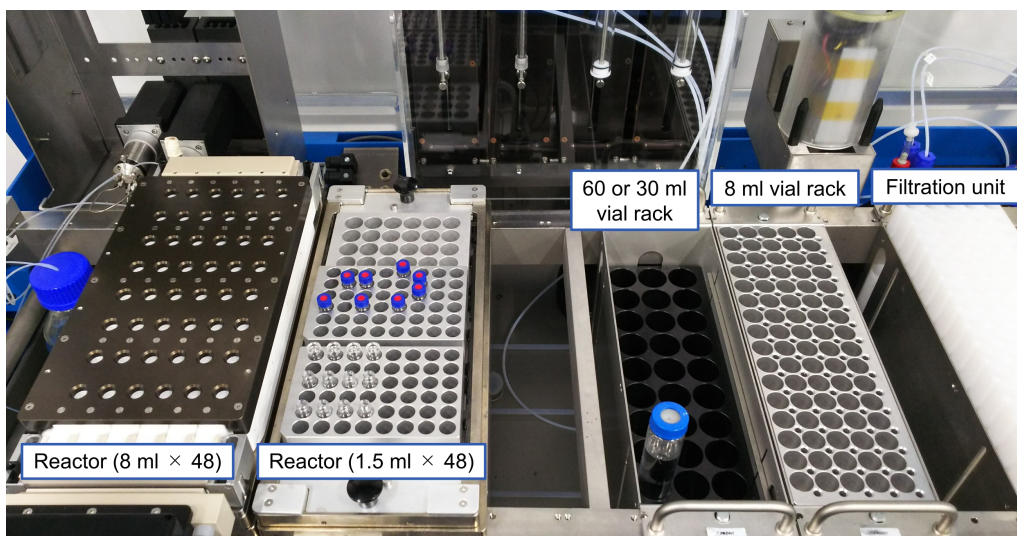


Figure 2.12. Robot synthesizer.

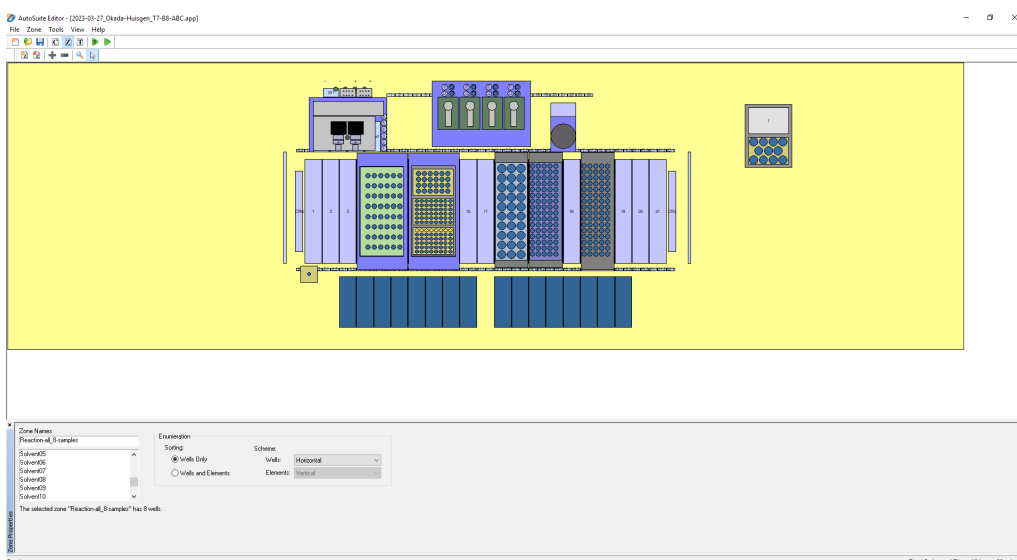


Figure 2.13. Zone for ISYNTH.

2.5.2.1 Zone

Zone is a two-dimensional coordinate parameter in the robot synthesizer. Zone allows us to specify the storage and transport locations of substrates, solvents, catalysts, and so on, as well as the location of reactors where synthesis is performed. All zones used in experiments should be defined before defining Task. An example of Zone is shown in Figure 2.13.

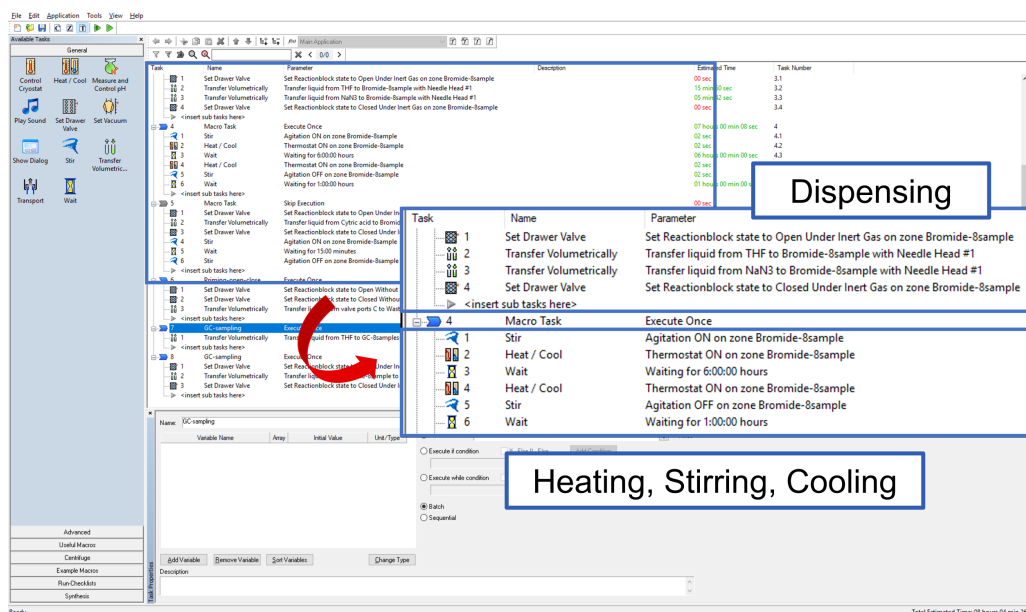


Figure 2.14. Task for ISYNTH.

2.5.2.2 Task

Task is a script that describes experimental commands that refer to specific experiment operation tasks performed by the robot. The robot executes the experiment according to the created Task file. Therefore, creating a Task file is similar to the process of a chemist writing an experimental procedure in experimental notebooks in advance. The correctness of the created file can be verified in advance by simulation on a computer. An example Task is shown in Figure 2.14.

References

- [1] S. Maeda, K. Morokuma, *The Journal of Chemical Physics* **2010**, *132*, 241102.
- [2] S. Maeda, Y. Harabuchi, *WIREs Computational Molecular Science* **2021**, *11*, e1538.
- [3] K. Fukui, *Accounts of Chemical Research* **1981**, *14*, 363–368.
- [4] C. Choi, R. Elber, *The Journal of Chemical Physics* **1991**, *94*, 751–760.

- [5] Y. Sumiya, S. Maeda, *Chemistry Letters* **2020**, *49*, 553–564.
- [6] E. I. Ioannidis, T. Z. H. Gani, H. J. Kulik, *Journal of Computational Chemistry* **2016**, *37*, 2106–2117.
- [7] J. H. Holland, *Adaptation in Natural and Artificial Systems*, **1975**.
- [8] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, N. de Freitas, *Proceedings of the IEEE* **2016**, *104*, 148–175.
- [9] D. P. Kingma, J. Ba, *Adam: A Method for Stochastic Optimization*, **2017**.
- [10] L. Breiman, *Machine Learning* **2001**, *45*, 5–32.
- [11] F. Hutter, H. H. Hoos, K. Leyton-Brown in *Learning and Intelligent Optimization*, Vol. 6683, C. A. C. Coello (Ed.), Springer Berlin Heidelberg, Berlin, Heidelberg, **2011**, pp. 507–523.
- [12] H. J. Kushner, *Journal of Basic Engineering* **1964**, *86*, 97–106.
- [13] J. Mockus, V. Tiesis, A. Zilinskas in *Towards Global Optimization*, Vol. 2, **2014**, pp. 117–129.
- [14] N. Srinivas, A. Krause, S. Kakade, M. Seeger in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, Omnipress, Madison, WI, USA, of *ICML'10*, pp. 1015–1022.
- [15] B. J. Shields, J. Stevens, J. Li, M. Parasram, F. Damani, J. I. M. Alvarado, J. M. Janey, R. P. Adams, A. G. Doyle, *Nature* **2021**, *590*, 89–96.
- [16] K. Kandasamy, A. Krishnamurthy, J. Schneider, B. Póczos in *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, PMLR, pp. 133–142.

Chapter 3

A Dataset of Computational Reaction Barriers for the Claisen Rearrangement: Chemical and Numerical Analysis

3.1 Introduction

For achieving desired reactivity and/or selectivity, optimization of a combination of substrate, catalyst, solvent, temperature, and so on is required. Depending on the number of possible combinations, it can be difficult to investigate all of the combinations in terms of resources and costs. Therefore, an efficient screening technique to find an optimum condition has been desired. Thanks to recent advances in chemoinformatics, successful applications of optimization algorithms such as genetic algorithm (GA) and Bayesian optimization (BO) to experimental reaction screening have been actively reported [1–10].

In experimental screening, quantities to maximize using an optimization algorithm are reaction yield and/or selectivity. On the other hand, these quantities can be estimated theoretically by evaluating Gibbs energy barriers. Quantum chemical calculation has grown to be a useful tool to assess Gibbs energy barriers. There have been numerous studies that discussed reactivity and selectivity utilizing quantum chemical calculations

[11–19]. Quantum chemical calculations have also been conducted for predicting reactivity and selectivity [20–23]. Therefore, reaction screening can also be done by evaluating Gibbs energy barriers theoretically for various combinations.

When identifying a reaction with a desired reactivity or selectivity, evaluating Gibbs energy barriers for all possible combinations is the most reliable way [24]. On the other hand, evaluating all possible combinations can be difficult, even in theoretical screening. Let me think about a case of an unimolecular rearrangement in which the substrate has five substitutable sites. In this case, the number of molecules that can be generated by substituting these five sites with ten substitution groups is 100,000, and evaluating Gibbs energy barriers for all of the 100,000 reactions based on accurate quantum chemical calculations could be computationally impractical.

Optimization algorithms such as GA and BO can also be useful to reduce the number of combinations actually to evaluate theoretically. It is noted that the purpose here is not to identify the best combination but to find one that falls into a specified number of top-ranked combinations. For example, all reactions having Gibbs energy barriers lower than ~ 80 kJ/mol provide the desired product nearly quantitatively at room temperature, and finding one of the good combinations having Gibbs energy barriers lower than ~ 80 kJ/mol is the actual subject when the purpose is to find a reaction that proceeds quantitatively at room temperature. Optimization algorithms such as GA and BO assist in finding top-ranked combinations with the least number of trials. GA, for example, finds a combination that falls into the top 0.1% among $\sim 100,000$ combinations with a 98.2% success rate after evaluating ~ 200 combinations, in an example shown below in this chapter. Evaluating 200 combinations by accurate quantum chemical calculations is feasible. Therefore, combining such an optimization algorithm with theoretical reaction screening would be promising.

Various implementations are possible for an optimization algorithm such as GA and

BO. Hence, this study focuses on generating a dataset of reaction barriers that can be used in testing various implementations of an optimization algorithm. The targeted size of the present dataset is set to 100,000, and the Claisen rearrangement for 100,000 allyl vinyl ether molecules with different substitutions is considered. Since performing highly accurate calculations for 100,000 reactions is not feasible, this study adopts a semiempirical theory PM7 as a method to evaluate the barriers. To verify whether the resultant dataset shows chemically reasonable trends and is useful in testing various implementations of an optimization algorithm, some features in the dataset are discussed. As a result, variations of barriers were found to be well interpretable based on physical quantities such as electronegativity of substituent groups, indicating that the dataset would be chemically reasonable. To further evaluate the dataset numerically, a simple GA program that optimizes the combination of substituents systematically is applied. GA achieved a much higher success rate of finding a top 0.1% molecule than the random search (RS), indicating that there would be a clear correlation between barriers and substitution groups at each substitution site in the dataset. These verification results suggest that the present dataset is useful in testing various implementations of optimization algorithms to accelerate theoretical reaction screening.

3.2 Methods

3.2.1 Barrier Calculation

The Claisen rearrangement is a [3,3]-sigmatropic rearrangement of an allyl vinyl ether, as illustrated in Figure 3.1. 100,000 allyl vinyl ether molecules were generated by substituting five hydrogen atoms of the unsubstituted molecule with 10 substituents, as illustrated in Figure 3.2.

Substituted reactant structures were generated using the unsubstituted reactant

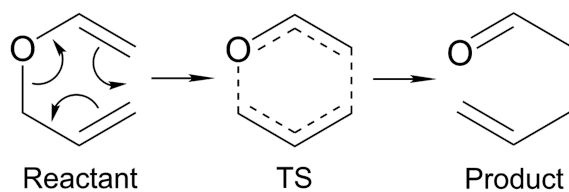


Figure 3.1. A schematic of the Claisen rearrangement.

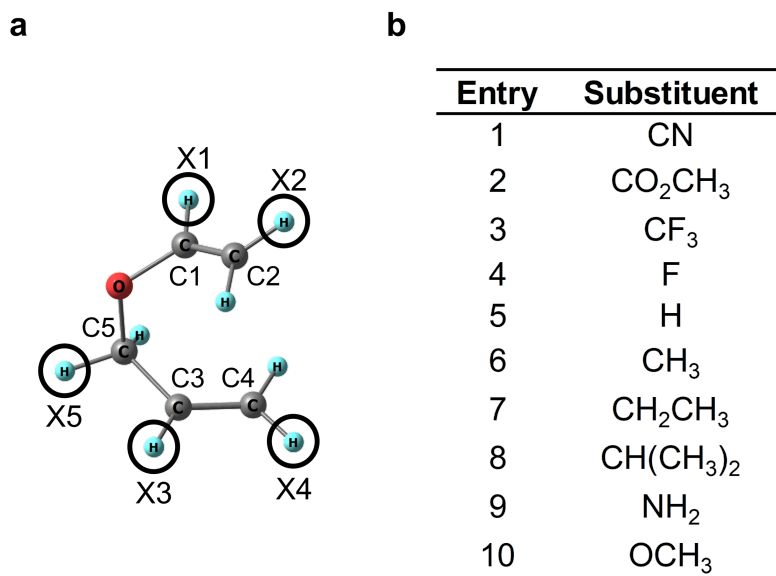


Figure 3.2. (a) Five substitution sites of allyl vinyl ether and (b) 10 substituents considered in this study.

structure shown in Figure 3.2a as a template. Pre-optimized substituents substituted to the five substitutable sites were rotated around the newly generated covalent bonds to reduce steric hindrances. These structures were subsequently optimized without imposing any constraints. Transition state (TS) structures were generated using the artificial force induced reaction (AFIR) method [25,26]. The force-induced path was relaxed by the locally updated planes (LUP) method [27], and the highest energy point along the relaxed path was further optimized to an actual TS using a quasi-Newton method. Intrinsic reaction coordinate (IRC) calculations were then performed from all obtained TSs to check whether they connect the right reactant and product pair [28,29]. All these calculations

were done at a semiempirical PM7 level [30] using an in-house GRRM program [31] interfaced with Gaussian 16 [32]. It is emphasized again that the present purpose is not to discuss accurate energetics in this system but to generate an extensive dataset that can be useful in testing various implementations of an optimization algorithm. Qualitative and quantitative evaluations below would justify the use of PM7 for this purpose.

In this study, the following two quantities were used as importance measures,

$$\Delta\Delta G_i^L = \Delta G_i^{TS} - \Delta G_i^{Reactant} \quad (3.1)$$

$$\Delta\Delta G_i^G = \Delta G_i^{TS} - \Delta G_i^{Parent} \quad (3.2)$$

where ΔG_i^{TS} and $\Delta G_i^{Reactant}$ are relative Gibbs energy values of TS and reactant, respectively, in the i th reaction. ΔG_i^{Parent} is given as,

$$\Delta G_i^{Parent} = \Delta G_i^{uAVE} + \sum_{k=1}^5 \Delta G^{R_kH} - 5\Delta G^{H_2} \quad (3.3)$$

where ΔG_i^{uAVE} is a relative Gibbs energy value of the unsubstituted allyl vinyl ether, ΔG^{R_kH} is a relative Gibbs energy value of an R_kH molecule where R_k is one of five substituents introduced to the i th reaction's reactant, and ΔG^{H_2} is a relative Gibbs energy value of a hydrogen molecule. In short, $\Delta\Delta G_i^L$ and $\Delta\Delta G_i^G$ correspond to the reaction barrier of each reaction and the relative stability of TS to the unsubstituted (parent) system, respectively. In other words, $\Delta\Delta G_i^L$ is a local measure of reactivity of the i th reaction and $\Delta\Delta G_i^G$ is a global measure of stability of the i th TS.

3.2.2 Genetic Algorithm

In this study, trials of finding a top 0.1% reaction in terms of $\Delta\Delta G_i^L$ or $\Delta\Delta G_i^G$ were conducted by GA to see if the dataset has any numerical trend. My implementation, which almost follows a standard one [33], is described below in brief.

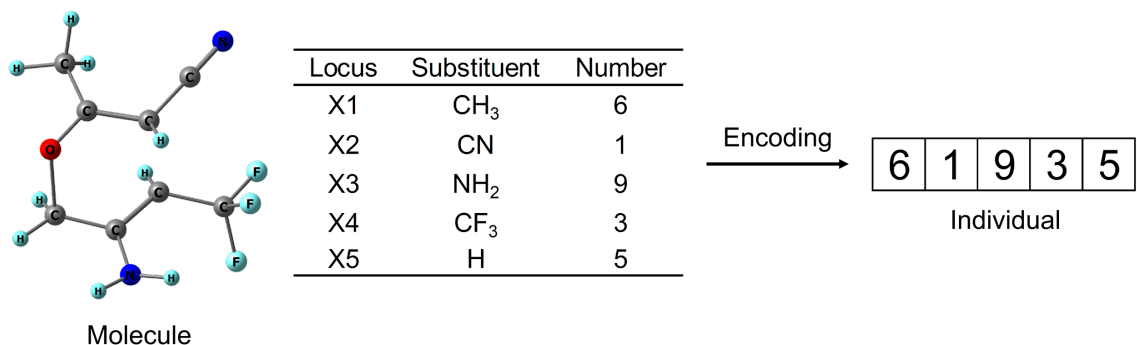


Figure 3.3. An example of a reactant molecule and its representation in GA.

As illustrated in Figure 3.3, each reaction is encoded by a combination of five numbers representing the substituent groups listed in Figure 3.2b. These numbers correspond to genes, and a set of these numbers represents an individual. Then, an individual possessing a high score, i.e., those having a low $\Delta\Delta G_i^L$ or $\Delta\Delta G_i^G$ value, is searched through the following optimization cycle.

1. The current population is generated by selecting *pop_size* individuals by the greedy method, and N_{trial} is set to 0. The greedy method generates individuals sequentially while choosing substituents that appear in individuals generated earlier with the least frequency.
2. Scores, i.e., $\Delta\Delta G_i^L$ or $\Delta\Delta G_i^G$ values, for individuals in the current population are obtained either from the record of the previous GA iteration or from the dataset, where N_{trial} is incremented by the number of times the dataset is accessed.
3. The best *elite_size* individuals (elite group) in the current population are copied to the next generation population.
4. While allowing duplication, *pop_size* individuals (parent individuals) are chosen from the current population by repeating a procedure electing the best one among *tournament_size* individuals randomly chosen from the current population.

5. ($2 \times pop_size$) children individuals are generated through ($2 \times pop_size$) times crossovers between two randomly chosen parent individuals, where their chromosomes, i.e., a set of genes, are divided into two fragments and one of them is swapped.
6. Random mutation is applied to some genes of the children individuals, where the ratio of a child to be chosen as a candidate of applying mutation is *ind_mutate_rate* and the ratio of a gene to be mutated is *gene_mutate_rate*.
7. From the ($2 \times pop_size$) children individuals, ($pop_size - elite_size$) individuals are chosen taking account of the gene diversity and added to the next generation population including the *elite_size* individuals from the elite group.
8. If N_{trial} reaches 200, then the cycle is terminated. Otherwise, the cycle is continued by copying the next generation population to the current population and going back to step 2.

The five hyper-parameters, i.e., *pop_size*, *elite_size*, *tournament_size*, *ind_mutate_rate*, and *gene_mutate_rate*, were determined to maximize the performance, as described in the results and discussion section.

3.3 Results and Discussion

3.3.1 Barrier Distribution

The above computational procedure found TSs for Claisen rearrangement for 99,072 cases out of the 100,000. In some of the 928 unsuccessful cases, TSs for different chemical transformations were obtained. In the other unsuccessful cases, the structures were dissociated into two parts during TS optimization. There were also cases where the calculations did not converge. In cases where reactant’s conformers and/or metastable intermediates

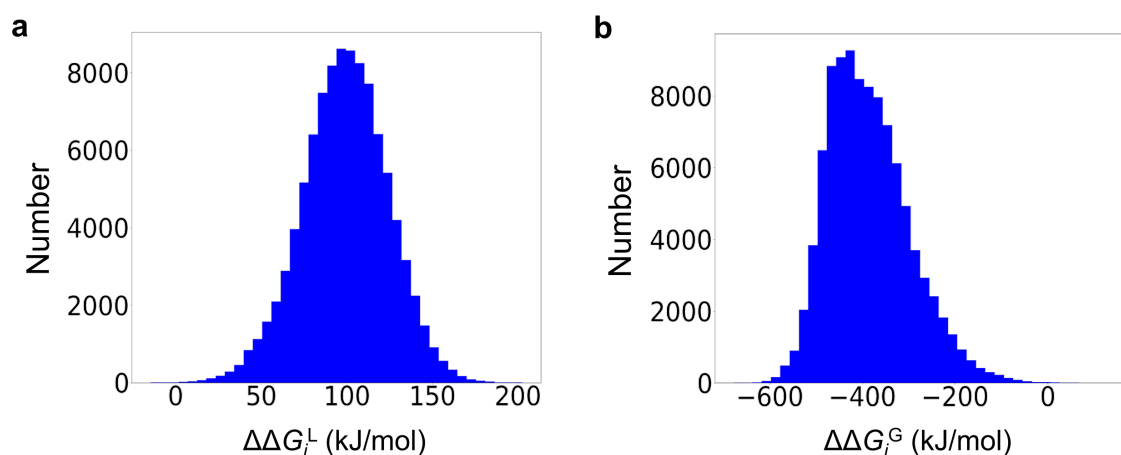


Figure 3.4. Distribution of (a) $\Delta\Delta G_i^L$ and (b) $\Delta\Delta G_i^G$ in the present dataset.

were found along the IRC path, the highest TS among those obtained along the path and the lowest reactant’s conformer were used to evaluate the Gibbs energy barrier. Although there were 12 cases where $\Delta\Delta G_i^L$ values were slightly negative due to strong destabilization of reactant molecules caused by inappropriate substitution combinations, these values were used without taking any special care in the analyses below.

Figure 3.4 shows histograms of distributions of $\Delta\Delta G_i^L$ and $\Delta\Delta G_i^G$ for the 99,072 reactions. As seen in these histograms, the present dataset includes reactions having various $\Delta\Delta G_i^L$ ranging from ~ 0 to ~ 200 kJ/mol. The distribution of $\Delta\Delta G_i^G$ is wider than that of $\Delta\Delta G_i^L$, indicating that the stability of the reactant molecules also has a wide distribution. It would be interesting to see how these distributions have arisen.

3.3.2 Chemical Analysis

Tabel 3.1 lists the number of appearances of substituent groups at different substitution sites in the top 0.1% (top 100) reactions in terms of $\Delta\Delta G_i^L$. As seen in Tabel 3.1, electron-donating groups (EDGs) such as NH_2 and OCH_3 frequently ranked in at sites X1, X4, and X5, and electron-withdrawing groups (EWGs) such as CO_2CH_3 , CF_3 , and

Table 3.1. The number of appearances of substituent groups at different substitution sites in the top 0.1% (top 100) reactions in terms of $\Delta\Delta G_i^L$.

X1		X2		X3		X4		X5	
Substituent	Number	Substituent	Number	Substituent	Number	Substituent	Number	Substituent	Number
OCH ₃	41	CO ₂ CH ₃	41	CO ₂ CH ₃	63	OCH ₃	32	NH ₂	92
F	22	CN	23	CF ₃	7	NH ₂	28	CH ₃	3
NH ₂	15	OCH ₃	11	CN	7	F	10	H	2
CF ₃	13	CF ₃	10	CH(CH ₃) ₂	6	CH ₃	10	CH(CH ₃) ₂	1
CN	4	F	7	H	4	H	9	CN	1
CO ₂ CH ₃	2	NH ₂	5	CH ₂ CH ₃	4	CH(CH ₃) ₂	3	CH ₂ CH ₃	1
CH(CH ₃) ₂	1	H	3	OCH ₃	3	CO ₂ CH ₃	3		
CH ₃	1			F	3	CH ₂ CH ₃	3		
CH ₂ CH ₃	1			NH ₂	2	CN	2		
				CH ₃	1				

CN ranked in at sites X2 and X3. As shown in Figure 3.5, the TS structure of Claisen rearrangement is composed of two substructures: the enolate moiety bearing a negative partial charge and the allyl moiety bearing a positive partial charge. Considering the resonance structure of these moieties, there is a negative charge on C2 and a positive charge on C4 and C5. Hence, EWG at X2 and EDG at X4 and X5 stabilize the partial charge and decrease the reaction barrier [34,35]. On the other hand, substituting EDG such as NH₂ and OCH₃ to X1 enhances the stability of the product by delocalizing a lone pair of electrons to the carbonyl π^* orbital. In particular, NH₂ forms an amide bond with the carbonyl group and significantly stabilizes the product [34,35]. It is known that reaction barriers often correlate with products' stability [36,37]. Although there is no clear explanation based on charge distributions on the reason why X3 prefers EWG, this trend was confirmed by DFT calculations at ω B97X-D [38]/def2-SVP [39] level to be correct as shown in Figure 3.6.

Table 3.2 presents the number of appearances of substituent groups at different substitution sites in the top 0.1% (top 100) reactions in terms of $\Delta\Delta G_i^G$. By definition,

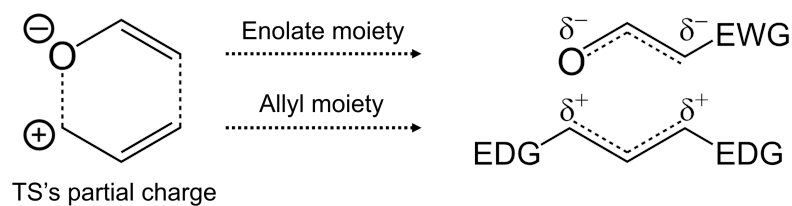


Figure 3.5. Partial charges at TS and preferred locations of EDG and EWG.

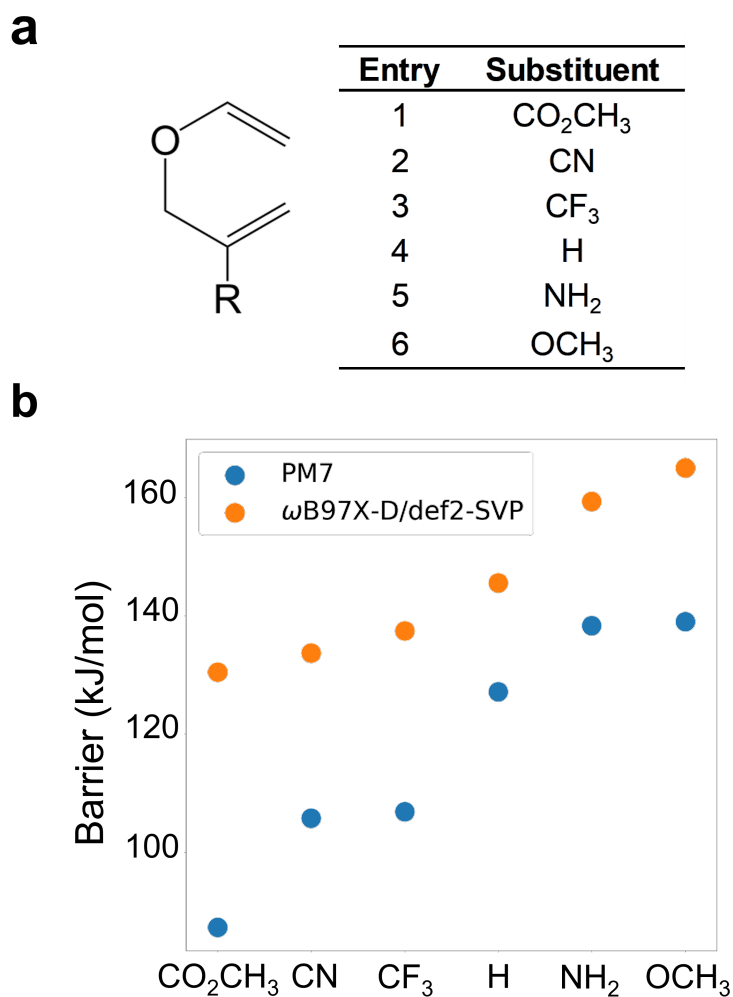


Figure 3.6. Substituent dependences of $\Delta\Delta G_i^L$ values at X3 by substitution groups listed in (a), at different computational levels (b).

Table 3.2. The number of appearances of substituent groups at different substitution sites in the top 0.1% (top 100) reactions in terms of $\Delta\Delta G_i^G$.

X1		X2		X3		X4		X5	
Substituent	Number	Substituent	Number	Substituent	Number	Substituent	Number	Substituent	Number
NH ₂	77	CO ₂ CH ₃	44	CO ₂ CH ₃	61	NH ₂	60	NH ₂	51
CF ₃	9	CF ₃	34	CN	20	CH ₃	7	CF ₃	11
OCH ₃	8	CN	18	CF ₃	18	CO ₂ CH ₃	7	CH ₂ CH ₃	9
CH(CH ₃) ₂	4	CH ₂ CH ₃	2	CH ₂ CH ₃	1	CH(CH ₃) ₂	6	CH(CH ₃) ₂	8
CH ₃	1	CH ₃	1			CH ₂ CH ₃	6	CH ₃	7
CO ₂ CH ₃	1	CH(CH ₃) ₂	1			OCH ₃	5	OCH ₃	5
						CF ₃	5	CN	4
						CN	4	CO ₂ CH ₃	4
								F	1

this ranking reflects the stability of the TS more directly. The main difference from Table 3.1 is seen at sites X1 and X4. At these sites, the most preferred substituent changed from OCH₃ to NH₂. This would be related to the preference of the CO₂CH₃ group at the adjacent X2 and X3 sites in the top 0.1% reactions. The presence of the NH₂ and CO₂CH₃ groups on adjacent sp² carbons allow the reactant molecule to take the resonance structures shown in Figure 3.7.

The stabilization in the reactant molecule due to the existence of these resonance structures increased $\Delta\Delta G_i^L$ for reactions taking X1 = NH₂ and X4 = NH₂ and consequently lowered the rankings of NH₂ at X1 and X4 in Table 3.1.

In summary, it can be said that the present dataset at the PM7 level catches the trends concerning the preferences of substitution groups at the five substitution sites well. Therefore, it is expected that the present dataset is useful in tuning implementations of an optimization algorithm like GA, as shown below.

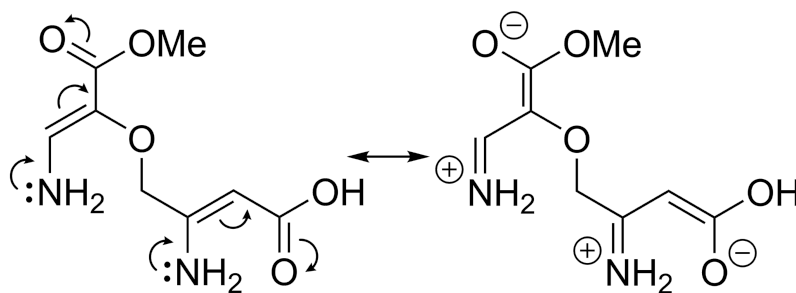


Figure 3.7. Resonance structures in systems having neighboring NH_2 and CO_2CH_3 .

3.3.3 Numerical Analysis

To check the performance of GA, the success rate of finding one of the top 0.1% (top 100) reactions within 200 trials was evaluated. The success rate was determined by running the GA search 10,000 times and counting the number of successful runs out of the 10,000. The five hyperparameters were determined by a grid search to maximize the success rate, where *pop_size* was changed in the range from 10 to 30 with step size 5, *elite_size* was changed in the range from 0 to 9 with step size 1, *tournament_size* was changed in the range from 2 to 9 with step size 1, *ind_mutate_rate* was changed in the range from 0.05 to 0.50 with a step size 0.05, and *gene_mutate_rate* was changed in the range from 0.05 to 0.50 with a step size 0.05. The success rate was evaluated at the 48,000 grid points, and the best parameters were determined as *pop_size* = 25, *elite_size* = 8, *tournament_size* = 6, *ind_mutate_rate* = 0.25, and *gene_mutate_rate* = 0.05.

Figure 3.8 (a) and (b) show the performances of the GA searches applied to the datasets of $\Delta\Delta G_i^L$ and $\Delta\Delta G_i^G$, respectively. The success rates increase depending on the number of trials and reach 95.9% and 98.2% after 200 trials for $\Delta\Delta G_i^L$ and $\Delta\Delta G_i^G$, respectively. As a reference, the performance of a simple RS is also shown in these figures. The success rate of the RS increases linearly. The GA searches show clear advantages above trials of 50 times. These results indicate that the target quantities, i.e., $\Delta\Delta G_i^L$ and $\Delta\Delta G_i^G$, have a clear correlation with groups and sites of substitutions in the present

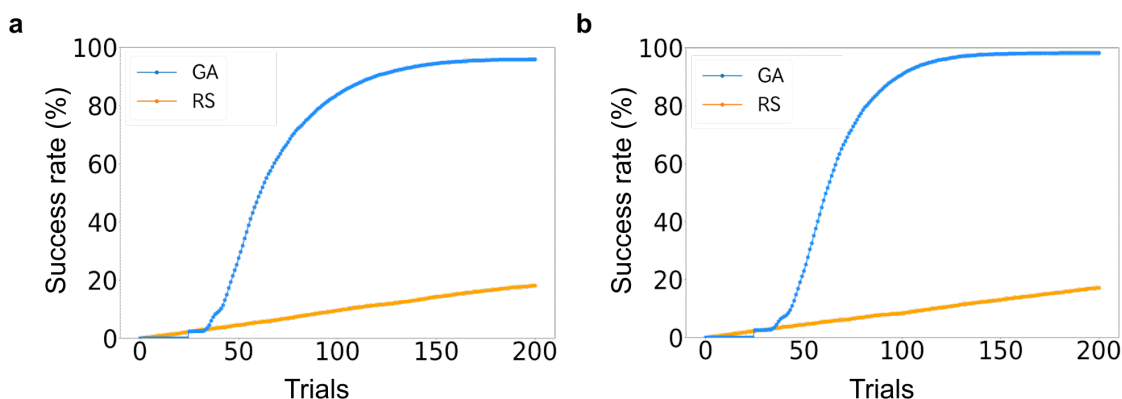


Figure 3.8. Performances of GA and RS using (a) $\Delta\Delta G_i^L$ and (b) $\Delta\Delta G_i^G$ as importance measures. The success rate at each point was evaluated by conducting GA or RS 10,000 times and taking the average.

datasets, which proves that the present datasets are useful in testing performances of computational reactions search/optimization algorithms.

3.4 Conclusion

Theoretical datasets of reaction barriers of Claisen rearrangement were generated. These reaction barriers were computed for $\sim 100,000$ allyl vinyl ether molecules with different substitutions at a semiempirical PM7 level. Best substitution patterns were well interpreted based on chemical features of the reactant, TS, and product and electronic properties of substitution groups such as electron withdrawing/donating abilities. The genetic algorithm with a simple implementation dramatically improved the success rates of finding one of the top 0.1% reactions, compared to a simple random search. These results suggest that the present datasets are useful in tuning parameters and implementations of computational reaction search algorithms.

In this study, the datasets were used in testing a simple GA algorithm in which each reaction was encoded by a set of integers representing each substituent. The present

datasets would also be useful in testing the more complicated algorithms such as Bayesian optimization utilizing geometric and electronic features of substituent groups as descriptor variables.

3.5 Data Availability

The present datasets of the reaction barrier of the Claisen rearrangement are available in the supplementary material of doi: 10.1002/minf.202100216.

References

- [1] R. P. Sheridan, S. K. Kearsley, *Journal of Chemical Information and Computer Sciences* **1995**, *35*, 310–320.
- [2] S. R. M. Pereira, F. Clerc, D. Farrusseng, J. C. van der Waal, T. Maschmeyer, C. Mirodatos, *QSAR & Combinatorial Science* **2005**, *24*, 45–57.
- [3] U. Rodemerck, M. Baerns, M. Holena, D. Wolf, *Applied Surface Science* **2004**, *223*, 168–174.
- [4] L. Weber, S. Wallbaum, C. Broger, K. Gubernator, *Angewandte Chemie International Edition in English* **1995**, *34*, 2280–2282.
- [5] L. Weber, *Drug Discovery Today* **1998**, *3*, 379–385.
- [6] F. Häse, L. M. Roch, C. Kreisbeck, A. Aspuru-Guzik, *ACS Central Science* **2018**, *4*, 1134–1145.
- [7] T. Ueno, T. D. Rhone, Z. Hou, T. Mizoguchi, K. Tsuda, *Materials Discovery* **2016**, *4*, 18–21.

- [8] B. J. Shields, J. Stevens, J. Li, M. Parasram, F. Damani, J. I. M. Alvarado, J. M. Janey, R. P. Adams, A. G. Doyle, *Nature* **2021**, *590*, 89–96.
- [9] B. Burger, P. M. Maffettone, V. V. Gusev, C. M. Aitchison, Y. Bai, X. Wang, X. Li, B. M. Alston, B. Li, R. Clowes, N. Rankin, B. Harris, R. S. Sprick, A. I. Cooper, *Nature* **2020**, *583*, 237–241.
- [10] M. Christensen, L. P. E. Yunker, F. Adedeji, F. Häse, L. M. Roch, T. Gensch, G. dos Passos Gomes, T. Zepel, M. S. Sigman, A. Aspuru-Guzik, J. E. Hein, *Communications Chemistry* **2021**, *4*, 1–12.
- [11] Nobuaki. Koga, Keiji. Morokuma, *Chemical Reviews* **1991**, *91*, 823–842.
- [12] K. N. Houk, P. H.-Y. Cheong, *Nature* **2008**, *455*, 309–313.
- [13] W. Thiel, *Angewandte Chemie International Edition* **2014**, *53*, 8605–8613.
- [14] W. M. C. Sameera, S. Maeda, K. Morokuma, *Accounts of Chemical Research* **2016**, *49*, 763–773.
- [15] K. N. Houk, F. Liu, *Accounts of Chemical Research* **2017**, *50*, 539–543.
- [16] S. Maeda, K. Ohno, *Journal of the American Chemical Society* **2008**, *130*, 17228–17229.
- [17] P. J. Donoghue, P. Helquist, P.-O. Norrby, O. Wiest, *Journal of the American Chemical Society* **2009**, *131*, 410–411.
- [18] E. Hansen, A. R. Rosales, B. Tutkowski, P.-O. Norrby, O. Wiest, *Accounts of Chemical Research* **2016**, *49*, 996–1005.
- [19] Y. Guan, V. M. Ingman, B. J. Rooks, S. E. Wheeler, *Journal of Chemical Theory and Computation* **2018**, *14*, 5249–5261.

- [20] S. Ahn, M. Hong, M. Sundararajan, D. H. Ess, M.-H. Baik, *Chemical Reviews* **2019**, *119*, 6509–6560.
- [21] A. R. Rosales, J. Wahlers, E. Limé, R. E. Meadows, K. W. Leslie, R. Savin, F. Bell, E. Hansen, P. Helquist, R. H. Munday, O. Wiest, P.-O. Norrby, *Nature Catalysis* **2019**, *2*, 41–45.
- [22] M. Foscato, V. R. Jensen, *ACS Catalysis* **2020**, *10*, 2354–2377.
- [23] T. Mita, Y. Harabuchi, S. Maeda, *Chemical Science* **2020**, *11*, 7569–7577.
- [24] G. F. von Rudorff, S. N. Heinen, M. Bragato, O. A. von Lilienfeld, *Machine Learning: Science and Technology* **2020**, *1*, 045026.
- [25] S. Maeda, K. Morokuma, *The Journal of Chemical Physics* **2010**, *132*, 241102.
- [26] S. Maeda, Y. Harabuchi, *WIREs Computational Molecular Science* **2021**, *11*, e1538.
- [27] C. Choi, R. Elber, *The Journal of Chemical Physics* **1991**, *94*, 751–760.
- [28] K. Fukui, *Accounts of Chemical Research* **1981**, *14*, 363–368.
- [29] S. Maeda, Y. Harabuchi, Y. Ono, T. Taketsugu, K. Morokuma, *International Journal of Quantum Chemistry* **2015**, *115*, 258–269.
- [30] J. J. P. Stewart, *Journal of Molecular Modeling* **2013**, *19*, 1–32.
- [31] S. Maeda, K. Ohno, K. Morokuma, *Physical Chemistry Chemical Physics* **2013**, *15*, 3683–3701.
- [32] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V.

Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, D. J. Fox, *Gaussian 16 Revision C.01*, **2016**, Gaussian Inc. Wallingford CT.

- [33] J. H. Holland, *Adaptation in Natural and Artificial Systems*, **1975**.
- [34] V. Aviyente, H. Y. Yoo, K. N. Houk, *The Journal of Organic Chemistry* **1997**, *62*, 6121–6128.
- [35] V. Aviyente, K. N. Houk, *The Journal of Physical Chemistry A* **2001**, *105*, 383–391.
- [36] R. P. Bell, C. N. Hinshelwood, *Proceedings of the Royal Society of London. Series A - Mathematical and Physical Sciences* **1997**, *154*, 414–429.
- [37] M. G. Evans, M. Polanyi, *Transactions of the Faraday Society* **1936**, *32*, 1333–1360.
- [38] J.-D. Chai, M. Head-Gordon, *Physical Chemistry Chemical Physics* **2008**, *10*, 6615–6620.
- [39] F. Weigend, R. Ahlrichs, *Physical Chemistry Chemical Physics* **2005**, *7*, 3297–3305.

Chapter 4

On Accelerating Substrate Optimization Using Computational Gibbs Energy Barriers: A Numerical Consideration Utilizing a Computational Dataset

4.1 Introduction

Substrate optimization is a time- and resource-consuming step in organic synthesis. Since the number of possible conditions increases exponentially depending on the number of substrates that need to be optimized, it is challenging to assess all of the combinations experimentally. It is generally done based on a chemist’s intuition and experience. Such an empirical process often takes a long time and consumes considerable resources. Thus, an efficient and reliable optimization method has been desired.

Informatics techniques such as linear and nonlinear regression, machine learning, data clustering, and black-box optimization, among others, are currently used in chemistry for a variety of purposes. These include predicting reaction yields and selectivities, elucidating reaction mechanisms, planning synthetic routes, developing descriptors, sug-

gesting improved reaction conditions, and designing materials [1–18]. Bayesian optimization (BO) is a powerful tool that balances exploration and exploitation [19, 20]. In BO, Gaussian process (GP) regression predicts an objective variable and estimates its variance. Then, the next candidate to be evaluated is proposed via an acquisition function. BO has been applied to various problems in chemistry [21–38], including reaction optimization in organic synthesis [39–42].

On the other hand, Gibbs energy barriers, calculated using quantum chemical calculations, have been widely used to interpret and explain known reactivities and selectivities of organic reactions [43–53]. In addition, their ability to predict results in advance of experiments has also been successful in many examples [8, 54–61]. In one-step reactions, the Gibbs energy difference between the transition state (TS) and the reactant determines the rate constant and thus represents the reactivity [62]. Even in multistep reactions, the reactivity is represented by the Gibbs energy difference between the TS of the rate-determining step and the lowest intermediate before the rate-determining step, under the generalized pre-equilibrium approximation [63, 64]. Therefore, this study regards Gibbs energy barriers as experimental and computational reactivities.

This study considers the use of theoretical Gibbs energy barriers to accelerate BO-assisted substrate optimization and, consequently, reduce the number of required experiments. In such a BO, data of two kinds, i.e., experimental data and computational data, are used, as illustrated in Figure 4.1. To do that, one must consider the deviation between computational and experimental results. Relying too much on the computational data misleads optimization due to computational errors. At the same time, suppressing the use of computational data too much does not help to reduce the number of experiments. Moreover, the deviation varies depending on the computational levels. In Gibbs energy barriers of organic reactions, the deviation typically is ~ 5 kJ/mol by CCSD(T), 10–20

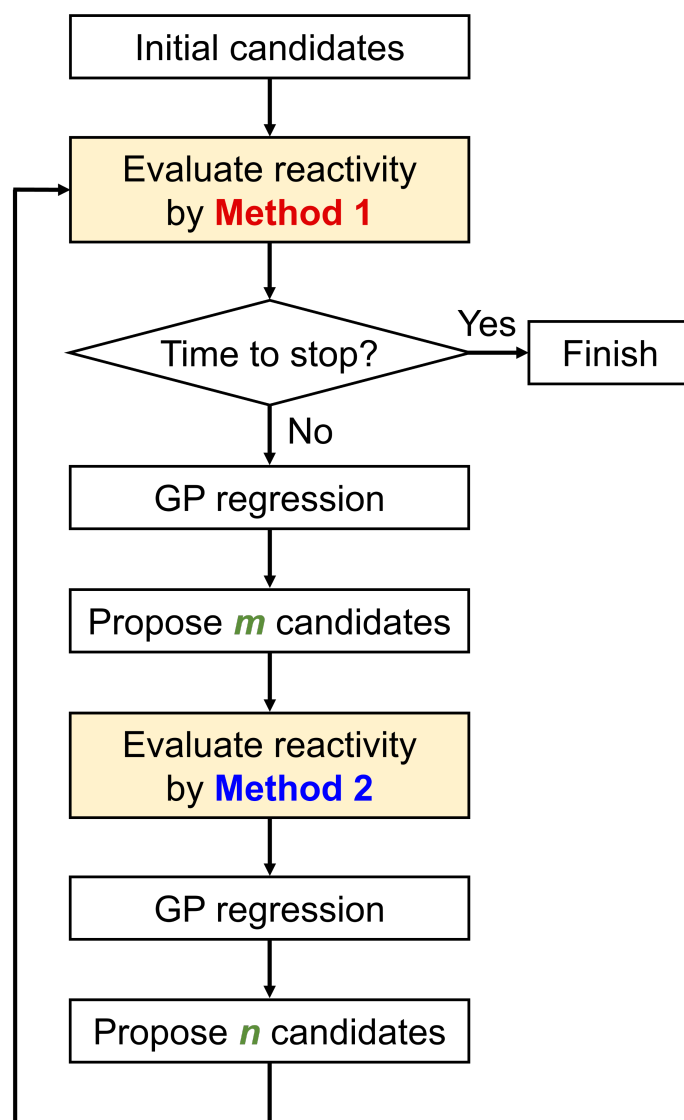


Figure 4.1. Flow of the BO-assisted substrate optimization using experimental and computational data. Method 1 and Method 2 correspond to the experiment and barrier calculation, respectively.

kJ/mol by density functional theory (DFT), and even more by semiempirical methods.

The main focus of this study, which examines the algorithm in Figure 4.1, is how to deal with the deviation between the two datasets of different nature (i.e., experimental vs computational). Figure 4.2 illustrates a similar situation where two measurement methods are combined to optimize a common value, one with high accuracy and high cost

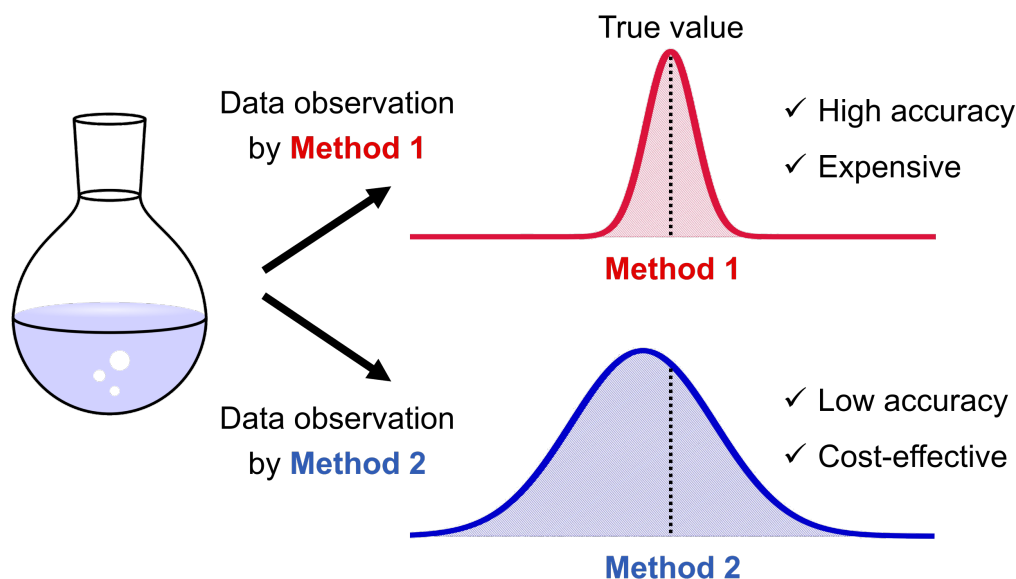


Figure 4.2. Two types of data observation approaches: Method 1 is a high accuracy and expensive observation; Method 2 is a low accuracy and cost-effective observation.

and one with low accuracy and low cost. The low-accuracy, low-cost measurement is also considered to be a biased and noisy but cost-effective source. Bayesian optimization that accesses biased and noisy but cost-effective sources of information along with accurate and expensive sources is called multi-information source Bayesian optimization [65, 66]. In the field of chemistry, multi-information source Bayesian optimization was employed to reduce the required cost of optimization and accelerate materials discovery [67], where it was demonstrated to be effective in the study of three optimization problems: global minimization of Rosenbrock function, geometry optimization of CO molecule, and maximization of intermolecular binding energies between hybrid organic-inorganic perovskites and solvent.

To investigate the performance of the algorithm in Figure 4.1 with sufficient statistical accuracy, avoiding enormous computational and experimental effort, I propose a procedure to do this virtually with practical effort instead of actually performing ex-

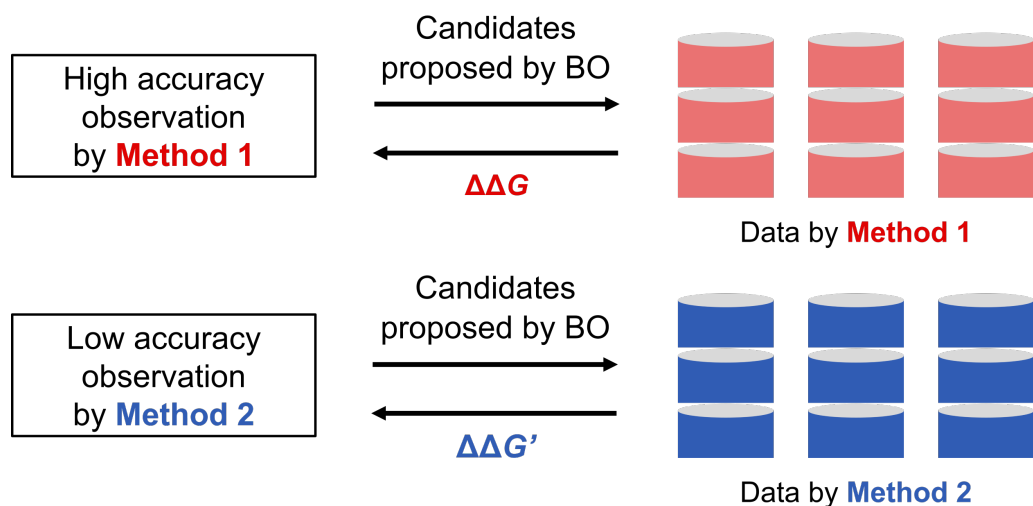


Figure 4.3. Virtual experiment (Method 1) and virtual computation (Method 2) in this study.

periments and computations. Such a virtual investigation requires two sets of reactivity data. As mentioned above, the Gibbs energy barrier is considered as an index of reactivity throughout this paper. As discussed above, computational data based on quantum chemical calculations have errors due to the approximations used to solve the electronic wave equations, and thus can be considered a biased and noisy but cost-effective source of multi-information source Bayesian optimization. In this study, it is assumed that the experimental error bar is much smaller than the errors in quantum chemical calculations. With these assumptions, I construct a model, as shown in Figure 4.3. In this model, I first adopt a large set of Gibbs energy barrier data (*vide infra*). Another dataset is created by replicating the original one and then modifying its values by adding random and systematic errors. The search for the optimal (low barrier) data in the original dataset is then performed with reference to these two datasets. In such a virtual study, the original dataset without errors is considered to be the virtual experimental dataset, and the replicated and modified dataset with errors is considered to be the virtual computational dataset. It should be emphasized that the purpose of this study is not to find a good

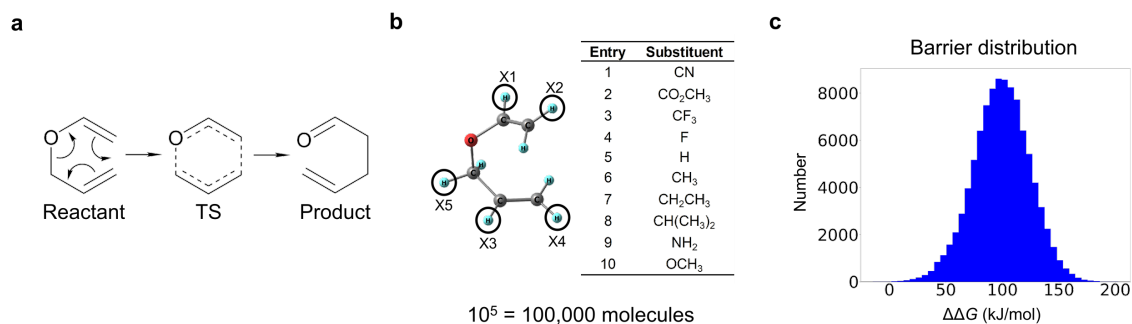


Figure 4.4. (a) Claisen rearrangement, (b) substitution sites and substituents considered in the Claisen rearrangement barrier dataset [68], and (c) the barrier distribution in the Claisen rearrangement barrier dataset.

chemical reaction in a particular system. The purpose is to study the algorithm in Figure 4.1 using data of two different accuracies and to see how effective the algorithm is in reducing the number of references to the higher accuracy data.

As a case study, this study focuses on the Claisen rearrangement and uses its theoretical Gibbs energy barrier dataset based on semiempirical PM7 theory reported in the literature [68]. The dataset contains Gibbs energy barriers of the Claisen rearrangement of allyl vinyl ether (Figure 4.4a) with $\sim 100,000$ different substitutions (Figure 4.4b). Even though all of the barriers are for the Claisen rearrangement, these barriers have a wide distribution, as illustrated in 4.4c. It is emphasized again that the purpose of this study is not to predict a new reaction using the Gibbs barrier dataset as descriptors, nor to provide an accurate dataset for any particular system; the purpose is to examine the Bayesian optimization algorithm that combines two datasets with different accuracies. For this purpose, these datasets can be anything as long as they are not too chemically unrealistic. Although the dataset I employed is based on semiempirical PM7 theory, it was carefully examined in the former publication (ref [68]) and was shown to well describe the substituent dependence of the reactivity of the Claisen rearrangement. Therefore,

the following numerical experiments using the PM7 dataset are sufficient for the present purpose.

This study conducts an extensive numerical investigation regarding these Gibbs energy barriers as virtual experimental data, while these Gibbs energy barriers plus systematic and random noises are virtual computational data, as illustrated in Figure 4.3. Such a numerical investigation avoids additional experimental and computational efforts, enabling massive trials to check statistical behaviors while adjusting various parameters. The present numerical investigation shows that even computational data including systematic and random errors accelerates BO-assisted substrate optimization and helps reduce the number of experiments.

4.2 Method

4.2.1 Barrier Dataset Utilized

To perform virtual experiments and computations, I use the Gibbs energy barrier dataset of the Claisen rearrangement [68]. The Claisen rearrangement is a [3, 3] sigmatropic rearrangement of an allyl vinyl ether, as illustrated in Figure 4.4a. This dataset includes theoretical Gibbs energy barriers of the Claisen rearrangement for 100,000 molecules computed based on a semiempirical theory PM7 [69]. Each molecule is generated by substituting 5 hydrogen atoms of the unsubstituted allyl vinyl ether molecule with 10 substituents. Five substitution sites and 10 substituents considered in the dataset are summarized in Figure 4.4b. Although there are 928 missing values in the dataset, these values were filled by random forest regression. It should be noted that the deviation between the PM7 dataset and the more accurate computational dataset obtained by, e.g., DFT is not essential in the present numerical investigation, which focuses purely on the deviations between the original dataset and the replicated and modified dataset. The only

requirement for the dataset used in this study is that it is chemically reasonable. The PM7 dataset has been carefully studied in the literature and has been shown to reproduce well the chemical trends discussed in previous DFT studies [70, 71]. Furthermore, after extensive numerical analysis, it has been shown that the dataset reflects well the chemical trends of different substitutions and is useful for testing different implementations of an optimization algorithm [68].

In BO and random forest regression, each substituent is represented by a descriptor variable. As descriptor variables, eight parameters, HOMO energy, LUMO energy, HOMO-LUMO gap, NBO charge, Hammett constant [72], and Sterimol parameters (substituent length: L , minimum radius: B_1 , and maximum radius: B_5) of each substituent capped by a hydrogen atom are used in this study. The reactivity of the Claisen rearrangement depends on the partial charge of the transition state, the stability of the product, and the interaction of the frontier orbitals. Previous studies have reported that these factors are controlled by substituent effects [70, 71]. Therefore, it is expected that the energies of the substituent frontier orbitals and the NBO charge will correlate with the reaction barrier. The Hammett constant, which represents the electron-donating or electron-withdrawing property of a substituent, and the Sterimol parameter, which represents the steric hindrance of a substituent, are also thought to affect reactivity.

HOMO energy, LUMO energy, HOMO-LUMO gap, and NBO charge were calculated using the Gaussian 16 software package [73] by DFT at the ω B97X-D/def2-SVP [74, 75] level of theory. Hammett constants σ_p were taken from the literature. Sterimol parameters L , B_1 , and B_5 were calculated using Winmoster [76]. Since each aryl-vinyl-ether molecule in the dataset has 5 substitution sites, each molecule is represented by 40 descriptor variables. Values of these parameters are listed in Table 4.1.

Table 4.1. Lists of descriptor variables for each substituent.

Substituent	HOMO (eV)	LUMO (eV)	Gap (eV)	NBO charge	σ_p	L (Å)	B ₁ (Å)	B ₅ (Å)
CN	-12.27	2.58	14.85	0.22152	0.66	3.7810	1.7000	1.7000
CO ₂ CH ₃	-10.13	2.25	12.38	0.13372	0.45	4.6592	1.7901	3.7086
CF ₃	-13.11	4.38	17.49	0.12820	0.54	3.0327	2.0928	2.7158
F	-12.99	3.80	16.79	0.53020	0.06	2.3889	1.4700	1.4700
H	-13.86	4.25	18.11	0.00000	0.00	1.9578	1.2000	1.2000
CH ₃	-12.86	4.27	17.13	0.21486	-0.17	2.7979	1.7171	2.2351
CH ₂ CH ₃	-11.45	4.01	15.45	0.21158	-0.15	3.9584	1.7620	3.3698
CH(CH ₃) ₂	-11.00	3.81	14.81	0.21332	-0.15	3.9102	2.1241	3.3814
NH ₂	-9.20	3.85	13.04	0.35184	-0.66	2.5695	1.5500	2.1817
OCH ₃	-9.56	3.84	13.40	0.45253	-0.27	3.7011	1.5200	3.2946

4.2.2 Input and Output

BO is a sequential global optimization strategy to find $\mathbf{x}^*$ that maximizes or minimizes output $f(\mathbf{x})$. In substrate optimization, input $\mathbf{x}$ is a set of descriptor variables representing a substrate combination. The output $f(\mathbf{x})$ is an objective value such as reaction yield, rate constant, and Gibbs energy barrier. In this study, $\mathbf{x}$ is a set of descriptor variables representing an allyl vinyl ether molecule, and $f(\mathbf{x})$ is the Gibbs energy barrier of their Claisen rearrangement.

In the present numerical investigation, a virtual experiment returns the Gibbs energy barrier for the corresponding allyl vinyl ether, i.e., $\Delta\Delta G$, from the dataset. On the other hand, a virtual computation returns the Gibbs energy barrier for the corresponding allyl vinyl ether from the dataset after adding systematic and random noises ϵ , i.e., $\Delta\Delta G + \epsilon$, where ϵ is the Gaussian noise following the normal distribution $\mathcal{N}(\mu, \sigma^2)$ with mean μ and variance σ^2 . In quantum chemical calculations, a specific computational method often systematically underestimates or overestimates the Gibbs energy barriers for a particular chemical transformation. The former μ represents such a systematic deviation. The latter σ^2 expresses deviations caused by factors having no systematic trend. It is noted that pre-

cisely representing a deviation for a specific computational level is not the purpose of this study. Instead, this numerical investigation aims to show that such virtual computational results can help reduce the number of virtual experiments.

4.2.3 Implementation of Bayesian Optimization

Figure 4.1 presents my implementation of BO utilizing both virtual experimental (Method 1) and computational (Method 2) results. First, Gibbs energy barriers are evaluated by virtual experiments for n allyl vinyl ether molecules, where these are randomly chosen in the initial cycle or otherwise are ones proposed in the last cycle. The Gibbs energy barrier data are standardized to have a mean of 0 and a standard deviation of 1.

Second, a posterior mean and variance for the reaction barrier, representing its expected value and the uncertainty of the prediction, respectively, are estimated by GP regression. In the implementation of GP, a Python package GPyTorch is used [77]. As a kernel function, this study employs the radial basis function kernel:

$$k_{\text{RBF}}(\mathbf{x}_1, \mathbf{x}_2) = \theta_1 \exp\left(-\frac{1}{2}(\mathbf{x}_1 - \mathbf{x}_2)^\top \Theta^{-2}(\mathbf{x}_1 - \mathbf{x}_2)\right) \quad (4.1)$$

where θ_1 is the output scale parameter and Θ_2 is the length scale parameter. These hyperparameters determine the shape of the covariance function and need to be optimized to fit the observed data. This operation is performed via the type-II maximum a posteriori estimation. In this step, a γ prior is assigned over each hyperparameter. The γ distribution is a type of continuous probability distribution and has two parameters called a shape parameter α and a rate parameter β . In my implementation, α and β for length scale parameter and output scale parameter of radial basis function kernel, and noise were specified as shown in Table 4.2. Then, m allyl-vinyl-ether molecules to be evaluated next by virtual computations are proposed via an acquisition function. As the acquisition func-

Table 4.2. Lists of shape parameter and rate parameter for each hyperparameter.

Parameter name	α	β
Length scale parameter	1.5	0.1
Output scale parameter	7.0	1.0
Noise	1.5	0.5

tion, I adopt the expected improvement (EI) function. The best candidate is estimated by evaluating EI values for all possible allyl vinyl ether molecules, where candidates that a virtual experiment or computation has already done are excluded to avoid duplication. In my initial tests, the EI function outperformed the other acquisition functions such as the probability improvement function and the upper confident bound function for the present system. Adopting the Kriging believer algorithm [20], the GP regression and the subsequent candidate proposal are iterated m times. In the first iteration, all previously evaluated virtual experimental and computational results are used with the same weight as data points for the GP regression. In the k th iteration, in addition to the virtual experimental and computational values, $(k - 1)$ GP predicted mean values obtained in the first to $(k - 1)$ -th iterations for the candidates proposed in the corresponding iterations are additionally used with the same weight as data points for the GP regression. In my initial tests, this Kriging believer algorithm outperformed the other multiple candidate suggestion schemes, such as the ϵ -greedy and Thompson sampling for the present system [78, 79].

Third, Gibbs energy barriers are evaluated by virtual computations for the proposed m allyl vinyl ether molecules. Then, n candidate molecules are proposed based on the GP regression through the same procedure taken in the second step, where candidates that virtual experiment has already done are excluded to avoid duplication. The cycle is terminated after acquiring 100 virtual experimental results in this study.

4.3 Results and Discussion

4.3.1 Performance Comparison

To clarify whether computational data can be used to reduce the number of experiments, I benchmarked the performance of the algorithm in Figure 4.1. The success rate of finding one of the 10 most reactive allyl vinyl ether (one of those having the 10 lowest Gibbs energy barriers) within 100 virtual experiments was evaluated to benchmark the performance. The numerical investigations were done for the case without virtual computation ($m = 0$) and for all of the possible combinations of the virtual computational batch size $m = 5, 10, 15, 20, 25$, and 30 , mean deviation $\mu = -20.0, -10.0, 0.0, 10.0$, and 20.0 kJ/mol, and standard deviation $\sigma = 0.0, 5.0, 10.0, 15.0$, and 20.0 kJ/mol. In all cases, the virtual experimental batch size n was fixed to 5. I performed these numerical investigations 200 times for each combination, given different initial candidates and Gaussian noise ϵ , and discussed their averages as statistically (nearly) converged results. In total, 30,000 numerical investigations were thus performed. Using the literature dataset enabled such a massive test requiring millions of experiments.

Figure 4.5 shows the changes in performance for different combinations of m , μ , and σ . In each panel of Figure 4.5, the blue line denoted by $m = 0$ corresponds to the reference obtained without the use of virtual computational results. All of the other lines in Figure 4.5 show better performance than the reference. Even the worst $\mu = 20.0$ and $\sigma = 20.0$ kJ/mol case in panel o shows a better success rate at the 20th batch (100th virtual experiment) than the reference. Knowing that even virtual computational results including such a significant error help improve the success rate and consequently reduce the number of virtual experiments is encouraging.

Figure 4.5 also shows that the performance changes, depending on m and σ . When

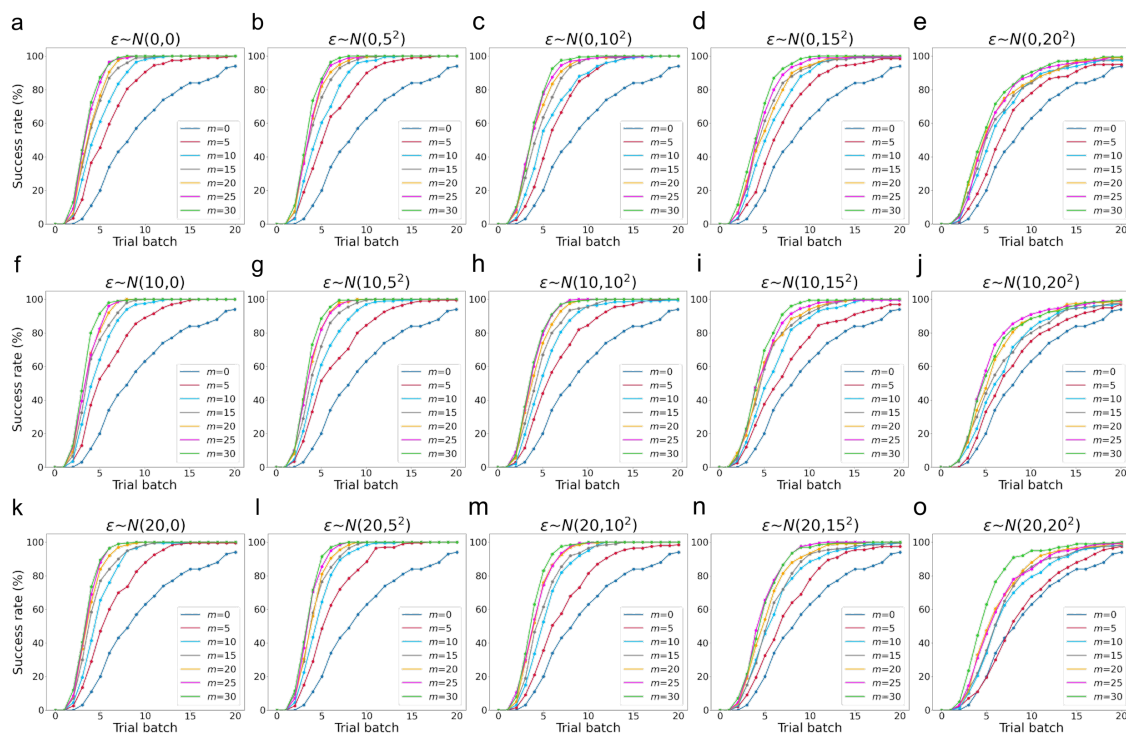


Figure 4.5. Rise in the success rate depending on the increase in the number of batches under each combination of m , negative μ , and σ .

m increases, the success rate rises rapidly. This is an intuitively reasonable result that using more virtual computational data improves performance. When σ increases from the left panel to the right panel in Figure 4.5, the success rate decreases. In contrast, μ does not affect the performance much compared to m and σ . This implies that the GP regression easily recognizes and learns the systematic deviation. Similar trends are also found for $\mu = -10$ and -20 kJ/mol cases, as shown in Figure 4.6.

It was also found that the success rate decreases compared to the reference when μ and σ exceed the range $|\mu| \leq 20$ kJ/mol and $\sigma \leq 20$ kJ/mol. Figure 4.7 shows the changes in performance for the case without the use of virtual computation ($m = 0$) and for all of the possible combinations of the virtual computational batch size $m = 5, 10, 15, 20, 25$, and 30 , mean deviation $\mu = 20.0, 30.0$, and 40.0 kJ/mol, and standard deviation $\sigma = 20.0, 25.0, 30.0, 35.0$, and 40.0 kJ/mol. As shown in Figure 4.7, when the computational error

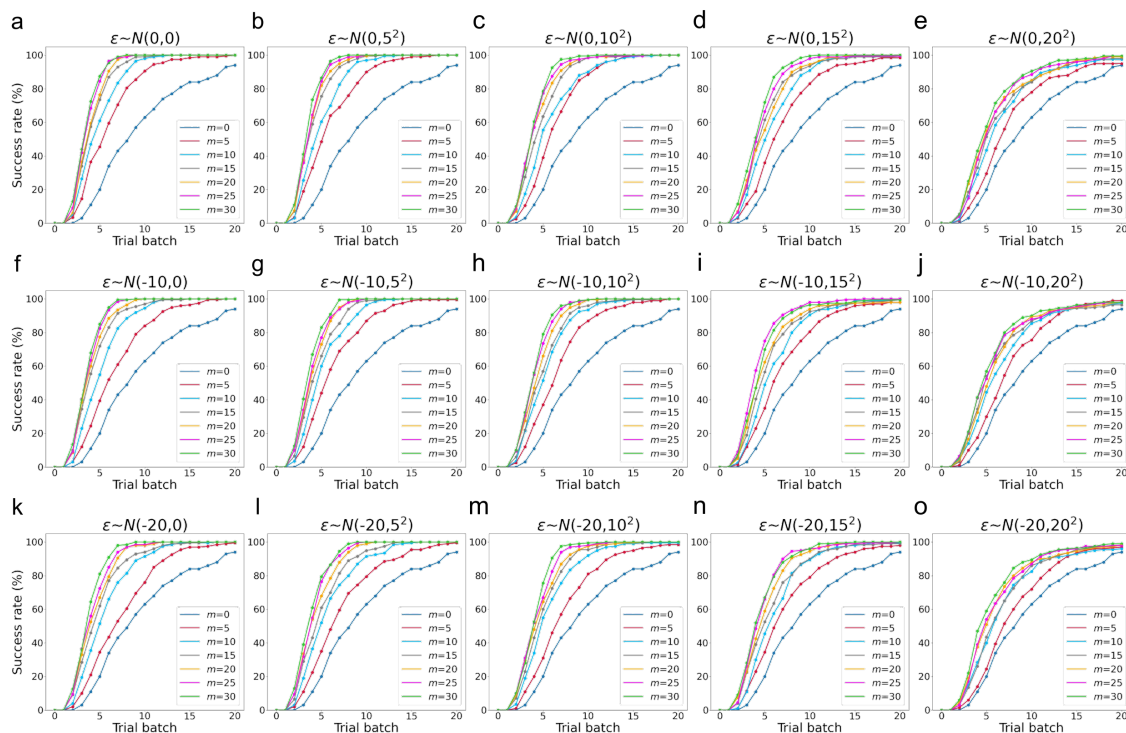


Figure 4.6. Rise in the success rate depending on the increase in the number of batches under different combinations of m , μ , and σ , when μ is negative.

is too large, my method cannot accelerate the convergence of Bayesian optimization and requires more virtual experiments than general Bayesian optimization. Therefore, when the computational error is too large, e.g. exceeding the range $|\mu| \leq 20$ kJ/mol and $\sigma \leq 20$ kJ/mol in the present case, my method does not work well.

4.3.2 Heatmap Analysis

These trends are further visualized on the heatmaps in Figure 4.8 for different virtual computational batch sizes. Each heatmap presents the number of virtual experiments required to reach the 95% success rate. As m increases, the blue area in which the 95% success rate can be achieved by virtual experiments less than 50 times increases. Such a trend indicates that the number of experiments can be reduced using more computational results. As seen in Figure 4.5a, the improvement saturates when m is increased over a

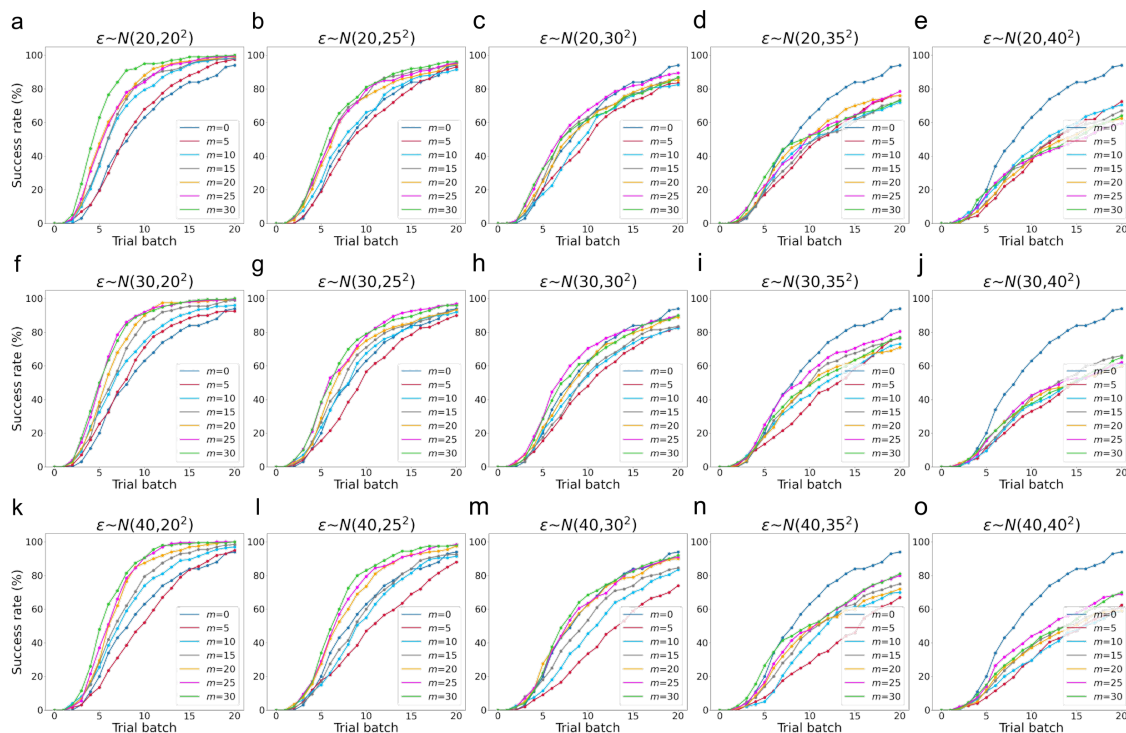


Figure 4.7. Rise in the success rate depending on the increase in the number of batches under different combinations of m , μ , and σ , when μ , and σ are large.

certain level, even with $\mu = \sigma = 0.0$ kJ/mol. This would be due to the Kriging believer algorithm, which suggests multiple candidates using a GP model and emphasizes errors in the GP model.

Figure 4.8 also shows that the change in performance is highly dependent on σ , but not so much on μ . However, all of the heatmaps are asymmetric and show better performance when μ is positive. This is because the virtual computational results bias the GP model to accept molecules like those already evaluated by virtual computation when μ is negative (underestimating their barriers), especially in the early optimization stage. On the other hand, the virtual computational results bias the GP model to avoid molecules like ones already evaluated by virtual computation when μ is positive (overestimating their barriers). Such a difference in the early stage explains the asymmetric heatmaps in Figure 4.8. Although numerical investigations were performed only using BO in this study, I

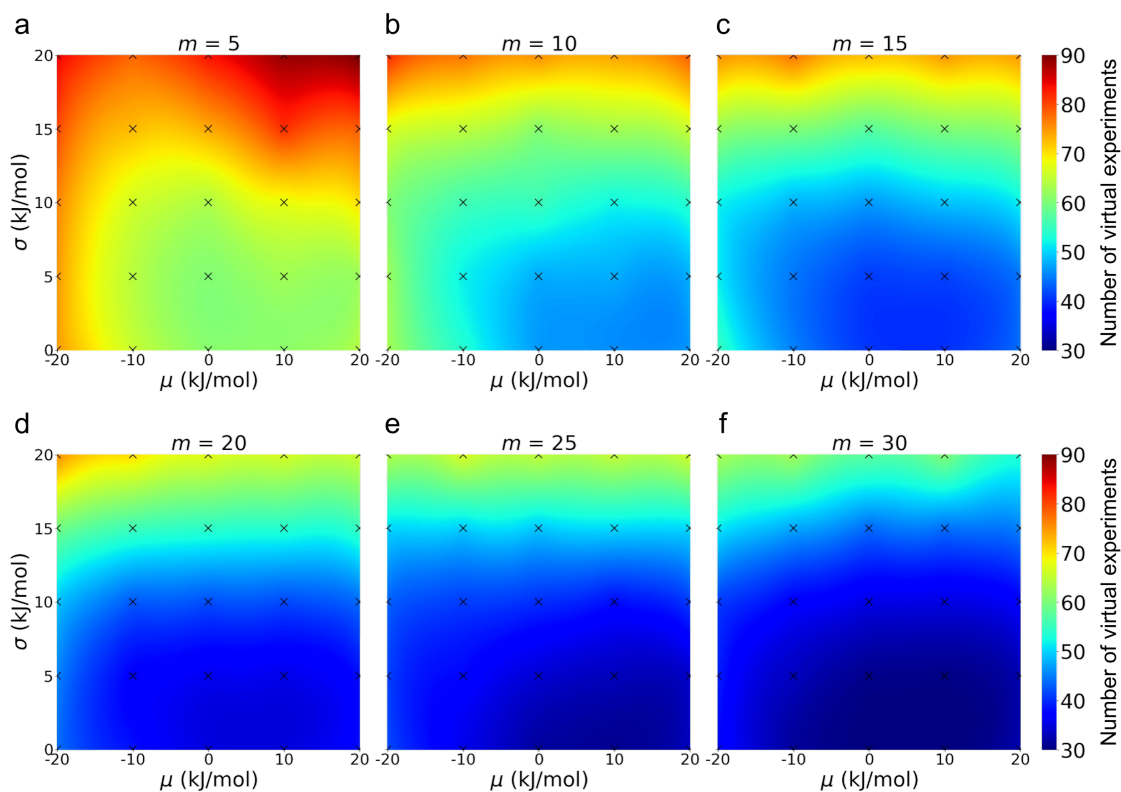


Figure 4.8. Heatmap of the number of virtual experiments required to reach the 95% success rate of finding the top 10 reactions from the Claisen barrier dataset.

expect similar results to be obtained even with other optimization algorithms such as the genetic algorithm and particle swarm optimization algorithm.

4.3.3 Limitation

Finally, I discuss issues one needs to consider when one uses the algorithm in Figure 4.1. One is how to determine the rate-determining step from a multistep reaction profile and compute an overall reactivity [80]. Another is how to accurately determine the ensemble of various conformations at the rate-determining TS [45–52, 81]. There also are methods to directly determine the reaction yield through automatic reaction path network explorations considering formations of various resting and byproduct states [82, 83]. Evaluating the reactivity accurately, considering an entire reaction profile or reaction path

network, could often be more time-consuming than a single experiment. On the other hand, focusing only on a few elementary steps could mislead BO due to a wrong kinetic description based on an oversimplified model. In a similar vein, the choice of the computational level is also crucial. In general, cost and efficiency are a trade-off in quantum chemical calculations. However, using computational methods with lower computational costs can result in larger computational errors and weaken the superiority of my method. Therefore, when using my method, it is important to carefully determine the appropriate computational level for barrier calculations, taking into consideration the available computational resources and costs and taking note of a balance between the time and cost required for a single experiment and a single barrier calculation. While my method may reduce the number of experiments, it may not be as effective for reactions that require a significant computational effort to compute the reaction barrier. On the other hand, reactions explainable with a relatively simple mechanism possessing a clear rate-determining step would be suitable targets for the algorithm shown in Figure 4.1.

4.4 Conclusion

In this study, I virtually validated the impact of using computational Gibbs energy barriers in BO-assisted substrate optimization and showed that it potentially accelerates BO-assisted substrate optimization and reduces the number of required experiments. To perform a thorough numerical investigation, Gibbs energy barriers in the literature dataset were employed virtually regarding the Gibbs energy barriers as experimental results and those with systematic and random noises as computational ones. Such a numerical model allowed me to examine as many as 30,000 cases and acquire statistically (nearly) converged outputs. The present numerical validation suggested that the use of computational Gibbs energy barriers considerably accelerates substrate optimization, especially when a

computational method affected by a small random error is adopted. It was also found that systematic errors are not a serious matter. Moreover, it was shown that even a computational method affected by a relatively sizable random error of 20 kJ/mol, more significant than the typical error by DFT, is sufficiently helpful in reducing the number of experiments. These findings encourage the utilization of computational Gibbs energy barriers in actual substrate optimizations. Although this study has focused on the optimization of the combination of substituents in the reactants of the Claisen rearrangement, my method can easily be applied to reaction optimizations of different types, varying temperatures, solvents, ligands on organometallic complexes, etc. In these cases, barrier calculations are also carried out by changing these conditions. A further study combining real experimental and computational results is in progress in my group.

4.5 Data Availability

All of the barrier data used in this study were taken from the Supporting Information of ref 68.

4.6 Code Availability

The Python code used for implementing Bayesian optimization and benchmarking its performance can be found at: <https://github.com/Hiroaki-Okada/bo-comp-barrier>.

References

- [1] A. Milo, E. N. Bess, M. S. Sigman, *Nature* **2014**, *507*, 210–214.
- [2] Z. L. Niemeyer, A. Milo, D. P. Hickey, M. S. Sigman, *Nature Chemistry* **2016**, *8*, 610–617.

- [3] Y. Park, Z. L. Niemeyer, J.-Q. Yu, M. S. Sigman, *Organometallics* **2018**, *37*, 203–210.
- [4] A. Yada, K. Nagata, Y. Ando, T. Matsumura, S. Ichinoseki, K. Sato, *Chemistry Letters* **2018**, *47*, 284–287.
- [5] D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher, A. G. Doyle, *Science* **2018**, *360*, 186–190.
- [6] M. K. Nielsen, D. T. Ahneman, O. Riera, A. G. Doyle, *Journal of the American Chemical Society* **2018**, *140*, 5004–5008.
- [7] A. F. Zahrt, J. J. Henle, B. T. Rose, Y. Wang, W. T. Darrow, S. E. Denmark, *Science* **2019**, *363*, eaau5631.
- [8] P. Friederich, G. Dos Passos Gomes, R. De Bin, A. Aspuru-Guzik, D. Balcells, *Chemical Science* **2020**, *11*, 4584–4601.
- [9] C. W. Coley, R. Barzilay, T. S. Jaakkola, W. H. Green, K. F. Jensen, *ACS Central Science* **2017**, *3*, 434–443.
- [10] J. M. Granda, L. Donina, V. Dragone, D.-L. Long, L. Cronin, *Nature* **2018**, *559*, 377–381.
- [11] C. W. Coley, W. H. Green, K. F. Jensen, *Accounts of Chemical Research* **2018**, *51*, 1281–1289.
- [12] M. H. S. Segler, M. Preuss, M. P. Waller, *Nature* **2018**, *555*, 604–610.
- [13] H. Gao, T. J. Struble, C. W. Coley, Y. Wang, W. H. Green, K. F. Jensen, *ACS Central Science* **2018**, *4*, 1465–1476.
- [14] K. C. Harper, E. N. Bess, M. S. Sigman, *Nature Chemistry* **2012**, *4*, 366–374.

- [15] C. B. Santiago, J.-Y. Guo, M. S. Sigman, *Chemical Science* **2018**, *9*, 2398–2412.
- [16] J. P. Reid, M. S. Sigman, *Nature* **2019**, *571*, 343–348.
- [17] J. N. Wei, D. Duvenaud, A. Aspuru-Guzik, *ACS Central Science* **2016**, *2*, 725–732.
- [18] K. Wu, A. G. Doyle, *Nature Chemistry* **2017**, *9*, 779–784.
- [19] P. I. Frazier, *A Tutorial on Bayesian Optimization*, **2018**.
- [20] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, N. De Freitas, *Proceedings of the IEEE* **2016**, *104*, 148–175.
- [21] A. M. Schweidtmann, A. D. Clayton, N. Holmes, E. Bradford, R. A. Bourne, A. A. Lapkin, *Chemical Engineering Journal* **2018**, *352*, 277–282.
- [22] B. Burger, P. M. Maffettone, V. V. Gusev, C. M. Aitchison, Y. Bai, X. Wang, X. Li, B. M. Alston, B. Li, R. Clowes, N. Rankin, B. Harris, R. S. Sprick, A. I. Cooper, *Nature* **2020**, *583*, 237–241.
- [23] P. M. Attia, A. Grover, N. Jin, K. A. Severson, T. M. Markov, Y.-H. Liao, M. H. Chen, B. Cheong, N. Perkins, Z. Yang, P. K. Herring, M. Aykol, S. J. Harris, R. D. Braatz, S. Ermon, W. C. Chueh, *Nature* **2020**, *578*, 397–402.
- [24] C. B. Wahl, M. Aykol, J. H. Swisher, J. H. Montoya, S. K. Suram, C. A. Mirkin, *Science Advances* **2021**, *7*, eabj5505.
- [25] T. S. Frisby, Z. Gong, C. J. Langmead, *Bioinformatics* **2021**, *37*, i451–i459.
- [26] Y. Xie, C. Zhang, H. Deng, B. Zheng, J.-W. Su, K. Shutt, J. Lin, *ACS Applied Materials & Interfaces* **2021**, *13*, 53485–53491.
- [27] R. Liang, X. Duan, J. Zhang, Z. Yuan, *Reaction Chemistry & Engineering* **2022**, *7*, 590–598.

- [28] R. V. Krems, *Physical Chemistry Chemical Physics* **2019**, *21*, 13392–13410.
- [29] S. Langner, F. Häse, J. D. Perea, T. Stubhan, J. Hauch, L. M. Roch, T. Heumueller, A. Aspuru-Guzik, C. J. Brabec, *Advanced Materials* **2020**, *32*, 1907801.
- [30] Y. Zhang, D. W. Apley, W. Chen, *Scientific Reports* **2020**, *10*, 4924.
- [31] R. A. Vargas-Hernández, *The Journal of Physical Chemistry A* **2020**, *124*, 4053–4061.
- [32] T. Ueno, T. D. Rhone, Z. Hou, T. Mizoguchi, K. Tsuda, *Materials Discovery* **2016**, *4*, 18–21.
- [33] F. Häse, L. M. Roch, C. Kreisbeck, A. Aspuru-Guzik, *ACS Central Science* **2018**, *4*, 1134–1145.
- [34] F. Häse, M. Aldeghi, R. J. Hickman, L. M. Roch, A. Aspuru-Guzik, *Applied Physics Reviews* **2021**, *8*, 031406.
- [35] L. Chan, G. R. Hutchison, G. M. Morris, *Physical Chemistry Chemical Physics* **2020**, *22*, 5211–5219.
- [36] K. C. Felton, J. G. Rittig, A. A. Lapkin, *Chemistry–Methods* **2021**, *1*, 116–122.
- [37] M. Aldeghi, F. Häse, R. J. Hickman, I. Tamblyn, A. Aspuru-Guzik, *Chemical Science* **2021**, *12*, 14792–14807.
- [38] Y. Wang, T.-Y. Chen, D. G. Vlachos, *Journal of Chemical Information and Modeling* **2021**, *61*, 5312–5319.
- [39] B. J. Shields, J. Stevens, J. Li, M. Parasram, F. Damani, J. I. M. Alvarado, J. M. Janey, R. P. Adams, A. G. Doyle, *Nature* **2021**, *590*, 89–96.

- [40] M. Christensen, L. P. E. Yunker, F. Adediji, F. Häse, L. M. Roch, T. Gensch, G. Dos Passos Gomes, T. Zepel, M. S. Sigman, A. Aspuru-Guzik, J. E. Hein, *Communications Chemistry* **2021**, *4*, 112.
- [41] N. Sugisawa, H. Sugisawa, Y. Otake, R. V. Krems, H. Nakamura, S. Fuse, *Chemistry-Methods* **2021**, *1*, 484–490.
- [42] N. V. Faurschou, R. H. Taaning, C. M. Pedersen, *Chemical Science* **2023**, *14*, 6319–6329.
- [43] Nobuaki. Koga, Keiji. Morokuma, *Chemical Reviews* **1991**, *91*, 823–842.
- [44] K. N. Houk, P. H.-Y. Cheong, *Nature* **2008**, *455*, 309–313.
- [45] W. Thiel, *Angewandte Chemie International Edition* **2014**, *53*, 8605–8613.
- [46] W. M. C. Sameera, S. Maeda, K. Morokuma, *Accounts of Chemical Research* **2016**, *49*, 763–773.
- [47] K. N. Houk, F. Liu, *Accounts of Chemical Research* **2017**, *50*, 539–543.
- [48] S. Maeda, K. Ohno, *Journal of the American Chemical Society* **2008**, *130*, 17228–17229.
- [49] P. J. Donoghue, P. Helquist, P.-O. Norrby, O. Wiest, *Journal of the American Chemical Society* **2009**, *131*, 410–411.
- [50] M. Hatanaka, S. Maeda, K. Morokuma, *Journal of Chemical Theory and Computation* **2013**, *9*, 2882–2886.
- [51] E. Hansen, A. R. Rosales, B. Tutkowski, P.-O. Norrby, O. Wiest, *Accounts of Chemical Research* **2016**, *49*, 996–1005.

- [52] Y. Guan, V. M. Ingman, B. J. Rooks, S. E. Wheeler, *Journal of Chemical Theory and Computation* **2018**, *14*, 5249–5261.
- [53] W. Matsuoka, Y. Harabuchi, S. Maeda, *ACS Catalysis* **2022**, *12*, 3752–3766.
- [54] S. Ahn, M. Hong, M. Sundararajan, D. H. Ess, M.-H. Baik, *Chemical Reviews* **2019**, *119*, 6509–6560.
- [55] A. R. Rosales, J. Wahlers, E. Limé, R. E. Meadows, K. W. Leslie, R. Savin, F. Bell, E. Hansen, P. Helquist, R. H. Munday, O. Wiest, P.-O. Norrby, *Nature Catalysis* **2019**, *2*, 41–45.
- [56] M. Foscato, V. R. Jensen, *ACS Catalysis* **2020**, *10*, 2354–2377.
- [57] T. Mita, Y. Harabuchi, S. Maeda, *Chemical Science* **2020**, *11*, 7569–7577.
- [58] M. Mukai, K. Nagao, S. Yamaguchi, H. Ohmiya, *Bulletin of the Chemical Society of Japan* **2022**, *95*, 271–277.
- [59] S. Yamaguchi, *Organic & Biomolecular Chemistry* **2022**, *20*, 6057–6071.
- [60] S. Gallarati, R. Fabregat, R. Laplaza, S. Bhattacharjee, M. D. Wodrich, C. Corminboeuf, *Chemical Science* **2021**, *12*, 6879–6889.
- [61] S. M. Maley, D.-H. Kwon, N. Rollins, J. C. Stanley, O. L. Sydora, S. M. Bischof, D. H. Ess, *Chemical Science* **2020**, *11*, 9665–9674.
- [62] A. Fernández-Ramos, J. A. Miller, S. J. Klippenstein, D. G. Truhlar, *Chemical Reviews* **2006**, *106*, 4518–4584.
- [63] J. S. Shaffer, A. K. Chakraborty, *Macromolecules* **1993**, *26*, 1120–1136.
- [64] M. Rae, M. N. Berberan-Santos, *Journal of Chemical Education* **2004**, *81*, 436.

- [65] K. Swersky, J. Snoek, R. P. Adams in *Advances in Neural Information Processing Systems*, Curran Associates, Inc.
- [66] M. Poloczek, J. Wang, P. Frazier in *Advances in Neural Information Processing Systems*, Curran Associates, Inc.
- [67] H. C. Herbol, M. Poloczek, P. Clancy, *Materials Horizons* **2020**, *7*, 2113–2123.
- [68] H. Okada, S. Maeda, *Molecular Informatics* **2022**, *41*, 2100216.
- [69] J. J. P. Stewart, *Journal of Molecular Modeling* **2013**, *19*, 1–32.
- [70] V. Aviyente, H. Y. Yoo, K. N. Houk, *The Journal of Organic Chemistry* **1997**, *62*, 6121–6128.
- [71] V. Aviyente, K. N. Houk, *The Journal of Physical Chemistry A* **2001**, *105*, 383–391.
- [72] Corwin. Hansch, A. Leo, R. W. Taft, *Chemical Reviews* **1991**, *91*, 165–195.
- [73] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W.

- Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, D. J. Fox, *Gaussian 16 Revision C.01*, **2016**, Gaussian Inc. Wallingford CT.
- [74] J.-D. Chai, M. Head-Gordon, *Physical Chemistry Chemical Physics* **2008**, *10*, 6615–6620.
- [75] F. Weigend, R. Ahlrichs, *Physical Chemistry Chemical Physics* **2005**, *7*, 3297–3305.
- [76] Winmostar, *Winmostar*, <https://winmostar.com/jp/>, **2023**, Accessed on November 26, 2023.
- [77] J. Gardner, G. Pleiss, K. Q. Weinberger, D. Bindel, A. G. Wilson in *Advances in Neural Information Processing Systems*, Curran Associates, Inc.
- [78] K. Kandasamy, A. Krishnamurthy, J. Schneider, B. Póczos in *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, PMLR, pp. 133–142.
- [79] J. M. Hernández-Lobato, J. Requeima, E. O. Pyzer-Knapp, A. Aspuru-Guzik in *Proceedings of the 34th International Conference on Machine Learning*, PMLR, pp. 1470–1479.
- [80] S. Kozuch, S. Shaik, *Accounts of Chemical Research* **2011**, *44*, 101–110.
- [81] S. Chen, T. Nielson, E. Zalt, B. B. Skjelstad, B. Borough, W. J. Hirschi, S. Yu, D. Balcells, D. H. Ess, *Topics in Catalysis* **2022**, *65*, 312–324.
- [82] Y. Sumiya, Y. Harabuchi, Y. Nagata, S. Maeda, *JACS Au* **2022**, *2*, 1181–1188.
- [83] H. Hayashi, H. Katsuyama, H. Takano, Y. Harabuchi, S. Maeda, T. Mita, *Nature Synthesis* **2022**, *1*, 804–814.

Chapter 5

Regioselectivity Maximization of the Thermal Azide-alkyne Cycloadditions Based on Bayesian Substrate Optimization

5.1 Introduction

The azide-alkyne cycloaddition between organic azide and alkyne is a powerful organic reaction that allows for the direct synthesis of 1,2,3-triazoles. This reaction is a type of 1,3-dipolar addition reaction proposed by Smith and generalized by Rolf Hüsken, in which a 1,3-dipolar azide and a dipolarophile alkyne undergo pericyclic cyclization [1–3]. In the azide-alkyne cycloadditions, both 1,4- and 1,5-disubstituted isomers with different relative positions of substituents can be produced, typically resulting in a mixture of the two isomers (Figure 5.1). Extensive research has been conducted to control the regioselectivity of this reaction.

In 2002, the copper(I)-catalyzed azide-alkyne cycloaddition was reported, demonstrating the selective synthesis of 1,4-disubstituted isomers [4–7]. This reaction is consistent with the concept of "click chemistry" proposed by Sharpless and co-workers and is a process that yields a single isomer in high yield under mild conditions [8]. Subsequently,

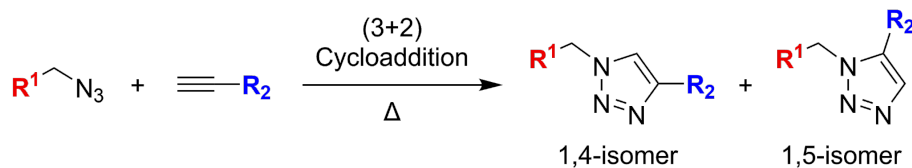


Figure 5.1. Azide-alkyne cycloaddition and two isomers.

in 2005, the ruthenium(II)-catalyzed azide-alkyne 1,3-cycloaddition was introduced, enabling the regioselective formation of 1,5-disubstituted isomers [9–12]. These discoveries made it possible to control the selectivity of the azide-alkyne cycloadditions in the presence of catalysts. Despite these remarkable successes, controlling the selectivity of the azide-alkyne cycloadditions under non-catalytic conditions remains a topic of interest [13–17]. In this study, I have focused on maximizing the regioselectivity of the catalyst-free thermal azide-alkyne cycloaddition through reaction screening.

In recent years, informatics and quantum chemical calculations have emerged as promising tools for reaction screening and reducing experimental trial and error, and minimizing cost. Informatics techniques, such as machine learning and optimization algorithms, have proven successful in various applications, including the prediction of reaction yields, reaction selectivities, forward synthesis designs, retrosynthetic analyses, and optimizing reaction conditions [18–29]. Notably, Bayesian optimization is an algorithm that efficiently explores optimal solutions by balancing "exploitation" which focuses on areas where optimal solutions are expected, and "exploration" which investigates areas where data collection is inadequate [30,31]. Bayesian optimization consists of two main processes: (1) predicting the objective variable and assessing prediction uncertainty using a surrogate model, such as a Gaussian process (GP), and (2) proposing candidate solutions for the next iteration via an acquisition function. Bayesian optimization is widely employed in the field of chemistry and has demonstrated success in many problems including optimizing

reaction conditions [32–41].

On the other hand, Gibbs free energy changes of chemical reactions calculated by quantum chemical calculations have been used to analyze and interpret experimental reactivity, and selectivity [42–52]. Some studies have even directly predicted the reaction outcome prior to experiments [53–63]. Thus, quantum chemical calculation can predict desirable substrates and reaction conditions by evaluating Gibbs free energy barrier of a reaction without experiments. Especially in one-step reactions with a clear transition state (TS) structure, the height of the reaction barrier and its difference directly determine reactivity and selectivity, making it easy to predict experimental results.

Here, I screened a combination of substrates of an azide and an alkyne using substrate optimization algorithms with Bayesian optimization and quantum chemical calculations to maximize the regioselectivity of the thermal azide-alkyne cycloadditions, especially the regioselectivity of 1,5-disubstituted isomers. Three different optimization algorithms were implemented and run independently to compare the experimental and computational costs required for optimization. The optimization algorithms propose combinations of substrates whose regioselectivity is examined from a search space of 256 combinations of substrates, based on available experimental and theoretical regioselectivity data. The regioselectivity of the thermal azide-alkyne cycloaddition in the proposed combinations was evaluated using an automated synthesis robot and quantum chemical calculations.

5.2 Method

5.2.1 Reaction Space

In experiments, azide and terminal alkyne substrates were dissolved in *N,N*-dimethylformamide and stirred at 100 °C for 18 h (Figure 5.2a). I defined the reaction space with a combinatorial set of 16 azides (Figure 5.2b) and 16 alkynes (Figure 5.2c), resulting in a total

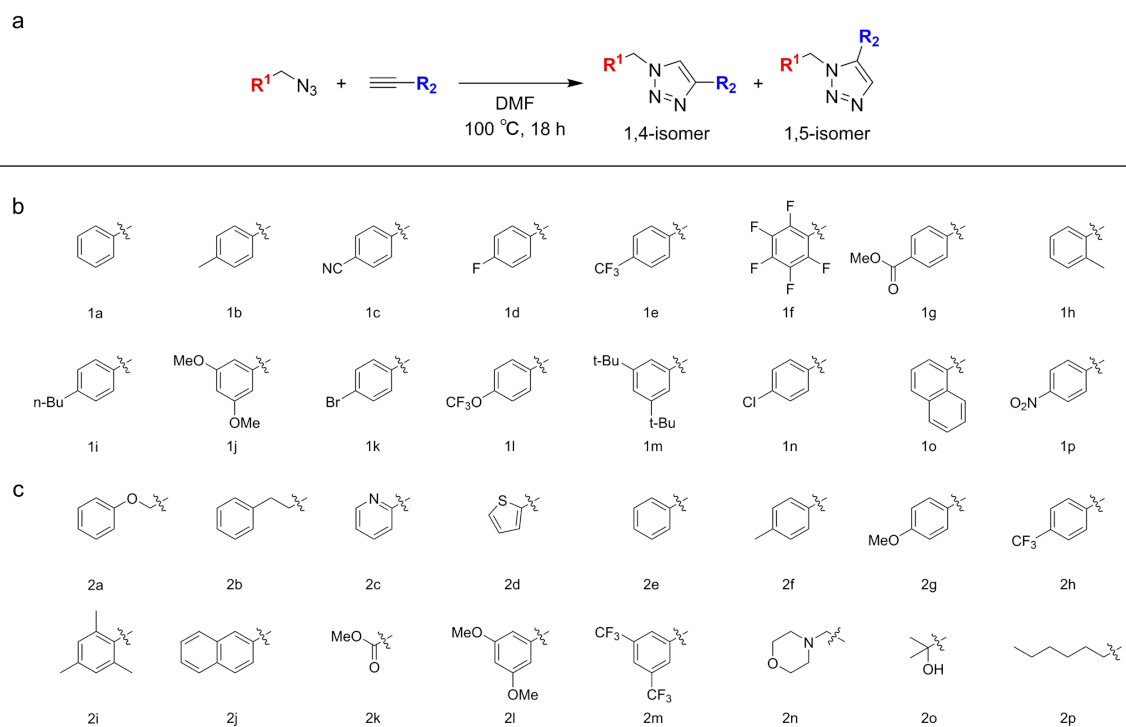


Figure 5.2. (a) Reaction condition for the thermal cycloaddition between an azide and terminal alkyne. (b) 16 candidate azides. (c) 16 candidate alkynes.

of 256 combinations of substrates within the search space. Each candidate substrate was chosen to have substituents with various electronic and steric properties, such as electron-donating, electron-withdrawing, bulky, and small. Here, I have also confirmed that the solvent does not affect the regioselectivity of the thermal azide-alkyne cycloaddition reaction by another experiment.

5.2.2 Substrate Optimization Scheme

Three substrate optimization algorithms were run independently to maximize the regioselectivity of the reaction. The experimental and computational costs of optimization for each scheme were compared.

1. **Bayesian Optimization Using Experimental Selectivity:** This algorithm uses Bayesian optimization and experimental selectivity data to search for optimal com-

binations of substrates. In Bayesian optimization, the GP regression is performed based on the available experimental data to evaluate the posterior mean and variance of the selectivity for each substrate combination. Then, the next batch of the experiments is proposed via an acquisition function (Figure 5.3a). This is the most common implementation of Bayesian substrate optimization, using only experimental results as a source of information. Thus, this scheme is a reference to other algorithms that utilize the computational results as a second source of information. Details of the implementation of Bayesian optimization are described in Chapter 5.2.3.

2. Bayesian Optimization Using Experimental and Theoretical Selectivity:

This algorithm also searches for the best substrate combination using Bayesian optimization. However, it differs from the first algorithm in that it uses not only the experimental selectivity data but also the theoretical selectivity data in the GP regression (Figure 5.3a). This algorithm aims to reduce the number of experiments required for Bayesian substrate optimization by utilizing computational results as a second source of information. In general, observational errors in experimental results are sufficiently smaller than computational errors in computational results. In such cases, experimental data can be considered accurate but expensive data, while theoretical data can be considered biased and noisy but cost-effective data, in terms of experimental cost. In Bayesian optimization, the approach of using cheap but cost-effective data as the second source of information to suppress the use of accurate but expensive data and consequently reduce the optimization cost is called multi-information source Bayesian optimization [64, 65]. In the field of chemistry, multi-information source Bayesian optimization was employed to reduce the required cost of solving optimization problems such as geometry optimization, binding en-

ergy maximization, and substrate optimization [66, 67]. Hence, multi-information source Bayesian optimization using the theoretical selectivity data as cheap but cost-effective data is expected to suppress the use of the experimental data and thus reduce the number of experiments required for substrate optimization.

3. **Greedy Algorithm Based on Theoretical Selectivity:** In this algorithm, Bayesian optimization is not used. Instead, the theoretical selectivity of all of the combinations of substrates is determined by quantum chemical calculations in advance of experiments. Combinations of substrates are ranked according to their theoretical selectivity, and experiments are carried out greedily, starting with the substrate combination with the highest theoretical selectivity (Figure 5.3b). This algorithm is a computation-driven reaction screening method in which experiments are conducted based on the theoretical results. Since the search space contains only 256 combinations in my case, it is not difficult to calculate the regioselectivity in all of the combinations of substrates. It is expected that this algorithm can discover the optimal substrate combination with a small number of experiments if the deviation between computational and experimental selectivity is small.

5.2.3 Implementation of Bayesian Optimization

In Bayesian optimization, a GP was used as a surrogate model. The Python package GPyTorch was used to implement the GP regression [68]. As an acquisition function, the upper confident bound (UCB) function was employed [69].

Electronic, steric, and structural features extracted from azide, alkyne, and triazole were used as candidate descriptor variables for the GP regression. Since the selection of descriptor variables is essential in machine learning, I briefly evaluated the importance of each candidate descriptor variable for each iteration of Bayesian optimization. Specifically,

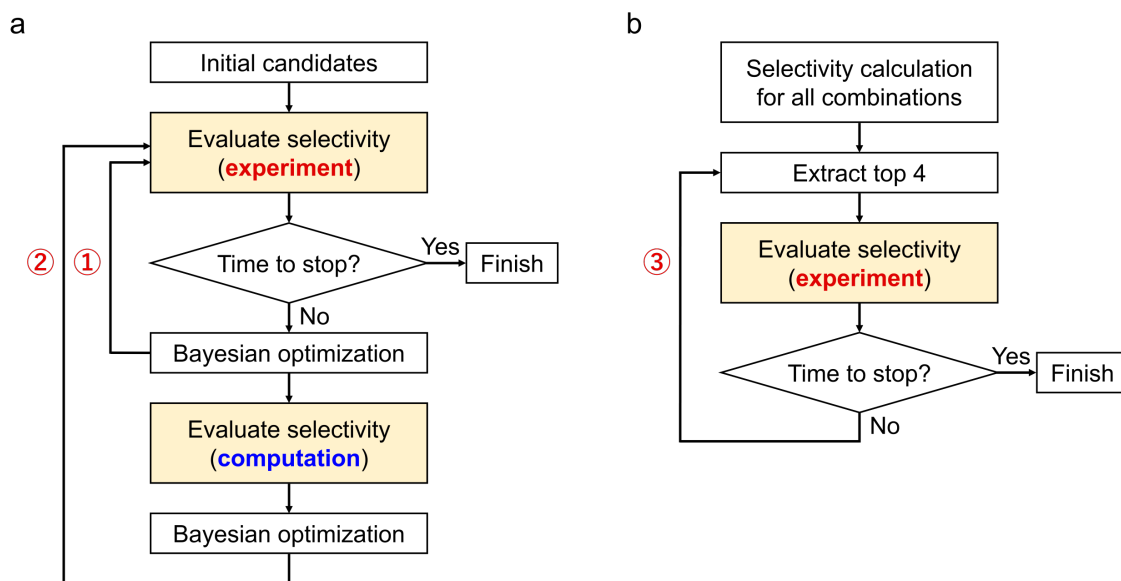


Figure 5.3. Flow of three substrate optimization algorithms. (a) Scheme 1 is a Bayesian substrate optimization using the experimental regioselectivity data, and Scheme 2 is a Bayesian substrate optimization that utilizes both the experimental and theoretical regioselectivity data. (b) Scheme 3 is a greedy algorithm based solely on the theoretical regioselectivity. The batch size of experiments and computations was always set to four.

the correlation coefficient and the maximum information coefficient [70] between each candidate descriptor and the objective variable (i.e., the selectivity of the reaction) were evaluated for each iteration, and only variables with either value greater than 0.5 were regarded as important and used as descriptor variables in the GP regression. Note that all selected descriptor variables are standardized to have a mean of 0 and a standard deviation of 1.

The kernel function and the gamma prior to its hyperparameters were grid searched and the combination with the highest cross validation score was selected as the best. The kernel function is a function that represents the covariance between input points $\mathbf{x}_i$ and

$\mathbf{x}_j$. The candidate kernel functions were chosen from the following functions:

$$k_{\text{linear}}(\mathbf{x}_i, \mathbf{x}_j) = v \mathbf{x}^\top \mathbf{x} \quad (5.1)$$

$$k_{\text{matérn1}}(\mathbf{x}_i, \mathbf{x}_j) = \theta_1 \exp(-r) \quad (5.2)$$

$$k_{\text{matérn1+linear}}(\mathbf{x}_i, \mathbf{x}_j) = \theta_1 \exp(-r) + v \mathbf{x}^\top \mathbf{x} \quad (5.3)$$

$$k_{\text{matérn3}}(\mathbf{x}_i, \mathbf{x}_j) = \theta_1 (1 + \sqrt{3}r) \exp(-\sqrt{3}r) \quad (5.4)$$

$$k_{\text{matérn3+linear}}(\mathbf{x}_i, \mathbf{x}_j) = \theta_1 (1 + \sqrt{3}r) \exp(-\sqrt{3}r) + v \mathbf{x}^\top \mathbf{x} \quad (5.5)$$

$$k_{\text{matérn5}}(\mathbf{x}_i, \mathbf{x}_j) = \theta_1 \left(1 + \sqrt{5}r + \frac{5}{3}r^2 \right) \exp(-\sqrt{5}r) \quad (5.6)$$

$$k_{\text{matérn5+linear}}(\mathbf{x}_i, \mathbf{x}_j) = \theta_1 \left(1 + \sqrt{5}r + \frac{5}{3}r^2 \right) \exp(-\sqrt{5}r) + v \mathbf{x}^\top \mathbf{x} \quad (5.7)$$

$$k_{\text{RBF}}(\mathbf{x}_i, \mathbf{x}_j) = \theta_1 \exp\left(-\frac{1}{2}r^2\right) \quad (5.8)$$

$$k_{\text{RBF+linear}}(\mathbf{x}_i, \mathbf{x}_j) = \theta_1 \exp\left(-\frac{1}{2}r^2\right) + v \mathbf{x}^\top \mathbf{x} \quad (5.9)$$

where v is the variance parameter, θ_1 is the output scale parameter, and r^2 is the square of Euclidean distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ scaled by the length scale parameter Θ . If Θ is a constant diagonal matrix, one length scale parameter is applied to all dimensions of input. If Θ is a non-constant diagonal matrix, each dimension of input has its own length scale parameter. This type of kernel function is called an automatic relevance determination (ARD) kernel. In this study, I tried both kernels with and without ARD for all kernels except the linear kernel. Thus, the total number of kernel functions considered was 17. In cross validation, available data were split into training data and test data, and GP regression model was trained via the type-II maximum a posteriori estimation using the training data. Adam was used as the optimizer. Finally, the combination of the kernel function and the gamma prior that maximized the coefficient of determination was used for optimization.

The batch size of the experiments and computations was set to four, and Bayesian

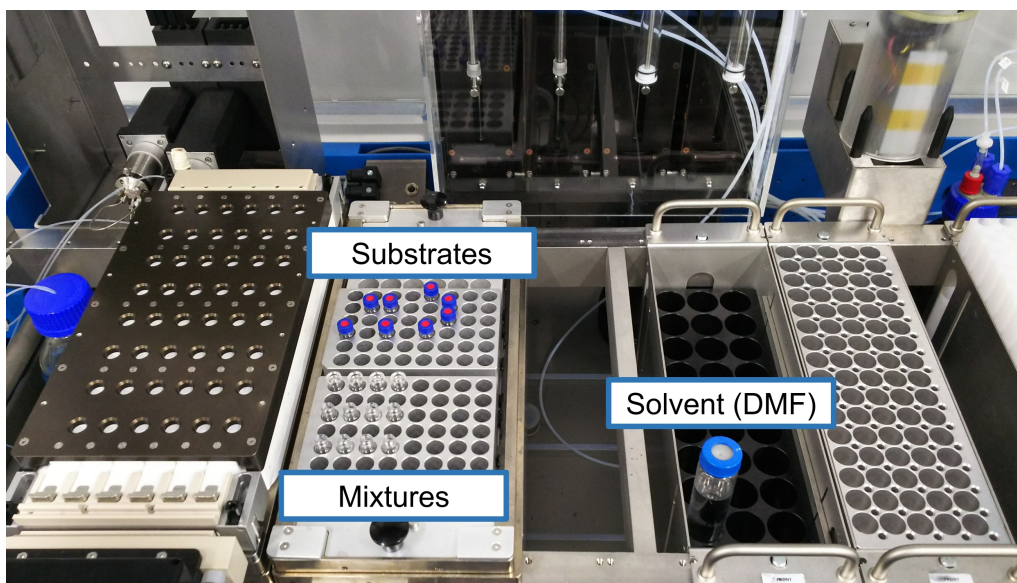


Figure 5.4. Synthesis of the 1,2,3-triazoles using the automated synthesis robot. Azide and terminal alkyne substrates were dissolved in *N,N*-dimethylformamide and transferred to a vial. The reaction mixtures were stirred at 100 °C for 18 h.

optimization was parallelized using the Kriging believer algorithm so that four combinations of substrates could be proposed simultaneously [31].

5.2.4 Experimental Details

An automated synthesis robot (Chemspeed ISYNTH robotic synthesis platform) was employed to carry out the thermal azide-alkyne cycloaddition reaction. Most operations such as mixing substrates, adding solvents, and heating, stirring, and cooling reaction mixtures were automated and parallelized using the robot (Figure 5.4). This is suitable for batch experiments in which multiple experiments are conducted simultaneously.

The thermal azide-alkyne cycloaddition reaction generally produces a mixture of 1,4- and 1,5-disubstituted isomers. To determine the signals of 1,4-substituted isomers on a gas chromatography (GC) and NMR, the copper(I)-catalyzed azide-alkyne cycloaddition reactions were carried out in parallel with the thermal azide-alkyne cycloaddition for the

combinations of substrates proposed by each algorithm.

In most cases, the selectivity of the reaction was evaluated from the area ratio of GC signals of the two isomers since the S/N ratio of GC spectrum is smaller than that of NMR spectrum. In some cases, however, GC signals of two isomers were too close to each other, making it difficult to evaluate the selectivity. In such cases, selectivity was evaluated from NMR spectra instead of GC spectra. Fortunately, there were no cases where the signals of the two isomers were inseparable, either by GC or NMR.

5.2.5 Computational Details

In the thermal azide-alkyne cycloaddition, the energy difference between the 1,4- and 1,5-TSs in the two reaction pathways producing the 1,4- and 1,5-disubstituted isomers determines the theoretical regioselectivity. To evaluate the theoretical selectivity for each substrate combination, the non-substituted TS structure of the thermal azide-alkyne cycloaddition of hydrogen azide with acetylene was determined in advance. Previous theoretical studies have shown that the reaction proceeds through a concerted process [71,72]. Therefore, I focused on the pathway where an azide and an alkyne directly form a triazole in a single step. In other words, I searched for TSs where two covalent bonds are generated simultaneously.

The substituents of azide and alkyne substrates proposed by Bayesian optimization were introduced into the azide and alkyne moieties of the unsubstituted TS to generate model structures for 1,4- and 1,5-disubstituted TSs. Each substituent moiety was rotated around the newly generated covalent bonds to reduce steric hindrances and was subsequently optimized at the semiempirical GFN2-xTB [73] level interfaced with Orca 4.2 [74]. The conformation of the substituents in each model TS was systematically explored using the single-component (-SC) AFIR method implemented in an in-house version

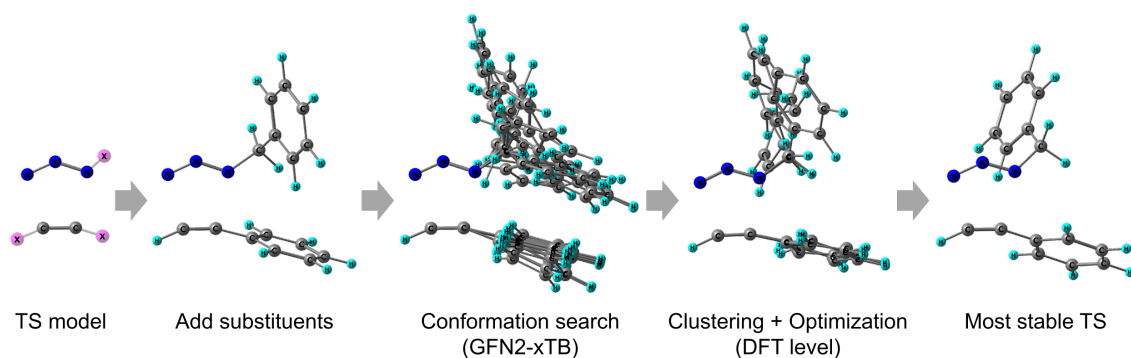


Figure 5.5. Concise scheme to obtain the most stable 1,5-disubstituted TS structure. The same procedure, with different substituent addition positions, also provides the most stable 1,4-disubstituted TS structure.

of the GRRM program [75,76]. The model collision parameter GAMMA was set to 100. SC-AFIR calculations were also done at the GFN2-xTB level.

Among the obtained 1,5-disubstituted approximate TS structures, structures with electron energy ranging from the most stable to +30 kJ/mol were roughly clustered based on the distance matrix. The most stable structure in each cluster was then extracted and optimized to the exact TSs using a quasi-Newton method. The intrinsic reaction coordinate (IRC) calculations [77,78] were performed from all of the obtained TSs to check whether they connected the appropriate reactants and products. These calculations were conducted using the Gaussian 16 software package [79] with the B3LYP-D3/def2TZVPP//B3LYP-D3/def2SVP [80] level of theory. The solute electron density model (SMD) method [81] was employed to incorporate the solvent effect of *N,N*-dimethylformamide. The same procedure was applied to the 1,4-disubstituted approximate TS structures. The flow of searching for the most stable TS structure is summarized in Figure 5.5. Finally, the Boltzmann distribution was calculated using the difference in Gibbs free energies of the most stable 1,4- and 1,5-disubstituted TSs to evaluate the theoretical selectivity for the thermal azide-alkyne cycloaddition.

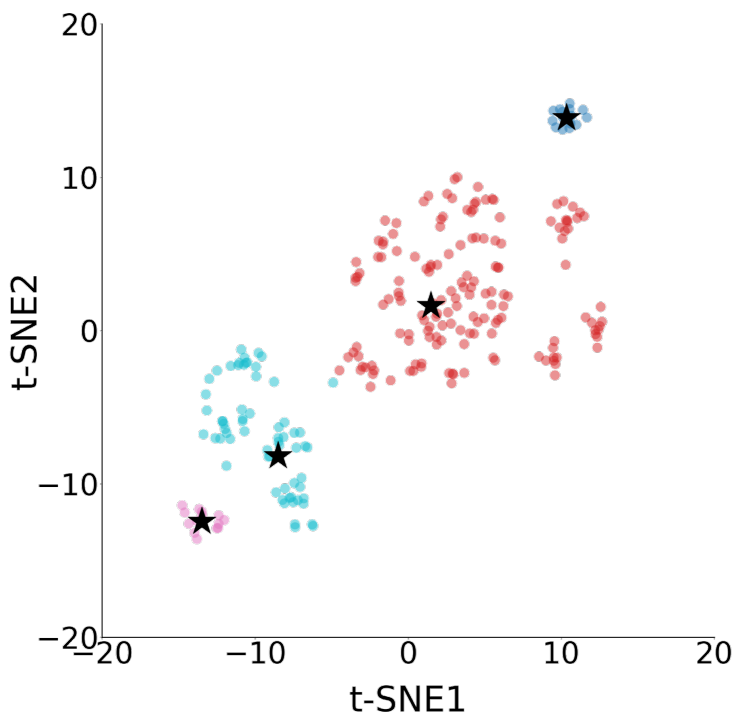


Figure 5.6. The search space clustered by k -means. 43-Dimensional descriptor variables were compressed into a 2-dimensional space with t-distributed stochastic neighbor embedding (t-SNE). The data with the smallest Euclidean distance from the centroid of each cluster is marked with a black star.

5.3 Results and Discussion

5.3.1 Optimization Results

The search space was partitioned into four clusters using k -means clustering (Figure 5.6). The initial batch for Schemes 1 and 2 was determined by selecting the substrate combination with the smallest Euclidean distance from the centroid of each cluster. The initial batch for Scheme 3 was the four combinations of substrates with the highest theoretical 1,5-regioselectivity out of 256 combinations of substrates. The combinations of substrates that contained **2k** and **2o** formed small clusters, each containing 16 data points (pink and dark blue clusters), while the remaining combinations were roughly partitioned into two

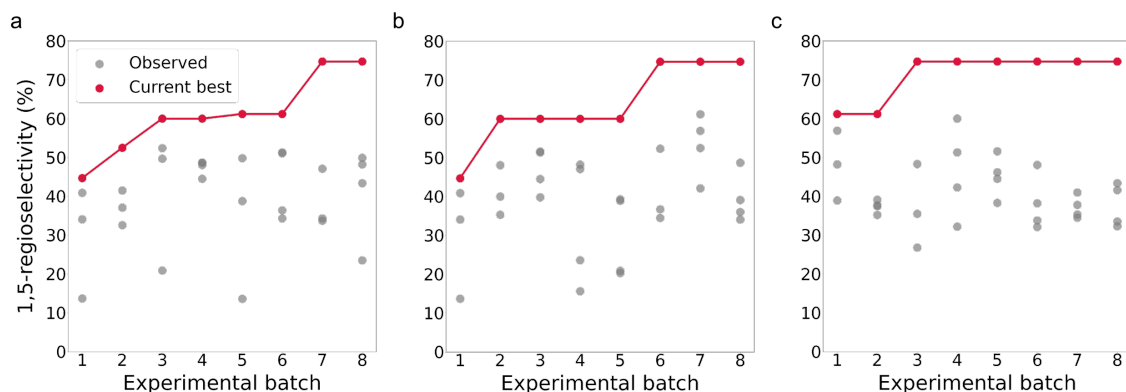


Figure 5.7. History of substrate combination optimization in (a) Scheme 1, (b) Scheme 2, and (c) Scheme 3. The horizontal axis represents the number of experimental batches executed, while the vertical axis denotes the experimental regioselectivity of the 1,5-disubstituted isomer. Gray circles depict the four experimental results obtained in each batch, and red circles represent the maximum regioselectivity of the 1,5-disubstituted isomer achieved up to the previous iteration.

large clusters (red and light blue clusters). The formation of distinct small clusters by **2k** and **2o** may be attributed to their notably large structural specificity.

Eight batches of experiments were conducted in each optimization scheme. Since the experimental batch size was set to four, the total number of experiments performed was calculated as $3 \text{ (optimization schemes)} \times 4 \text{ (batch size)} \times 8 \text{ (executed batches)} = 96$. Some combinations of substrates were overlapped among the optimization schemes. The number of combinations for which the regioselectivity was investigated was 71. Experimental regioselectivity observed in each optimization scheme was shown in Figure 5.7 as a function of the executed experimental batch. The maximum experimental regioselectivity of 1,5-disubstituted isomer observed was 74.7%. All of the three optimization schemes succeed in finding the corresponding combination of substrates.

The distribution of experimental 1,5-regioselectivity is shown in Figure 5.8. Only 10

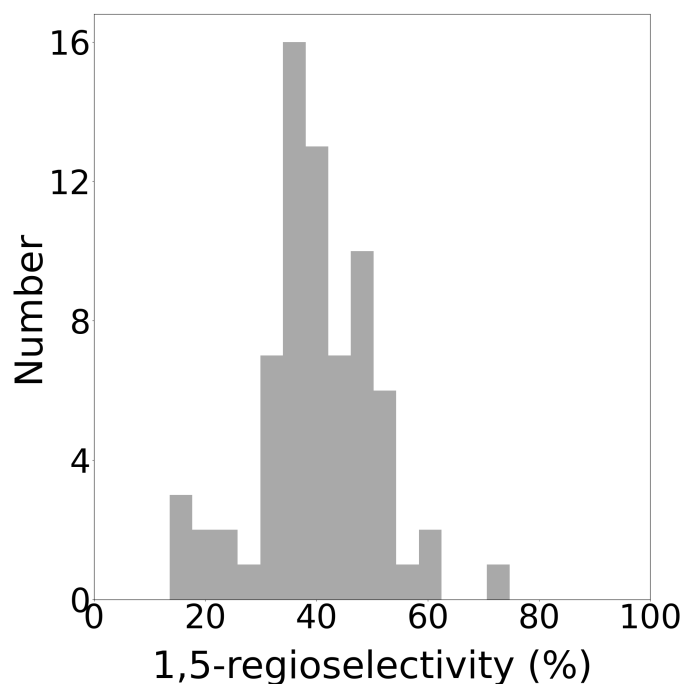


Figure 5.8. Distribution of experimental 1,5-disubstituted isomer selectivity. The mean and the standard deviation of the selectivity were 39.9% and 10.7%, respectively.

combinations of substrates showed more than 50.0% 1,5-regioselectivity, indicating that the thermal azide-alkyne cycloaddition predominantly favors 1,4-regioselectivity.

5.3.2 Substituent Effects Analysis

The top five combinations of substrates with the highest 1,5-regioselectivity are shown in Table 5.1. The electron-rich, slightly bulky **2i** specifically enhanced 1,5-regioselectivity. A theoretical study by Houk and co-workers revealed that 1,3-cycloaddition of an azide with an electron-rich alkyne proceeds via the interaction of the LUMO of the azide with the HOMO of the alkyne [82]. The LUMO of the azide has a slightly larger coefficient on the unsubstituted nitrogen atom, while the HOMO of the electron-rich alkyne has a larger coefficient on the unsubstituted carbon atom. Consequently, the interaction between the LUMO of the azide and the HOMO of the electron-rich alkyne favors the preferential production of the 1,5-disubstituted isomer. To validate this preference, I calculated the

Table 5.1. Top five combinations of substrates with the highest regioselectivity for 1,5-disubstituted isomers. $\Delta E_1 = (E_{\text{azide-LUMO}} - E_{\text{alkyne-HOMO}})$ and $\Delta E_2 = (E_{\text{alkyne-LUMO}} - E_{\text{azide-HOMO}})$. The unit is eV.

Rank	Azide	Alkyne	ΔE_1	ΔE_2	$\Delta E_2 - \Delta E_1$	1,5-regioselectivity (%)
1	1f	2i	5.09	6.15	1.06	74.7
2	1c	2i	5.38	5.90	0.52	61.2
3	1p	2i	5.34	6.13	0.79	60.0
4	1k	2i	5.35	6.14	0.78	56.9
5	1f	2c	5.78	5.70	-0.08	52.5

energy difference between the azide LUMO and alkyne HOMO, as well as between the alkyne LUMO and azide HOMO, for the top five combinations of substrates. The results confirmed my expectations, as the energy difference between the azide LUMO and alkyne HOMO was consistently smaller than the energy difference between the alkyne LUMO and azide HOMO in the ranking 1-4 combinations. It’s worth noting that the energy difference between the alkyne LUMO and azide HOMO was smaller than that between the azide LUMO and alkyne HOMO for the rank five combination. However, this result is not surprising, considering that the regioselectivity is influenced by not only frontier orbital interactions but also various factors, including charge and steric hindrance, and that the 1,5-regioselectivity of the rank five combination is low at 52.5%.

The top five combinations of substrates with the lowest 1,5-regioselectivity (i.e., the highest 1,4-regioselectivity) are summarized in Table 5.2. Although Bayesian optimization is configured to search for combinations of substrates that maximize 1,5-regioselectivity, it is not surprising that, in the process of balancing exploration and exploitation, combinations with opposite selectivities are proposed and discovered. As shown in Table 5.2, **2k** with an acetyl group notably enhances the 1,4-regioselectivity. This substituent effect is likely due to the distribution of partial charges between the azide and the alkyne. The

Table 5.2. Top five combinations of substrates with the highest regioselectivity for 1,4-disubstituted isomers. $\Delta E_1 = (E_{\text{azide-LUMO}} - E_{\text{alkyne-HOMO}})$ and $\Delta E_2 = (E_{\text{alkyne-LUMO}} - E_{\text{azide-HOMO}})$. The unit is eV.

Rank	Azide	Alkyne	ΔE_1	ΔE_2	$\Delta E_2 - \Delta E_1$	1,4-regioselectivity (%)
1	1p	2k	6.97	5.77	-1.20	86.4
2	1k	2k	6.98	5.77	-1.21	86.3
3	1f	2k	6.72	5.79	-0.93	84.4
4	1c	2d	5.48	5.58	0.10	79.7
5	1b	2d	5.51	5.80	0.28	79.1

nitrogen atoms at both ends of the azide group are negatively charged, but the nitrogen adjacent to the substituent exhibits a larger negative partial charge. On the other hand, **2k** has a relatively positive charge on the unsubstituted carbon due to the resonance. Consequently, strong electrostatic interactions between the substituted nitrogen and the unsubstituted carbon are believed to preferentially lead to the formation of the 1,4-disubstituted isomer (Figure 5.9).

Frontier molecular orbital energy analysis also explains why **2k** improves the selectivity of the 1,4-substituted isomers. According to the theoretical work of Houk and co-workers, the reaction between an azide and an electron-deficient alkyne proceeds via an interaction between the HOMO of the azide and the LUMO of the alkyne [82]. The HOMO of the azide has a large coefficient on the substituted nitrogen atom, while the LUMO of the electron-deficient alkyne has a large coefficient on the unsubstituted carbon atom. Thus, the interaction of these two orbitals preferentially produces 1,4-disubstituted isomers. Indeed, as shown in Table 5.2, the energy difference between the alkyne-LUMO and azide-HOMO is smaller than that between the azide-LUMO and alkyne-HOMO for the ranking 1-3 combinations using **2k**.

I then compared the theoretical and experimental selectivities for 71 experimentally

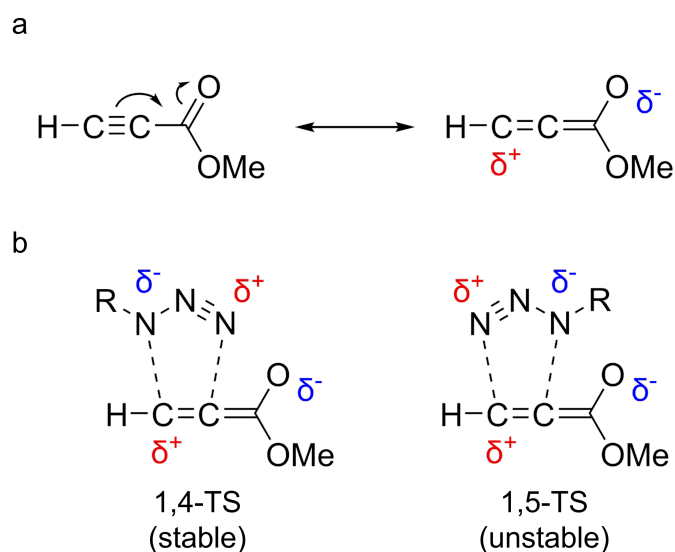


Figure 5.9. (a) Resonance structures and partial charge distribution of the alkyne with the acetyl group. (b) Partial charge distribution of the 1,4-TS and 1,5-TS with the alkyne with the acetyl group. The 1,4-TS is stabilized by strong electrostatic interactions, resulting in preferential production of the 1,4-isomer.

validated combinations of substrates. Despite a systematic search for the most stable TS structures, a substantial deviation between theoretical and experimental selectivity is observed (Figure 5.10). While the trend of qualitative experimental selectivity is predicted, quantitative experimental selectivity prediction is not successful. I also computed selectivity with other functions (ω B97X-D [83] and M06L [84], M06 [85], M062X [85]) and basis sets (6-31G++(2df,2pd) [86] and aug-cc-pVTZ [87]), however, the deviation did not improve essentially. This significant deviation may adversely affect the prediction accuracy of GP regression and the optimization efficiency of Bayesian optimization based on experimental and theoretical selectivity.

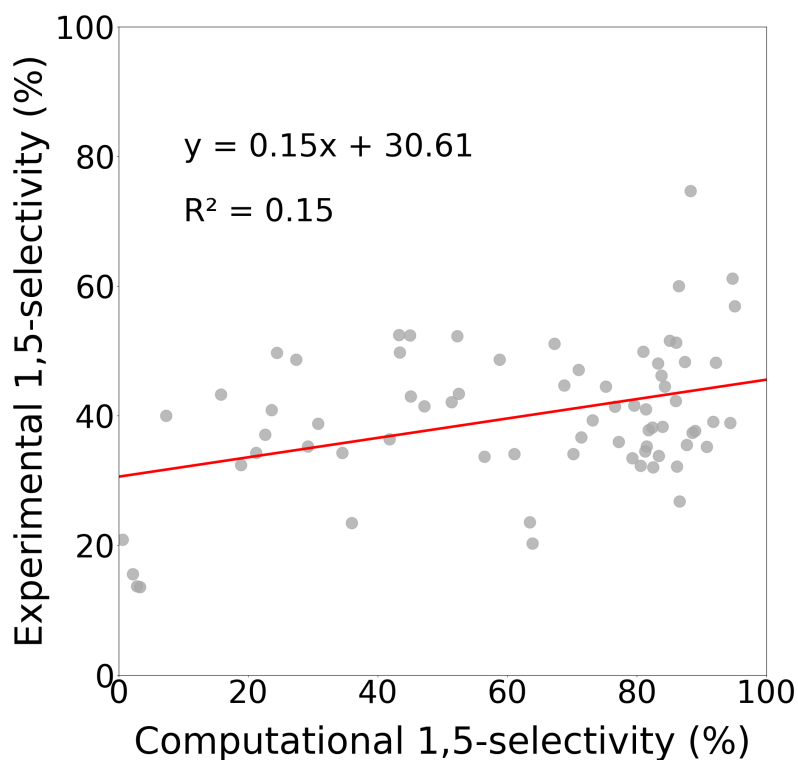


Figure 5.10. Comparison of experimental and theoretical 1,5-regioselectivity of the thermal azide-alkyne cycloaddition.

5.3.3 Performance Comparison

In this study, one batch of experiments or selectivity calculations took approximately 1-2 days. Hence, the time cost required for the experiments and the calculations was almost the same. Although it is difficult to directly compare the human and financial costs required for experiments and computations, considering the risks and hazards associated with the synthesis of explosive azides, and the human costs involved in performing experiments that are difficult to automate, reducing the number of experiments by investing a certain amount of computational cost is meaningful. Despite the significant computational error (Figure 5.10), the greedy method (Scheme 3) efficiently found the optimal combination with the fewest experiments (Table 5.3). This result suggests that, in the case where a score of an objective variable for all possible combinations can be calculated at a

Table 5.3. Lists of the required number of experiments and selectivity calculations in each substrate optimization scheme.

Scheme	The number of experiments	The number of calculations
1	28	0
2	24	20
3	12	256

low cost, computation-driven reaction screening can significantly reduce the experimental costs required for substrate optimization. Table 5.3 also shows that Bayesian optimization using both experimental and theoretical selectivity data (Scheme 2) optimized combination of substrates at relatively low computational cost and with fewer experiments than conventional Bayesian optimization using only experimental selectivity data (Scheme 1). This result shows that even the computational selectivity, which is affected by computational error, helps to reduce the required number of experiments for Bayesian substrate optimization.

Finally, I discuss the scope of application of the obtained results. First, it should be noted that the use of computational results does not always make reaction screening more efficient when the computational cost is very large. For example, to predict the theoretical reactivity or selectivity of a multi-step reaction, it is necessary to evaluate the Gibbs free energies of all of the TSs which may be a rate-determining step. In addition, chemical reactions involving additives such as catalysts and bases increase the number of possible conformations and require a more systematic search to determine the most stable conformation. In such cases, the cost of a single calculation can significantly exceed the cost of a single experiment, reducing screening efficiency. It has also been reported that if the computational error is too large, the utilization of computational results as a second source of information decreases the performance of Bayesian optimization [67]. There-

fore, it is necessary to correctly estimate the computational cost and the computational error when performing reaction screening using theoretical data. In addition, Bayesian optimization results exhibit randomness based on the initial sample. Therefore, starting Bayesian optimization from different initial samples can lead to different outcomes.

5.4 Conclusion

In this study, I have optimized the combinations of azide and terminal alkyne substrates to maximize the regioselectivity of the thermal azide-alkyne cycloaddition. From a search space of 256 combinations of substrates (16 azides and 16 alkynes), three optimization algorithms (Bayesian optimization using experimental selectivity, Bayesian optimization utilizing experimental and theoretical selectivity, and a greedy method relying solely on theoretical selectivity) using different data were utilized to search for the optimal substrate combination. The regioselectivities of the combinations of substrates proposed by the optimization algorithms were evaluated experimentally using an automated synthesis robot with quantum chemical calculations. The maximum 1,5-regioselectivity achieved was 74.7%, while the maximum 1,4-regioselectivity was 86.4%. The chemical interpretation for these substrates enhancing the regioselectivity of the thermal azide-alkyne cycloaddition could be explained by frontier orbitals and partial charges.

Among the three algorithms, the greedy algorithm has succeeded in finding the best substrate combination with the fewest number of experiments. This result supports the superiority of computation-driven reaction screening in substrate optimization. Furthermore, Bayesian optimization utilizing both experimental and theoretical selectivity data was able to find optimal combinations of substrates with fewer experiments compared to conventional Bayesian optimization using only experimental selectivity data, despite a significant computational error. This result suggests that the theoretical selectivity data

can be useful in reducing the experimental cost of Bayesian substrate optimization. Further studies on the application of the method developed in this study to maximize the regioselectivity in various chemical transformation reactions are under investigation in my laboratory, together with studies on the enhancement of the optimization algorithms.

References

- [1] *Proceedings of the Chemical Society* **1961**, 357–396.
- [2] R. Huisgen, *Angewandte Chemie International Edition in English* **1963**, *2*, 565–598.
- [3] R. Huisgen, *Angewandte Chemie International Edition in English* **1963**, *2*, 633–645.
- [4] V. V. Rostovtsev, L. G. Green, V. V. Fokin, K. B. Sharpless, *Angewandte Chemie International Edition* **2002**, *41*, 2596–2599.
- [5] C. W. Tornøe, C. Christensen, M. Meldal, *The Journal of Organic Chemistry* **2002**, *67*, 3057–3064.
- [6] F. Amblard, J. H. Cho, R. F. Schinazi, *Chemical Reviews* **2009**, *109*, 4207–4220.
- [7] J. E. Hein, V. V. Fokin, *Chemical Society Reviews* **2010**, *39*, 1302–1315.
- [8] H. C. Kolb, M. G. Finn, K. B. Sharpless, *Angewandte Chemie International Edition* **2001**, *40*, 2004–2021.
- [9] L. Zhang, X. Chen, P. Xue, H. H. Y. Sun, I. D. Williams, K. B. Sharpless, V. V. Fokin, G. Jia, *Journal of the American Chemical Society* **2005**, *127*, 15998–15999.
- [10] L. K. Rasmussen, B. C. Boren, V. V. Fokin, *Organic Letters* **2007**, *9*, 5337–5339.
- [11] B. C. Boren, S. Narayan, L. K. Rasmussen, L. Zhang, H. Zhao, Z. Lin, G. Jia, V. V. Fokin, *Journal of the American Chemical Society* **2008**, *130*, 8923–8930.

- [12] J. R. Johansson, T. Beke-Somfai, A. Said Stålsmeden, N. Kann, *Chemical Reviews* **2016**, *116*, 14726–14768.
- [13] Z.-X. Wang, H.-L. Qin, *Chemical Communications* **2003**, 2450–2451.
- [14] A. Krasieński, V. V. Fokin, K. B. Sharpless, *Organic Letters* **2004**, *6*, 1237–1240.
- [15] F. Kloss, U. Köhn, B. O. Jahn, M. D. Hager, H. Görls, U. S. Schubert, *Chemistry – An Asian Journal* **2011**, *6*, 2816–2824.
- [16] G. Cheng, X. Zeng, J. Shen, X. Wang, X. Cui, *Angewandte Chemie International Edition* **2013**, *52*, 13265–13268.
- [17] A. Kayet, T. Pathak, *The Journal of Organic Chemistry* **2013**, *78*, 9865–9875.
- [18] K. Wu, A. G. Doyle, *Nature Chemistry* **2017**, *9*, 779–784.
- [19] J. M. Granda, L. Donina, V. Dragone, D.-L. Long, L. Cronin, *Nature* **2018**, *559*, 377–381.
- [20] A. M. Żurański, J. I. Martinez Alvarado, B. J. Shields, A. G. Doyle, *Accounts of Chemical Research* **2021**, *54*, 1856–1865.
- [21] A. Milo, E. N. Bess, M. S. Sigman, *Nature* **2014**, *507*, 210–214.
- [22] Z. L. Niemeyer, A. Milo, D. P. Hickey, M. S. Sigman, *Nature Chemistry* **2016**, *8*, 610–617.
- [23] S. Yamaguchi, T. Nishimura, Y. Hibe, M. Nagai, H. Sato, I. Johnston, *Journal of Computational Chemistry* **2017**, *38*, 1825–1833.
- [24] A. F. Zahrt, J. J. Henle, B. T. Rose, Y. Wang, W. T. Darrow, S. E. Denmark, *Science* **2019**, *363*, eaau5631.

- [25] J. P. Reid, M. S. Sigman, *Nature* **2019**, *571*, 343–348.
- [26] J. N. Wei, D. Duvenaud, A. Aspuru-Guzik, *ACS Central Science* **2016**, *2*, 725–732.
- [27] C. W. Coley, R. Barzilay, T. S. Jaakkola, W. H. Green, K. F. Jensen, *ACS Central Science* **2017**, *3*, 434–443.
- [28] M. H. S. Segler, M. Preuss, M. P. Waller, *Nature* **2018**, *555*, 604–610.
- [29] H. Gao, T. J. Struble, C. W. Coley, Y. Wang, W. H. Green, K. F. Jensen, *ACS Central Science* **2018**, *4*, 1465–1476.
- [30] P. I. Frazier, *A Tutorial on Bayesian Optimization*, **2018**.
- [31] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, N. de Freitas, *Proceedings of the IEEE* **2016**, *104*, 148–175.
- [32] A. M. Schweidtmann, A. D. Clayton, N. Holmes, E. Bradford, R. A. Bourne, A. A. Lapkin, *Chemical Engineering Journal* **2018**, *352*, 277–282.
- [33] B. Burger, P. M. Maffettone, V. V. Gusev, C. M. Aitchison, Y. Bai, X. Wang, X. Li, B. M. Alston, B. Li, R. Clowes, N. Rankin, B. Harris, R. S. Sprick, A. I. Cooper, *Nature* **2020**, *583*, 237–241.
- [34] P. M. Attia, A. Grover, N. Jin, K. A. Severson, T. M. Markov, Y.-H. Liao, M. H. Chen, B. Cheong, N. Perkins, Z. Yang, P. K. Herring, M. Aykol, S. J. Harris, R. D. Braatz, S. Ermon, W. C. Chueh, *Nature* **2020**, *578*, 397–402.
- [35] S. Langner, F. Häse, J. D. Perea, T. Stubhan, J. Hauch, L. M. Roch, T. Heumueller, A. Aspuru-Guzik, C. J. Brabec, *Advanced Materials* **2020**, *32*, 1907801.
- [36] T. Ueno, T. D. Rhone, Z. Hou, T. Mizoguchi, K. Tsuda, *Materials Discovery* **2016**, *4*, 18–21.

- [37] F. Häse, L. M. Roch, C. Kreisbeck, A. Aspuru-Guzik, *ACS Central Science* **2018**, *4*, 1134–1145.
- [38] B. J. Shields, J. Stevens, J. Li, M. Parasram, F. Damani, J. I. M. Alvarado, J. M. Janey, R. P. Adams, A. G. Doyle, *Nature* **2021**, *590*, 89–96.
- [39] M. Christensen, L. P. E. Yunker, F. Adedeji, F. Häse, L. M. Roch, T. Gensch, G. dos Passos Gomes, T. Zepel, M. S. Sigman, A. Aspuru-Guzik, J. E. Hein, *Communications Chemistry* **2021**, *4*, 1–12.
- [40] N. Sugisawa, H. Sugisawa, Y. Otake, R. V. Krems, H. Nakamura, S. Fuse, *Chemistry–Methods* **2021**, *1*, 484–490.
- [41] N. V. Faurschou, R. H. Taaning, C. M. Pedersen, *Chemical Science* **2023**, *14*, 6319–6329.
- [42] Nobuaki. Koga, Keiji. Morokuma, *Chemical Reviews* **1991**, *91*, 823–842.
- [43] K. N. Houk, P. H.-Y. Cheong, *Nature* **2008**, *455*, 309–313.
- [44] W. Thiel, *Angewandte Chemie International Edition* **2014**, *53*, 8605–8613.
- [45] W. M. C. Sameera, S. Maeda, K. Morokuma, *Accounts of Chemical Research* **2016**, *49*, 763–773.
- [46] K. N. Houk, F. Liu, *Accounts of Chemical Research* **2017**, *50*, 539–543.
- [47] S. Maeda, K. Ohno, *Journal of the American Chemical Society* **2008**, *130*, 17228–17229.
- [48] P. J. Donoghue, P. Helquist, P.-O. Norrby, O. Wiest, *Journal of the American Chemical Society* **2009**, *131*, 410–411.

- [49] M. Hatanaka, S. Maeda, K. Morokuma, *Journal of Chemical Theory and Computation* **2013**, *9*, 2882–2886.
- [50] E. Hansen, A. R. Rosales, B. Tutkowski, P.-O. Norrby, O. Wiest, *Accounts of Chemical Research* **2016**, *49*, 996–1005.
- [51] Y. Guan, V. M. Ingman, B. J. Rooks, S. E. Wheeler, *Journal of Chemical Theory and Computation* **2018**, *14*, 5249–5261.
- [52] W. Matsuoka, Y. Harabuchi, S. Maeda, *ACS Catalysis* **2022**, *12*, 3752–3766.
- [53] P. Friederich, G. d. P. Gomes, R. D. Bin, A. Aspuru-Guzik, D. Balcells, *Chemical Science* **2020**, *11*, 4584–4601.
- [54] S. Ahn, M. Hong, M. Sundararajan, D. H. Ess, M.-H. Baik, *Chemical Reviews* **2019**, *119*, 6509–6560.
- [55] A. R. Rosales, J. Wahlers, E. Limé, R. E. Meadows, K. W. Leslie, R. Savin, F. Bell, E. Hansen, P. Helquist, R. H. Munday, O. Wiest, P.-O. Norrby, *Nature Catalysis* **2019**, *2*, 41–45.
- [56] M. Foscato, V. R. Jensen, *ACS Catalysis* **2020**, *10*, 2354–2377.
- [57] T. Mita, Y. Harabuchi, S. Maeda, *Chemical Science* **2020**, *11*, 7569–7577.
- [58] M. Mukai, K. Nagao, S. Yamaguchi, H. Ohmiya, *Bulletin of the Chemical Society of Japan* **2022**, *95*, 271–277.
- [59] S. Yamaguchi, *Organic & Biomolecular Chemistry* **2022**, *20*, 6057–6071.
- [60] S. Gallarati, R. Fabregat, R. Laplaza, S. Bhattacharjee, M. D. Wodrich, C. Corminboeuf, *Chemical Science* **2021**, *12*, 6879–6889.

- [61] S. M. Maley, D.-H. Kwon, N. Rollins, J. C. Stanley, O. L. Sydora, S. M. Bischof, D. H. Ess, *Chemical Science* **2020**, *11*, 9665–9674.
- [62] Y. Sumiya, Y. Harabuchi, Y. Nagata, S. Maeda, *JACS Au* **2022**, *2*, 1181–1188.
- [63] H. Hayashi, H. Katsuyama, H. Takano, Y. Harabuchi, S. Maeda, T. Mita, *Nature Synthesis* **2022**, *1*, 804–814.
- [64] K. Swersky, J. Snoek, R. P. Adams in *Advances in Neural Information Processing Systems*, Curran Associates, Inc.
- [65] M. Poloczek, J. Wang, P. Frazier in *Advances in Neural Information Processing Systems*, Curran Associates, Inc.
- [66] H. C. Herbol, M. Poloczek, P. Clancy, *Materials Horizons* **2020**, *7*, 2113–2123.
- [67] H. Okada, S. Maeda, *ACS Omega* **2024**, *9*, 7123–7131.
- [68] J. Gardner, G. Pleiss, K. Q. Weinberger, D. Bindel, A. G. Wilson in *Advances in Neural Information Processing Systems*, Curran Associates, Inc.
- [69] N. Srinivas, A. Krause, S. M. Kakade, M. Seeger, *IEEE Transactions on Information Theory* **2012**, *58*, 3250–3265.
- [70] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, P. C. Sabeti, *Science* **2011**, *334*, 1518–1524.
- [71] J. J. W. McDouall, M. A. Robb, U. Niazi, F. Bernardi, H. B. Schlegel, *Journal of the American Chemical Society* **1987**, *109*, 4642–4648.
- [72] M. T. Nguyen, A. K. Chandra, S. Sakai, K. Morokuma, *The Journal of Organic Chemistry* **1999**, *64*, 65–69.

- [73] *GFN2- x TB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions — Journal of Chemical Theory and Computation*, <https://pubs.acs.org/doi/10.1021/acs.jctc.8b01176>.
- [74] F. Neese, F. Wennmohs, U. Becker, C. Riplinger, *The Journal of Chemical Physics* **2020**, *152*, 224108.
- [75] S. Maeda, K. Morokuma, *The Journal of Chemical Physics* **2010**, *132*, 241102.
- [76] S. Maeda, Y. Harabuchi, *WIREs Computational Molecular Science* **2021**, *11*, e1538.
- [77] K. Fukui, *Accounts of Chemical Research* **1981**, *14*, 363–368.
- [78] S. Maeda, Y. Harabuchi, Y. Ono, T. Taketsugu, K. Morokuma, *International Journal of Quantum Chemistry* **2015**, *115*, 258–269.
- [79] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W.

- Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, D. J. Fox, *Gaussian 16 Revision C.01*, **2016**, Gaussian Inc. Wallingford CT.
- [80] F. Weigend, R. Ahlrichs, *Physical Chemistry Chemical Physics* **2005**, *7*, 3297–3305.
- [81] A. V. Marenich, C. J. Cramer, D. G. Truhlar, *The Journal of Physical Chemistry B* **2009**, *113*, 6378–6396.
- [82] K. N. Houk, Joyner. Sims, C. R. Watts, L. J. Luskus, *Journal of the American Chemical Society* **1973**, *95*, 7301–7315.
- [83] J.-D. Chai, M. Head-Gordon, *Physical Chemistry Chemical Physics* **2008**, *10*, 6615–6620.
- [84] Y. Zhao, D. G. Truhlar, *The Journal of Chemical Physics* **2006**, *125*, 194101.
- [85] Y. Zhao, D. G. Truhlar, *Theoretical Chemistry Accounts* **2008**, *120*, 215–241.
- [86] *AB INITIO Molecular Orbital Theory* — Wiley, <https://www.wiley.com/en-us/AB+INITIO+Molecular+Orbital+Theory-p-9780471812418>.
- [87] T. H. Dunning, *The Journal of Chemical Physics* **1989**, *90*, 1007–1023.

Chapter 6

General Conclusion

In this study, I developed a method to reduce the number of experiments required for reaction screening in organic synthesis.

In Chapter 2, an overview of the methods used in this study was provided. Initially, the artificial force induced reaction (AFIR) method was introduced. The details of the geometry generation algorithm, the genetic algorithm, and Bayesian optimization were also presented. Additionally, the Chemspeed ISYNTH robotic platform, designed for automated execution of general experimental manipulations, was introduced.

In Chapter 3, the reaction barrier datasets that can be used to benchmark the performance of optimization algorithms were created. The reaction barrier of the Claisen rearrangement for 100,000 allyl vinyl ether molecules with different substitutions was evaluated at a semiempirical PM7 level using the AFIR method. After evaluating its chemical and numerical features, it was found that the dataset well reflects the chemical trends of various substitutions. Using the created datasets, the performance of the genetic algorithm and the random search were compared. The obtained results suggested that the reaction barrier datasets of the Claisen rearrangement can be used to evaluate the performance of optimization algorithms.

In Chapter 4, a method to reduce the number of synthetic experiments required for

reaction screening by combining Bayesian optimization and reaction barrier calculations was developed. In this method, the reactivity or selectivity of chemical reactions is evaluated not only through experiments but also via Gibbs energy barrier calculations. To validate the impact of using computational Gibbs energy barrier data in Bayesian optimization for reaction screening, an extensive numerical investigation using the reaction barrier dataset of the Claisen rearrangement was conducted. The results showed that even the computational reactivity affected by noises as much as 20 kJ/mol helps reduce the number of required experiments.

In Chapter 5, the developed method was applied to the thermal azide-alkyne cycloaddition. To maximize the regioselectivity of the 1,5-disubstituted isomer, various combinations of azides and alkynes were explored with three optimization schemes that utilize Bayesian optimization and selectivity calculations. The results indicated that Bayesian optimization, utilizing both experimental and theoretical selectivity data, can effectively reduce the experimental cost required for substrate optimization and accelerate reaction screening. It was also found that certain combinations of azides and alkynes showed high 1,5-regioselectivity (74.7%) or 1,4-regioselectivity (86.4%).

The method developed in this study simultaneously handles two types of data with different levels of observation accuracy and costs. Generally, obtaining accurate data requires a lot of cost, whereas acquiring cost-effective data yields low-accurate, biased, and noised data. In reaction screening, accurate and expensive data correspond to experimental reactivity data, while the cost-effective, biased, and noised data correspond to computational reactivity data. The objective of the developed method is to minimize access to expensive experimental data by utilizing cost-effective computational data, despite its lower accuracy. This study demonstrates that utilizing computational data affected by computational errors can reduce the number of experiments and accelerate reaction

screening. The developed method is versatile and applicable to various reaction screening problems, thus contributing to the acceleration of reaction screening using quantum chemical calculations.

Future prospects include applications to more complex organic reactions and closed-loop reaction optimization by combining barrier calculations with Bayesian optimization and robotic synthesizers.

Acknowledgements

All the research was carried out under Professor Satoshi Maeda, Department of Chemistry, Faculty of Science, Hokkaido University. It was a great experience for me to see and respect what a talented researcher is like. I would like to express my deepest gratitude for his kind guidance, discussion, and support over the past six years.

I am grateful to Professors Tetsuya Taketsugu, Keisuke Takahashi, Hajime Ito, and Junya Hasegawa for serving as my thesis committee members. In particular, Professor Keisuke Takahashi provided valuable advice on the implementation of Bayesian optimization. His insights into data science have taught me the importance of not readily trusting available data. I express my sincere gratitude to him.

I would like to thank Specially Appointed Assistant Professor Yu Harabuchi for his guidance over the past six years. His close contact with students gave me a lot of insights and motivation. I express my gratitude for his kindness.

Specially Appointed Associate Professor Yuuya Nagata and Specially Appointed Assistant Professor Seiji Akiyama taught me the knowledge and techniques of organic experiments. I did not expect to work on organic synthesis again after being assigned to theoretical chemistry laboratory, but it was a very valuable experience for me. I would like to appreciate them.

I would like to express my gratitude to two of my seniors, Makito Takagi and Kanami Sugiyama, who provided all kinds of support in the laboratory. Their deep and broad

knowledge, skills, and kindness were a great support to me.

Takuma Ito and Hitoshi Nabata, my classmates in the Department of Chemistry, gave me hope for my daily life and inspiration for my research. During the past three years of my doctoral studies, I have come to deeply appreciate the presence of my peers in the same academic year.

I am also deeply grateful to the competitive programming platform, AtCoder. It is thanks to the unique contests and problems provided by AtCoder that I have found programming enjoyable. Without this experience, I might not have been able to implement the algorithms used in this research.

Finally, I would like to express my deepest gratitude to Yuko Kitamura for her love, kindness, and exploration of the world of chemistry with me during my long university life. She encouraged me and gave me the strength to overcome difficulties in my research. Thank you so much.