Theoretical studies on effects of protein-ligand/protein-protein interactions on protein functions

（タンパク質機能に対するタンパク質－リガンド/－タンパク質相互作用の理論解析）

YAMAMOTO YUTA

February 2023

Theoretical studies on effects of protein-ligand/protein-protein interactions on protein functions

（タンパク質機能に対するタンパク質－リガンド/－タンパク質相互作用の理論解析）

YAMAMOTO YUTA

Doctoral Program in Physics

Submitted to the

Degree Programs in Pure and Applied Sciences of the

Graduate School of Science and Technology

in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy in

Science

at the

University of Tsukuba

Contents

Chapter 1. General Introduction

Purpose of this Chapter

Information processing technology has become an indispensable part of modern society as computer performance has improved and the fields of use have expanded. The use of computers in research fields is no exception. Computational analysis methods such as fragment molecular orbital (FMO) calculations and molecular dynamics (MD) simulations are used in basic and applied research in a wide range of fields, including physics, chemistry, and pharmacology. The advantage of computational analysis methods is that they can efficiently analyze molecular-level phenomena that cannot be obtained by experimental analysis. MD simulations and FMO calculations are very effective methods for computational analysis of large-scale systems such as proteins, membranes, and polymeric materials. However, the huge volume of data obtained by these methods often makes manual analysis difficult. In recent years, the approach to scientific research has begun to shift from working hypothesis-driven science to data-driven science, and how to extract useful information from big data has become important. Against this background, M. Sultan *et al.* reported a case study of analysis of protein conformational changes using Random Forest,[1] a type of machine learning, and the development of analysis tools using machine learning has attracted attention.

In this doctoral thesis, computational analyses of several proteins were performed to reveal new scientific knowledge about these proteins and to establish new analytical methods. This chapter provides basic backgrounds about the study.

1-1. Protein and its function

Proteins play a leading role in almost all biological processes, whether at the molecular, cellular, or tissue level. Therefore, it is important to understand their functions at the atomic level in order to understand biological phenomena. Therefore, this section describes proteins.

1-1-1. Structures of proteins

Proteins are linear polymers consisting of repeating units of amino acid residues. The carboxyl group of each amino acid residue in the proteins is linked to the amino group of the next amino acid residue by a peptide bond (**Figure 1a**).

There are twenty amino acids that are specified by the genetic codes. There are twenty amino acids that are specified by the genetic codes. The amino acids in the proteins have various properties such as charges, hydrophobicity, volumes, and shapes. The property of each amino acid residue is determined by the side chain. Because possible conformations of the protein backbone and possible patterns of hydrogen bonding interactions between the backbone atoms are limited, a few characteristic local structures of the protein backbone are seen frequently within protein structures.[2] These local structures are called secondary structures. The alpha helices (α-helices) and beta sheets (β-sheets) are the most common secondary structures (**Figure 1b**). The hydrogen bonds stabilizing the alpha helix structures are formed between *i*-th and (*i*+4)-th residues (**Figure 1c**). On the other hand, in the beta sheets, the backbone hydrogen bonds are formed between two sheets (**Figure 1d**).

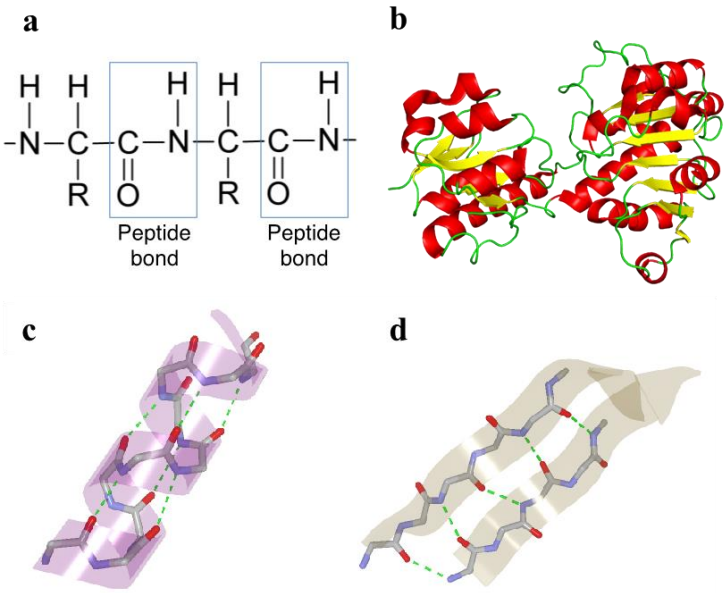


Figure 1. Primary and secondary structures of protein. (**a**) Schematic structure of the protein backbone. R is an amino acid side chain. (**b**) A crystal structure of C-terminal binding protein (PDB ID: 2OME). The α-helices and β-sheets are colored red and yellow, respectively. Large parts of this protein adopt the α-helix or β-sheet structures, and the alpha helices and beta sheets are connected by loops colored light green. The backbone hydrogen bonding interactions that stabilize the secondary structures are shown for (**c**) an alpha helix and (**d**) a beta sheet.

1-1-2. Functional expression and conformational change of proteins

A protein changes its own structure either globally or locally, when it acts a biologically essential function (**Figure 2**). Thus, it is most essential to understand these structural changes dynamically at the atomic level in order to understand biological phenomena. However, this is not easy study both experimentally and theoretically/computationally, and Scientists are currently working on various approaches.

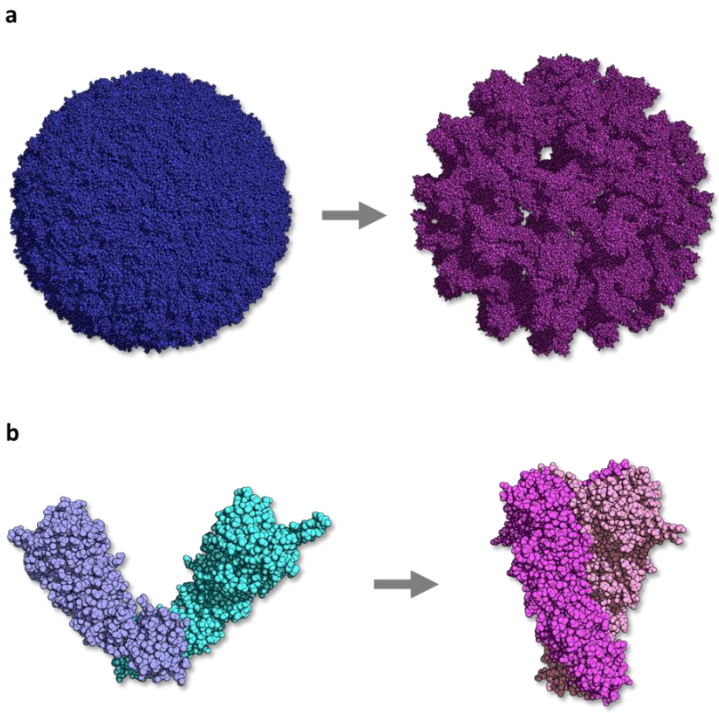


Figure 2. Conformational changes when a protein expresses its function. (**a**) Dengue virus: Dengue virus makes a conformational change from the inactive state on the left side to the active state on the right side during infection. (**b**) Hsp90: a chaperone protein that functions by changing its conformation from the inactive state (left-side) to the active state (right-side).

1-2. Paradigm shift from hypothesis-driven research to data-driven research

In recent years, the research approach has been changing from conventional hypothesis-driven science, i.e., science in the form of testing hypotheses assumed by researchers through experiments, to data-driven science, i.e., finding data to explain them based on experimental data obtained in various forms. This section provides background on the paradigm shift in research approaches.

1-2-1. Growth of computational science methods

It has not yet been half a century since the importance of computational science was first recognized and was able to provide some analysis for experiments. First principles calculations based on quantum mechanics were initially used only in the fundamental fields of quantum

chemistry and condensed matter physics. More recently, they have provided reliable analysis in applied sciences and have become indispensable in a variety of research fields. The history of computing has seen processing power continue to increase at an exponentially rapid rate from the birth of the electronic computer to the present. This is evidenced by the remarkable advances in personal computers and smart phones. Clearly, the reason for the astonishingly large progress in computational science is due to these advances in computers and the device technologies that support them. On the other hand, computational theory has also made great progress. In order to fully utilize large computational resources, an excellent computational theory is necessary, and it is fair to say that the synergistic effect of the two together has led to the development of computational science. Typical examples are density functional theory (DFT) calculations. This method makes it possible to calculate the stable atomic configuration structure, electronic states, and physical properties of a system from the principles of quantum mechanics itself, without the need for empirical knowledge. Many previous calculation methods have been described using this first-principles method, which is extremely precise and reliable. However, this method requires more computation time and storage capacity than the method using empirical potentials, and the scale of computation grows explosively as the system size increases. This limitation makes it difficult to describe systems much larger than the atomic scale in practical terms. Systems in which phenomena occurring at the atomic scale are intertwined with phenomena ranging from meso to macroscopic scales are the most challenging subjects for computational physics. Such calculations are required, for example, when attempting to elucidate the physiological functions of biological materials such as proteins and molecular motors from the elementary processes of individual atoms and molecules, or in the design of fuel cells from the atomic scale. The systems targeted in these studies are, of course, important as applied technologies, but they are also challenging problems whose principals have not yet been fully elucidated from a molecular theoretical viewpoint. Such problems cannot be solved by blindly using the current computational methods on a giant computer but require the development of new hybrid analytical methods. The QM/MM method and Fragment

molecular orbital (FMO) method are examples of such hybrid-type analysis methods. In order to promote computational science, which opens the way to such dramatic advances in science and technology, it is important to construct innovative computational theories, not to mention the development of computers themselves. [3]

1-2-2. Importance of data mining for scientific databases

Due to the rapid progress in genome science in recent years, sophisticated genome sequencing devices have been developed one after another, and the genome sequences of a vast number of organisms are being determined on a daily basis. Furthermore, in addition to genomes, it has become possible to comprehensively collect a vast amount of data, called omics, on gene expression, proteins, metabolites, and so on. Therefore, scientific databases with guaranteed data quality and quantity are being prepared. In parallel with the development of scientific databases, computational technologies are also advancing, and methods for extracting information from huge databases, such as clustering[4] and deep learning[5], are being developed. Therefore, it is expected that new information can be obtained by data mining against life science genome databases accumulated so far, such as GenBank[6]. Therefore, data mining efforts for various life science databases are important.

1-2-3. Data-driven research

Data mining of databases that accumulate data from experiments has recently become a popular practice, as exemplified by the search for the shortest pathway in organic synthesis[7, 8] and the prediction of new strains of influenza[9]. However, data mining using data obtained from computational analysis methods, such as MD simulations and FMO calculations, is still a relatively rare approach. The establishment of a data mining method for data obtained by computational simulations is expected to be a breakthrough that will lead to a dramatic growth in scientific research based on new approaches.

1-3. Computational Analysis Methods

This section will give a background of the computational/theoretical analysis methods used in this study.

1-3-1. Molecular dynamics (MD) simulation

Molecular dynamics (MD) simulation is a method of computer simulation of the physical motion of atoms and molecules.[10, 11] MD simulations take track with changes of energy, position, velocity and other parameters while time. Generally, when the N-body system simulated, the time evolution of the state is obtained by calculating the Newton equations of motion (**Equation 1**).

$$m_i \frac{d^2}{dt^2}\mathbf{r}_i = -\nabla_i U(\mathbf{r}_1 \cdot \cdot \cdot \mathbf{r}_N) \quad i = 1 \sim N \qquad \textbf{(Equation 1)}$$

m_i and $\mathbf{r}_i$ are the position and the mass of particle i, t is time, and U is the potential energy of the whole system. The potential energy was calculated by the sum of four energy components: (C1) stretched bond energy in the molecule, (C2) transformational angular energy of the angle between two bonds, (C3) dihedral torsion energy of the torsion angle, and (C4) non-bonding interaction energy (van der Waals (vdW) interaction and coulomb interaction). In the classical MD simulation of proteins, the potential energy of the system is modeled to represent the stable structure of the protein, represented by the molecular force field. There are several types of the molecular force fields in the world. They are used depending on the application which users used. It is specified as a group of empirical parameters according to the application. Typical types of molecular force fields are AMBER, CHARMM and MMx (X = 2, 3, 4). For example, the AMBER force field[12] is defined at below (**Equation 2**).

$$U(\{\mathbf{r}_i\}) = \sum_{\substack{\text{bonds}\\(ij)}} k_{\mathrm{r}}(r_{ij} - r_{\mathrm{eq}})^2 + \sum_{\substack{\text{angle}\\(ijk)}} k_{\theta}(\theta_{ijk} - \theta_{\mathrm{eq}})^2 + \sum_{\substack{\text{dihedrals}\\(ijkl)}} \frac{k_{\varphi}}{2}\left[1 - \cos\left(n\phi_{ijkl} - \gamma\right)\right] + \sum_{\mathrm{i<j}} \left[\frac{A_{\mathrm{ij}}}{r_{ij}^{12}} - \frac{B_{\mathrm{ij}}}{r_{ij}^{6}}\right] + \sum_{i<j} \frac{q_i q_{\mathrm{j}}}{\varepsilon r_{ij}} \quad \textbf{(Equation 2)}$$

The first, second, and third terms are expressed as functions of the bond distance r between two atoms in the molecule, the bond angle θ between three atoms, and the dihedral angle φ consisting of four atoms, respectively. Here, k_{X} and X_{eq} (X= r, θ, φ) are parameters determined for each amino acid residue. n and γ in the third term are the identical rotational number and the phase for the torsion angle. On the other hand, the fourth and fifth terms represent the contribution of the van der Waals (vdW) interaction and the electrostatic interaction between two atoms, respectively, both of which are functions of the distance between them. In addition, A_{ij} and B_{ij} are parameters for each atomic pair obtained from the vdW radius, and q_i is the charge of the i-th atom.

MD simulation has become an indispensable tool for understanding the physical basis of the structure and function of biomacromolecules, as the development of well-known parameter sets, such as the AMBER force field, allows analysis over long periods of time and for large molecules.[13, 14]

1-3-2. Targeted molecular dynamics (TMD) simulation

Targeted molecular dynamics (TMD) simulation is a tool implemented in Nanoscale Molecular Dynamics[15] (NAMD; Not Another Molecular Dynamics Program), one of the programs that performs MD simulations.

TMD simulation is a calculation method that applies steering forces to each atom so that the Root Mean Square (RMS) of the initial structure and the target structure approach each other during the simulation. Force is added as a gradient of the potential energy of each atom (**Equation 3**).

$$U_{\mathrm{TMD}} = \frac{1}{2}\frac{k}{N}[\mathrm{RMS}(t) - \mathrm{RMS}^{*}(t)]^2 \qquad \textbf{(Equation 3)}$$

where RMS(t) is the RMS distance (RMSD) of the structure obtained by the simulation when it is closest to the target structure, and $\mathrm{RMS}^{*}(t)$ is the reference value of the RMSD required for structural change, calculated from the RMSD of the initial and target structures. This allows us to track the rare event of protein conformational change by MD simulation in a short period of time.

Structural changes that are important for protein function expression are "rare events" that are observed in MD simulations for even longer times than can be tracked in regular MD simulations. Therefore, to track large-scale conformational changes such as protein folding and domain motion, either long-time dynamics or efficient sampling methods such as multi-canonical or replica exchange methods were necessary. The former has been achieved only on supercomputers such as Anton, while the latter requires the acquisition of certain know-how for each method to achieve structural sampling, and not everyone can easily perform the calculations. TMD is one of the simpler and faster sampling methods.

1-3-3. Molecular orbital method

The molecular orbital (MO) method[16] is an approximate solution of the Schrödinger equation for the electronic state of a molecular system. The Schrödinger equation for multi-electron wave functions is reduced to the Schrödinger (HF; Hartree-Fock) equation for molecular orbitals (functions), which are one-electron wave functions, via the orbital approximation.[17] MO methods can be further classified into *ab initio* (non-empirical) MO methods, semi-empirical MO methods, and Hückel MO methods.

The *ab initio* MO method is based on the principle of not using any numerical data obtained from experiments, while the semi-empirical MO method uses parameters previously determined to reproduce the physical properties (experimental values) of a group of reference molecules in order

to reduce computational costs. In the 1960s and 1970s, calculations using the MO method began to be performed. Computer performance was very poor, and the scope of the *ab initio* MO method was limited to very small molecules. Therefore, some theoretical chemists developed semiempirical MO methods with the aim of developing practical MO methods applicable to larger molecules. Recently, the performance of the combinator has improved considerably, making it possible to apply the *ab initio* MO method to much larger systems, and at the same time, the semiempirical MO method has expanded its range of application to larger systems that cannot be handled by the *ab initio* MO method. On the other hand, when the system of interest is large or when dealing with a collection of molecules, it is difficult to perform MO calculations even with the current computer environment. In such cases, the structure, vibration, and reaction of the molecular system are sometimes discussed using analytical functions that approximate the potential energy surfaces.

The electron configuration of the He atom is said to form a closed shell $(1s)^2$, and the two electrons both occupy 1s orbitals. Therefore, when the He atoms are ionized by X-ray excitation and the energy of the emitted photoelectrons is measured (photoelectron spectrum), it is expected that only one peak corresponding to the ionization of 1s electrons is measured. However, when photoelectrons are measured with high sensitivity, photoelectrons emitted from He atoms are observed at 40-50 eV, on the low energy side of the main peak. Therefore, the concept of MOs in the one-electron picture is a kind of approximation, and the actual situation is described by a more complicated wave function. In next section, I will start with the Hartree-Fock (HF) approximation, which is an *ab initio* theory of the one-electron approximation, and touch upon the electron correlation problem in the next section.

1-3-4. Hartree-Fock (HF) method[17-19]

An atom with *n* electrons is a complex system in which electrons move in a field created by the nucleus and interact with each other. $\Psi(1,2,\cdots,N)$ If we focus on one of the *n electrons*, we can consider this electron to be in a stationary state $\phi_1(1)$ in the centrosymmetric field created by the

nucleus and the remaining electrons. The wave function $\Psi(1,2,\cdots,N)$ of an n-electron atom can be expressed as a product of one-electron wave functions $f_1(1)f_2(2)\cdots f_N(N)$ in the centrally symmetric field created by the nucleus and the remaining electrons. This is called the one-electron approximation. This is called the one-electron approximation. Within this one-electron approximation, each electron depends on the states of all the remaining electrons, so they must all be determined simultaneously. Hartree was the first to apply the variational method to the problem of constructing the wave function of a many-electron system using the set of one-electron functions $\{f_i\}$ in 1928. Hartree used the simple product of f_i (the Hartree product) to represent the wave function of a many-electron system. However, this wave function does not satisfy the fundamental property of the electron, the Fermi particle, which is pan-symmetry. To overcome this shortcoming, two years later in 1930, Fock expressed the wave function of a many-electron system using the Slater determinant, which takes the antisymmetry of the electron into account (HF method). These methods include the wave function in the operator. Therefore, we treat the problem by the method of self-consistent fields.

These two methods, i.e., Hartree's method and Hartree-Fock's method, have different energy expressions depending on the form of the function used. Specifically, owing to the antisymmetry of the electrons, HF method includes an integral called the exchange integral, which is a purely quantum mechanical effect, in the energy table. Also, the antisymmetry results in Pauli's exclusion law, which is not analogous to the classical picture.

In the electron configuration for the electron ground state (ground configuration), the electrons are packed two at a time, starting from the MO with the lowest orbital energy. When the system has an even number of electrons and all MOs are packed with two electrons each (closed-shell system), the number of electrons in the α spin (upward spin) and β spin (downward spin) are equal and symmetry is preserved, but when the number of electrons is odd or the number of electrons in one spin exceeds that of the other spin (open-shell system) Two possibilities arise depending on how the MOs are handled for the electrons of each spin. If the electron configuration

is a closed-shell system, the MOs for both spins will be the same, and the spin-restricted Hartee-Fock (RHF) method is applied.

In the case of the Seki shell system, on the other hand, the spin-restricted open-shell HF (ROHF) method restricts the MOs to be the same for the orbitals occupied by two electrons each, and the spin-unrestricted HF (UHF) method defines the MOs for each spin separately. The ROHF method satisfies spin symmetry, while the UHF method does not. On the other hand, the UHF method imposes no restriction on the MOs and thus gives lower energy values than the ROHF method; when the UHF method is applied, care must be taken when the spin symmetry deviations are large.

Next, we derive the HF equation based on the Linear Combination of Atomic Orbital (LCAO) method as an actual solution of the HF method. The electron-electron interaction can be roughly classified into the exchange interaction, which is caused by the fact that electrons are Fermions, and the so-called Hartree interaction, which is caused by the Coulomb repulsion force between two electrons. The exchange interaction is based on Pauli's exclusion law, which states that Fermions cannot occupy the same state at the same time, is treated explicitly in the HF method.

For an accurate treatment of the electronic state, the Slater determinant, a 3*n*-dimensional wave function satisfying Pauli's exclusion criterion,

$$\Psi(1,2,\cdots,n) = \frac{1}{\sqrt{n}} \begin{vmatrix} f_1(1) & f_2(1) & \cdots & f_n(1) \\ f_1(2) & f_2(2) & \cdots & f_n(2) \\ \vdots & \vdots & \ddots & \vdots \\ f_1(n) & f_2(n) & \cdots & f_n(n) \end{vmatrix} \qquad \textbf{(Equation 4)}$$

is treated as the trial function. The Hamiltonian of a given molecule is given by,

$$H = \sum_{i=1}^{n} \left(-\frac{\nabla_i^2}{2} + V_{Ne}(\mathbf{r}_i) \right) + \frac{1}{2} \sum_{i \neq j}^{n} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|} \qquad \textbf{(Equation 5)}$$

The energy evaluated for the Hamiltonian with the trial wave function,

$$E_0 = \mathrm{Min}\langle \Psi | H | \Psi \rangle = \mathrm{Min}\langle \Psi | \left[\sum_{i=1}^{n} \left(-\frac{\nabla_i^2}{2} + V_{Ne}(\mathbf{r}_i) \right) + \frac{1}{2} \sum_{i \neq j}^{n} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|} \right] | \Psi \rangle \quad \textbf{(Equation 6)}$$

results in the following HF energy,

$$E_{\mathrm{HF}} = \sum_{i=1}^{\mathrm{occ}} h_{ii} + \frac{1}{2} \sum_{i,j}^{\mathrm{occ}} \left((ii|jj) - \frac{1}{2}(ij|ji) \right) \quad \textbf{(Equation 7)}$$

where the matrix elements are

$$h_{ii} = \int f_i^*(\mathbf{r}) \left(-\frac{\nabla^2}{2} + V_{\mathrm{ne}}(\mathbf{r}) \right) f_i(\mathbf{r}) d\mathbf{r} \quad \textbf{(Equation 8)}$$

$$(ij|kl) = \int\int \frac{f_i^*(\mathbf{r}) f_j(\mathbf{r}) f_k^*(\mathbf{r}') f_l(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r}\, d\mathbf{r}' \quad \textbf{(Equation 9)}$$

The former is called the one-electron integral and the latter the two-electron integral, respectively. Now, the HF Lagrangian is constructed by adding the energy with the constraint of the orthonormalization on the orbital with Lagrange multipliers, ε_{ij}, as

$$L_{\mathrm{HF}} = \sum_{i=1}^{\mathrm{occ}} h_{ii} + \frac{1}{2} \sum_{i,j}^{\mathrm{occ}} \left((ii|jj) - \frac{1}{2}(ij|ji) \right) + \sum_{ij} e_{ij} \left(S_{ij} - d_{ij} \right) \quad \textbf{(Equation 10)}$$

where the overlap integral is defined as

$$S_{ij} = \int \Phi_i^*(\boldsymbol{r}) \Phi_j(\boldsymbol{r}) d\boldsymbol{r} \quad \textbf{(Equation 11)}$$

By considering the stationary solution for the variation of the orbital, f_i^* ($\frac{\partial L_{HF}}{\partial f_i^*} = 0$), we obtain the following HF equation,

$$\left[-\frac{\nabla^2}{2} + V_{\mathrm{Ne}} + V_{\mathrm{H}} - V_{\mathrm{X}} \right] f_i^{\mathrm{HF}}(\mathbf{r}) = \varepsilon_i^{\mathrm{HF}} f_i^{\mathrm{HF}}(\mathbf{r}) \qquad \textbf{(Equation 12)}$$

where

$$V_H(\mathbf{r}) = \sum_j \int \frac{f_j^{\mathrm{HF}*}(\mathbf{r}') f_j^{\mathrm{HF}}(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r}' \qquad \textbf{(Equation 13)}$$

$$V_{\mathrm{X}} f_i^{\mathrm{HF}}(\mathbf{r}) = \sum_j \left(\int \frac{f_j^{\mathrm{HF}*}(\mathbf{r}') f_i^{\mathrm{HF}}(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r}' \right) f_j^{\mathrm{HF}}(\mathbf{r}) \qquad \textbf{(Equation 14)}$$

Note here that the exchange term is represented by a nonlocal operator.

1-3-5. SCF method[16, 20]

This section explains how to solve the HF equation. For most cases, the HF equation can be accurately solved by numerical calculations. However, in molecular systems, there are many nuclei, which are the centers of attractive interactions, and it is difficult to obtain MOs numerically. Roothaan solved this difficulty. Instead of direct numerical calculations for the MOs, he used an analytical expansion method. In other words, he expanded the orbital functions by an appropriate set of the basis functions centered at the nuclei, χ_p.

$$\psi_i(\boldsymbol{r}) = \sum_{p=1}^{m} C_{pi} \chi_p(\boldsymbol{r}) \qquad \textbf{(Equation 15)}$$

where C_{pi} is the expansion coefficient, which is defined to minimize the energy. As a set of basis functions, atomic orbitals are employed. This is so-called Linear Combination of Atomic Orbitals

(LCAO) method. Since molecules are made of atoms, they will retain the properties of their constituent atoms. The basic idea is that the waves of a molecule may be represented by the superposition of the waves of atoms placed on the nucleus. While the extended Hückel method used Slater-type orbitals (STO), the *ab initio* MO method uses Gaussian-type orbitals (GTO) exclusively for practicality. The basic principle remains the same. the HF equation is replaced by an algebraic problem of matrix eigenproblems. This is known as the Roothaan SCF method.

Now, the vectors of the basic functions and expansion coefficients are

$$\boldsymbol{\chi} = \left(\chi_1, \chi_2, \cdots, \chi_m\right) \qquad \textbf{(Equation 16)}$$

$$\mathbf{C}_i = \begin{pmatrix} C_{1i} \\ C_{2i} \\ \vdots \\ C_{mi} \end{pmatrix} \qquad \textbf{(Equation 17)}$$

in any case

$$\psi_i = \boldsymbol{\chi}.\mathbf{C}_i \qquad \textbf{(Equation 18)}$$

is expressed as Substituting this basis into the HF equation, the Hartree-Fock-Roothaan equation for closed-shell molecules is finally given by the following secular equation as

$$\sum_{s=1}^{m} C_{is}(F_{rs} - \varepsilon_i S_{rs}) = 0 \ (r = 1, 2, \cdots, m) \qquad \textbf{(Equation 19)}$$

where m is the number of atomic orbitals in the molecular system, F_{rs} is the Fock matrix defined as

$$F_{rs} = h_{rs} + \sum_{t=1}^{m}\sum_{u=1}^{m} P_{tu}\left[\left(rs|tu\right) - \frac{1}{2}\left(ru|ts\right)\right] \qquad \textbf{(Equation 20)}$$

In the Fock matrix element given by the above equation, h_{rs} and $(rs|tu)$ are one- and two-electron matrix elements given respectively as

$$h_{rs} = \int c_r^*(\mathbf{r}) \left(-\frac{\nabla^2}{2} - \sum_A \frac{Z_A}{|\mathbf{R}_A - \mathbf{r}|} \right) c_s(\mathbf{r}) d\mathbf{r} \qquad \textbf{(Equation 21)}$$

$$(rs|tu) = \int \int \frac{\chi_r^*(\boldsymbol{r}) \chi_i^*(\boldsymbol{r}') \chi_u(\boldsymbol{r}') \chi_s(\boldsymbol{r})}{|\boldsymbol{r} - \boldsymbol{r}'|} d\boldsymbol{r} d\boldsymbol{r}' \qquad \textbf{(Equation 22)}$$

where P_{tu} is the density matrix given by

$$P_{tu} = 2 \sum_{j=1}^{n} C_{jt} C_{ju} \qquad \textbf{(Equation 23)}$$

S_{rs} is the overlap matrix as

$$S_{rs} = \int \chi_r^*(\boldsymbol{r}) \chi_s(\boldsymbol{r}) d\boldsymbol{r} \qquad \textbf{(Equation 24)}$$

1-3-6. Electron correlation[21, 22]

In the HF method, the electron-electron interaction is considered as a mean field using MOs, so the HF wave function is not an exact solution to the Schrödinger equation. The discrepancy between the exact electron-electron interaction and the that estimated by the mean field approximation, HF method, is called the electron correlation. Accurate estimation of electron correlation energies is the key to making quantitative comparisons with experimental data in first-principles calculations. There are two main methods for dealing with electron correlation. One is "wave function theory," which starts from the HF wave function and incorporates electron correlation at the wave function level, and the other is "density functional theory," which is computationally equivalent to the HF method but incorporates approximate electron correlation

effects derived from the model as a functional of density. Here, we will discuss these two electron correlation theories.

1-3-7. Wave function theory[16, 23]

The HF method describes the electronic ground state by a single Slater determinant, i.e. the HF ground state configuration. The electron configuration created by excitation of an electron from an occupied orbital to an unoccupied (virtual) orbital is called a singly excited configuration. Generally, one can construct the non-redundant *n*-multiply excitation configuration by removing *n* occupied orbitals and adding unoccupied virtual orbitals in each column of the Slater determinant. In the electron correlation method, the electron wave function expressing the accurate electron ground state takes the form of a linear combination of the ground state HF configuration and the excited configurations. The excitation configurations can be classified into single excitation, double excitation, triple excitation, and quadruple excitation configurations, depending on the number of electrons to be excited. The most important contribution to the estimation of electron correlation comes from the two-electron excitation (double excitation) configuration, and the second most important is the four-electron excitation (quadruple excitation) configuration. There are three methods for estimating electron correlations such as perturbation theory, variational method, and cluster expansion. A brief description of each method is given below.

1-3-8. Configuration interaction (CI) method

The most intuitive way to estimate electron correlations by a variational method is the Configuration Interaction (CI) method,[24, 25] in which the wave function is expanded using several electron configurations (Slater determinant and its linear combination). When a Hamiltonian matrix is constructed, the energies and the coefficients (called CI coefficients) for each configuration are estimated by diagonalizing the CI Hamiltonian matrix. The CI method that considers all excitation configurations from occupied orbitals to virtual orbitals is called the full CI (Full CI) method and gives the most accurate solution within the approximation of the basic functions used in the

calculation. However, it is not practical in terms of computational cost. In actual calculations, the CI method with wave function expansion terminated within a given excitation order is used. In the past, the CISD method, which considers up to single and double excitations, or the CID method, which considers only double excitations, were often used. However, the CI method with the limited order expansion has the defect that it does not satisfy "size-consistency", and is not often used now. Nonetheless, its basic idea is very important.

In the CI method, the wave function of type CI is given by

$$\Psi_{CI} = \sum_{I}^{0,S,D,\cdots} C_I \Phi_I \qquad \textbf{(Equation 25)}$$

The CI is defined as follows where 0, S, D, ... denote the HF ground state, one-electron excitation, two-electron excitation configuration, etc. By using the CI wave function, one can evaluate the energy as

$$E_{CI} = \frac{\langle \Psi_{CI} | \hat{H} | \Psi_{CI} \rangle}{\langle \Psi_{CI} | \Psi_{CI} \rangle} = \frac{\sum_{I,J}^{0,S,D,\cdots} C_I^* C_J H_{IJ}}{\sum_{I,J}^{0,S,D,\cdots} C_I^* C_J} \qquad \textbf{(Equation 26)}$$

where H_{IJ} is called the Hamiltonian matrix element given by

$$H_{IJ} = \langle \mathrm{F}_I | H | \mathrm{F}_J \rangle \qquad \textbf{(Equation 27)}$$

Note here that Φ_I is a set of all possible configurations. By adopting the variational principle with respect to C_I^*, we obtain the CI secular equation as

$$\sum_{J} (H_{IJ} - E_{CI} S_{IJ}) C_J = 0 \qquad \textbf{(Equation 28)}$$

It would seem that a tremendous number of Hamiltonian matrix elements must be computed to solve this equation. However, many matrix elements are actually zero owing to the fact that the Hamiltonian is described by one- and two-electron operators. Thus, the Hamiltonian matrix elements between excitation configurations consisting of more than three-electron excitations are zero. For example, elements such as H_{0T} and H_{0Q} are zero. This leads to the fact that the two-electron excitation configuration is the main correction to the ground-state energy.

1-3-9. Cluster expansion (CC) method

The wave function of the cluster expansion (CC) method[26, 27] is given by

$$\begin{aligned}\Psi_{CC} &= \exp\left(\hat{T}\right)\Phi_0 = \left(1+\hat{T}+\frac{1}{2}\hat{T}^2+\frac{1}{6}\hat{T}^3+\cdots\right)\Phi_0 \\ &= \left(1+\hat{T}_1+\left(\frac{1}{2}\hat{T}^2+\hat{T}_2\right)+\cdots\right)\Phi_0\end{aligned} \qquad \textbf{(Equation 29)}$$

It is characterized by the use of an exponential function as the excitation operator to create the wave function. Here, the excitation operator, *T*, is a sum of all possible excitation operators as

$$\hat{T} = \hat{T}_1 + \hat{T}_2 + \cdots \qquad \textbf{(Equation 30)}$$

where T_1, T_2, ... are operators that generate electron configurations for one-electron, two-electron, ... etc. excitations. Let us note that some of these operators contain expansion coefficients, which are variables of the CC method. In a similar manner, the CI wave function given by the above equation is

$$\Psi_{CI} = \left(1+\hat{C}\right)\Phi_0 \quad \hat{C} = \hat{C}_1 + \hat{C}_2 + \cdots \qquad \textbf{(Equation 31)}$$

C_1, C_2, ... are the creation operators for one-electron, two-electron, ... excitations including CI coefficients. The coupled cluster method with single and doubles with third-order perturbation correction (CCSD(T)) method, where the contribution of three-electron excitations is estimated from the perturbation method, gives highly accurate results comparable to chemical accuracy and is called the golden standard method. Comparing the Taylor expansion and the CI wave function in the second line of the equation, it can be seen that various extra terms appear in the CC method. For example, the CISD and CCSD have the same number of expansion coefficients for the CI and cluster expansion methods, which even consider the same electronic excitations for the *C* and *T* operators. In other words, the CC method can incorporate a larger number of excitation configurations with the same number of variables as the CI method. Therefore, it gives more accurate results than the CI method. Unlike the CI method, the equations to be solved are the following nonlinear equations

$$\Phi_0 e^{-T} H e^{T} \Phi_0 = E \qquad \textbf{(Equation 32)}$$

$$\Phi_I e^{-T} H e^{T} \Phi_0 = 0 \qquad \textbf{(Equation 33)}$$

It can be solved by an iterative method. Since it is not a variational method, it can take a value lower than the true energy.

1-3-10. Møller-Plesset perturbation (MP*n*) method

The simplest method for estimating electron correlation is the Møller-Plesset (MP2) method[28] based on second-order perturbation theory, which treats electron correlation as a perturbation and determines the energy and the electron wave function of a given system by perturbation expansion. As the order of the perturbation is increased, the accuracy of the calculation increases, but the computational cost increases rapidly. The MP4 method, which considers perturbation expansions up to the fourth order, is now often used because of the recent improvement in computer performance.

The MP method, like the HF method, satisfies "size consistency" and can be discussed based on relative energies for systems of different sizes when determining molecular binding energies, etc.

The MP method splits the Hamiltonian into a one-electron term and an interaction term as

$$\hat{H} = \widehat{H_0} + \hat{V}, \ \widehat{H_0} = \sum_i \hat{F}(i) \qquad \textbf{(Equation 34)}$$

where the one-electron term consists of the Fock operator and includes the mean field (part of the electron interaction) under the one-electron approximation. By applying second-order perturbation theory to this partitioned Hamiltonian yields

$$E^{(2)} = \sum_I^D \frac{V_{0I}V_{I0}}{E_0 - E_I}, V_{0I} = \langle \Phi_0 | \hat{V} | \Phi_I \rangle \qquad \textbf{(Equation 35)}$$

Using the set of MOs, we obtain

$$E^{(2)} = -\sum_{i,j}^{\text{occ}} \sum_{a,b}^{\text{vir}} \frac{(ai|bj)\left[2(ia|jb) - (ib|ja)\right]}{e_a + e_b - e_i - e_j} \qquad \textbf{(Equation 36)}$$

$$(ia|jb) = \int f_i^*(\mathbf{r}_1) f_a(\mathbf{r}) \frac{1}{r_{12}} f_j^*(\mathbf{r}_2) f_b(\mathbf{r}_2) \qquad \textbf{(Equation 37)}$$

where ε_p is the p-th orbital energy, (pq|rs) is the two-electron integral. Note that i, j label the indices representing the occupied orbitals, and a, b do the virtual orbitals, respectively.

1-3-11. Multi-configulation SCF (MCSCF) method

The HF method used a single Slater determinant to represent the ground state wave function. However, the HF method cannot correctly represent, for example, the electronic state of a hydrogen molecule dissociated into two hydrogen atoms. To solve this drawback, one can express the

wavefunction using several Slater determinants. This is the basic idea of the multiconfiguration (MC) SCF method.[17] The wavefunction of the MCSCF method is

$$\Psi_{MC} = \sum_I A_I \Phi_I \qquad \textbf{(Equation 38)}$$

The CI expansion is used as in the CI method, where A_I is the expansion coefficient. The difference from the CI method is that the orbital coefficients of each configuration function Φ_I (one-electron HF orbitals in the HF method) are also variationally optimized at the same time. As with the CI method, the expansion coefficients are determined using the variational method. In the MCSCF method, the MOs are also determined simultaneously by SCF calculations. Therefore, the calculation is much more difficult than the HF method. In the MCSCF method, there are many choices in how to select the configuration functions. Today, the most commonly used MCSCF method is the complete active space (CAS) SCF method, which uses MOs near the valence orbitals, and all possible configurations of electronic excitation in them as configuration functions. The space of the chosen MOs is called the active space (active space).

1-3-12. Density functional theory (DFT)

The traditional approach to electron correlations in the first-principles calculations considers electron correlations based on wave functions. In contrast, density functional theory (DFT)[29, 30] starts from the electron density, not the wave function. Behind the possibility of such a treatment is a fundamental theorem called the Hohenberg-Kohn (HK) theorem, which can be applied in a straightforward manner. A straightforward application of the HK theorem is the Thomas-Fermi (Dirac) theory, which estimates the total energy of atoms about 15% to 50% lower. It also fails to describe the formation of molecules, such as the fact that molecules are more stable when they dissociate into atoms in pieces. On the other hand, the approximation introduced by Kohn and Sham in 1965 made the density functional theory a practically useful method.

In the KS-DFT, the kinetic energy term is reduced to a form that can be computed accurately and easily, and the remaining part is all squeezed into the exchange correlation functional. In KS-DFT, the kinetic energy term is reduced to a form that can be calculated accurately and easily, and the remaining part is squeezed entirely into the exchange correlation functional. The kinetic energy term in the KS approximation can be approximated by the KS orbital $\{j_i\}$ and the

$$T_{\mathrm{KS}}\left[\{j_i\}\right] = \sum_{i=1}^{N} \langle j_i | \left(-\frac{\nabla^2}{2} \right) | j_i \rangle \qquad \textbf{(Equation 39)}$$

Let be the kinetic energy term of the HF method. which has the same form as the kinetic energy term in the HF method. The electron density can be calculated using the KS orbitals as

$$\rho(\boldsymbol{r}) = \sum_{i=1}^{occ} |\varphi_i(\boldsymbol{r})|^2 \qquad \textbf{(Equation 40)}$$

In the KS approximation, as with the HF orbitals, the KS spin orbitals can be divided into occupied spin orbitals occupied by a single electron and virtual spin orbitals containing no electrons.

$$E_{\mathrm{KS}}\left[r, \{j_i\}\right] = T_{\mathrm{KS}}\left[\{j_i\}\right] + V_{\mathrm{ne}}[r] + J[r] + E_{\mathrm{XC}}[r] \qquad \textbf{(Equation 41)}$$

The form of the equation is as follows. where V_{ne} is the electron-nuclear attraction term, J is the Coulomb term, and E_{xc} is the exchange-correlation term. Now, the exchange-correlation functionals are functional differentiated by the density as the exchange-correlation potential, as follows

$$\hat{v}_{\mathrm{XC}} = \frac{dE_{\mathrm{XC}}[r]}{dr(\mathbf{r})} \qquad \textbf{(Equation 42)}$$

defined as "KS", the KS equation is obtained in the same way as the HF method was derived.

$$\hat{F}_{\mathrm{KS}}\phi_i = \varepsilon_i \phi_i \qquad \hat{F}_{\mathrm{KS}} = \hat{h} + \hat{V}_H + \hat{v}_{\mathrm{XC}} \qquad \textbf{(Equation 43)}$$

where *h is the* one-electron operator (sum of the kinetic energy operator and the nuclear-electron attraction operator) as in the HF method, V_{H} is the inter-electron Coulomb operator, and v_{xc} is the exchange correlation potential operator.

1-3-13. Fragment molecular orbital (FMO) method

The fragment molecular orbital (FMO) method separates a molecular assembly or large molecule to some fragments to calculate at *ab initio* level. *Ab initio* calculations are executed on monomers (fragments) and dimer (fragment pairs). When execute *ab initio* calculation in FMO method, it performs the calculation under the electrostatic potential of the other fragments. Even large proteins can be calculated. This is because each fragment to be calculated is only a few dozen atoms.[8] E_{FMO} which is the total energy of the whole system was calculated by the later equation (**Equation 44**).

$$E_{\mathrm{FMO}} = \sum_{I}^{N} E_I + \sum_{I>J}^{N} \left(E_{IJ} - E_I - E_J\right) + \cdots (\text{higher body contribution}) \qquad \textbf{(Equation 44)}$$

E_I and E_{IJ} are the total energies of the monomer and dimer, respectively. In the FMO method, it is possible to directly calculate the interaction energy between fragment pairs due to the nature of the equation for calculating the total energy of the system. These are called the fragment interaction energy (IFIE) or pair interaction energy (PIE). PIE analysis is an effective analytical technique for comprehending the properties of molecular- molecular interactions. Especially, it reveals the interaction energy between the residues around the ligand and the ligand in protein/ligand

complexes. In drug design field, functional group optimization of drug candidate compounds use PIE analysis (**Figure 3a**).[31]

However, when conducting PIE analysis within or between proteins, the number of PIEs to be analyzed is huge (approximately [number of amino acid residues in protein]^2 / 2), making the analysis difficult (**Figure 3b**). Therefore, it is important to develop analysis tools that facilitate the analysis of PIEs within or between proteins.

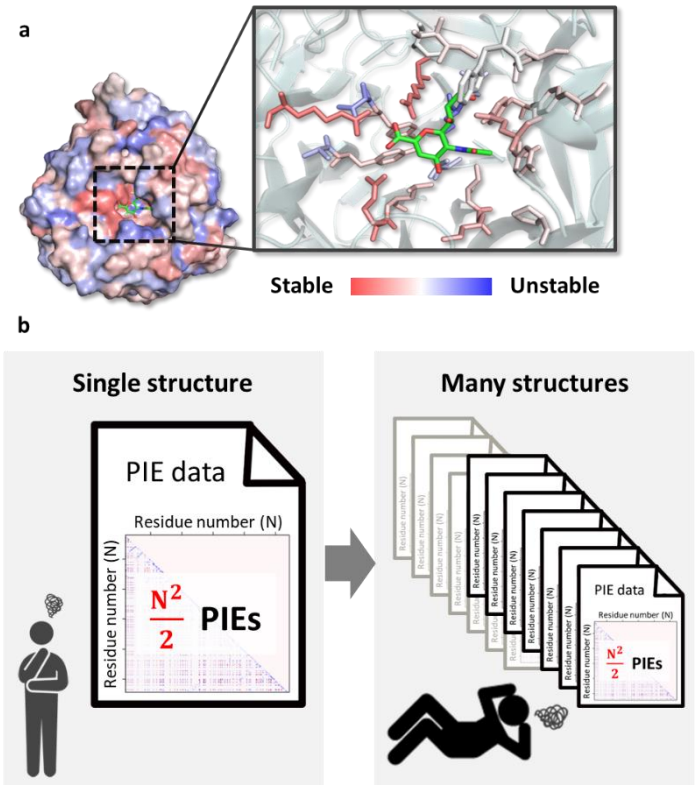


Figure 3. (**a**) FMO calculation results image. The FMO calculation can quantitatively evaluate and visualize the interaction between the ligand and the amino acid residues around the ligand as shown in the figure. (**b**) The FMO calculation is a very useful analysis method when the amino acid residues to be analyzed are specified, such as "residues around the ligand". However, when the number of analysis targets increases, such as "protein-protein interactions" or "all structures sampled from MD trajectory data", the analysis becomes difficult.

1-4. Machine learning

In this study, machine learning algorithms, which have been remarkably developed in recent years, were used to analyze a huge amount of data. This section provides background on the three machine learning algorithms used in this study, decision trees, bootstrap aggregating, and random forests, and provide an overview of these algorithms.

1-4-1. Decision tree

Decision trees are used as predictive models in the field of machine learning to draw conclusions about the target value of an item based on observations of that item (**Figure 4**). Decision trees are often used in data mining because the process of classification can be easily interpreted. In this case, the decision tree shows a tree structure in which the leaves (nodes) represent the classification and the branches represent the collection of features (features) that led to the classification.[32]

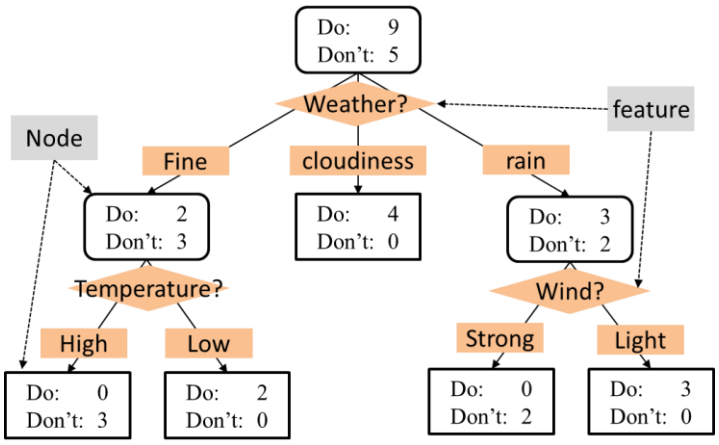


Figure 4. Diagram of a decision tree. A decision tree that determines the conditions for playing golf is shown. The black boxes are nodes and the orange ones are features. The decision tree has a hierarchical structure based on several features.

The values named the Gini index (Gini impurity; I_G) is calculated in the decision tree algorithm. It is used to measure importance of features. The Gini index is a value calculated for a node classified by a feature as the impurity of the data in that node (**Equation 45**).

$$I_G(t) = 1 - \sum_{i=1}^{c} \left(\frac{n_i}{n}\right)^2 \qquad \textbf{(Equation 45)}$$

t is the node number at which the Gini index was calculated, n is the number of all data in the node, and c is the number of data types. The closer the Gini index is to 0, the higher the impurity and the less important the features are.

1-4-2. Bootstrap aggregating (Bagging)

Bagging is a method in which the training data used for each classifier is obtained by boosted trap sampling*, and the prediction model (weak classifier) are used for prediction and finally combined in some way (majority vote, average, and output as features to build a new learner) (**Figure 5**).

* Boost trap sampling: a method of randomly extracting some data from a population, allowing for overlap, when there is data to serve as the population.

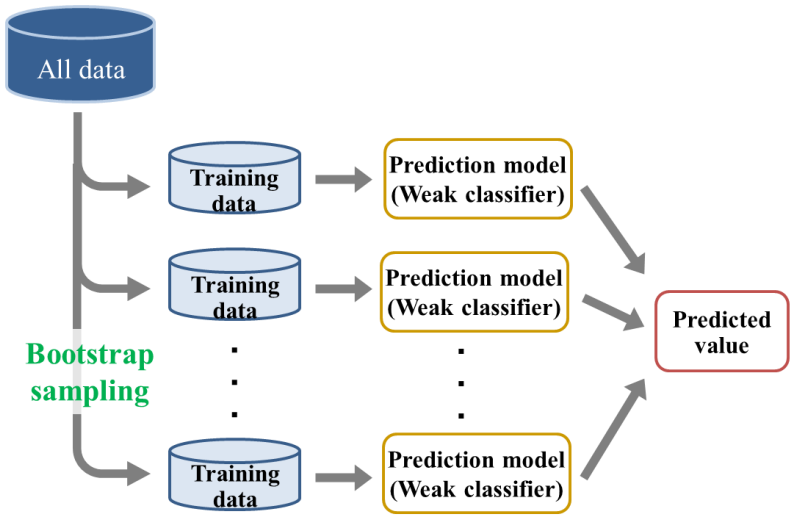


Figure 5. Diagram of Bootstrap aggregating. In bagging, each training (learning) data is obtained by boost strap sampling, a prediction model (weak classifier) is created, and finally the predictions are combined in some way to bring out the predictions.

1-4-3. Random forest (RF)

Random Forest is an ensemble learning algorithm that combines and bagging multiple decision trees to show improved generalization capability.[33, 34] The learning process is as follows.

1. For each of the *N* decision trees, do the following.
   i) Sampling from the training data set (sampling by boost trap sampling)
   ii) Create and train predictive models with decision trees
2. Completing the random forest when all decision trees have been trained

In decision trees, the values named the Gini index is calculated as an indicator of importance of features. It refers to the impurity of the data of a node among the nodes classified by a certain feature. The closer the Gini index is to 0 (1), the lower (higher) the purity of the data, indicating that the feature is unimportant (important). In the RF calculation, the values named the Gini importance is calculated by gathering the Gini indices calculated for all decision trees (**Equation 46**).

$$\text{Gini importance } (\theta) = \sum_{\tau} I_{G_\tau} \qquad \textbf{(Equation 46)}$$

θ indicates the number of features used in the analysis, and τ indicates the number of total nodes.

1-5. Scope of this doctoral thesis

In light of the above, this doctoral thesis aimed to reveal new scientific knowledge about proteins and to establish new analytical methods. Chapter 2 focused on the FMO method and analyzes protein-ligand interactions between a new drug and its receptor. In Chapter 3, protein-protein interaction analysis between canine distemper virus and lymphocyte activating molecules was performed using the FMO method, since the results of Chapter 2 suggested that the

FMO method is also useful for protein-protein interaction analysis. In Chapter 4, based on the findings in the previous chapter, a new analysis tool using the FMO method was constructed and validated. In Chapter 5, I used the analysis tools developed in Chapter 4 to analyze proteins for which detailed molecular mechanisms have not been elucidated.

Chapter 2. Theoretical elucidation of the molecular association model of PPARα and its novel ligand, pemafibrate

Purpose of this Chapter

In this chapter, I focused on the FMO method and performed protein-ligand interaction analysis for novel drugs and their receptors. The usefulness of the FMO method for protein-ligand interaction analysis was verified.

2-1. Introduction

2-1-1. Peroxisome proliferator-activated receptor (PPAR)

In recent years, research on nuclear receptors has remarkably advanced, and many so-called orphan receptors with unidentified ligands have been discovered and their functions are elucidated. Among them, the peroxisome proliferator-activated receptors (PPARs) are ones of the most dramatically studied nuclear receptors. PPARs have been shown to be involved in the regulation of the cellular response to steroid hormones, thyroid hormones, vitamin D3, and retinoic acid. In 1990, PPARs were first screened from a mouse liver cDNA library by Green *et al*. as a clone encoding a protein activated by the antihyperlipidemic drug peroxisome proliferator.[35] Subsequently, PPARs were found to be not only a proliferator of peroxisomes, an intracellular organelle involved in lipolysis, but also a ligand-inducible transcription factor that regulates various key genes involved in peroxisome proliferation. In other words, in addition to lipid metabolism such as neutral fat lowering, carbohydrate metabolism such as insulin sensitivity enhancement, and various physiological activities such as atherosclerosis, immune and inflammatory responses, cell proliferation, and malignant tumor control, the ligands have been applied to the treatment of many diseases.

It is known that PPARs consist of three subtypes, i.e., PPARα, PPARβ/δ, and PPARγ. PPARs form a heterodimer with the retinoid X receptor (RXR) and bind to the peroxisome proliferator responsive element (PPRE), a DNA-binding domain (DBD), to regulate gene expression in the downstream regions. These receptors share a common structural composition consisting of an N-terminal variable domain with ligand-independent activation function, a conserved DBD, and a ligand-binding domain (LBD) with ligand-dependent activation function.[36] Since PPARs are involved in the transcription of genes involved in cell proliferation and differentiation, immune response, and sugar and lipid metabolism, they are attracting attention as therapeutic agents for diabetes and metabolic diseases. Therefore, PPAR agonists are considered important tools for the treatment of diabetes and metabolic

syndrome.

PPARα is mainly found in the liver, heart, kidney, and gastrointestinal tract, which are organs with high utilization of fatty acids, and regulates fatty acid metabolism, especially fatty acid oxidation. When PPARα is activated by agonists, the C-terminal helix H12 (AF-2 interface) is activated, promoting heterodimerization with RXRα and recruitment of nuclear coactivators, which eventually interact with a DNA binding site in PPRE to regulate target gene transcription.[37] Several reports have shown that PPARα activated by various ligands has anti-inflammatory effects. However, PPARα agonists are believed to have not only the anti-inflammatory effects mentioned above but also inflammation-inducing effects, and their mechanism of action is complex. Thus, the detailed analysis on the interaction between PPARα and a specific agonist is important for the therapeutic targeting on PPARα.

2-1-2. Therapeutics targeting PPARs

A number of natural and synthetic PPAR ligands have been identified.[38] Among them, fibrate-type drugs for hyperlipidemia ubiquitously activate PPARα, which regulates lipid flux by modulating fatty acid transport and β-oxidation in the liver, and can lower triglyceride (TG) levels and increase high-density lipoprotein (HDL) cholesterol levels in patients with hyperlipidemia and type 2 diabetes, thereby improving plasma lipid profiles and potentially preventing coronary artery disease and stroke.[39] However, the efficacy of these fibrates is limited by their weak activity against PPARα and dose-dependent side effects.[40] Pemafibrate (**Figure 6**), on the other hand, is a novel, highly active, selective PPARα modulator (SPPARMα) that was found to potently and specifically enhance PPARα activity.[41] This modulator was found to exert beneficial effects on lipid metabolism, reverse cholesterol transport, and inflammation, resulting in anti-obesity effects and, overall, having a transcriptional effect that exceeds that of clinically used fibrates.[42] The modulator also exhibited potent TG-lowering effects in dyslipidemia subjects with high TG and low HDL

cholesterol, without increasing side effects.[43] Previous studies have shown that pemafibrate causes higher PPARα activation than other fibrate drugs, hence the name SPPARMα.[44] Recently, Takei *et al.* compared the effects of pemafibrate with those of classical PPARα agonists and found that pemafibrate activates PPARα transcriptional activity more effectively than classical agonists.[45]

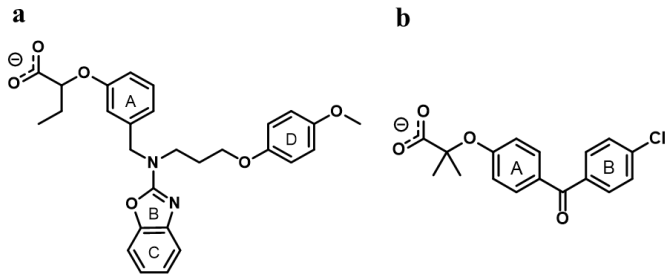


Figure 6. The structures of (**a**) pemafibrate and (**b**) fenofibrate.

2-1-3. PPARα and pemafibrate complex structure

At the time the study was conducted, the structure of the PPARα LBD complex with pemafibrate remained unknown. It was quite difficult to obtain the structure of PPARα experimentally due to a flexible loop in the LBD region which makes instability. However, knowing its complex structure is essential for understanding the structural basis of its mechanism of action and ultimately for designing better ligands with improved binding affinity and selectivity. Therefore, to elucidate the molecular basis for the regulation of PPARα activity by pemafibrate, I obtained the structure of pemafibrate bound PPARα using molecular simulations combined with quantum mechanics/molecular mechanics calculations. Then, using the FMO method[8] based on first principle calculations, I determined a new binding pattern for this modulator in the LBD of PPARα. Subsequently, the binding of the PPARγ coactivator 1α (PGC-1α) to the PPARα-pemafibrate complex was examined in detail.

2-2. Materials and Methods

2-2-1. Structural model construction

I first created the complex structure of the PPARα and coactivator PGC-1α using the Molecular Operating Environment (MOE) program[46]. Initial structure was taken from the X-ray structure of the GW409544 ligand bound to PPARα with coactivator motif. The X-ray structure of the LXXLL peptide derived from steroid receptor coactivator 1, SRC1 (PDB ID: 1K7L),[47] in complex with the PPARγ ligand rosiglitazone PPARγ coactivator 1α, PGC-1α (PDB ID: 3CS8) was constructed.[48] The ligand, which bounded to X-ray structure, was replaced by fenofibrate or pemafibrate using dock utility of MOE. I performed pre-optimization of the structures by using Amber10:EHT force field with solvation energies were calculated with the born model. After constructing complex structure, I was performed molecular dynamics (MD) simulations up to 100 ns to analyze the stability of the modeled structures. MD simulations were done by explicitly specifying water molecules as the solvent. In the all MD simulations, the AMBER ff14SB force field[49] and the TIP3P water model in AMBER 14[50] were used for the protein and the solvent, respectably. Calculations were performed at 300 K and 1 bar pressure using the NPT ensemble.

2-2-2. QM/MM Calculation

The structure of the PPARα-ligand complex was fully optimized at the QM/MM method using the NWChem program[51], with the B3LYP-D functional and 6-31G(d) basis set used for the QM part and the AMBER99 force field for the MM part. The ligand and its surrounding important residues were considered to be placed in the QM region, and other residues in the MM region.

2-2-3. Calculation of FMO

FMO calculations were performed using the PAICS program[52]. The correlated

Resolution-of-Identity second order Møller–Plesset (RI-MP2) level[53] and a correlation-matched double ζ basis set (cc-pVDZ) were used for all FMO calculations.

2-3. Results and Discussion

2-3-1. The complex structure of pemafibrate or fenofibrate bound to PPARα

The ligand binding pocket interface of the constructed model structure of PPARα is shown on the surface. The structure suggests that the ligand binding pocket is Y-shaped and located in the center of the LBD. There were some interesting differences between the structures of the QM/MM-optimized pemafibrate/PPARα complex and the fenofibrate/PPARα complex. Fenofibrate occupies only arm I of the cavity extending from the center of the ligand binding pocket toward the AF-2 helix (H12). It interacts with amino acid residues mainly through the polar head (COO-) of fenofibrate, forming an effective H-binding network. In contrast, pemafibrate, which has a Y-shaped molecular structure, occupied all regions of the ligand binding pocket, including arm II and arm III between helix 3 and β-sheet (**Figure 7**). This indicates that pemafibrate is not only involved in the hydrogen bond network formed between the polar head and SER280 (H4), TYR314 (H6), HIS440 (H11), TYR464 (H12), but also makes important interactions with other residues (**Figure 8a**). In particular, CYS276 and VAL332 interact with the aminobenzoxazole moiety (BC ring, **Figure 6a**) via CH-p. Similarly, THR279 (hydrogen bond), TYR334 (p-p interaction) and MET220 (hydrogen bond) interact with phenoxy-alkyl groups, including ring D, and GLN277 (hydrogen bond) interacts with the COO^- group of pemafibrate. These interactions are largely absent in fenofibrate-bound PPARα (**Figure 9a**). This difference in interaction mode is also related to the size of the LBPs, which were calculated to be 828 $Å^3$ for the pemafibrate-bound form and 1163 $Å^3$ for the fenofibrate-bound form. This indicates that pemafibrate has two pharmacophores, aminobenzoxazole and dimethoxybenzene, which strongly interact with PPARα.

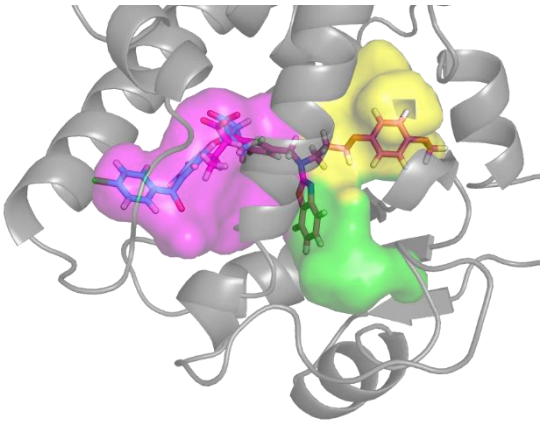


Figure 7. Binding patterns of pemafibrate and fenofibrate to human PPARα. Magenta is pemafibrate and light blue is fenofibrate. The binding pocket has three pharmacophore regions: red, yellow, and green.

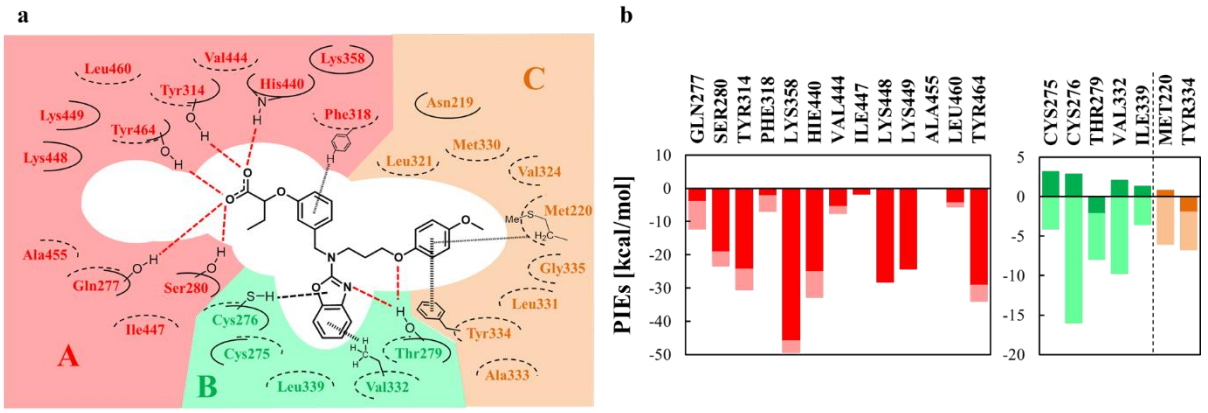


Figure 8. (**a**) Interaction between residues in the LBP of PPARα and pemafibrate. Hydrogen bonding in red dashed line, SH-p interaction in black dashed line, and p-p interaction in black dotted line. (**b**) PIE between amino acids and pemafibrate. Bars in red, green, cyan represents the interactions in region A, B, C, respectively. Dark colors denote electrostatic interaction and mild colors denote dispersion interactions.

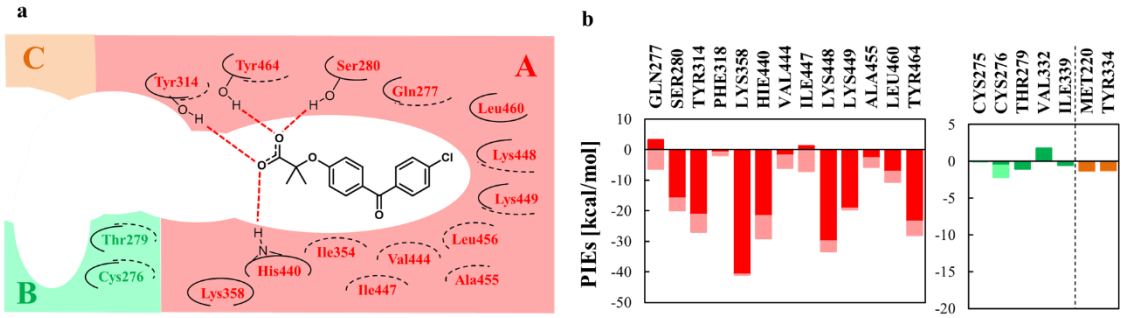


Figure 9. (**a**) Interaction between residues in the LBP of PPARα and fenofibrate. Hydrogen bonding in red dashed line. (**b**) PIE between amino acids and fenofibrate. Bars in red, green, cyan represents the interactions in region A, B, C, respectively. Dark colors denote electrostatic interaction and mild colors denote dispersion interactions.

2-3-2. Interaction between pemafibrate and PPARα

In both pemafibrate and fenofibrate bounded complexes, there were four strong hydrogen bonds formed between the COO^- group and SER280, TYR314, HIS440, and TYR464 by hydrogen bond network that anchors the ligand position within the LBP (**Figure 8a**, **Figure 9a**). Furthermore, this network should provide a stable tertiary structure for PPARα. This is because each of these four amino acids is derived from a different helix. SER280 is derived from helix H4, TYR314 from H6, HIS440 from H11, and TYR464 from terminal H12. This is consistent with previous experimental findings that ligands with COO^- groups co-crystallize with PPARs.[54] The fragment interaction obtained for these hydrogen bonds between the COO^- groups of the PPARα/pemafibrate complex and SER280, TYR314, HIS440, TYR464. From the FMO calculations, these hydrogen bond interactions appear to play a dominant role in ligand binding in the PPARα cavity. The FMO calculations indicate that these hydrogen bonding interactions appear to play a dominant role in ligand binding in the PPARα cavity. Other adjacent important amino acid variants were also estimated, indicating the importance of ILE272. a closer look at the FMO results shows that the COO^-

group of pemafibrate forms a large hydrogen bond with GLN277 (-12.3 kcal mol^{-1}), which is a PPARα/fenofibrate complex This is rarely seen in the PPARα/phenofibrate complex. In other words, the effect of SER280 in the latter may be shared by the combined effect of SER280 and GLN277 in the former. The calculated PIEs also confirm several other significant interactions (**Figure 8b**), with CYS276 (H4) interacting CH-π with amino benzoxazole (-13.1 kcal mol^{-1}) and VAL332 interacting CH-π with the same oxazole moiety (-7.6 kcal mol^{-1}). TYR334 on the loop near the entrance of the LBP mainly interacts π-π with the D ring (-6.7 kcal mol^{-1}). THR279 forms hydrogen bonds (-7.9 kcal mol^{-1}) with the methoxy phenoxy and amino benzoxazole moieties.

2-3-3. Interaction between fenofibrate and PPARα

The interaction pattern with the COO^- group is qualitatively similar to that of pemafibrate-bound PPARα, except for the involvement of GLN277 (**Figure 9a**). The calculated PIEs for hydrogen bonds in the H-bond network are -20.0 kcal mol^{-1} (SER280), -27.0 kcal mol^{-1} (TYR314), -29.1 kcal mol^{-1} (HIS440) and -28.1 kcal mol^{-1} (TYR464). The importance of these interactions was also confirmed by biochemical experiments. The terminal B ring (**Figure 6**) is involved in an important dispersal interaction (**Figure 9b**). It mediates the CH-π interaction between LEU456 and ILE447. The A ring of fenofibrate interacts primarily with ILE354 and partially with VAL444, which also interacts with the B ring; GLN277 interacts with the benzene ring of the A ring; and GLN277 interacts with the benzene ring of the B ring.

2-3-4. PGC-1α and PPARα/pemafibrate interaction

The concept of SPPARM is that the binding of such ligands leads to different conformational changes and different patterns of cofactor mobilization, promoting specific biological responses. In mutation experiments conducted by our group, the V306A mutant,

when co-transfected with PGC-1α, blunted only pemafibrate-induced PPARα activation, while fenofibrate increased transcriptional activity to the same extent as WT of PPARα.[55] This mutation experiment indicates that the effect of pemafibrate on PPARα transcriptional activation may be dependent on V306 affecting PPARα binding to PGC-1α. To understand this difference, I thoroughly analyzed the interaction between Val306 and its neighbors: coactivators with LXXLL motifs, such as PGC-1α, are recruited to nuclear receptors, such as PPARs, upon ligand binding to activate transcription of their cognate target genes. Val 306 interacts directly with several amino acids of PGC-1α via a dispersion-dominated van der Waals interaction (**Figure 10a**). The dispersion energies from FMO calculations indicate that Val306 binds PPARα and pemafibrate more strongly than fenofibrate, 4 kcal mol^{-1} strongly interacts with PGC-1α (**Figure 10b**). This may explain the experimental results that mutation of Val306 results in less activity of pemafibrate-bound PPARα than fenofibrate-bound PPARα.

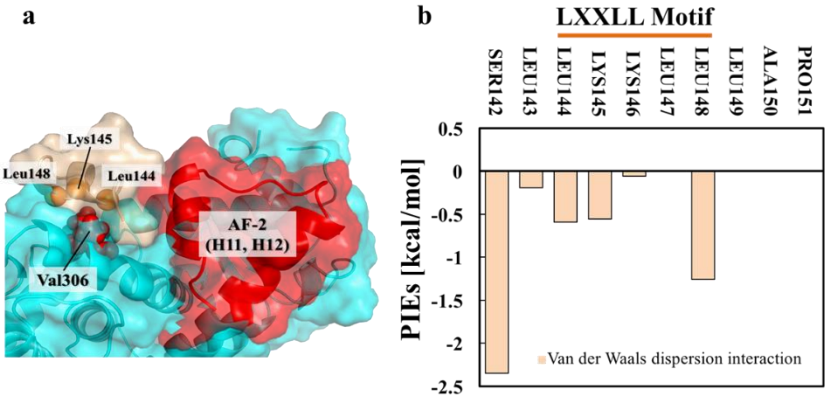


Figure 10. (**a**) Complex structure of AF-2 interfaces of PPARα and LXXLL peptide motif of a cofactor. Interaction mode of VAL306 with the LXXLL motif which including the interaction with the AF-2 interface. (**b**) van der Waals dispersion interaction energy between V306 in PPARα and amino acid residues of LXXLL motif. This figure shows that difference of the calculated FMO interactions of VAL306 with PGC-1α when fenofibrate/pemafibrate is bound in the LBD of PPARα (pemafibrate - fenofibrate).

2-4. Summary

In many cases, ligand binding changes the conformation of the protein so that it can successfully bind to the ligand. This is commonly referred to by the concept of induced fit. In fact, ligation of PPARα is caused by the stabilization of the entire structure of the LBD by the ligand, resulting in a more compact and rigid conformation. As a result, the LBD becomes more compact and rigid, the AF-2 interface is stabilized, and coactivators are mobilized through it.[56] Thus, enhanced induced conformability could increase the transcriptional activity of PPARs. Structural analysis reveals that pemafibrate ligation is triggered by induced fitting. Since the ligand covers the largest region of the cavity and interacts with residues in all regions of the cavity, it can efficiently activate induced fitting. This can be assessed from the change in the size of the LBP. The cavity sizes of pemafibrate- and fenofibrate-bound PPARα are estimated to be 828 $Å^3$ and 1163 $Å^3$, respectively; it is recalled that PPARα has a cavity of about 1300 $Å^3$.[56] This indicates that binding to pemafibrate promotes induced binding by causing a conformational change in the cavity region that allows the ligand to bind more easily.

FMO calculations show that the hydrogen bonding network between SER280, TYR314, HIS440, and TYR464 and the ligand is much stronger for pemafibrate-bound PPARα than for phenofibrate-bound PPARα. Interactions with the amino benzoxazole and dimethoxybenzene moieties of pemafibrate are completely absent in the fenofibrate-bound PPARα. The prominent interactions with these two pharmacophores are CYS276 and VAL332 for amino benzoxazole and MET220, THR279, and TYR334 for dimethoxybenzene. Also worth mentioning is the p-p interaction of about -7.0 kcal mol between the D ring of pemafibrate and TYR334 on the loop located at the entrance of the LBP. This particular interaction may stabilize the flexible loop region. The sum of all PIEs involving the ligand can reflect the overall binding of the ligand to the protein. I found that pemafibrate interacts more strongly with PPARα than fenofibrate by -69.5 kcal mol, suggesting that pemafibrate may be a more

potent ligand for PPARα than fenofibrate. These results indicate that the enhanced activity of PPARα due to ligand binding of this novel pemafibrate is mainly due to its ability to interact with the largest number of residues from throughout the cavity region. This Y-shaped ligand matches well with the Y-shaped LBP of PPARα. This lock-and-key nature makes pemafibrate a novel and potent modulator, Hennuyer *et al.* concluded from GST pull-down experiments that pemafibrate appears to promote the recruitment of coactivators, including PGC-1α, more efficiently than fenofibrate I conclude.[42] Our *in silico* studies show that overall LBP is more stabilized by pemafibrate than by fenofibrate through ligand binding. This major conformational change in LBP would then be reflected in the stabilized AF-2 interface, and our analysis of FMO calculations and in vitro experiments indicate that PPARα bound to pemafibrate interacts efficiently with its cofactor, PGC-1α. This confirms the assumption made in previous experiments that VAL306 plays an important role in enhancing the interaction of AF-2 with the PGC-1α cofactor. In other words, pemafibrate has an advantage over the smaller fenofibrate in binding of PPARα to the LBD by three specifically tailored pharmacophores. The molecular basis for the increased activity of pemafibrate-bound PPARα, investigated in an *in silico* approach based on *ab initio* calculations, reveals a clearly novel pattern of binding mode. This novel SPPARMα modulator interacts with all regions of LBP. All three pharmacophores, polar COO^- head, amino benzoxazole, and methoxy phenoxy, play equally important roles. Compared to fenofibrate, pemafibrate was shown to stabilize LBP more efficiently and to stimulate a stronger interaction between the AF-2 interface and the PGC-1α coactivator; VAL306 plays a pivotal role in enhancing this interaction, and the interaction between the AF-2 interface and the PGC-1α coactivator was also enhanced by pemafibrate. This result was also confirmed by in vitro experiments.

From this study the FMO method proved to be useful for protein-ligand interaction analysis. Because of the nature of the calculation method, the FMO method divides the protein into amino acids and performs the calculation, not only protein-ligand interactions but

also protein-protein Therefore, it is possible to analyze not only protein-ligand interactions, but also protein-protein interactions. Therefore, I decided to conduct protein-protein interaction analysis in the next chapter.

Chapter 3. Computational study of interspecies transmission of CDV/SLAM protein-protein interactions

Purpose of this Chapter

The FMO method has proven to be useful for the analysis of protein-ligand interactions, since the FMO method, due to the nature of its calculation method, divides the protein into amino acid units and performs calculations for each amino acid. Therefore, it is possible to analyze protein-protein interactions as well as protein-ligand interactions. Therefore, in this chapter, I conducted protein-protein interaction analysis and verified the usefulness of the FMO method for protein-protein interaction analysis.

3-1. Introduction

3-1-1. Morbillivirus (MoV)

Morbillivirus (MoV) belong to the *Paramyxoviridae* family and infect animals systemically, causing high mortality and morbidity.[57] During the infection cycle, the viral hemagglutinin (H) protein interacts with signaling lymphocyte-activated molucur (SLAM) and poliovirus receptor-like 4 (nectin-4) expressed on host immune cells and epithelial cells, respectively.[57] sequence is highly conserved among species, but the amino acid sequence of SLAM is not, suggesting that the interaction between H-protein and SLAM defines morbillivirus host selectivity. Currently, seven species of viruses in the morbillivirus genus have been isolated, including measles virus (MV), which infects humans.[57] Because canine morbilliviruses (canine distemper virus, CDV) cause severe infections in carnivores, and CDV in particular causes fatal outbreaks in non-human primates, there is interest in understanding the interspecies transmission of morbilliviruses. Accumulating evidence indicates that CDV infects animals of the genus Macaca, but not humans.[58-60] It has also been shown that macaca SLAM, but not human SLAM, functions as a receptor for CDV.[60, 61] Furthermore, a slight variation in the CDV H-protein allows this protein to interact with other primate SLAMs of the genus Saguinus (e.g., cottontail tamarin) and Homo (human).[62-64] The molecular mechanism of interspecies transmission of CDV in primates may be elucidated by analyzing differences in SLAMs of these species, and structural data will play an important role in elucidating this mechanism.

3-1-2. Importance of the N-terminal region in the MoV and SLAM complex structure

Crystal structures of morbillivirus H proteins have been attempted by many research groups,[65-68] and several structures of complexes with receptor proteins have been reported.[66, 68] Hashiguchi *et al*. reported the crystal structure of MV H protein (MV-H) in a complex with the known receptor for MV, cottonwood tamarin (SLAM). This structure and the analysis of

the cottonwood vitamarin SLAM mutant indicated that residues such as N72, V74, E75, and K77 in the CC'-loop region of cottonwood vitamarin SLAM form an interaction with MV-H.[66] Binding to SLAM changes the oligomeric state of MV-H, and this change in the oligomeric state of MV-H MV-H oligomeric state upon binding to SLAM, suggesting that this change triggers MV fusion.[69, 70] On the other hand, the role of the N-terminal region of SLAM in viral infection was unclear. The functional importance of this region has been demonstrated by combined analyses using viral infection assays and SLAM mutants. Seki *et al.* reported that the M29S mutant of human SLAM does not interact with MV-H, suggesting that structural analysis of the N-terminal region of SLAM is necessary to fully elucidate the molecular mechanism of H-protein-SLAM complex formation.[71]

Thus, the flexible N-terminal region of SLAM may be important for facilitating morbillivirus infection. However, since there are no structural data on the N-terminal region of CDV-H and SLAM, an approach other than X-ray crystallography is needed to elucidate the function of the N-terminal region of SLAM in CDV infection. Therefore, I constructed a complex model of CDV-H and the N-terminal region of SLAM and analyzed the interaction energy between CDV-H and the N-terminal region of SLAM. Using the constructed model, I analyzed the interaction energy between SLAM and CDV-H by computational chemistry method. As a result, it was inferred how the residues in the N-terminal region of SLAM affect the interaction between CDV-H and SLAM at the molecular level.

3-2. Materials and Methods

3-2-1. Homology modeling

The structure of the complex was constructed using the MOE program[46]. The crystal structure of the MV-H-SLAM complex (PDB ID: 3ALW) was used as the first template.[66] This structure consists of an MV-H head (amino acids 184-607) and a SLAM-V (amino acids 30-140) domain, and these two domains are connected by a flexible linker

(Gly-Gly-Gly-Ser)3 of 12 residues. Thus, the structure lacked the following two elements. (1) the N-terminus of SLAM (including the critical 28 residues) and (2) part of the N-terminus of MV-H. The incomplete N-terminal regions of MV-H and SLAM were modeled using MOE's Loop Modeler utility to obtain a complete MV-H-SLAM complex structure. For this purpose, I used the sequence of the MV IC-B strain (GenBank accession number NC_001498) for the MV-H protein part, part of the N-terminal region of the published MV-H structure (PDB ID: 2ZB6[65]), and for the SLAM part I used macaca slam (GenBank accession number XM_ 001117605) and the first MV-H- SLAM template (PDB ID: 3ALW[66]) for the SLAM part. This complete model, compMV-macaca SLAM, was used as a template for structural modeling of the CDV-H-macSLAM complex: the CDV protein portion of the CDV Ac96I strain sequence (GenBank accession number AB753775) and the newly constructed compMV-macSLAM structure was used as a template to model the structure of the CDV-H-macSLAM complex. The missing hydrogen atoms were added with the Pro-tonate 3D utility of MOE using the AMBER10:EHT force field, and the solvation energy was determined with the Born model. The resulting structures were fully optimized using the AMBER10:EHT force field; the structures of the H28R and M29S SLAM mutants complexed with CDV-H were modeled using the constructed WT CDV-H-SLAM structure and MOE's Protein Builder utility. All structures were visualized by PyMOL[72].

3-2-2. Molecular dynamics simulation

Initial setup for the MD simulations was performed using AMBER14[50] and ff14SB force fields[49]. The constructed complex structures were solvated using the TIP3P water model in a 110 × 90 × 90 Å3 cubic box. Neutralizing counterions were added to each system. Topology files created in AMBER were converted to GROMACS format using the acpype.py script.[73] All MD simulations were performed using the GROMACS package.[74] Bonds with H atoms in the constructed structures were treated as rigid bodies using the LINCS algorithm.[75]

In order to equilibrate the whole system, I was done 800 ps NPT simulations by using the Nose-Hoover thermostat at 300 K with keeping heavy atoms constrained.[76, 77] After equilibration, a 100 ns NPT simulation was performed with the Parrinello–Rahman method at 1 bar and 300 K.[78, 79] As nonlocal interactions, electrostatic interactions were calculated using the particle mesh Ewald method with a real space cutoff of 10 Å.

3-2-3. Coupling free energy calculation

Using the Molecular Mechanics Generalized Born Surface Area (MM-GB/SA) method[80] implemented in AMBER14, binding free energies were calculated for all simulated systems included in the MD calculations; a total of 100 conformations were extracted from the last 20 ns of the MD simulations. MM-GBSA calculations were performed after removal of water molecules and counterions. The enthalpy term (H) was calculated using the modified GB model developed by A. Onufriev *et al*.[81] The concentration of 1-1 mobile counterions in solution was set to 0.15 M.

3-2-4. Calculation of RMSD and RMSF

The root mean square deviation (RMSD) and root mean square fluctuation (RMSF) were calculated with the AMBER14 cpptraj analysis tool.[82] The structures were sampled at 10 ps intervals. Before each calculation, external translational and rotational motions were removed by minimizing the RMSD distance of the Cα atom relative to the equivalent atom in the first frame of the orbitals The RMSD and RMSF values were calculated for the Cα atom.

3-2-5. Fragment molecular orbital (FMO) calculations

The FMO calculations were performed using the PAICS program.[52] Correlation Resolution-of-Identity 2-order Moller Plesset (RI-MP2) level of theory and correlation matching double-zeta basis set cc-pVDZ were used for the calculations. Fragment assignment

and PAICS input generation were performed using PaicsView.[83] Output was analyzed using RbAnalysisFMO[84].

Briefly, the FMO method, a quantum mechanical method based on first principles calculations, is a powerful theoretical tool for the reliable study of protein-ligand interactions. The ligand is also considered a fragment, and the properties of the fragments are combined to derive the properties of the entire system in a many-body expansion. By considering two body types (fragment pairs) in this way, it is possible to calculate the fragment interaction energy (IFIE), which is an important physical quantity in understanding protein-ligand binding. In this study, I used FMO to study protein-protein interactions; FMO is now being used as a valuable tool to describe protein-ligand interactions.[55]

3-3. Results and Discussion

3-3-1. Comparison of protein sequences of human SLAM and macaca SLAM

Because CDV-H has been shown to interact with macaque SLAMs but not with human SLAMs, I initially wanted to identify the residues that trigger cross-species transmission of CDV among primates.[60, 61, 63] This suggested that differences in the amino acid sequences of the two SLAMs were responsible for their differential affinity for CDV-H. The sequence alignment of human and macaca SLAMs is shown in **Figure 11**, suggesting that they differ from each other by only 11 residues. In general, SLAM can be divided into five domains: signal peptide (shown in pink in **Figure 11**), V domain (blue), C2 domain (yellow), transmembrane domain (green), and cytoplasmic domain (purple). structural data from the complex of MV-H and cottonwood marine SLAM indicate that the V domain only contributes to the interaction with CDV-H. Only residues 28 and 48 of the V-domain differ in amino acid type between macaque and human SLAMs (**Figure 11**). Residue 48 is located distal to the interaction interface between MV-H and SLAM (**Figure 12b**), suggesting that this residue is not important for the interaction. In addition, the structure of the N-terminal region of SLAM

(red square in **Figure 11**) by Hashiguchi *et al*. lacks residue 28, so the potential role of residue 28 in complex formation cannot be determined.[66] Furthermore, modeling of the structure of the N-terminal region of human SLAM complexed with MV-H suggests that residue 28 plays only a minor role in the interaction with MV-H.[71] Therefore, an approach that predicts the structure in the N-terminal region of SLAM is needed to determine the role of residue 28 in the interaction with CDV-H.

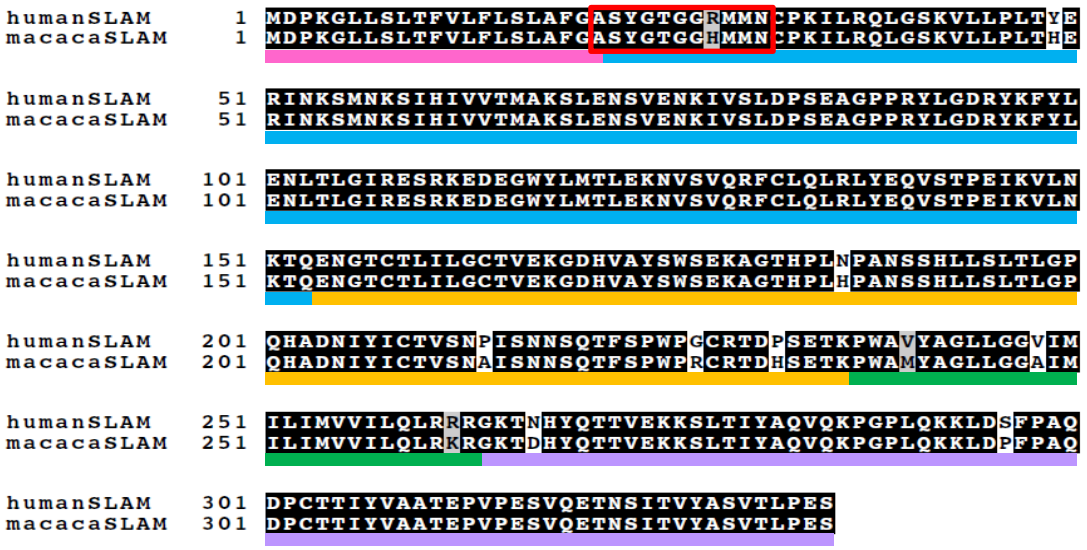


Figure 11. Sequence alignment of human SLAM and macaca SLAM. Sequences were obtained from GenBank database, accession numbers for human SLAM and macaca SLAM were NP_003028 and XM_001117605, respectively. The alignment was done by ClustalW. There are five domains in SLAM: signal peptide (pink), V domain (blue), C2 domain (yellow), transmembrane domain (green) and cytoplasmic domain (purple).

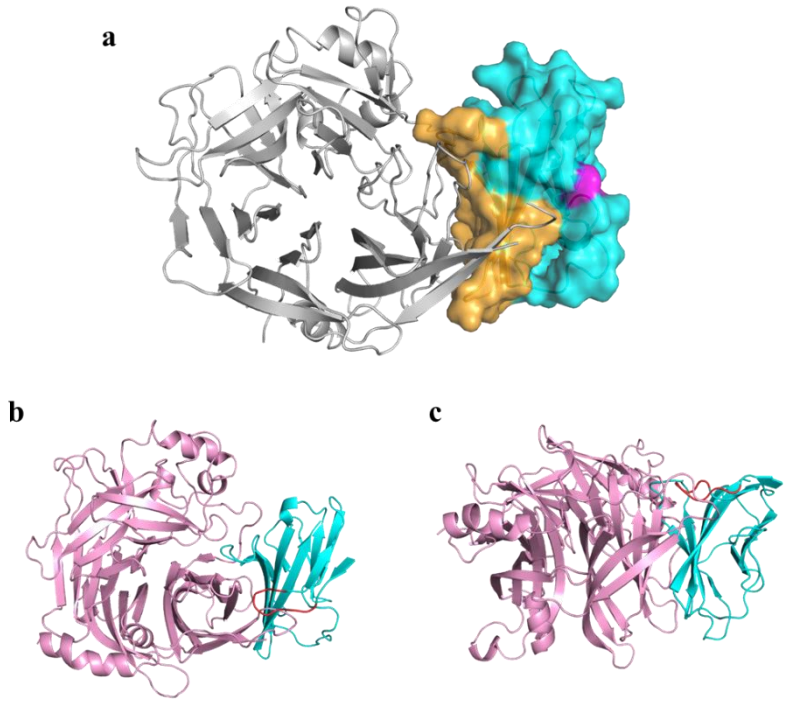


Figure 12. (**a**) Complex structure of MV-H (gray) and macaca SLAM (surface representation). The interaction surface bound to MV-H is shown in orange and others are in cyan. Residue of position 48 (magenta) is not located at the interaction surface (orange). Constructed complex structure of CDV-H (pink) and macaca SLAM (cyan). (**b**) front view and (**c**) side view.

3-3-2. Interaction energy analysis between CDV-H and SLAM by fragment molecular orbital (FMO) analysis

The potential functional roles of residues in the N-terminal region of SLAM were determined by using a computational chemistry approach to construct a homology model of macaca SLAM complexed with CDV-H. The complex model structure of CDV-H and macaca SLAM was constructed using Molecular Operating Environment (MOE) software, using the crystal structure of the complex of MV-H and cottonwood marine SLAM as a template; the sequence identities of MV-H and CDV-H and cottonwood marine SLAM and macaca slam are 35% and 83%, respectively, and modeling revealed suggested that a highly accurate model structure of CDV-H and macaca slam could be obtained. For the modeled CDV-H, I

confirmed that the structure did not unfold after MD simulation. Next, I performed complementation of the residues in the N-terminal region using the structural geometry of this region and the scoring calculated by MOE. The structure with the highest score was selected from the generated model. I also confirmed that important interactions are conserved not only in the crystal structure but also in the modeled structure.

At the interaction interface between the N-terminal region of macaca SLAM (red in **Figure 13a**) and CDV-H (pink in **Figure 13a**), two residues of macaca SLAM (His28 and Met29) were shown to interact with three residues of CDV-H (Tyr186, Arg543, Thr544, and Phe600). This observation was supported by quantitative FMO analysis of interaction energies: the energies of His28 and Met29 were -36.9 and -32.2 kcal/mol, respectively, and these values were more than 15 kcal/mol lower than those of the other residues (**Figure 13b**). The calculated fragment interaction energies (PIE) of the residues Tyr186, Arg543, and Phe600 of CDV-H formed interactions with macaca SLAM (**Figure 13c**). The IFIEs of Tyr186, Arg543, and Phe600 were -5.6, -19.8, and -2.7 kcal/mol, respectively.

In summary, His28 is a major contributor to the formation of the CDV-H-macacaca SLAM complex. In particular, the interaction between His28 and three CDV-H residues (Tyr186, Arg543, and Phe600) is a major contributor to the formation of this complex.

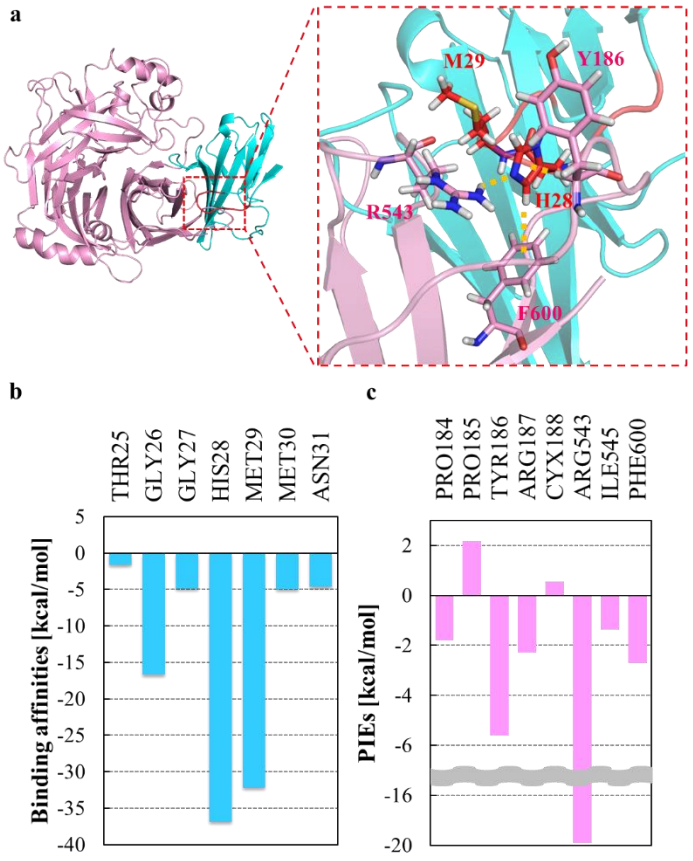


Figure 13. (**a**) Residues in CDV-H (pink) interacted with N-terminal domain (red) in macaca SLAM (cyan). (**b**) Binding affinities of the residues of N-terminal domain in macaca SLAM by FMO analysis. (**c**) pair interaction energies (PIEs) of the interacting residues of CDV-H by FMO analysis.

3-3-3. Molecular dynamics simulation of the complex of CDV-H and macaca SLAM

The above static structural analysis indicated that His28 and Met29 of macaca SLAM play an important role in the stable complex formation with CDV-H. To verify this observation, the dynamics of the CDV-H-macacaca SLAM complex was investigated by molecular dynamics (MD) simulations of the CDV-H complex with wild-type (WT), macaca SLAM H28R mutant, and M29S mutant. Human SLAM has arginine at position 28.

First, root mean square deviation (RMSD) values of the Cα atoms were calculated, and it was found that the structure equilibrated at an RMSD value of approximately 2.2 Å during the 100 ns simulation (**Figure 14a**). Next, the structural changes at the N-terminus (residues 25-31) of macaca SLAM during the simulation were analyzed by calculating the root mean square fluctuation (RMSF) values of the Cα atoms (**Figure 14b**). The analysis revealed that only the H28R mutant of macaca SLAM exhibited increased flexibility in this region: the RMSF values for residues 27-29 of the H28R mutant were >2.0 Å (**Figure 14b**, red), while those for WT (black) and M29S (blue) were <1.3 Å (**Figure 14b**). Analysis of the orbital structure also supported this observation. As shown in **Figure 14c**, the flexibility of the N-terminal region of maca SLAM H28R was clearly higher than that of the WT and M29S

mutants (**Figure 14c**). MMGBSA analysis showed that the H28R mutation, compared to the WT and M29S mutations, interacted between maca SLAM and CDV-H energy was found to be reduced by ~20 kcal/mol (

Table 1).

The above quantitative analysis of the static and dynamic structures of CDV-H and macaca SLAM indicates that the interaction formed between His28 of macaca SLAM and the residues of CDV-H is essential for the formation of a stable macaca SLAM-CDV-H complex. Specifically, the formation of an interaction between side chain H28 and each of Y186, R543, and Y600 is thought to be responsible for the high stability of macaca SLAM-CDV-H (**Figure 13c**); mutation of H28R abolishes this interaction, resulting in a highly flexible N-terminal region. Here, I performed a computational analysis of the H28R mutant to show the difference between human SLAM and macaca SLAM, and a similar phenomenon may be observed in the H28K mutant.

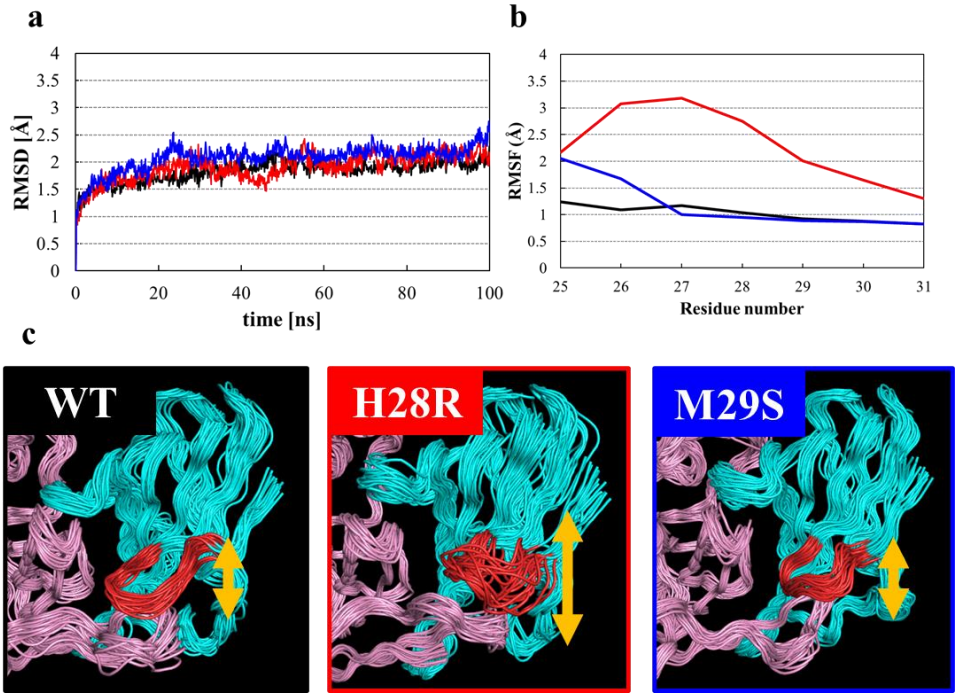


Figure 14. Results of MD simulations about the CDV-H/macaca SLAM complex. (**a**) RMSD values of Cα atoms during the 100 ns MD simulations. This figure shows that the RMSD

values of WT, H28R and M29S colored black, red, and blue line, respectively. (**b**) RMSFs of the residues in N-terminal region (from 25 to 31) in macaca SLAM. The displayed color is the same in (a). (**c**) Trajectory structures obtained from MD simulation of the CDV-H/macaca SLAM complex.

Table 1. Binding energies between macaca SLAM and CDV-H calculated by MMGBSA analysis.

	ΔG (kcal/mol)
WT	−31.8
H28R	−7.2
M29S	−30.9

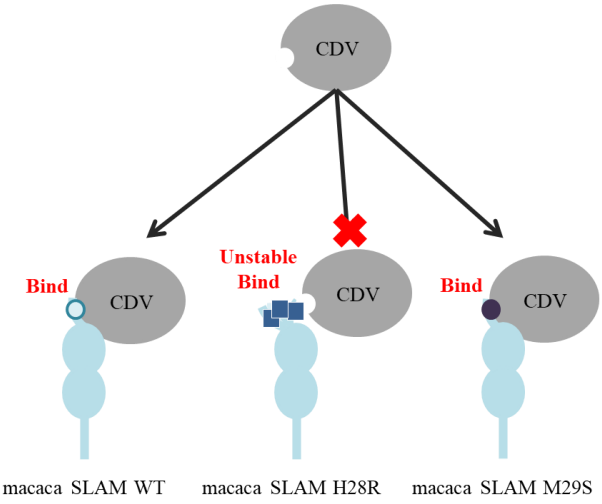
3-4. Summary

The molecular basis for the differentiation of receptor function between macaca SLAM and human SLAM in response to CDV infection has not been studied in detail. In this report, I propose a detailed and specific model for the molecular mechanism by which the N-terminal region of macaca SLAM forms a stable interaction with CDV-H. The results obtained in this study are summarized in **Figure 15**, which shows that only the H28R mutation destabilizes the interaction. I have previously demonstrated the effectiveness of the simulation approach in analyzing the interaction between human SLAM and MV-H.[71] The combination of calculations and experiments revealed that Met29 in the N-terminal region of human SLAM is essential for the formation of the interaction with MV-H. Since the protein regions that form the interaction are very flexible, this computational approach is expected to facilitate the analysis of intermolecular interactions that cannot be revealed by crystallographic analysis.

This chapter proves that the FMO method is also useful for protein-protein interaction

analysis. However, when intra-protein or protein-protein interaction analysis is performed, the number of interaction pairs to be analyzed is huge (about [number of amino acid residues in protein]^2 / 2), which makes the analysis difficult. Therefore, it is important to develop an analysis tool that facilitates the analysis of interactions within or between proteins, and in the next chapter, I construct a new analysis tool using the FMO method and validate the method.

Figure 15. This figure showing that the key role of important residues in N-terminal region of macaca SLAM play in interaction with CDV-H.



Chapter 4. Development of Random Forest-Fragment Molecular Orbital (RF-FMO) Method for Dynamic Protein Interaction Analysis and Application to Src Tyrosine Kinase

Purpose of this Chapter

In this chapter, I establish a valuable tool (RF-FMO) for extracting important residues and interactions between amino acid residues by combining the Random Forest (RF) method which is one of the machine learning algorithms, and PIE analysis based on the fragment molecular orbital (FMO) calculation, based on the knowledge obtained in the previous chapter. The RF-FMO will be benchmarked against Src-Kinase, for which the functional mechanism has been elucidated at the molecular level through a wide range of previous studies.

4-1. Introduction

4-1-1. Src-tyrosine kinase

Protein tyrosine phosphorylation is found primarily in multi-cellular organisms. Differentiation, development and metabolism were controlled by protein tyrosine phosphorylation. And these regulates need tight regulation of inter-cellular signaling. The first to be discovered were non-receptor tyrosine kinases (Src) in the various tyrosine kinases that phosphorylate proteins tyrosine. Currently, a family of tyrosine kinases have been reported in the paper. It named Src family tyrosine kinases (SFKs), and has nine members (Src, Fyn, Yes, Lck, Lyn, Hck, Fgr, Blk). The intra-cellular domains which including GPCRs, cytokine receptors, growth factors and integrins were interacted with SFKs and SFKs have common structural domain.[85]

Src tyrosine kinase is well known as an important regulator of signaling specific to cell proliferation. This protein exists exclusively in the plasma membrane. It has important role to activate protein synthesis systems and cell proliferation. That activation was done by transducing signaling molecules from various protein receptors.[86, 87] There are strongly correlation between tumor growth and steady activation of transcription factors by Src tyrosine kinase.[88] Thus, molecules which control the switching between inactive and active states of Src tyrosine kinase are regard to be candidates for anti-tumor drugs. Anti-tumor effects are exerted by indirect and direct inhibition of the phosphorylation of Src tyrosine kinase. In recent years, Drugs such as Iressa which inhibit its activity are clinically used to some malignancies. It also known about the inhibitors of Src tyrosine kinase can prevent acute inflammatory reactions such as lung injury. From the above information, to understand the conformational changes of Src tyrosine kinase at the atomic level is important. It may play an important role in clinical applications and inhibitor design to know the molecular basis of it. If you want to know the general issues related to Src tyrosine kinase, it may help you to see some reviews and references paper.[89-91]

4-1-2. Structure of Src-tyrosine kinase

Src tyrosine kinase has five domains. These are named a kinase domain (green), a flexible binding region (Linker, yellow), an SH2 domain (light orange), SH3 domain (orange) and an anchor segment (anchor), in order from the C-terminal end of the protein (**Figure 16a**). When Src tyrosine kinase regulates its function, it makes a series of conformational changes which bend large joints. In the active state, the protein takes on an extended structure. On the other hand, the protein has a small folded structure in inactive state (**Figure 16a**).[92] The conformational changes in proteins that change the two states described above are known to be induced by conformational changes in the A-loop in the kinase domain. A-loop is residues 404-424 of Src tyrosine kinase. Therefore, A-loop is particularly important for inducing this large conformational change (**Figure 16b, c**).[93]

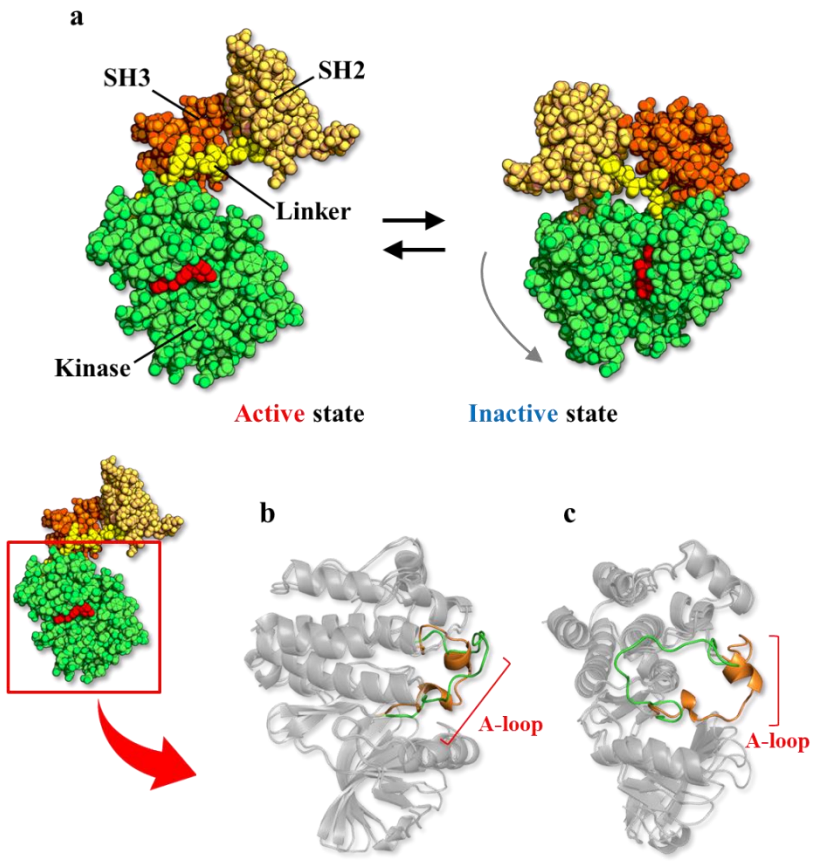


Figure 16. (**a**) Active state structure (left) and inactive state structure (right) of Src-Kinase. Src-Kinase has four domains: SH2 domain (light orange), and SH3 domain (orange), Linker domain (yellow) and Kinase domain (green). When Src-Kinase express its function, it makes conformational change. SH2 domain and Kinase domain are open in active state structure. When the protein is inactive state structure, it is folded into a compact shape. The structures (**b**) front and (**c**) side view which active state structure and inactive state structure superimposed on each other. A-loop undergoes an interesting conformational change between two states. It forms α-helix in the inactive state (orange) and extended in the active state (green).

4-1-3. Efforts for Protein Structure Change Analysis

Many biological phenomena are carried out by proteins. In general, proteins perform important biological functions by changing their structure, either globally or locally. Therefore, to understand the details of life phenomena, it is crucial to understand these

structural changes of proteins at the molecular level. However, from experimental, theoretical, and computational perspectives, understanding protein conformational changes at the molecular level is a challenging task. Researchers are currently addressing this challenge by developing several approaches. From a theoretical and computational perspective, these include molecular dynamics (MD) simulations and quantum chemistry (QC) calculations such as the fragment molecular orbital (FMO) methods[8] were used for it. Each of these calculation methods has its advantages and disadvantages. The amount of data obtained from MD and FMO calculations is large, and at first glance it may seem that there is a lot of information to be gained from the analysis. However, it is difficult to obtain physicochemical knowledge about protein function from such a large amount of computational data.

In current years, the methods which analyze huge amount of data have been used to analyze detail of torsion angles of amino acid residues in important functional proteins. Machine learning method which is one of the data science methods was used to analyze the dihedral angles of amino acid residues obtained MD trajectory analysis in Sultan *et al.*.[1] In their study, they first performed a basic experiment using alanine dipeptides, using the φ-ψ angles of peptides obtained by MD simulations as training data to train a random forest algorithm. The learning results were used to analyze MD trajectory data, and it was found that the conformational states of the peptides were neatly clustered. Furthermore, they performed the same analysis using the φ, ψ, and χ angles of the Src tyrosine kinase as training data. The Gini importance obtained from the random forest analysis allowed identify the amino acid residues and domains that are important for the conformational change of the protein. These studies demonstrate the effectiveness of an analysis tool that combines the random forest method with computational chemistry.

However, conformational changes in proteins are mainly caused by interaction changes between amino acid residues. In addition, interactions between amino acid residues are often formed between atoms outside the main chain. Therefore, I think that the analysis based on

the interaction network between amino acid residues can provide a more detailed understanding of the mechanism of protein conformational change than the analysis based on the dihedral angle of the main chain. In this study, we developed a powerful analysis tool called Random Forest-FMO (RF-FMO). After the development of RF-FMO, the Src tyrosine kinase was used as a target protein for the analysis.

4-2. Random forest-fragment molecular orbital (RF-FMO) method

4-2-1. Disadvantages of clustering analysis of trajectory data using random forests developed by Sultan

Sultan *et al.*[1] reported a trajectory data analysis method using random forests, but two shortcomings were predicted for this tool. The first is the number of estimator and max of depth settings used when training with random forests. The number of estimator and max of depth settings used when training in a random forest, where number of estimators is the number of decision trees to be created and max of depth is the number of node layers to be created for a single decision tree. The second is whether the analysis can be performed on proteins with more amino acid residues than those used in the benchmark calculations of Sultan *et al*.[1]

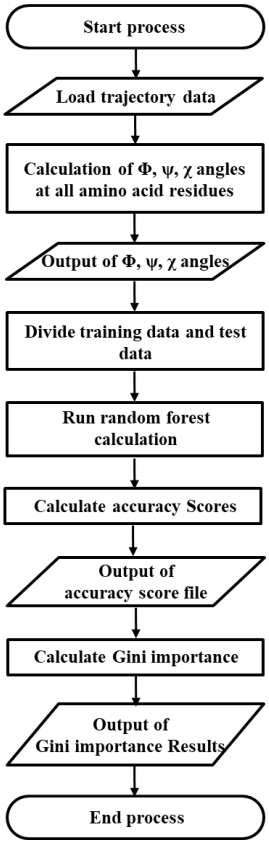
These two shortcomings may reduce the accuracy of the calculation and make it difficult to adapt the method to proteins with more than about 300 residues. Therefore, we have developed a new tool (RF-MD) that verifies and improves on these shortcomings, and will provide insight into the development of RF-FMO.

4-2-2. Validation of the analysis method developed by Sultan

The trajectory analysis tool reported in the Sultan *et al*.[1] That tool was not available on the web. Therefore, we created the trajectory analysis tool of Sultan *et al*. based on the description in the paper (**Schema 1**). The tool was developed entirely using the Scikit-learn library implemented in Python 2.7.[94] The MD Traj library was used for reading and writing

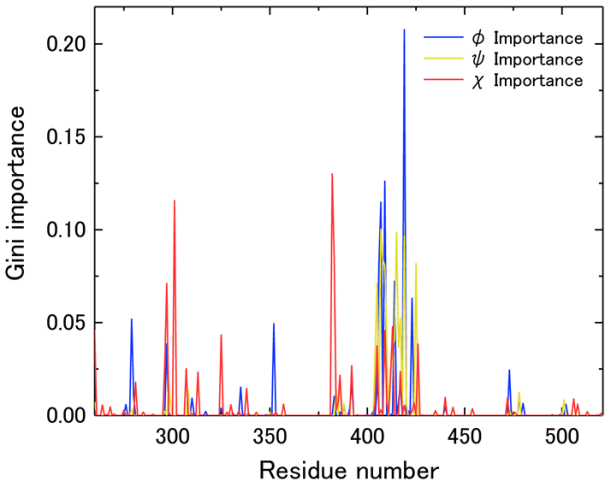
trajectory data.[95] This tool automatically extracts amino acid residues that are important for conformational changes of proteins with two conformational states by using MD simulation trajectories of the two conformational states as input. The number of estimator and max depth settings used in the random forest calculation in this tool were 30 and 4, respectively.

Schema 1. Analysis flowchart for the Trajectory Analysis Tool based on the paper by Sultan *et al*



First, in order to confirm that the tool works as well as the one of Sultan *et al*. we performed automatic extraction of amino acid residues using Src-Kinase trajectory data generated by MD simulation as input. The Gini importance values of φ, ψ, and χ obtained from this analysis were plotted for each amino acid residue (**Figure 17**), and similar to Sultan et al. Gini importance peaks were observed. Therefore, the tool created in this study is considered to be almost similar to the tool of Sultan *et al*.[1]

Figure 17. Results of the analysis of the author's trajectory analysis tool for Src-Kinase. The obtained Gini importance values of φ, ψ, and χ are depicted graphically for each amino acid residue. The graph suggests that Lys295, Glu310, His384, and A-loop are important, similar to the results of Sultan *et al*.



To verify the accuracy, we repeated the same calculation five times and compared the results (**Figure 18a**). To compare the Gini importance errors, the mean value of the Coefficient of Variation (***CV***) of the Gini importance ($\overline{\boldsymbol{CV}}$) was obtained as follows ((**Equation** 47 and **(Equation 48**)

$$\boldsymbol{CV} = \frac{\boldsymbol{\sigma}}{\overline{\boldsymbol{x}}} \qquad \textbf{(Equation 47)}$$

$$\overline{\boldsymbol{CV}} = \frac{\sum \boldsymbol{CV_i}}{\boldsymbol{i}} \qquad \textbf{(Equation 48)}$$

σ is the standard deviation of Gini importance at each amino acid residue and $\overline{\boldsymbol{x}}$ is the mean value of Gini importance at each amino acid residue. This value indicates that the larger the value, the larger the error, and the smaller the value, the smaller the error. This result indicates that the Sultan *et al*. tool[1] is inaccurate due to the variation in the values calculated for each calculation.

Furthermore, when the Gini importance values were calculated five times using the tool created with the CtBP2 trajectory as input, it was found that there was an even larger variation in the Gini importance values from calculation to calculation (**Figure 19a**). This is because the number of amino acid residues in hCtBP2 is nearly twice as large as in Src-Kinase.

These results indicate that the Sultan *et al*. tool[1] is inaccurate and that it is difficult to extract important amino acid residues because the calculated Gini importance values vary from calculation to calculation.

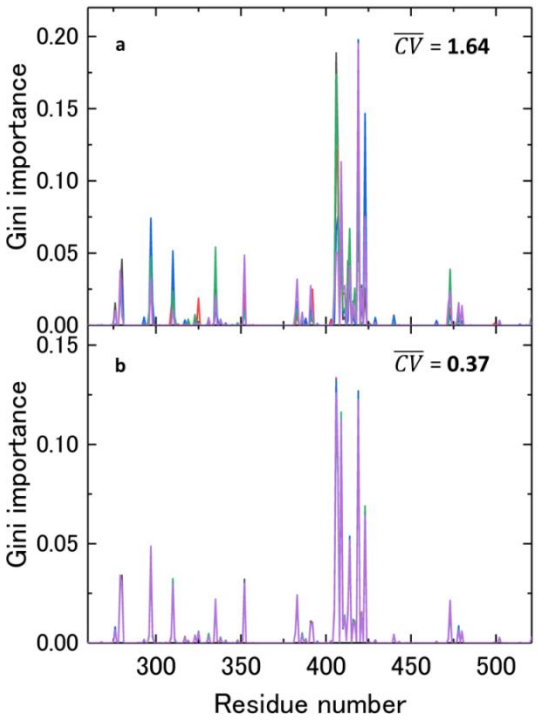


Figure 18. Accuracy validation of Sultan's tool using Src-Kinase. (**a**) Src-Kinase trajectory data were analyzed five times using the tool of Sultan *et al*. Gini importance values obtained from each calculation were plotted. The plotted curves do not overlap, indicating that the results varied from calculation to calculation. (**b**) The same procedure as in a was performed using RF-MD. The five curves overlapped nicely and the value of $\overline{\boldsymbol{CV}}$ was reduced, indicating that the accuracy of the calculation was successfully improved.

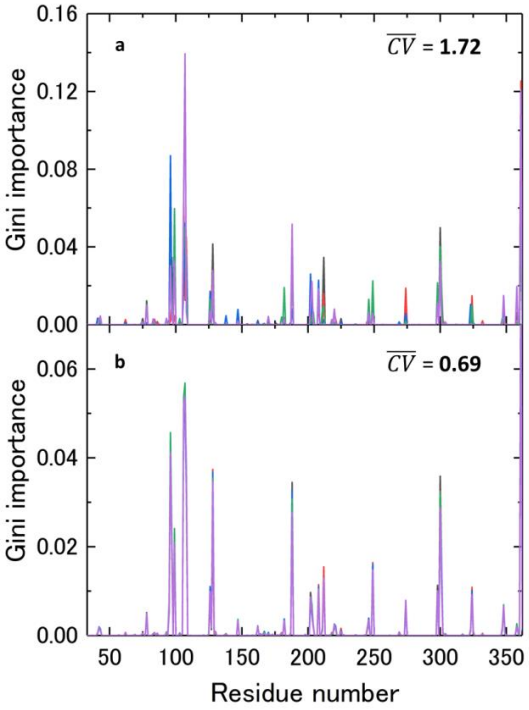
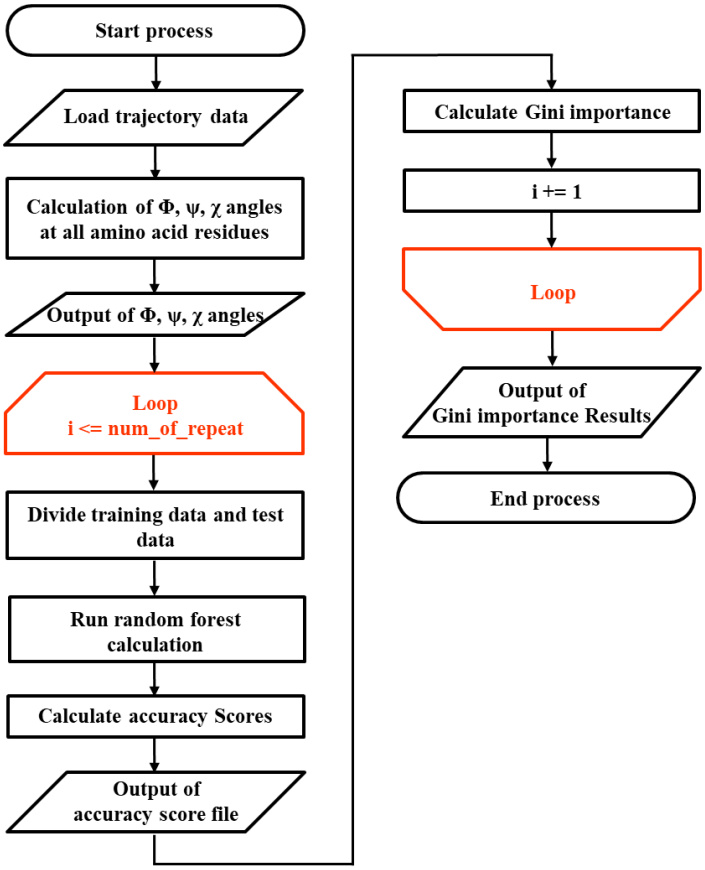


Figure 19. Accuracy validation of Sultan et al.'s tool using CtBP2. (a) CtBP2 trajectory data were analyzed five times using the tool of Sultan *et al*. Gini importance values obtained from each calculation are plotted. The plotted curves do not overlap, indicating that the results varied from calculation to calculation. Furthermore, the increase in $\overline{\boldsymbol{CV}}$ values compared to Src-Kinase suggests that the accuracy will further decrease as the number of data increases. (b) The same operation as in a was performed using RF-MD. The five curves overlap nicely and the value of $\overline{\boldsymbol{CV}}$ has decreased, suggesting that the accuracy of the calculation has been successfully improved.

4-2-3. Development of RF-MD and consideration of accuracy improvement

RF-MD was created to overcome the weaknesses found in the tool of Sultan *et al*. Basically, to increase the accuracy of the random forest calculation performed in the tool, it was said that the number of estimator and max depth settings should be increased. However, when the number of estimators was increased, the computation time became very long due to the large amount of data, and there were concerns about over-learning for the max depth. Therefore, we decided to repeat the calculation to obtain the Gini importance value and calculate the average value of the obtained Gini importance (**Schema 2**). This would allow for easy parallelization of each calculation and avoid over-learning.

Schema 2. Flowchart of the RF-MD analysis



First, we compared the accuracy with the tool of Sultan *et al*. using Src-Kinase trajectory data.[1] As in the previous section, we performed the same calculation five times and compared the Gini importance values calculated for each calculation (**Figure 18b**). As a result, the variation in Gini importance values was almost eliminated and the accuracy problem was successfully solved without loss of computation time (**Table 2**).

Next, we performed the same operation using hCtBP2 trajectory data (**Figure 19b**). No variation in Gini importance values was observed for this result as well. This suggests that high accuracy can be maintained even as the number of data increases. This suggests that the RF-MD can maintain high accuracy even when the number of data increases, which was a problem with the trajectory analysis tool of Sultan *et al*.[1] Based on the findings obtained here, we developed the RF-FMO in the next section.

Table 2. Comparison of analysis time [in Sec] between the tools of Sultan et al. and RF-FMO. The time required for analysis by each trajectory is shown. Single Processor is the time required for the analysis with the tool of Sultan *et al*. and multi processors is the time required for the analysis with RF-MD. The RF-MD tool repeats the same calculation 1000 times to improve accuracy and averages the obtained values, so the time required to repeat the calculation 1000 times was used for comparison. As a result, it is clear that RF-MD is faster.

	Single Processor		Multi Processors	
Number of cycle	Src-Kinase	CtBP2	Src-Kinase	CtBP2
1	4.73	11.33	-	-
1000	4726.63	11328.33	211.97	379.05

4-2-4. Development RF-FMO

The RF-FMO takes the trajectory data of the protein in two states from MD simulations as inputs (denoted Form A and Form B). The analysis process of the RF-FMO method is shown in **Scheme 3**. At the first, Root Mean Square Deviation (RMSD) was calculated using the inputs of Form A and Form B trajectories. RF-FMO used the initial structure of each trajectory for the reference structure of the RMSD at each trajectory. After calculated RMSD for each trajectory, RF-FMO crate two type of histograms by using the RMSD calculated, one is calculated by all structures, another one is calculated by randomly sampled up structures that were distributed within a range of standard deviation values to the left and right with respect to the median. This extraction method allowed us to extract only the major structures in each state. Then, RF-FMO calculate the interaction energy between amino acid residues by FMO method after energy optimization using GROMACS was performed on each sampled structure. Finally, RF-FMO method calculates the Gini importance value. To calculate the Gini importance values, a random forest calculation was performed using the interaction energy values between amino

acid residues calculated in the FMO calculation as training data. In addition, the betweenness centrality of each amino acid residue was calculated by graph theory analysis using the interaction energy values between amino acid residues. The resulting betweenness centrality of each amino acid residue in each structure was used as training data to calculate its Gini importance value.

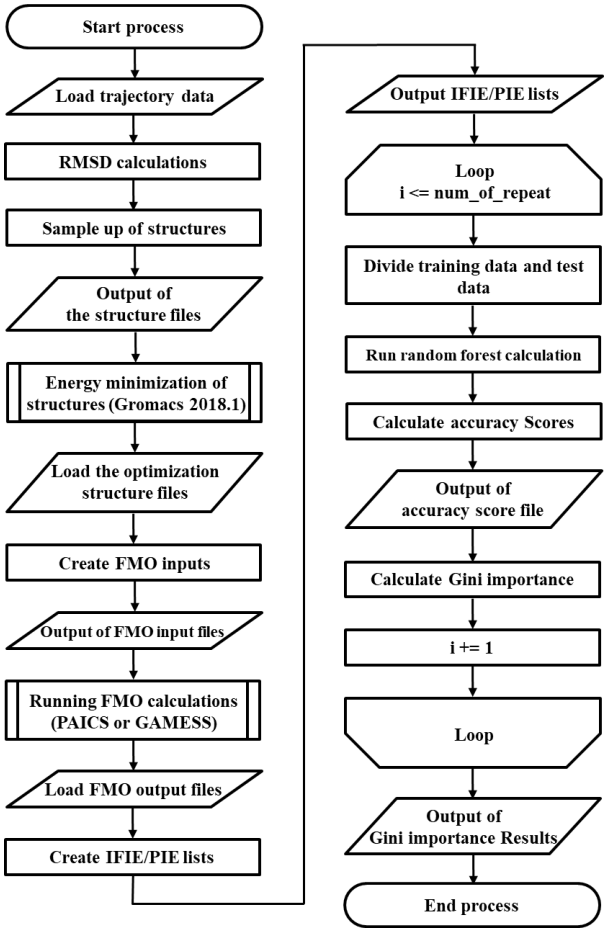
Graph theory is a mathematical theory about graphs consisting of a set of nodes and a set of edges. The theory defines each node and the edges that connect the nodes, and analyzes the characteristics of each node and edge. The graph theory analysis performed in RF-FMO is as follows. Each fragment (amino acid residues) used in the FMO calculation is defined as a node, and the interaction energy between amino acid residues calculated by the FMO calculation is defined as an edge. Values of amino acid residue interaction energy were used normalized between 0.1 and 1.1. This mean that the analysis was performed by creating a graph in which the nodes of each amino acid residue were connected to each other by the presence or absence of interaction energy. The connection between each node is such that the higher the interaction energy, the stronger the connection, and amino acid residues with no interaction are not connected by edges and are not related to each other. Betweenness Centrality B_j,[96, 97] is a numerical measure of how much of the node of interest is in the shortest path between two other nodes. The higher the value, the more it is always involved in an interaction network with somewhere else. For example, a highway junction or an approver in a paper submission are one example of cases where influence is generated because information is frequently transmitted there because it has to pass through there. It is represented by the following equation.

$$B_j = \sum_{i<k} \frac{g_{ijk}}{g_{ik}} \qquad \textbf{(Equation 49)}$$

where g_{ik} is the number of the shortest paths between nodes *i* and k that include node *j*. g_{ijk} is the number of shortest paths between nodes *i* and *k*.

The RF-FMO analysis tool was developed using the Scikit learn library implemented in Python 2.7,[94] and the md-traj library was used to read and write trajectory data.[95] The RF-FMO takes as input trajectory data from MD simulations of proteins in two conformational states. From those trajectory data, it samples the conformations and performs an FMO calculation for each sampled conformation, using the amino acid interaction energy calculated by the FMO calculation to perform random forest analysis and graph theory analysis. The RF-FMO automatically determined important amino acid residues for conformational change from the output of the FMO calculations performed for each of the sampled structures by computational processing.

Scheme 3. Flow chart of RF-FMO analysis.



4-3. Materials and Methods

4-3-1. Construction of initial structure

The structure obtained from the X-ray crystal structures were used for construction of initial structure of Src tyrosine kinase. Two X-ray crystal structures was used. The complex structures of MPZ bound to Src tyrosine kinase (PDB ID: 1Y57[98]) which functions as the active state was used for active state structure. The complex structure of AMP-PNP bound to Src tyrosine kinase (PDB ID: 2SRC[92]) which functions as the inactive state was used for inactive state structure. MOE 2016.08 was used for construction of all initial structures.[46] All constructed structure were optimized with Amber10:EHT force field for the protein. To take account of solvent effects used the Born implicit solvent model. To equalize the amino acid

residues length of each protein, the residues from Trp260 to Thr521 were extracted in each X-ray crystal structures. Molecules other than Src tyrosine kinase that were placed in the PDB were removed except for those necessary for analysis. Crystal water and molecules bound to the protein were removed, except for the ligands MPZ and AMP-PNP. Both terminals of the cleaved protein were capped by methylation. All missing hydrogen atoms were added using the Protonate 3D utility. Finally, energy minimization of the complex structure was performed. Energy minimization was performed in two stages. First, only the added hydrogen atoms were minimized with the energy minimization utility until a threshold value of 0.00001 kcal mol^{-1} $Å^{-2}$ was set. Finally, the protein complex was minimized to a threshold-set value of 0.01 kcal mol^{-1} $Å^{-2}$.

4-3-2. Preparation of MD simulations

Using the complex structure constructed in section 4-3-1, I constructed a simulation system for MD simulations. To create topology files, I used the LEaP function of AMBER14[50]. The molecular force field of ff14SB[49] was used for the protein, and the custom force field was made by Walker *et al.*[99, 100] used for the NADH/NAD^+. The periodic boundary of the MD simulation was set at a distance of 12 Å from the complex. The TIP3P water model was placed in the solvation box.[101] In order to neutralize the target system, I added counter ions and physiological conditions were set to 150 mM KCl. The topology files created by the LEaP function are in AMBER format. It were converted to GROMACS format using the acpype.py script because the MD simulation requires GROMACS format.[73]

4-3-3. MD simulation

I was used the GROMACS package 2018 for all MD simulations.[74] First, constraints were imposed on the entire system and steep minimization was performed. Energy minimization was performed step-by-step. First, energy minimization was performed only for the missing

hydrogen atoms given by Protonate 3D utility. A constraint of 1000 kJ mol^{-1} nm^{-2} was imposed on the protein atoms except for the hydrogen atoms, which were minimized in 10,000 steps. Next, since the orientation of side chains may not be stable in X-ray crystal structures due to packing and other effects. Energy minimization of side chains of protein was performed. A constraint of 1000 kJ mol^{-1} nm^{-2} was imposed on the main chain atoms and minimized in 10,000 steps. Finally, the minimization was performed in 10,000 steps without any constraints (total 30,000 steps). The threshold for energy minimization was set at 10 kJ mol^{-1} nm^{-2}. Next, annealing of water molecules automatically placed by the Leap function was performed. The simulation was repeated 10 times, with constraints placed on atoms other than water molecules, and the entire system was heated from 0 to 300 K during 100 ps. Next, the entire system was heated from 0 to 300 K during 300 ps. The NVT ensemble (T = 300 K) was employed in the annealing and heating simulations. For protein atoms except hydrogen atoms, the temperature was controlled using the v-rescale method with a constraint constant of 1000 kJ mol^{-1} nm^{-2}. [102] Then, an equilibrium calculation of 700 ps was performed using the NPT ensemble (T = 300 K and P = 1 bar) with the Nosé-Hoover[76, 77] and Parrinello-Rahman methods[78, 79] for temperature and pressure control. The constraining constant, excluding hydrogen atoms in the protein, decreased gradually every 100 ps from 1000 kJ mol^{-1} nm^{-2} to 10 kJ mol^{-1} nm^{-2}. In addition, an equilibration calculation of 500 ps was performed with the temperature set to 300 K in the NPT ensemble. After a series of equilibration simulations were completed, a final production calculation of 100 ns was performed in the NPT ensemble. All simulations were performed with a time step of 2 fs. All hydrogen atoms were constrained by applying the LINCS method.[75] The thermostat-barostat coupling time was set to 1.0 ps^{-1}. A particle mesh Ewald algorithm was used for the long-range electrostatic interaction.[103] The cutoff value was set to 10 Å. The compression ratio was set to 4.5×10^{-5} bar.

4-3-4. FMO Calculation

All FMO calculations were performed by using the GAMESS program package.[104] I used the density functional tight coupling (DFTB) method[105] with the 3ob parameter set[106, 107] for the FMO calculations. In order to account for solvent effects, I used the PCM method.[108] The DFTB calculations used Grimme's DFT-D3 dispersion correction to account for dispersion interactions.[109, 110]

4-4. Results and Discussion

4-4-1. Evaluation of structural stability of active and inactive states of Src-Kinase

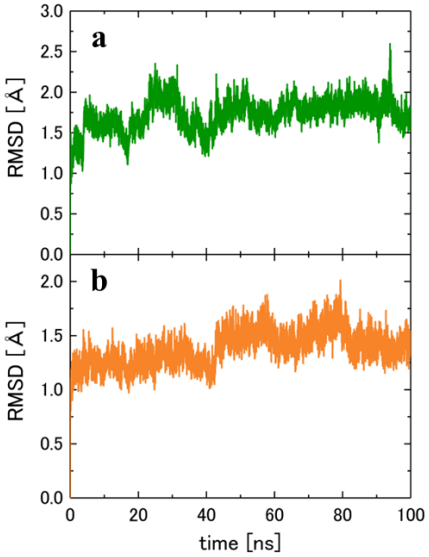
I confirmed by MD simulation that the structures of the active and inactive states obtained by X-ray crystal structure are stable structures. The modeled structures of the active and inactive states were sampled by 100ns MD simulation as the initial structure. RMSD values were calculated from the calculated trajectory data of both structures with the initial structure in the inactive state as the reference structure. The average RMSD values were 3.8 Å for the active state and 1.3 Å for the inactive state. The difference between the average RMSD values of the two structures suggests that the inactive and active states exist in different and stable conformations as protein conformations. Furthermore, the respective RMSD values during the simulation varied stably with respect to the average value, indicating that both structures are thermally stable. This allowed verify the stability of the modeled structures. As a result, we have decided to use these trajectory data for future analyses of interactions between amino acid residues using the RF-FMO method.

4-4-2. Extraction of amino acid residues important for the conformational change of Src tyrosine Kinase between active and inactive states by RF-FMO

RF-FMO analysis was performed using trajectory data obtained from MD simulations as

input. The results obtained by this RF-FMO analysis are the amino acid residues that are important for the conformational change of Src tyrosine kinase between the active and inactive states. The results of the RF-FMO analysis are shown in below. The RMSD was calculated based on the trajectory data for the active and inactive states (**Figure 20a** and **b**). The respective initial structures were used as reference structures for the RMSD. From these RMSD values, histograms for both states were calculated (**Figure 21a** and **c**). Based on these histograms, we sampled 50 structures for each of the active and inactive states (100 structures in total). The structures sampled up were randomly and sampled from structures within a standard deviation of the median value. The RMSD histograms of all sampled structures are showed in **Figure 21b** and **d**. FMO calculations were performed for all sampled up structures. Finally, Gini importance values of amino acid residues and pair interaction energies between the amino acid residues were calculated using the flow shown in **Scheme 3**. The Gini importance values for the amino acid residues and interaction energy of each amino acid residue were calculated with the following parameters. The number of estimators was set to 30, the maximum depth to 4, and the number of iterations to 1000. The calculated Gini importance values were sorted in order of increasing value and plotted on a graph (**Figure 22**). The features with the highest Gini importance values (the area with red background in the graph) were considered as the features extracted by RF-FMO. Finally, 32 of the 623 amino acid residues were automatically extracted, yielding an extraction rate of 5.1% (**Table 3**). In addition, 61 of the 34453 interaction energies were extracted, yielding an extraction rate of 0.2% (**Table 4**). From this 34453 x 100 = 3445300 interaction data set, RF-FMO automatically detected significant interaction pairs. This analysis is difficult to perform manually. The development of RF-FMO allow us to perform this analysis. The RF-FMO analysis is also based on the dynamic interaction energy sampled by the MD simulation. This is different from the static interaction analysis of a single structure by conventional FMO, which is also an advantage of this method over conventional methods.

Figure 20. RMSD calculated from MD simulation results of (**a**) active state and (**b**) inactive state.



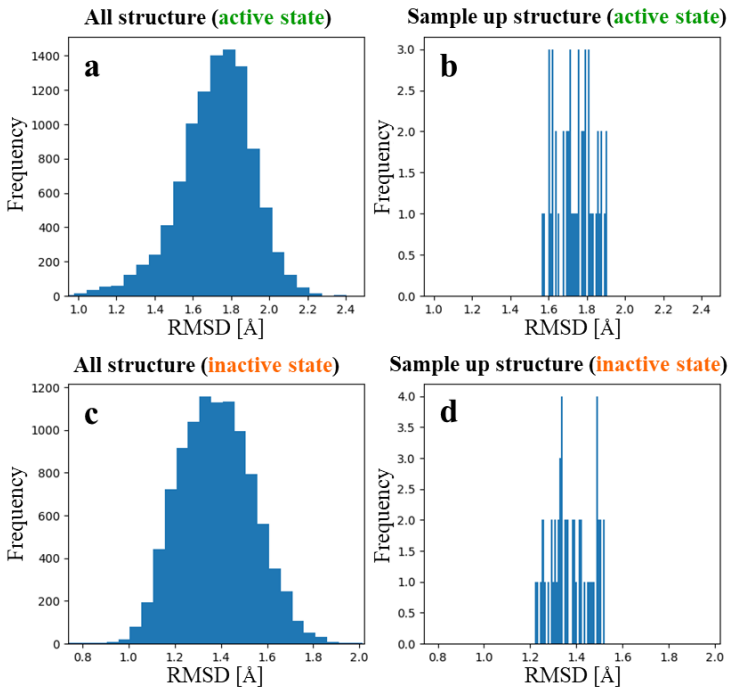


Figure 21. The histogram of RMSD in (**a**) active state and (**c**) inactive state structures. The RMSD histogram of the sampled structures from (**b**) active state trajectory and (**d**) inactive state trajectory.

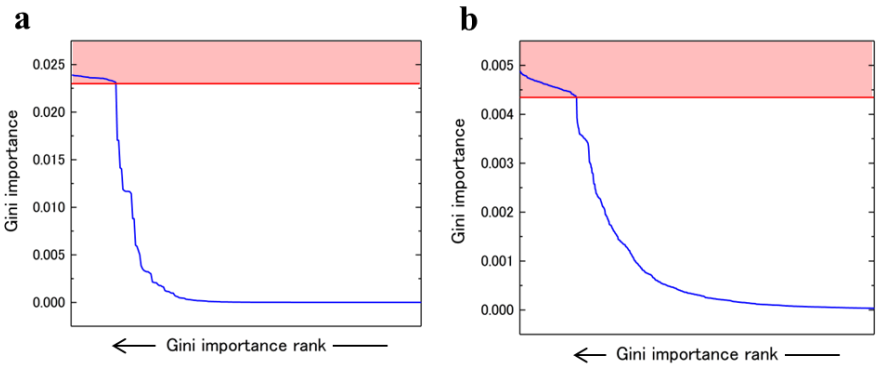


Figure 22. (**a**) The Gini importance values of the betweenness centrality of amino acid residues are sorted and plotted in order of high value. Amino acid residues located in the area where the plot is painted red are considered as amino acid residues extracted by RF-FMO. (**b**) The Gini importance values of the pair interaction energy are sorted and plotted in high order of value. Pair interaction energy located in the area where the plot is painted red are considered as pair interaction energy extracted by RF-FMO.

Table 3. List of the amino acid residues detected by RF-FMO analysis.

LYS272	GLY274	PHE278	THR289	THR301	MET302	PHE307	GLU310	GLY421	PHE424
GLU332	ILE336	SER345	ARG379	ASN381	TYR382	VAL383	ASP386	PRO425	TRP428
ALA389	GLY406	ARG409	LEU410	ILE411	GLU412	ASP413	GLU415	ILE441	SER443
TYR416	GLN420								

Table 4. List of the interacting amino acid residue pairs detected by RF-FMO analysis.

THR301-GLY279	PHE307-LEU297	**GLU310-LYS295**	PHE424-TYR416	LYS427-PRO425
VAL328-PHE307	GLU332-GLY300	GLU332-LEU297	PRO425-ASP386	ILE441-ASN381
ILE336-ILE294	ILE336-ALA311	ILE336-GLU310	PRO425-TYR416	TRP428-ASP386
ILE336-LEU297	PHE349-SER345	VAL383-VAL377	LYS423-ASP413	SER447-ASP386
ALA389-SER345	ALA389-LEU350	GLY406-HIE384	PHE424-ALA422	THR429-PHE424
GLY406-VAL383	LEU407-ASP386	ALA408-GLY406	PHE424-GLN420	GLU454-TRP428
ALA408-ASP386	ALA408-GLU310	ARG409-TYR382	ALA418-TYR416	GLU510-ARG379
ARG409-GLU310	ARG409-VAL383	ARG409-GLY406	GLN420-GLU415	LYS423-GLU415
ARG409-ARG385	LEU410-ALA408	LEU410-PHE307	TYR416-ALA408	ASP413-ILE411
LEU410-TYR382	LEU410-GLY406	LEU410-GLU310	ASN414-GLU412	TYR416-ILE411
LEU410-ASN381	ILE411-TYR382	ILE411-VAL383	THR417-ASP413	ALA418-GLU415
ILE411-MET302	ILE411-PHE278	ILE411-ARG409	GLU412-THR301	GLU412-ARG409
GLU412-MET302				

The amino acid residues and PIE extracted by RF-FMO were used for structural analysis. The extracted amino acid residues were projected onto the X-ray crystal structure in the inactive state (**Figure 23a and b**). These figures suggest that the automatically extracted amino acid residues are in or around the A-loop. This suggests that the A-loop is important for the conformational change between the active and inactive states. Furthermore, the structural relationship between the two amino acid residues was analyzed for the interaction energies extracted by RF-FMO analysis. The analysis revealed that the interaction network of Lys295, Glu310, and Arg409 was important (**Table 4**). Comparing these three residues in the active and inactive states, it was found that Lys295 and Glu310 form a strong salt bridge in the active state. On the other hand, Glu310 and Arg409 formed salt bridges in the inactive state (**Figure 23c** and **d**). These results suggest that the conformational change between the active and inactive states

of Src tyrosine kinase is due to a change in the interaction pattern of Glu310. The results of RF-FMO analysis are consistent with Sultan *et al.*[1] and Ozkirimli and Post,[111] and support the validity of the RF-FMO analysis constructed in this study. These suggest that RF-FMO is a useful tool for automatically analyzing the results of large numbers of FMO calculations and easily extracting important amino acid residues.

In summary, in the active form, Lys295 and Glu310 interact, and Arg409 is no longer immobilized by interaction with Glu310. This changes the A-loop into a flexible, fully extended loop structure. In contrast, in the inactive form, Glu310 and Arg409 on the A-loop form salt bridges, suggesting that the A-loop folds compactly to form a helical structure (red).

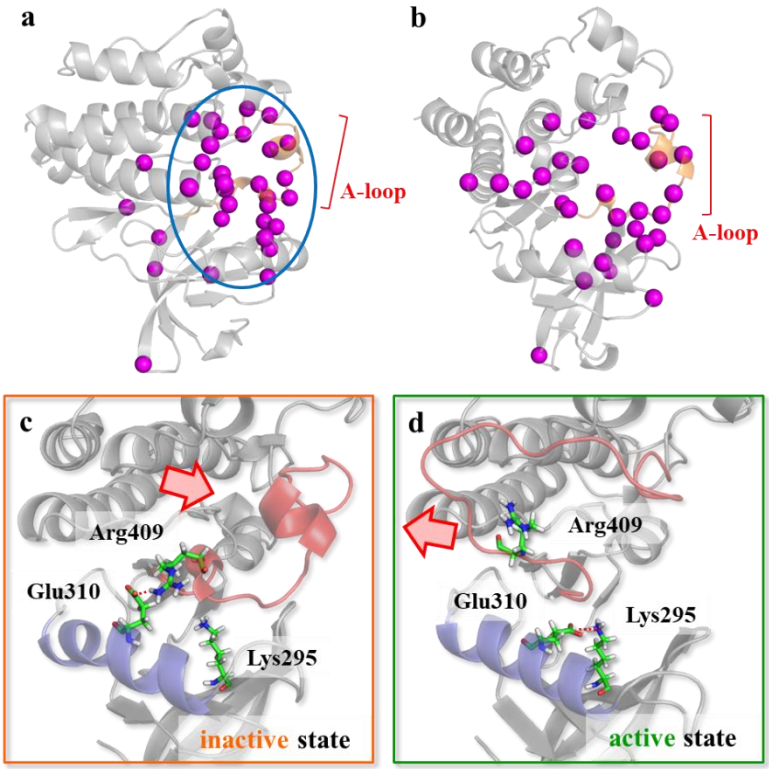


Figure 23. Extracted residues by RF-FMO were showing in (**a**) front view and (**b**) side view. The amino acid residues (magenta ball) extracted by RF-FMO are projected onto the crystal structure of the inactive state (A-loop is denoted in orange). The interaction network with Glu310, Lys295 and Arg409 in the (**c**) inactive and (**d**) active state (A-loop is denoted in red).

4-5. Summary

In this study, I developed a novel analysis tool (RF-FMO) which automatically extracts amino acid residues important for conformational changes based on interaction energy values calculated by FMO calculations. To validate the developed method, RF-FMO analysis was performed on Src tyrosine kinase. As a result, 32 amino acid residues were successfully extracted automatically from 623 amino acid residues. The extraction rate was 5.1%. I was also able to automatically extract 61 residue pairs out of 34453 interaction energies between amino acid residues. The extraction rate was 0.2%. These results indicate that RF-FMO is a method that can automatically extract interactions important for conformational changes from huge amount of interaction energy data. RF-FMO was very effective due to the amount of data that could not be analyzed by artificially performed FMO interaction analysis. This analysis method may also be useful for QM/MM calculations when ligand or cofactor dynamics are relevant to the enzyme reaction process.[112] In the next chapter, we will attempt to apply this method to proteins for which the mechanism of conformational change at the molecular level is unknown.

Chapter 5. Theoretical Study on the Control Mechanism of hCtBP2 Open-to-Close Transition

Purpose of this Chapter

In this chapter, I analyzed CtBP2, a protein for which the mechanism of conformational change at the molecular level is unknown, using the RF-FMO constructed in the previous chapter, in order to clarify the molecular mechanism of conformational change.

5-1. Introduction

5-1-1. C-terminal Binding Protein 2 (CtBP2)

The C-terminal binding protein (CtBP) family is present in a variety of organisms and plays multiple biological roles.[113] Although the CtBP family is known to include CtBP1 and CtBP2, it is difficult to completely distinguish their functional roles in mammals.[33] It is well known that CtBPs in mammals act as corepressors that form complexes with transcriptional repressors associated with cancer and metabolic diseases.[114] In particular, human CtBP2 (hCtBP2) is known to exhibit dehydrogenase activity.[115] It has been suggested that the complex formation between hCtBP2 and NADH during this process depends on the conformational changes of hCtBP2, its association state, and its intracellular NADH/NAD^+ status. I have previously investigated the effect of NADH ligand binding on the stability of dimer formation by molecular dynamics (MD) simulations.[116] However, the detailed molecular mechanism by which CtBP2 forms the dimer and the types of atomic-level interactions that regulate it have not been elucidated.

5-1-2. Structure of CtBP2

CtBP has a Rossmann fold-type structure consisting of three domains: an NADH-binding domain (NBD), a substrate-binding domain (SBD), and a dimerization loop domain (DLD) (**Figure 1) The** NBD domain binds to NADH/NAD^+ and is critical for CtBP function The NBD domain binds NADH/NAD^+ and plays an important role in CtBP function. Kumar *et al.* also suggested that the conformational change in the SBD, which opens and closes upon binding of NADH/NAD^+ to the NBD, is important for the activity of CtBP.[115] Therefore, in this letter, I would like to further identify which amino acid interactions are involved in the conformational changes between the open and closed states.

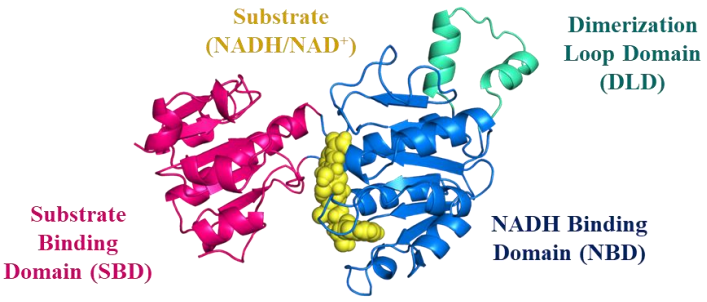


Figure 24. The structure of CtBP with NADH/NAD^+ (yellow). CtBP has three domains called NADH binding domain (NBD, blue), substrate binding domain (SBD, pink) and dimerization loop domain (DLD, green).

5-2. Materials and Methods

5-2-1. Construction of initial structure

The X-ray crystal structure of the hCtBP2/NAD(H) complex (PDB ID: 2OME) was used to construct the initial structure. The A and B chains of the X-ray crystal structure were extracted and hydrogenated using the Protonate 3D function. The C- and N-termini of the proteins were capped with methyl groups; all ligands in the X-ray crystal structure were hydrogenated to NADH. Finally, the protein complexes were energy minimized with a threshold of 0.01 kcal mol^{-1} $Å^{-2}$. All initial structures were constructed using MOE 2016.08.[46] Amber10:EHT force field and Born implicit solvation model were used for minimization.

5-2-2. MD Simulation

Using the model structure constructed in 5-2-1, I constructed a simulation system using the LEaP function of the AMBER14[50] LEaP function. Protein and NADH were simulated using the ff14SB force fields[49] and the force field created by Walker *et al.* respectively.[99, 100] A solvation box was constructed so that the distance from the complex to the periodic boundary was 12 Å. The TIP3P water model was constructed.[101] For neutralization, 150 mM KCl was added as a counter ion. The constructed AMBER topology files were converted to NAMD

format and to GROMACS format using ParmEd[117] and GROMACS format using the acpype.py script[73].

5-2-3. TMD

Since the open state structure of hCtBP2 has not been previously obtained by X-ray crystallography, targeted MD (TMD) calculations were performed using NAMD 2.11[15] to induce the structural transition from the closed state to the open state. Target MD (TMD) calculations were performed to induce a structural transition from the closed state to the open state using the model structure. First, the entire system was equilibrated. Next, a conventional 10 ns MD calculation was performed in the NPT ensemble (300 K, 1 bar), followed by a 9 ns TMD simulation with a force constant of 30 kcal mol^{-1} $Å^{-2}$. Simulations were performed with a time step of 2 fs and the SHAKE method[118] was applied to constrain all interatomic bonds. For water molecules, the SETTLE method[119] was used to constrain every 2 fs. The Langevin method was used for temperature control and the Langevin piston Nose-Hoover method for pressure control.[76, 120] Short-range van der Waals and electrostatic interactions were cut off at 9 Å. Long-range electrostatic interactions were controlled by the particle mesh Ewald algorithm.[103] All conventional MD simulations for 100 ns were performed using the GROMACS package 2018.1.[74] The structure constructed from the crystal structure was used as the initial structure of the closed state, and the final structure obtained from the TMD simulations was used as the initial structure of the open state. After stepwise equilibration, 100 ns of generation was performed in the NPT ensemble (300 K, 1 bar), using the Nosé-Hoover method[76, 77] for temperature control and the Parrinello-Rahman method [78, 79] for pressure control. In a series of simulations, the time increment was set to 2 fs and the LINCS method[75] was applied to constrain all hydrogen atoms. For further long-range electrostatic interactions, the particle mesh Ewald algorithm was used.[103]

5-2-4. FMO Calculation

The GAMESS program package was used for the FMO calculations.[104] The density functional tight binding (DFTB) method[105] was employed and the calculations were performed using the 3OB parameter set.[106, 107] In the DFTB calculations, Grimme's DFT-D3 dispersion correction was used to account for dispersion interactions.[109, 110] To account for solvent effects, the polarizable continuum model (PCM) was applied to account for solvent effects.[108]

5-2-5. RF-FMO Analysis

The RF-FMO method developed by the authors was used to analyze the large number of pair interactions obtained from the FMO calculations and to find important amino acid residues associated with open/closed conformational transitions. RF-FMO is a machine learning based method that is used to determine the importance of a particular pair interaction from all pairs in all snapshot structures and all amino acid residues in all snapshot structures.[121] This may reflect the dynamics of the protein in the interaction analysis. All RF-FMO analyses were performed in Python 2.7[94] using the Scikit-learn library. The MD Traj library was used to read and write trajectory data.[95]

5-3. Results and Discussion

5-3-1. hCtBP2 dynamic structure change between Open state ↔ Closed state

MD simulations were performed for 100 ns using the open and closed structures of hCtBP2 generated from the TMD and the crystal structure, respectively, as initial structures. The average RMSD was found to be 2.1 Å for the closed structure and 3.0 Å for the open structure. This suggests that the final structure of the TMD simulation did not return to the closed state, but remained in another stable state. To further confirm the reliability of the obtained open state structure, it was superimposed on the open structure of D-glycerate dehydrogenase (PDB ID: 1GDH), a related protein with the same Rossmann fold, and the

RMSD value for both was 1.57 Å. This result suggests that TMDs may give hCtBP2 an open This result suggests that TMD may give the open structure of hCtBP2. In fact, the initial structure was a closed structure with the NBD and SBD closed, whereas the sampled structure was an open structure with the NBD and SBD open (**Figure 25**). These results suggest that hCtBP2 can take both open and closed conformations.

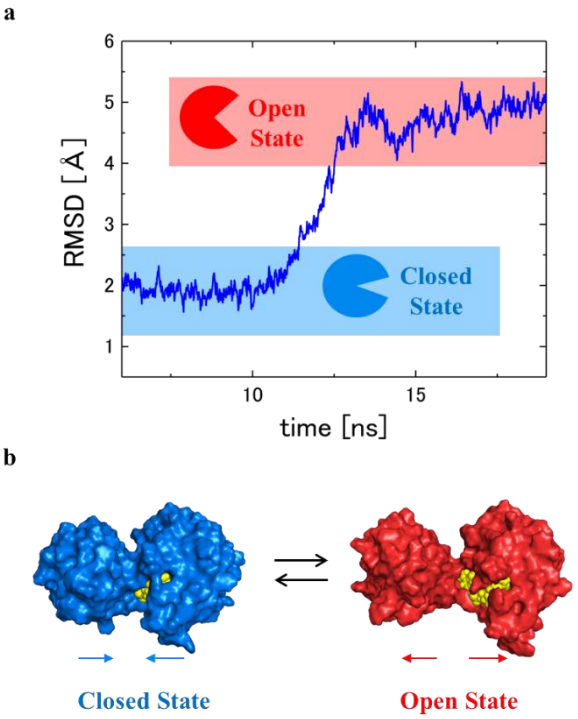


Figure 25. Results of TMD simulation. (**a**) RMSD is calculated from the trajectories which generated from TMD simulation. RMSD was calculated based on the closed state initial structure. (**b**) This Figure shows that two structures; the initial structure (blue) and the final structure of the TMD simulation (red).

5-3-2. Extraction of amino acid residues important for the conformational change of hCtBP2 between Open and Closed states by RF-FMO

Amino acid residues important for the conformational change between the open and closed states of hCtBP2 were automatically extracted in RF-FMO. The RMSD values confirm that the open and closed states are well equilibrated for 100 ns (**Figure 26a** and **b**). The reference structure for calculating RMSD was the initial structure of each structure. A total of 400 structures were used in the RF-FMO interaction analysis by randomly sampling 200 structures from both the open and closed states. Their distributions are plotted in **Figure 27**.

For each amino acid residue, a Gini importance value for interaction energy (estimated number = 50, maximum depth = 5, number of repetitions = 100,000) and a Gini importance value for intersex center per amino acid residue (estimated number = 50, maximum depth = 5, number of repetitions = 10,000) were calculated. The calculated Gini importance values were sorted in order of increasing value, and the top 5% were considered important features. Finally, 25 out of 1,324 amino acid residues (extraction rate: 1.9%, **Table 5**) and 57 out of 875,826 interaction pairs (extraction rate: 0.007%, **Table 6**) were automatically extracted.

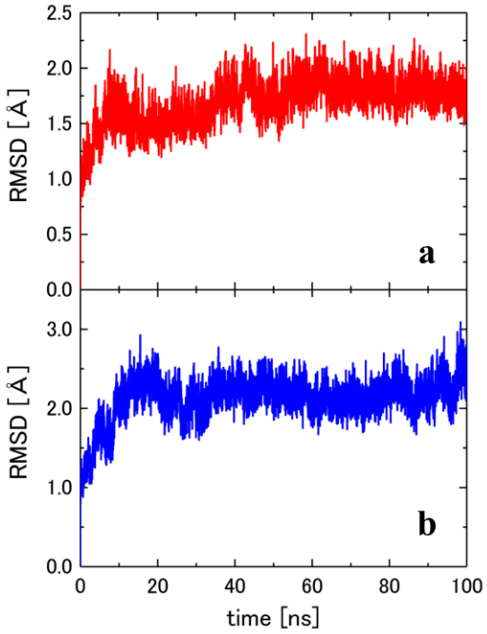


Figure 26. RMSD calculated from MD simulation results of (**a**) open state and (**b**) closed state.

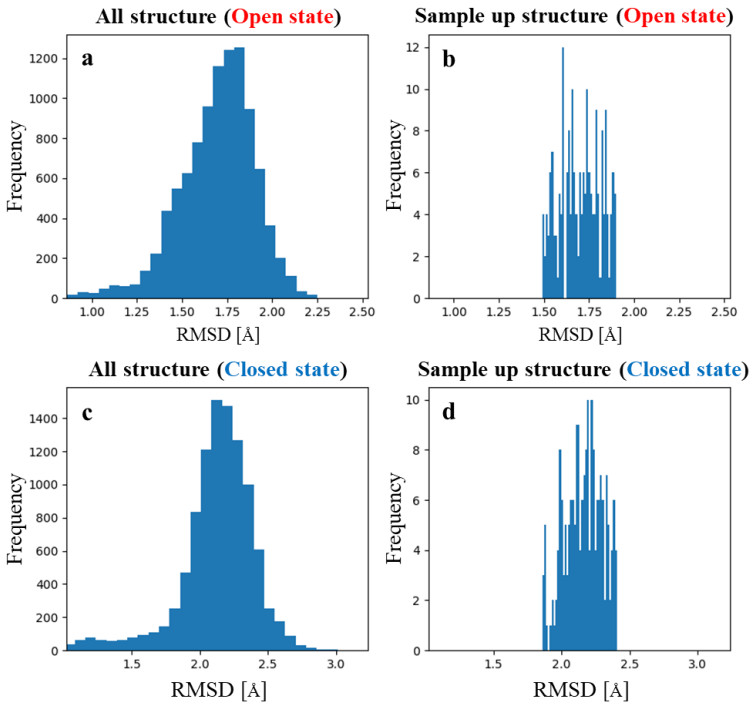


Figure 27. The histogram of RMSD in (**a**) open state and (**c**) closed state structures. The RMSD histogram of the sampled structures from (**b**) open state trajectory and (**d**) closed state trajectory.

The critical amino acid residues extracted by RF-FMO were projected onto the structure in a closed state (**Figure 28a** and **b**). From this figure, no specific domain was identified and the dynamic important residues were found to span all domains. Therefore, I next selected important residues based on pair interaction energies and found that His321 was extracted because it interacts with multiple amino acid residues (**Table 6**). These results suggest that the His321-mediated interaction network is important for hCtBP2. Therefore, I further analyzed the interaction network formed by His321. **Figure 28c, d, e** and **f** show the local structures of the closed and open states around His321. From these figures, it is clear that the switching between the open and closed states is mediated by changes in the interaction network formed by His83, Glu301, and His321. In the closed state, His83 and His321 form hydrogen bonds (possibly π-π or CH-π interactions) (**Figure 28c**). In the open state, on the other hand, His321 is oriented and forms a hydrogen bond with Glu301 (**Figure**

28d); when His83 and His321 interact, the loop (80-Loop) in which His83 is located shifts toward His321 to form the closed structure, but when His321 interacts with Glu 301 interacts with His321, the 80-Loop shifts toward the SBD domain and stabilizes the open conformation, thus failing to form a stable interaction with His83 (**Figure 28e** and **f**). These results indicate that the switching between the open and closed states is the result of a change in the interaction network formed by His83, Glu301, and His321.

Table 5. List of the amino acid residues of hCtBP2 detected by RF-FMO analysis.

ARG42	ASP43	ASP53	SER64	GLN66	VAL72	VAL72	GLY78
GLU93	ASP109	ASN110	ASP112	LEU119	CYS124	CYS124	THR138
LEU145	ARG169	ALA201	PRO211	LEU241	GLY264	ASN269	ARG272
GLN283	HIS298	ASP312	HIS321	GLU340	GLY347	GLU351	

Table 6. List of the interacting amino acid residue pairs in hCtBP2 detected by RF-FMO analysis.

ALA172-ARG169	ALA270-LEU241	ALA323-THR138	LEU119-LYS96
ALA344-GLY78	ALA201-ALA199	ALA201-VAL197	LYS200-ALA201
ALA323-**HIS321**	ALA344-GLY78	ARG354-GLU340	NAI365-ARG272
ARG148-LEU145	ARG169-GLY41	ARG169-TYR82	THR84-SER64
ARG198-ALA201	ARG272-ASP109	ARG354-GLU340	VAL72-GLU70
ARG42-ARG169	ARG42-GLU166	ASN110-HIS83	LEU353-GLU351
ASN269-SER240	ASN315-LEU145	ASN74-VAL72	LEU353-GLU351
ASP112-LEU87	ASP109-SER106	ASP312-ALA282	MET80-GLY78
ASP43-ARG169	ASP43-GLU166	ASP43-LEU38	SER173-ARG169
ASP53-ARG33	CYS44-ARG42	GLU47-ASP43	VAL268-LEU241
GLY347-ARG343	GLY78-LEU35	HIS242-THR138	ILE68-GLN66
HIS321-HIS298	**HIS321**-SER106	**HIS321**-HIS83	LEU213-PRO211
HIS321-THR319	HIS141-THR138	HIS298-GLU278	LYS96-GLU93
ILE68-GLN66	ILE85-SER64	ILE126-CYS124	PRO302-HIS298
TRP151-LEU145			

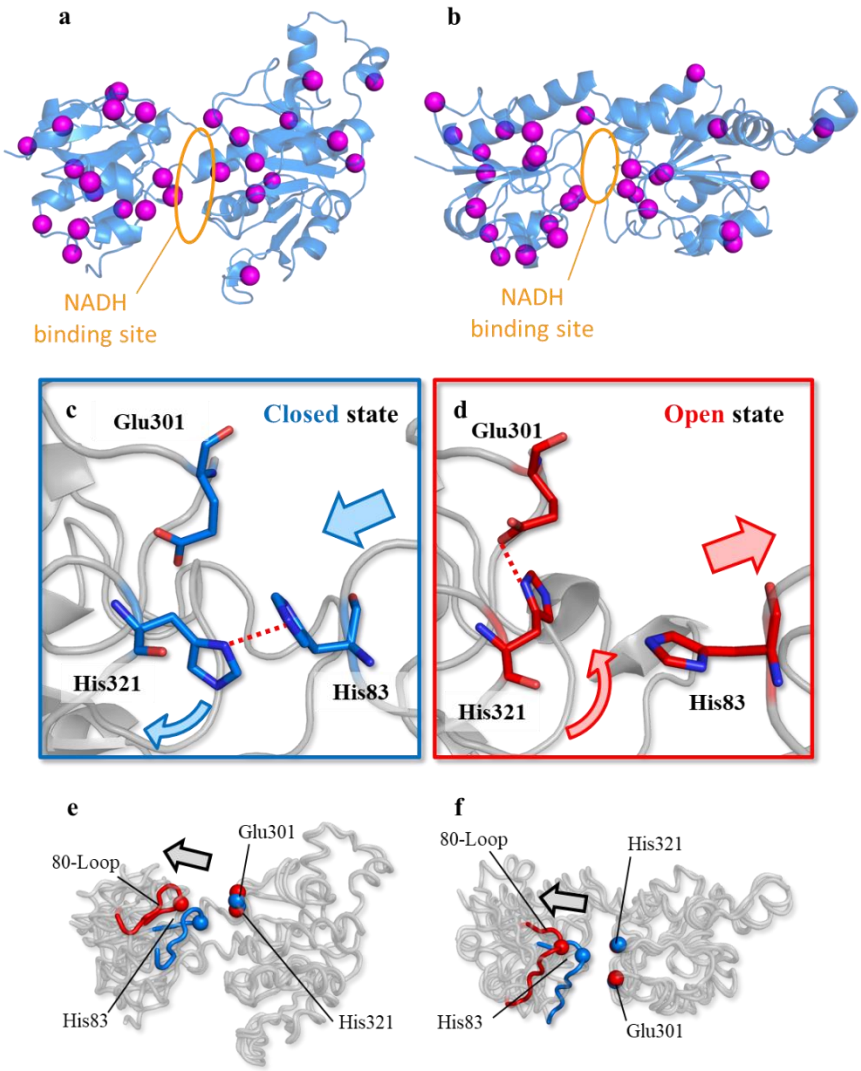


Figure 28. Extracted residues by RF-FMO were showing in (**a**) front view and (**b**) top view. The amino acid residues (magenta ball) extracted by RF-FMO are projected onto the crystal structure of the closed state. Changed the interaction network in His83, Glu301 and His321 cause Open state ↔ Closed state conformation change. (**c**) His83 and His321 form a stable interaction in the closed state. (**d**) In the open state, His321 flips toward the Glu301, and Glu301 and His321 form an stable interaction. Overall structures of hCtBP2; (**e**) front view and (**f**) top view.

5-3-3. Verification of the Open state ↔ Closed state conformational change pathway of hCtBP2

To determine the structural transition between the two states, the distance between Glu301 and His321 and the distance between His83 and His321 were used as population variables (CV). Trajectory data for the open and closed states were mapped to the CVs (**Figure 29a**). It can be seen that the upper left cluster represents the open state and the lower right region represents the closed state. This means that the open and closed states are well characterized by these CVs. Potential.

Evaluation of the energy surface of the hCtBP2 complex yielded open and closed potential energy minima (**Figure 29b**). This again confirms the formation of a switching interaction network between His83, Glu301, and His321. This figure also suggests the existence of a metastable intermediate state between the open and closed states. This suggests that hCtBP2 undergoes an open ↔ closed conformational change via the intermediate state. This was accompanied by changes in the interaction network formed by His83, Glu301, and His321.

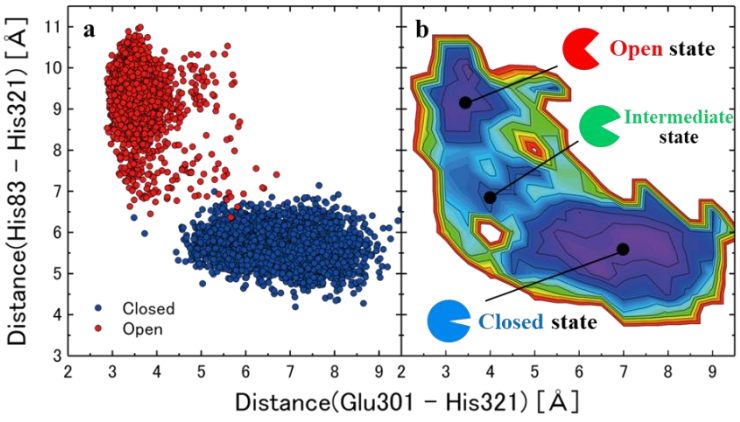


Figure 29. Conformational change pathway between open state and closed state of hCtBP2. (**a**) Trajectory profiles obtained and (**b**) potential energy surface obtained from MD simulations.

5-4. Summary

The results obtained by RF-FMO revealed that the switch between the open and closed states is caused by a change in the interaction network formed by His83, Glu301, and His321, and succeeded in elucidating the detailed mechanism of the structural change undergone by hCtBP2. Furthermore, the potential surface energy of hCtBP2 reveals the existence of an intermediate state in the open ↔ closed conformational change. This suggests that the open state ↔ closed state conformational change of hCtBP2 is caused by a change in the interaction network formed by His83, Glu301, and His321, leading to a conformational change from the open state to the closed state via an intermediate state. Bioinformatics such as sequence alignment and full consensus design may be useful in the search and design of CtBP family proteins without His83, Glu301, and His321. Indeed, His83, Glu301, and His321 identified in this RF-FMO have been found to be conserved in many sequences of a wide variety of CtBP families. This suggests that these amino acid residues may be involved in CtBP function. The present RF-FMO analysis clearly indicates that these conserved amino acids are important for structural changes at the molecular level of hCtBP2.

Chapter 6. Concluding remarks

In this doctoral thesis, I performed computational analysis of several proteins by interaction analysis based on the FMO method and revealed new scientific knowledge about these proteins.

Chapter 2 presented a theoretical analysis of the complex structure of PPARα, a ligand-dependent transcription factor involved in the regulation of lipid homeostasis and known to ameliorate hypertriglyceridemia, and its novel ligand, pemafibrate using FMO calculations. The findings of the interaction analysis of pemafibrate bound to PPARα were applied to luciferase assay experiments with mutants, whereby results supporting the computational predictions were observed. The unique binding mode of pemafibrate revealed a novel recognition pattern for nuclear receptor ligands, suggesting a new basis for ligand design, improving ligand binding affinity and selectivity, and providing clues for better clinical results. This demonstrates the usefulness of the FMO method for protein-ligand interaction analysis.

In Chapter 3, I used the properties of the FMO method to analyze protein-protein interactions. I studied the molecular mechanism of host recognition by morbilliviruses, which have high host specificity. FMO calculations were performed on the CDV-H/SLAM complex structure constructed by homology modeling, and the FMO results revealed that the interaction between the N-terminal portion of SLAM and CDV-H is important for host recognition. This finding has provided a steppingstone for the development of therapeutics against morbilliviruses, including the measles virus. This study also demonstrated the usefulness of the FMO method for analyzing protein-protein interactions. However, when intra-protein or protein-protein interaction analysis is performed, the number of interaction pairs to be analyzed is huge (about [number of amino acid residues in protein]^2 / 2), which makes the analysis difficult. Therefore, it was important to develop analytical tools that

facilitate the analysis of interactions within or between proteins.

In Chapter 4, based on the findings in Chapter 3, I established and validated a method for efficiently discovering important interactions by combining the Random Forest (RF) method, a machine learning algorithm, with interaction analysis based on the FMO method (RF-FMO). RF-FMO method, an analysis method that performs FMO calculations on all snapshot structures extracted from MD simulations and automatically extracts specific critical pair interactions that are important for structural changes. Using the developed RF-FMO method by the author, I analyzed Src-Kinase, whose functional expression mechanism has been clarified at the molecular level by the wide variety of studies that have been conducted. Finally, from a total of 623 residues and 34453 amino acid residue interaction energies, I succeeded in automatically extracting 32 (5.1% extraction rate) and 61 residue pairs (0.2% extraction rate) that are important in regulating the active-inactive transition. This result shows that this method can automatically extract interactions important for conformational changes from a huge amount of interaction energy data, replacing the FMO interaction analysis that has been performed manually so far. The results of the analysis in this chapter are consistent with previous experimental results, demonstrating the usefulness of the presented method. In the next chapter, I applied RF-FMO to proteins for which the mechanism of conformational change at the molecular level is unknown.

In Chapter 5, I used RF-FMO to analyze hCtBP2, a protein for which the mechanism of conformational change at the molecular level is unknown. hCtBP2 was analyzed by RF-FMO, and 25 out of 1324 residues (extraction rate: 1.9%) were automatically extracted, and the interaction energy was 57 out of 875826 (extraction rate: 0.007%). Analysis of the extracted amino acid residue interaction energy pairs revealed that a large number of interactions with His321 were extracted. This result suggests that the His321-mediated interaction network is important in hCtBP2. Further analysis of the interaction network of His321 in the open and closed states revealed that the switching between the open and closed states is mediated by

changes in the interaction network formed by His83, Glu301, and His321. state by changes in the interaction network formed by His83, Glu301, and His321, and succeeded in clarifying the detailed mechanism of the structural change.

The RF-FMO method developed in this study can be used to analyze not only Src-Kinase and CtBP2 but also various other proteins. It is expected that the RF-FMO method will be used in the future to elucidate the functions of proteins for which detailed molecular mechanisms have not yet been elucidated.

References

1. M. M. Sultan, G. Kiss, D. Shukla and V. S. Pande, *J Chem Theory Comput* **2014**, 10, 5217-5223.
2. W. Kabsch and C. Sander, *Biopolymers* **1983**, 22, 2577-2637.
3. T. Masaru, *Applied Physics* **2005**, 74, 1023-1023.
4. B. S. Everitt, S. Landau, M. Leese and D. Stahl, *Cluster Analysis*, Wiley, 5th edn., 2011.
5. G. E. Hinton and R. R. Salakhutdinov, *Science* **2006**, 313, 504.
6. D. A. Benson, I. Karsch-Mizrachi, K. Clark, D. J. Lipman, J. Ostell and E. W. Sayers, *Nucleic acids research* **2012**, 40, D48-D53.
7. P. Raccuglia, K. C. Elbert, P. D. F. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier and A. J. Norquist, *Nature* **2016**, 533, 73.
8. K. Kitaura, E. Ikeo, T. Asada, T. Nakano and M. Uebayasi, *Chemical Physics Letters* **1999**, 313, 701-706.
9. X. Qiang, Z. Kou, G. Fang and Y. Wang, *Molecules* **2018**, 23.
10. M. Karplus and J. A. McCammon, *Nature Structural Biology* **2002**, 9, 646-652.
11. R. Salomon-Ferrer, D. A. Case and R. C. Walker, *WIREs Computational Molecular Science* **2013**, 3, 198-210.
12. W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell and P. A. Kollman, *Journal of the American Chemical Society* **1995**, 117, 5179-5197.
13. M. Karplus and J. A. McCammon, *Nature Structural Biology* **2002**, 9, 646.
14. G. Zhao, J. R. Perilla, E. L. Yufenyuy, X. Meng, B. Chen, J. Ning, J. Ahn, A. M. Gronenborn, K. Schulten, C. Aiken and P. Zhang, *Nature* **2013**, 497, 643-646.
15. J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R.

D. Skeel, L. Kale and K. Schulten, *J Comput Chem* **2005**, 26, 1781-1802.

16. C. C. J. Roothaan, *Reviews of Modern Physics* **1951**, 23, 69-89.
17. A. Szabo and N. S. Ostlund, *Modern quantum chemistry: introduction to advanced electronic structure theory*, Courier Corporation, 2012.
18. D. R. Hartree, *Mathematical Proceedings of the Cambridge Philosophical Society* **1928**, 24, 426-437.
19. V. Fock, *Zeitschrift für Physik* **1930**, 61, 126-148.
20. C. C. J. Roothaan, *Reviews of Modern Physics* **1960**, 32, 179-185.
21. R. J. Bartlett, *Annual review of physical chemistry* **1981**, 32, 359-401.
22. S. Saebo and P. Pulay, *Annual Review of Physical Chemistry* **1993**, 44, 213-236.
23. W. Kohn, *Reviews of Modern Physics* **1999**, 71, 1253-1266.
24. C. David Sherrill and H. F. Schaefer, in *Advances in Quantum Chemistry*, eds. P.-O. Löwdin, J. R. Sabin, M. C. Zerner and E. Brändas, Academic Press, 1999, vol. 34, pp. 143-269.
25. C. J. Cramer, *Essentials of computational chemistry: theories and models*, John Wiley & Sons, 2013.
26. F. Coester, *Nuclear Physics* **1958**, 7, 421-424.
27. F. Coester and H. Kümmel, *Nuclear Physics* **1960**, 17, 477-485.
28. C. Møller and M. S. Plesset, *Physical Review* **1934**, 46, 618-622.
29. P. Hohenberg and W. Kohn, *Physical Review* **1964**, 136, B864-B871.
30. W. Kohn and L. J. Sham, *Physical Review* **1965**, 140, A1133-A1138.
31. A. Heifetz, G. Trani, M. Aldeghi, C. H. MacKinnon, P. A. McEwan, F. A. Brookfield, E. I. Chudyk, M. Bodkin, Z. Pei, J. D. Burch and D. F. Ortwine, *Journal of Medicinal Chemistry* **2016**, 59, 4352-4363.
32. T. Menzies and Y. Hu, *Computer* **2003**, 36, 22-29.
33. J. D. Hildebrand and P. Soriano, *Molecular and Cellular Biology* **2002**, 22, 5296.

34. L. Breiman, *Machine Learning* **2001**, 45, 5-32.

35. I. Issemann and S. Green, *Nature* **1990**, 347, 645-650.

36. Y. Brélivet, N. Rochel and D. Moras, *Molecular and Cellular Endocrinology* **2012**, 348, 466-473.

37. B. P. Kota, T. H.-W. Huang and B. D. Roufogalis, *Pharmacological Research* **2005**, 51, 85-94.

38. W. Bourguet, P. Germain and H. Gronemeyer, *Trends in Pharmacological Sciences* **2000**, 21, 381-388.

39. H. B. Rubins, J. Davenport, V. Babikian, L. M. Brass, D. Collins, L. Wexler, S. Wagner, V. Papademetriou, G. Rutan and S. J. Robins, *Circulation-Hagertown* **2001**, 103, 2828-2833.

40. A. L. Catapano, I. Graham, G. De Backer, O. Wiklund, M. J. Chapman, H. Drexel, A. W. Hoes, C. S. Jennings, U. Landmesser, T. R. Pedersen, Ž. Reiner, G. Riccardi, M.-R. Taskinen, L. Tokgozoglu, W. M. M. Verschuren, C. Vlachopoulos, D. A. Wood and J. L. Zamorano, *Atherosclerosis* **2016**, 253, 281-344.

41. S. Raza-Iqbal, T. Tanaka, M. Anai, T. Inagaki, Y. Matsumura, K. Ikeda, A. Taguchi, F. J. Gonzalez, J. Sakai and T. Kodama, *Journal of Atherosclerosis and Thrombosis* **2015**, 22, 754-772.

42. N. Hennuyer, I. Duplan, C. Paquet, J. Vanhoutte, E. Woitrain, V. Touche, S. Colin, E. Vallez, S. Lestavel, P. Lefebvre and B. Staels, *Atherosclerosis* **2016**, 249, 200-208.

43. S. Ishibashi, S. Yamashita, H. Arai, E. Araki, K. Yokote, H. Suganami, J.-C. Fruchart and T. Kodama, *Atherosclerosis* **2016**, 249, 36-43.

44. J.-C. Fruchart, *Cardiovascular Diabetology* **2013**, 12, 82.

45. K. Takei, S.-i. Han, Y. Murayama, A. Satoh, F. Oikawa, H. Ohno, Y. Osaki, T. Matsuzaka, M. Sekiya, H. Iwasaki, S. Yatoh, N. Yahagi, H. Suzuki, N. Yamada, Y. Nakagawa and H. Shimano, *Journal of Diabetes Investigation* **2017**, 8, 446-452.

46. Molecular Operating Environment (MOE), 2013.08 Chemical Computing Group ULC, 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7.

47. H. E. Xu, M. H. Lambert, V. G. Montana, K. D. Plunket, L. B. Moore, J. L. Collins, J. A. Oplinger, S. A. Kliewer, R. T. Gampe Jr and D. D. McKee, *Proceedings of the National Academy of Sciences* **2001**, 98, 13919-13924.

48. Y. Li, A. Kovach, K. Suino-Powell, D. Martynowski and H. E. Xu, *Journal of Biological Chemistry* **2008**, 283, 19132-19139.

49. J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser and C. Simmerling, *J Chem Theory Comput* **2015**, 11, 3696-3713.

50. D. A. Case, V. Babin, J. Berryman, R. Betz, Q. Cai, D. Cerutti, T. Cheatham Iii, T. Darden, R. Duke, H. Gohlke, A. W. Goetz, S. Gusarov, N. Homeyer, P. Janowski, J. Kaus, I. Kolossváry, A. Kovalenko, T. S. Lee, S. LeGrand, T. Luchko, R. Luo, B. Madej, K. M. Merz, F. Paesani, D. R. Roe, A. Roitberg, C. Sagui, R. Salomon-Ferrer, G. Seabra, C. L. Simmerling, W. Smith, J. Swails, Walker, J. Wang, R. M. Wolf, X. Wu and P. A. Kollman, **2014**.

51. M. Valiev, E. J. Bylaska, N. Govind, K. Kowalski, T. P. Straatsma, H. J. J. Van Dam, D. Wang, J. Nieplocha, E. Apra, T. L. Windus and W. A. de Jong, *Computer Physics Communications* **2010**, 181, 1477-1489.

52. T. Ishikawa and K. Kuwata, *Chemical Physics Letters* **2009**, 474, 195-198.

53. D. E. Bernholdt and R. J. Harrison, *Chemical Physics Letters* **1996**, 250, 477-484.

54. A. Bernardes, P. C. T. Souza, J. R. C. Muniz, C. G. Ricci, S. D. Ayers, N. M. Parekh, A. S. Godoy, D. B. B. Trivella, P. Reinach, P. Webb, M. S. Skaf and I. Polikarpov, *Journal of Molecular Biology* **2013**, 425, 2878-2893.

55. Y. Yamamoto, K. Takei, S. Arulmozhiraja, V. Sladek, N. Matsuo, S. I. Han, T. Matsuzaka, M. Sekiya, T. Tokiwa, M. Shoji, Y. Shigeta, Y. Nakagawa, H. Tokiwa and H. Shimano, *Biochem Biophys Res Commun* **2018**, 499, 239-245.

56. P. Cronet, J. F. W. Petersen, R. Folmer, N. Blomberg, K. Sjöblom, U. Karlsson, E.-L. Lindstedt and K. Bamberg, *Structure* **2001**, 9, 699-706.

57. M. Takeda, F. Seki, Y. Yamamoto, N. Nao and H. Tokiwa, *Curr Opin Virol* **2020**, 41, 38-45.

58. W. Qiu, Y. Zheng, S. Zhang, Q. Fan, H. Liu, F. Zhang, W. Wang, G. Liao and R. Hu, *Emerging infectious diseases* **2011**, 17, 1541.

59. Z. Sun, A. Li, H. Ye, Y. Shi, Z. Hu and L. Zeng, *Veterinary Microbiology* **2010**, 141, 374-378.

60. K. Sakai, N. Nagata, Y. Ami, F. Seki, Y. Suzaki, N. Iwata-Yoshikawa, T. Suzuki, S. Fukushi, T. Mizutani, T. Yoshikawa, N. Otsuki, I. Kurane, K. Komase, R. Yamaguchi, H. Hasegawa, M. Saijo, M. Takeda and S. Morikawa, *Journal of Virology* **2013**, 87, 1105.

61. N. Feng, Y. Liu, J. Wang, W. Xu, T. Li, T. Wang, L. Wang, Y. Yu, H. Wang, Y. Zhao, S. Yang, Y. Gao, G. Hu and X. Xia, *BMC Veterinary Research* **2016**, 12, 160.

62. F. Seki, N. Ono, R. Yamaguchi and Y. Yanagi, *Journal of Virology* **2003**, 77, 9943-9950.

63. M. Bieringer, J. W. Han, S. Kendl, M. Khosravi, P. Plattet and J. Schneider-Schaulies, *PLOS ONE* **2013**, 8, e57488.

64. K. Sakai, T. Yoshikawa, F. Seki, S. Fukushi, M. Tahara, N. Nagata, Y. Ami, T. Mizutani, I. Kurane, R. Yamaguchi, H. Hasegawa, M. Saijo, K. Komase, S. Morikawa and M. Takeda, *Journal of Virology* **2013**, 87, 7170-7175.

65. T. Hashiguchi, M. Kajikawa, N. Maita, M. Takeda, K. Kuroki, K. Sasaki, D. Kohda, Y. Yanagi and K. Maenaka, *Proceedings of the National Academy of Sciences* **2007**, 104, 19535-19540.

66. T. Hashiguchi, T. Ose, M. Kubota, N. Maita, J. Kamishikiryo, K. Maenaka and Y. Yanagi, *Nature Structural & Molecular Biology* **2011**, 18, 135-141.

67. C. Santiago, M. L. Celma, T. Stehle and J. M. Casasnovas, *Nature Structural & Molecular Biology* **2010**, 17, 124-129.

68. X. Zhang, G. Lu, J. Qi, Y. Li, Y. He, X. Xu, J. Shi, C. W. H. Zhang, J. Yan and G. F. Gao, *Nature Structural & Molecular Biology* **2013**, 20, 67-72.

69. H. Tatsuo, N. Ono, K. Tanaka and Y. Yanagi, *Nature* **2000**, 406, 893-897.

70. F. Kobune, H. Sakata and A. Sugiura, *Journal of Virology* **1990**, 64, 700-705.

71. F. Seki, Y. Yamamoto, H. Fukuhara, K. Ohishi, T. Maruyama, K. Maenaka, H. Tokiwa and M. Takeda, *Front Microbiol* **2020**, 11, 1830.

72. *Journal*.

73. A. W. Sousa da Silva and W. F. Vranken, *BMC Research Notes* **2012**, 5, 367.

74. M.J. Abraham, D. van der Spoel, E. Lindahl and B. Hess, GROMACS User Manual version 2018, www.gromacs.org.

75. B. Hess, H. Bekker, H. J. C. Berendsen and J. G. E. M. Fraaije, *Journal of Computational Chemistry* **1997**, 18, 1463-1472.

76. S. Nosé, *Molecular Physics* **1984**, 52, 255-268.

77. W. G. Hoover, *Physical Review A* **1985**, 31, 1695-1697.

78. M. Parrinello and A. Rahman, *Journal of Applied Physics* **1981**, 52, 7182-7190.

79. S. Nosé and M. L. Klein, *Molecular Physics* **1983**, 50, 1055-1076.

80. B. R. Miller, T. D. McGee, J. M. Swails, N. Homeyer, H. Gohlke and A. E. Roitberg, *Journal of Chemical Theory and Computation* **2012**, 8, 3314-3321.

81. A. Onufriev, D. Bashford and D. A. Case, *The Journal of Physical Chemistry B* **2000**, 104, 3712-3720.

82. D. R. Roe and T. E. Cheatham, *Journal of Chemical Theory and Computation* **2013**, 9, 3084-3095.

83. T. Ishikawa, Paics View, http://www.paics.net/paics_view_e.html.

84. T. Tokiwa, S. Nakano, Y. Yamamoto, T. Ishikawa, S. Ito, V. Sladek, K. Fukuzawa, Y.

Mochizuki, H. Tokiwa, F. Misaizu and Y. Shigeta, *Journal of Chemical Information and Modeling* **2018**, DOI: 10.1021/acs.jcim.8b00649.

85. T. Hunter, *Current Opinion in Cell Biology* **2009**, 21, 140-146.

86. M. T. Brown and J. A. Cooper, *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer* **1996**, 1287, 121-149.

87. C. L. Abram and S. A. Courtneidge, *Experimental Cell Research* **2000**, 254, 1-13.

88. R. Garcia, T. L. Bowman, G. Niu, H. Yu, S. Minton, C. A. Muro-Cacho, C. E. Cox, R. Falcone, R. Fairclough, S. Parsons, A. Laudano, A. Gazit, A. Levitzki, A. Kraker and R. Jove, *Oncogene* **2001**, 20, 2499-2513.

89. R. Roskoski, *Pharmacological Research* **2015**, 94, 9-25.

90. Z. Yao, K. Darowski, N. St-Denis, V. Wong, F. Offensperger, A. Villedieu, S. Amin, R. Malty, H. Aoki, H. Guo, Y. Xu, C. Iorio, M. Kotlyar, A. Emili, I. Jurisica, B. G. Neel, M. Babu, A.-C. Gingras and I. Stagljar, *Molecular cell* **2017**, 65, 347-360.

91. J. Rivera-Torres and E. San José, *Frontiers in Pharmacology* **2019**, 10.

92. W. Xu, A. Doshi, M. Lei, M. J. Eck and S. C. Harrison, *Molecular Cell* **1999**, 3, 629-638.

93. D. Shukla, Y. Meng, B. Roux and V. S. Pande, *Nat Commun* **2014**, 5, 3397.

94. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss and V. Dubourg, *the Journal of machine Learning research* **2011**, 12, 2825-2830.

95. Robert T. McGibbon, Kyle A. Beauchamp, Matthew P. Harrigan, C. Klein, Jason M. Swails, Carlos X. Hernández, Christian R. Schwantes, L.-P. Wang, Thomas J. Lane and Vijay S. Pande, *Biophysical Journal* **2015**, 109, 1528-1532.

96. L. C. Freeman, *Sociometry* **1977**, 40, 35-41.

97. N. R. Taylor, *Computational and Structural Biotechnology Journal* **2013**, 5, e201302006.

98. S. W. Cowan-Jacob, G. Fendrich, P. W. Manley, W. Jahnke, D. Fabbro, J. Liebetanz and T. Meyer, *Structure* **2005**, 13, 861-871.

99. J. J. Pavelites, J. Gao, P. A. Bash and A. D. Mackerell Jr, *Journal of Computational Chemistry* **1997**, 18, 221-239.

100. R. C. Walker, M. M. de Souza, I. P. Mercer, I. R. Gould and D. R. Klug, *The Journal of Physical Chemistry B* **2002**, 106, 11658-11665.

101. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, *The Journal of Chemical Physics* **1983**, 79, 926-935.

102. G. Bussi, F. L. Gervasio, A. Laio and M. Parrinello, *J Am Chem Soc* **2006**, 128, 13435-13441.

103. T. Darden, D. York and L. Pedersen, *The Journal of Chemical Physics* **1993**, 98, 10089-10092.

104. D. G. Fedorov and K. Kitaura, *The Journal of Chemical Physics* **2004**, 120, 6832-6840.

105. M. Gaus, Q. Cui and M. Elstner, *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2014**, 4, 49-61.

106. M. Gaus, A. Goez and M. Elstner, *Journal of Chemical Theory and Computation* **2013**, 9, 338-354.

107. M. Gaus, X. Lu, M. Elstner and Q. Cui, *Journal of Chemical Theory and Computation* **2014**, 10, 1518-1537.

108. H. Li, D. G. Fedorov, T. Nagata, K. Kitaura, J. H. Jensen and M. S. Gordon, *Journal of Computational Chemistry* **2010**, 31, 778-790.

109. S. Grimme, J. Antony, S. Ehrlich and H. Krieg, *The Journal of Chemical Physics* **2010**, 132, 154104.

110. S. Grimme, S. Ehrlich and L. Goerigk, *Journal of Computational Chemistry* **2011**, 32, 1456-1465.

111. E. Ozkirimli and C. B. Post, *Protein Sci* **2006**, 15, 1051-1062.

112. M. Shoji, T. Murakawa, S. Nakanishi, M. Boero, Y. Shigeta, H. Hayashi and T. Okajima, *Chemical Science* **2022**, 13, 10923-10938.

113. T. R. Stankiewicz, J. J. Gray, A. N. Winter and D. A. Linseman, *Biomolecular Concepts* **2014**, 5, 489-511.

114. J. Turner and M. Crossley, *BioEssays* **2001**, 23, 683-690.

115. V. Kumar, J. E. Carlson, K. A. Ohgi, T. A. Edwards, D. W. Rose, C. R. Escalante, M. G. Rosenfeld and A. K. Aggarwal, *Molecular Cell* **2002**, 10, 857-869.

116. T. Aoyagi, R. Yoshino, Y. Mitsuta, R. Morita, R. Harada and Y. Shigeta, *Chemistry Letters* **2021**, 51, 1-4.

117. M. R. Shirts, C. Klein, J. M. Swails, J. Yin, M. K. Gilson, D. L. Mobley, D. A. Case and E. D. Zhong, *Journal of Computer-Aided Molecular Design* **2017**, 31, 147-161.

118. J.-P. Ryckaert, G. Ciccotti and H. J. C. Berendsen, *Journal of Computational Physics* **1977**, 23, 327-341.

119. S. Miyamoto and P. A. Kollman, *Journal of Computational Chemistry* **1992**, 13, 952-962.

120. W. G. Hoover, K. Aoki, C. G. Hoover and S. V. De Groot, *Physica D: Nonlinear Phenomena* **2004**, 187, 253-267.

121. Y. Yamamoto, S. Nakano and Y. Shigeta, *Bulletin of the Chemical Society of Japan* **2022**, submitted.

Acknowledgements

I would like to take this opportunity to express my best gratitude to Professor Yasuteru Shigeta and Associate professor Ryuhei Harada for their assistance throughout my doctoral studies. I also acknowledge the collaborators, Prof. Hiroaki Tokiwa, Prof. Makoto Takeda, Prof. Hitoshi Shimano, Associate Prof. Shogo Nakano, Associate Prof. Takashi Ikawa and Dr. Vladimir Sladek for their assistance throughout my doctoral studies.

Publication List

◎: Doctoral thesis's paper, ○: First author.

・Konishi, H.; Matsubara, M.; Mori, K.; Tokiwa, T.; Arulmozhiraja, S.; Yamamoto, Y.; Ishikawa, Y.; Hashimoto, H.; Shigeta, Y.; Tokiwa, H.; Manabe, K., Mechanistic Insight into Weak Base-Catalyzed Generation of Carbon Monoxide from Phenyl Formate and Its Application to Catalytic Carbonylation at Room Temperature without Use of External Carbon Monoxide Gas. *Advanced Synthesis & Catalysis* **2017**, 359 (20), 3592-3601.

・Motoyama, T.; Nakano, S.; Yamamoto, Y.; Tokiwa, H.; Asano, Y.; Ito, S., Product Release Mechanism Associated with Structural Changes in Monomeric l-Threonine 3-Dehydrogenase. *Biochemistry* **2017**, 56 (43), 5758-5770.

・Akai, S.; Ikawa, T.; Kaneko, H.; Yamamoto, Y.; Arulmozhiraja, S.; Tokiwa, H., 3-(Triflyloxy)benzynes Enable the Regiocontrolled Cycloaddition of Cyclic Ureas to Synthesize 1,4-Benzodiazepine Derivatives. *Synlett* **2018**, 29 (07), 943-948.

◎ Yamamoto, Y.; Takei, K.; Arulmozhiraja, S.; Sladek, V.; Matsuo, N.; Han, S. I.; Matsuzaka, T.; Sekiya, M.; Tokiwa, T.; Shoji, M.; Shigeta, Y.; Nakagawa, Y.; Tokiwa, H.; Shimano, H., Molecular association model of PPARalpha and its new specific and efficient ligand, pemafibrate: Structural basis for SPPARMalpha. *Biochem Biophys Res Commun* **2018**, 499 (2), 239-245.

・Tokiwa, T.; Nakano, S.; Yamamoto, Y.; Ishikawa, T.; Ito, S.; Sladek, V.; Fukuzawa, K.; Mochizuki, Y.; Tokiwa, H.; Misaizu, F.; Shigeta, Y., Development of an Analysis Toolkit, AnalysisFMO, to Visualize Interaction Energies Generated by Fragment Molecular Orbital Calculations. *Journal of Chemical Information and Modeling* **2018**, 59(1), 25-30.

・Matsuzaka, T.; Kuba, M.; Koyasu, S.; Yamamoto, Y.; Motomura, K.; Arulmozhiraja, S.; Ohno, H.; Sharma, R.; Shimura, T.; Okajima, Y.; Han, S.-i.; Aita, Y.; Mizunoe, Y.; Osaki, Y.;

Iwasaki, H.; Yatoh, S.; Suzuki, H.; Sone, H.; Takeuchi, Y.; Yahagi, N.; Miyamoto, T.; Sekiya, M.; Nakagawa, Y.; Ema, M.; Takahashi, S.; Tokiwa, H.; Shimano, H., Hepatocyte Elovl6 determines ceramide acyl-chain length and hepatic insulin sensitivity in mice. *Hepatology* **2019**, doi.org/10.1002/hep.30953.

・Kawasaki, M.; Kambe, A.; Yamamoto, Y.; Arulmozhiraja, S.; Ito, S.; Nakagawa, Y.; Tokiwa, H.; Nakano, S.; Shimano, H., Elucidation of Molecular Mechanism of a Selective PPARalpha Modulator, Pemafibrate, through Combinational Approaches of X-ray Crystallography, Thermodynamic Analysis, and First-Principle Calculations. *Int J Mol Sci* **2020**, 21 (1).

・Takeda, M.; Seki, F.; Yamamoto, Y.; Nao, N.; Tokiwa, H., Animal morbilliviruses and their cross-species transmission potential. Curr Opin Virol 2020, 41, 38-45.

・Seki, F.; Yamamoto, Y.; Fukuhara, H.; Ohishi, K.; Maruyama, T.; Maenaka, K.; Tokiwa, H.; Takeda, M., Measles Virus Hemagglutinin Protein Establishes a Specific Interaction With the Extreme N-Terminal Region of Human Signaling Lymphocytic Activation Molecule to Enhance Infection. *Front Microbiol* **2020**, 11, 1830.

◎ Yamamoto, Y.; Nakano, S.; Seki, F.; Shigeta, Y.; Ito, S.; Tokiwa, H.; Takeda, M., Computational Analysis Reveals a Critical Point Mutation in the N-Terminal Region of the Signaling Lymphocytic Activation Molecule Responsible for the Cross-Species Infection with Canine Distemper Virus. *Molecules* **2021**, 26 (5), 1262.

○ Ikawa, T.; Yamamoto, Y.; Heguri, A.; Fukumoto, Y.; Murakami, T.; Takagi, A.; Masuda, Y.; Yahata, K.; Aoyama, H.; Shigeta, Y.; Tokiwa, H.; Akai, S., Could London Dispersion Force Control Regioselective (2 + 2) Cyclodimerizations of Benzynes? YES: Application to the Synthesis of Helical Biphenylenes. *Journal of the American Chemical Society* **2021**, 143 (29), 10853-10859.

・Sladek, V.; Yamamoto, Y.; Harada, R.; Shoji, M.; Shigeta, Y.; Sladek, V., pyProGA—A PyMOL plugin for protein residue network analysis. *PLOS ONE* **2021**, 16 (7), e0255167.

○ Yuta Yamamoto, Tomoharu Motoyama, Chiharu Ishida, Fumihito Hasebe, Yasuteru

Shigeta, Sohei Ito and Shogo Nakano, Biochemical and structural analysis of bona fide ancestral L-Lys a-oxidase to predict molecular evolution of substrate specificity. *ACS Omega* **2022**, 7 (48), 44407-44419.

◎ Yuta Yamamoto, Shogo Nakano and Yasuteru Shigeta, Dynamical interaction analysis of proteins by a random forest-fragment molecular orbital (RF-FMO) method and application to Src tyrosine kinase. *Bulletin of the Chemical Society of Japan* **2023**, doi.org/10.1246/bcsj.20220304.

◎ Yuta Yamamoto and Yasuteru Shigeta, Theoretical Study on the Regulating Mechanism of the Transition Between the Open-closed State of hCtBP2: A Combined Molecular Dynamics and Quantum Mechanical Interaction Analysis. *Chemistry Letters* **2023**, Accepted.