Computational studies of effects of surface charge on biomembrane and membrane permeability of molecules

Takahashi Teruyuki

February 2022

Computational studies of effects of surface charge on biomembrane and membrane permeability of molecules

Takahashi Teruyuki

Doctoral Program in Physics

Submitted to the Graduate School of

Pure and Applied Sciences

in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy in

Science

at the

University of Tsukuba

Table of contents

Chapter 6 Conclusion

General Conclusion

Acknowledgements

List of Publications

Chapter 1: Introduction

1-1 Transitions of drug development

Drugs that maintain human health have been of great interest in life science. Since ancient times, drugs have been used through medicinal herbs. For example, the antipyretic properties of willow bark have been known since the time of the ancient Greeks and Romans, and it has been widely used as a medicine in many parts of the world since then. In 1897, Felix Hoffmann of Bayer AG isolated the active ingredient from willow bark that was responsible for the antipyretic effect and named it salicin. The molecular structure of salicin was also elucidated. Later, in order to increase the efficacy of salicin, a derivative with a modified side chain was produced and named Aspirin **(Ref. 1-1)**. Aspirin is still a useful drug used worldwide as an antipyretic analgesic and rheumatism medicine. Since then, not only aspirin, but also other drugs have been developed by changing the side chains of existing drug molecules to enhance their efficacy, which has been a major research method for a long time.

Altering chemical groups toward lead compounds has been used mainly for small molecule drugs that are easy to synthesize and modify chemically, and these small molecule drugs are drugs with a molecular weight of 500Da or less. From the 1980's to the early 2000's, the development of these small molecule drugs was actively pursued mainly in the field of lifestyle-related diseases such as hypercholesterolemia, hypertension, and diabetes.

However, because they are the small molecules, there is a limit to the number of derivatives that can be made, and it was soon said that there was a limit to the development of new small-molecule drugs. Since the 2000s, antibody drugs with large molecular weights have played a leading role in the development of new drugs, and of course, because of their large molecular weights, it is possible to create a variety of drugs by modifying them suitable for target proteins. Nevertheless, the cost of development is enormous, leading to increased medical costs and disparities in medical care. On the other hand, middle-sized molecules, which are intermediate in size between small molecules and antibodies, receive much attention from experimental and theoretical viewpoints, but in any case, there are high expectations for new drugs for modification.

1-2 How to increase the efficacy of drugs

When enhancing the efficacy of a drug, the most important thing to focus on is to create a binding site that specifically binds to the target receptor. In the case of small molecule drugs, the goal is to create a key binding site that fits into the target receptor, which is often likened to the relationship between a key and a lock. For this purpose, a part of the drug molecule is modified to create a structure that better fits the binding site. The same is true for antibody drugs, where the objective is to create a binding site that specifically binds to the antigen.

On the other hand, there is another requirement for increasing the efficacy of drugs, i.e., increasing the permeability of the drug toward the cell membrane. This means that if there is a target in the cell, no matter how effective the drug is against the target, it will be completely useless if it cannot pass through the cell membrane. For example, insecticides are powerful enough to kill insects **(Ref.1-2)**, but they cannot pass through mammalian cell membranes, making them completely ineffective against mammals. In addition, the blood vessels leading to the human brain have a membrane called the blood-brain barrier, and if a drug is to act on the brain, for example, it must not only have an effect on the target but must also be able to pass through the blood-brain barrier to be effective. Although the above example is an extreme case, it is very important to increase the cell membrane permeability of drugs in order to enhance their efficacy.

1-3 Methods to improve membrane permeability

The following three methods can be used to evaluate membrane permeability.

(1) Synthesizing a number of partially modified derivatives of existing drugs, passing them through biological or artificial membranes, and evaluating them experimentally. **(Ref. 1-3)**

(2) To model the membrane permeability process, one needs to create a theory of membrane permeability by using several physical laws and evaluate the membrane permeability by applying the theoretical methods. **(Ref. 1-4, 1-5)**

(3) Evaluating the membrane permeability by using computer simulations and informatics. (**Refs. 1-5,1-6**)

Method (1) is the one that has been used for a long time. It is the most reliable method because it is based on actual synthesized drugs and is still the main method used today. However, it is a very inefficient method in terms of cost and time since it involves the actual synthesis of thousand millions of possible drugs.

The method described in (2) has also been developed by many researchers over the years. For example, Ernest Overton and others (**Ref. 1-4**) have attempted to investigate the permeability of cell membranes () by assuming that they are made of a single homogeneous plate, two layers, or several layers and applying the laws of physics to them. However, in physics, it is impossible to describe the motion of three or more particles interacting with each other algebraically, so several approximations must be used, which limits the ability to investigate membrane permeability in detail.

The method in (3) is said to be in between experiment and pure theory. That is, it is less accurate than the experiment, but it can be used to predict complicated phenomena that cannot be solved analytically. However, these computer-based methods are not perfect. Of course, if quantum mechanics and Newtonian mechanics could be completely correct and the amount of computation resources were infinite, the perfect prediction would be possible. Actually, the amount of computer resources on a computer is limited, and it is not possible to calculate everything. Therefore, it is important to select and use the best computational method within the limited computational resources to evaluate the membrane permeability of molecules, as described below.

1-4. Calculation Method and Scope of Application

As described before, in computer science, which calculates the movement of matter from the movement of microscopic atoms, the number of calculations is small if the number of atoms to be observed is small, and the movement and reaction can be investigated quite accurately. Therefore, in order to investigate biological molecules, we have to make some efforts to shorten the calculation time. The typical calculation methods are given as follows.

○ Quantum mechanical calculation (Ref. 1-8～1-10)

The molecules are comprised of nuclei and electrons. Therefore, in order to investigate the behavior of a given molecule accurately, one must use quantum mechanics to calculate the electronic structure of the molecules at the microscopic level. Therefore, it can only be applied to small molecules (molecules consisting of a few hundreds of atoms at most using ubiquitous computer resources) and requires a long computation time. Since the biomolecules in the living body interact with the surrounding environment, such as water, other biomolecules, and so on, one must model the surrounding environment as accurately and effectively as possible. For example, the solvent effects are inevitable in evaluating the solvation free energy of the molecule.

○ **Molecular dynamics (MD) simulations** (Newtonian equation of motion calculation) **(Ref. 1-11)**

In this method, a molecule is regarded as an assembly of atoms, where each atom is represented as a charged particle. Among atoms in a molecule, potential energies for bonds, angles, and torsions are modeled by analytic functions. Among the atoms between two molecules, the potentials are also modelled by Coulombic interaction and van der Waals interaction. Newton's equation of motion is applied to them for the calculation of their dynamics using predefined force field models. However, the accuracy of the calculation is lower than that of quantum mechanics because the electronic structure is not evaluated, and the defined force field parameters are used to model the molecules.

○ **Empirical method (Ref. 1-12)**

This method obtains results from a statistical point of view rather than calculations using physical laws. For example, when investigating the efficacy of a drug, the efficacy of a new drug is deduced from past data. It is especially used for dealing with vast numbers of molecules to be analyzed and screen out undesired ones. Nowadays, new methods such as neural networks (**Ref.1-13**) have appeared and are being actively studied.

For the phenomena around the cell membrane, the MD simulations can be used because it consists of more than hundred thousand of atoms. However, the membrane permeation of drugs is called a rare event, which is a phenomenon that exceeds the applicable time scale in the MD simulation. Empirical methods can be applied since there are no size or time constraints, but from a physical point of view, their accuracy may be compromised.

1-5 Water molecules around the cell membrane

In the world we live in, gravity plays a more important role than the electrostatic force between water molecules, which creates surface tension, etc. However, in the atomic size world, especially in water, the effects of electrostatic force and van der Waals force among atoms plays a more important role than gravity. For example, if one tries to pass water molecules through a glass hole with a diameter of a few nm, one will find that it is not easy to do so, because the water molecules bind to the polarity of the glass surface and also bind to each other in a network, so they cannot move at all. In this case, fluorine coating of the glass surface allows water molecules to pass through, but in this way, the existence of water molecules is significant in the microscopic world. Therefore, it is important to understand their physical properties in detail. It is also very

important to investigate the properties of the surface between water and membrane when considering the permeability of drugs to cell membranes.

However, it is difficult to study the behavior of water molecules on cell membranes and drug surfaces. It is impossible to observe them directly using either optical or electron microscopes, and it is also impossible to predict them analytically because they are many body systems composed of vast atoms. Therefore, computer simulations are the most effective research method. As I mentioned earlier, I thought it was essential to investigate the behavior of water molecules on the surface of the drug in order to study the permeability of the drug through the cell membrane, and I conducted my research prior to other studies.

1-6 Scope of this thesis

With the above in mind, I have studied the physical properties of molecules and cell membrane in aqueous solution at the atomic level. Because the computer resources are limited, I treat the cell membrane as a lipid bilayer model with periodic boundary condition. In order to show the modeling of the system and currently available computer simulation methods, I summarize the basic theories for the analytical and computer simulation methods and software in Sec. 2. In Sec. 3, I have performed the MD simulations of a single lipid layer with negative charge head group in solution, in order to understand the interaction of the charged ions interact with the lipid surface. In Sec. 4, I have evaluated the partition coefficient of molecules between water and *n*-octanol, $\mathrm{Log}P_{\mathrm{o/w}}$ which is often an indicator for the membrane permeability, by using quantum mechanical calculations combined with the empirical method. In Sec. 5, I tried to evaluate the membrane permeability of Bottromycin and its derivatives using several methodologies to get some knowledge of the current best available method to evaluate it. Finally, I summarize the thesis in Sec. 6.

1-7 Reference

1. Y. Nomura, K. Nomura, Y. Homma, *A Series of New Chemistry* (in Japanese), Suken Shuppan (2021).
2. T. Suzuki, T. Hongawa, I. Washitani, *Series of New Biology* (in Japanese), Suken Shuppan (2021).
3. C. Masungi, J. Mensch, A.Van Dijck, C. Borremans, B. Willems, C. Mackie, M. Noppe, E. Brewster, "Parallel Artificial Membrane Permeability Assay (PAMPA) combined with a 10-day multiscreen Caco-2 cell culture as a tool for assessing new drug candidates", *Pharmazie* **63**, 194-199 (2008).
4. D. Deamer, A. Kleinzeller, D. Fambrough, ed. "Membrane permeability: 100 years since Ernest Overton", New York: Elsevier (1999).
5. C. Hannesschlaeger, A. Horner, and P. Pohl, "Intrinsic Membrane Permeability to Small Molecules", *Chem. Rev*. **119**, 5922-5953(2019).
6. Y. Fukunishi, T. Mashimo, T. Kurosawa, Y. Wakabayashi, H. K. Nakamura, K. Takeuchi, "Prediction of Passive Membrane Permeability by Semi- Empirical Method Considering Viscous and Inertial Resistances and Different Rates of Conformational Change and Diffusion", *Mol. Inf.*, **39**, (2020).
7. W. Shinoda, "Permeability across lipid membranes", *Biochim. Biophys. Acta* **1858**, 2254-2265 (2016).
8. W. Kohn; L. J. Sham "Self-Consistent Equations Including Exchange and Correlation Effects". *Phys. Rev.* **140** (4A), A1133-1138 (1965).
9. R. G. Parr; W. Yang, "Density-Functional Theory of Atoms and Molecules" (International Series of Monographs on Chemistry Book 16), Oxford University Press, (1989).
10. T. Tsuneda, "Basics of Density Functional Thoery", (in Japanese), Kodansha Scientific books (2013).
11. B.J. Alder, T. E. Wainwright, "Studies in Molecular Dynamics. I. General Method". *J. Chem. Phys.* **31** (2), 459 (1959).
12. E. Spiga, M.T. Degiacomi, M. Dal Peraro, "New Strategies for Integrative Dynamic Modeling of Macromolecular Assembly", in T. Karabencheva-Christova, ed., Biomolecular Modelling and Simulations. *Adv. Protein Chem. Struct. Biol.* **96**. Academic Press 77-111 (2014).
13. Y. Saito, "Deep Learning from Scratch - Theory and Implementation of Deep Learning in Python", O'Reilly, Japan (2016).

Chapter 2 Target and Theoretical Background

2-1 Phospholipid bilayer model (Ref. 2-1)

Some biological membranes are composed of phospholipid bilayers, in which phospholipid molecules (**Fig. 2-1(a)**) are arranged in a double layer with the hydrophilic group on the outside and the hydrophobic group on the outside to form a single membrane (**Fig. 2-1(b)**). In reality, different species of organisms use different composition of several phospholipids.

When a substance passes through the phospholipid bilayer by diffusion, the way in which the particles pass through the phospholipid bilayer differs depending on the substance. Therefore, uncharged water, small hydrophobic oxygen, nitrogen, carbon dioxide, urea, and glycerol can easily pass through the phospholipid bilayer, while charged ions, large hydrophobic glucose, and proteins cannot pass through (**Fig. 2-1(c)**).

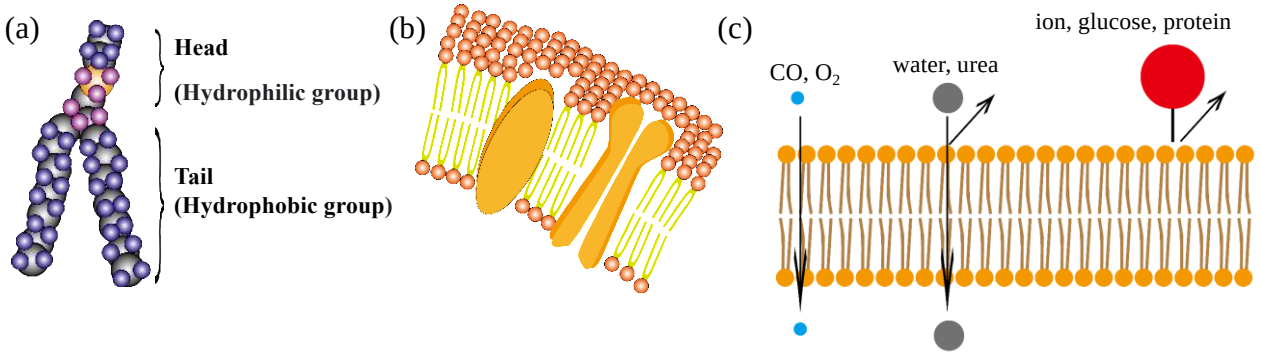


Figure 2-1 (a) molecular structure of phospholipid, (b) cell membrane, and (c) membrane permeation of molecules across the cell membrane.

2-2 Electric double-layer model (Refs. 2-2 and 2-3)

The electrical double layer is the structure of two layers formed when an electric potential is applied to the interface of an object in a fluid. In the past, it was used to describe the appearance of ions in a fluid on a metal surface given an electric potential in electrolysis, but it was also used to describe the appearance of ions in a solution in contact with an electrically charged biological membrane. The theory of this electric double layer has been constructed by Helmholtz, Gouy-Chapman, and Stern. According to their theory **(Ref. 2-3)**, the following two area were defined to distinguish the behavior in the ion concentration.

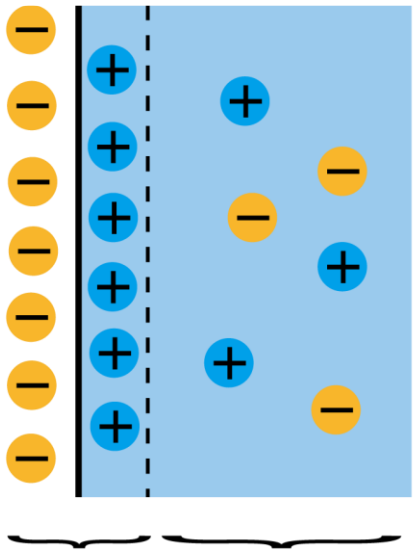


Figure 2-2. Electric double-layer model

① In a solution in contact with an electrically charged metal, counter ions line up to form a layer (electric double layer), which is called the Stern layer or Helmholtz layer as shown in **Fig. 2-2**.

② This layer is called the diffusion layer or the Gouy-Chapman layer, in which counter ions and co-ions diffuse with a difference in concentration. The thicknesses of the Stern and Gouy-Chapman layers are also predicted by the theory, but it is not known whether the theory is correct or not because many approximations are made during the calculation.

2-3 Debye-Hückel model for the electric bilayer (Refs. 2-2 and 2-3)

The behavior of the interaction of ions in electrolyte solutions was analyzed statistically by Peter Debye and Erich Hückel. The activity a_i of a certain component (ion) in the electrolyte and ion concentration c_i are related by the following relationship when the activity coefficient f_i as

$$a_i = f_i \cdot c_i \tag{2-1}$$

The activity coefficient f_i can be written as the following. This is called the extended Debye- Hückel equation:

$$\ln f_i = -\frac{z_i^2 e^2}{8\pi\varepsilon kT} \cdot \frac{\kappa}{1+\kappa r_i} \tag{2-2}$$

$$\kappa = \left(\frac{2N_A e^2 I}{\varepsilon kT}\right)^{\frac{1}{2}} \tag{2-3}$$

where r_i is the ionic radius, e is the elementary charge, and $\varepsilon(=\varepsilon_r\varepsilon_0)$ is the dielectric constant of water, k is Boltzmann's constant, T is temperature, and N_A is the Avogadro constant. The ionic strength I is characterized by c_i and z_i (the valence charge of the ion) as follows.

$$I = \frac{1}{2}\sum_i c_i z_i^2 \quad \cdots\cdot(2\text{-}4) \tag{2-4}$$

The activity coefficient f_i can also be written as

$$\log f_i = -\frac{Az_i^2\sqrt{I}}{1+Br_i\sqrt{I}} \tag{2-5}$$

$$A = \left(\frac{e^2}{4\varepsilon kT}\right)^{\frac{3}{2}}\left(\frac{2N_A}{\pi^2}\right)^{\frac{1}{2}}\frac{1}{\ln 10} \tag{2-6}$$

$$B = \sqrt{\frac{2N_A e^2}{\varepsilon kT}} \tag{2-7}$$

The inverse of κ

$$\kappa^{-1} = \sqrt{\frac{\varepsilon kT}{2N_A e^2 I}} \tag{2-8}$$

is a typical distance where the effect of electrolysis by the charge of ions becomes small due to the influence of surrounding ions, and is called the Debye length.

2-4 Middle-sized molecular drugs (Ref. 2-4～Ref. 2-6)

Middle-sized molecular drugs have molecular weights between 50 kDa and 150 kDa. In the past, many low-molecular-weight drugs with molecular weights of 50 kDa or less have been studied, but in recent years, the middle-sized molecular drugs with larger molecular weights have been attracting attention. Because of their large molecular weight, the number of possible structures (conformations) of these drugs is enormous compared to that of small drugs, and the experimental evaluation of their physical properties has not yet progressed. Therefore, the comparison with theoretical calculations, such as molecular simulations, has not been made to a large extent, and from the viewpoint of computational science, the subject is still in its infancy.

2-4-1 Special cyclic peptides (Ref. 2-6)

Special cyclic peptides are one of the most interesting systems among middle-sized molecular drugs. The classical (linear) peptide drugs or classical peptide compounds have several drawbacks as

(1) The structures of compound comprised of 10 to 20 amino acid residues are complex, and it is difficult to obtain structure-activity relationships in the pursuit of drug activity.

(2) It takes time to synthesize these peptides.

(3) Weak against gastric acid and intestinal enzymes and it cannot be administered orally.

(4) Almost no membrane permeability.

There were many weaknesses in such situation **(Ref. 2-7)**, special cyclic peptides have emerged as a new group of peptide compounds. One of the most famous special cyclic peptides is the naturally made cyclosporine. Cyclosporine is an immunosuppressive drug that inhibits the activation of calcineurin in T cells and suppresses the production of cytokines, which are low molecular weight proteins secreted by the cells **(Ref. 2-8)**. In addition to intravenous administration, cyclosporine is mainly used as an orally administered drug. The chemical structure of cyclosporine indicates that this kind of the special cyclic peptide have following characteristics:

(a) The peptide has a cyclic structure instead of a linear chain structure.

(b) The amide nitrogen in the main chain of the peptide backbone is sometimes methylated.

(c) Contains amino acids with non-natural type side chains and D-body amino acids

Recently, an artificial photosynthetic system has been established to synthesize special cyclic peptides with infinite chain length by mRNA display, and it has become possible to easily obtain structural diversity **(Ref. 2-9)**. Therefore, if (i) activity and substrate specificity comparable to those of antibodies obtained from

structural features, (ii) blood stability that enables oral administration, and (iii) low production cost comparable to that of small molecules can be realized, the possibility of middle-molecule drug discovery will increase. However, due to the aforementioned problems, a computational chemistry method for predicting and evaluating experimental data from synthesized peptides has not yet been established.

2-5 Parallel artificial membrane permeability assay (PAMPA)

There are several systems to evaluate lipid bilayer permeability. One of them is a method using cultured cells (Caco-2, MDCK, and so on). However, the labor cost is high due to the need for culture and pretreatment before measurement, and it is difficult to evaluate insoluble compounds. Therefore, this method is not suitable for routine measurement in the early stage of drug discovery, which is still under development. The Parallel Artificial Membrane Permeation Assay (PAMPA) was established by Kansy *et al.* in 1998 as a simple method for screening **(Ref. 2-10)**. In PAMPA, phospholipid membranes coated on membrane filters are considered as a substitute for biological membranes, and the membrane permeability is evaluated. By changing the lipid composition, organic solvent, amount of lipid solution, and buffer conditions, membranes with various properties can be prepared for various applications. The operation of PAMPA is very simple, and a robot that automates the data analysis is commercially available, making it suitable for routine measurements.

2-6 Electronic Structure Calculation

2-6-1 First-Principles Calculation (Ref. 2-11)

First-principles calculations is a method to predict the motion of electrons in a material by solving the Schrödinger equation, and is then used to predict the physical properties and chemical reactions. However, as a practical matter, it is difficult to solve the Schrödinger equation accurately for a target molecule, because of the many-body problem. Hartree-Fock (HF) method and density functional theory (DFT) are two major mean-field based approaches to solve the many-body problem, in which the wave function is represented by the Slater determinant of orbitals and the one-electron approximation is performed. Using these methods, the Schrödinger equation is solved for molecular systems (**Ref. 2-12**). Moreover, DFT takes the electron correlation into account which is missing in HF method, though the appropriate model exchange-correlation functional is selected in DFT calculations (**Ref. 2-13**). Here I give a brief explanation of the first-principles methods.

2-6-1-1 Hartree-Fock method (Ref. 2-12)

HF equation is an equation for finding a set of one-electron molecular orbitals such that it is the best approximation to the ground state when the Hamiltonian eigenfunctions (wave functions) representing the many-electron system are approximated by a single Slater determinant (HF approximation).

$$-\frac{1}{2m}\nabla^2\varphi_i(x) + V_{\mathrm{H}}(x)\varphi_i(x) - \int dy V_{\mathrm{X}}(x,y)\varphi_i(y) = \epsilon_i\varphi_i(x) \tag{2-9}$$

The above equation gives an approximate solution of $\{\varphi_i\}$. By collecting several terms, HF equation is rewritten as

$$\hat{F}\varphi = \epsilon_i\varphi \tag{2-10}$$

where $\hat{F}$ is referred to as Fock operator and include the eigenfunction φ. Thus, it can not be solved as an ordinary eigenvalue equation. By applying the solution obtained by this method as an approximate solution and solving it recursively, one can determine the mean field potential, and the wave function of the electrons moving in the mean field is obtained a self-conscious manner.

2-6-1-2 Density functional theory (Ref. 2-14)

DFT is based on the Hohenberg-Kohn theorem, which states that physical properties such as energy of an electronic system can be calculated from the electron density.

2-6-1-2-1 Hohenberg-Kohn's theorem (Ref. 2-14)

Consider the following system with *N* electrons in an external potential.

$$\mathrm{H} = \frac{\hbar^2}{2m}\sum_{i=1}^{N}\nabla_i^2 + \frac{e^2}{4\pi\varepsilon}\sum_{i=1}^{N}\sum_{j=1}^{i-1}\frac{1}{|r_i - r_j|} + \sum_{i=1}^{N}V(r_i) \tag{2-11}$$

In the following, for simplicity, we use the atomic unit system ($\hbar$=e=m=1) and the coefficient of the Coulomb interaction between electrons is also set to 1.

In the above system, the following two theorems hold.

Theorem 1: The ground state energy E_G is uniquely determined by the one-electron density $\rho(\mathrm{r})$.

Theorem 2: Ground State Energy $E_G(\rho)$ gives the minimum value by electron density $\rho_0(\mathrm{r})$ that is in ground state, and which is equal to the ground state energy of the system.

Therefore, it is possible to obtain the ground state electron density by searching for the electron density that gives the minimum energy.

2-6-1-2-2 Kohn-Sham method (Ref. 2-15)

We consider a system in which *N* electrons are moving independently without interacting with each other. However, although there is no direct interaction between the electrons, the interaction from the mean field exists as a potential. The equation derived by Kohn and Sham based on the above conditions is the Kohn-Sham equation shown below

$$\left[-\frac{1}{2}\nabla^2 + v_{\mathrm{eff}}(\mathbf{r})\right]\psi_i(\mathbf{r}) = \varepsilon_i\psi_i(\mathbf{r}) \tag{2-12}$$

where KS effective potential $v_{\mathrm{eff}}(\mathbf{r})$ consists of the external potential $v_{\mathrm{ext}}(\mathbf{r})$, the classical Coulomb potential $v_{\mathrm{Hartree}}(\mathbf{r})$ with electrons, and the exchange-correlation potential $v_{\mathrm{xc}}(\mathbf{r})$ as

$$v_{\mathrm{eff}}(\mathbf{r}) = v_{\mathrm{ext}}(\mathbf{r}) + v_{\mathrm{Hartree}}(\mathbf{r}) + v_{\mathrm{xc}}(\mathbf{r}) \tag{2-13}$$

The electron density $\rho(\mathbf{r})$ is expressed by using set of occupied orbitals $\{\psi_i(\mathbf{r})\}$ as

$$\rho(\mathbf{r}) = \sum_i^N \psi_i^*(\mathbf{r})\,\psi_i(\mathbf{r}) \tag{2-14}$$

Using this electron density, $v_{\mathrm{Hartree}}(\mathbf{r})$ is evaluated from the Poisson equation as

$$\nabla^2 v_{\mathrm{Hartree}}(\mathbf{r}) = -4\pi\rho(\mathbf{r}) \tag{2-15}$$

Using equation (2-15), $v_{\mathrm{Hartree}}(r)$ is again substituted into in $v_{\mathrm{eff}}(r)$ in equation (2-13), and this iteration is performed until the input and output of the electron density $\rho(\mathbf{r})$ will much. At that time, the total electron energy $E_{\mathrm{KS}}[\rho]$ takes the minimum value. Such an iterative calculation method is called Self-Consistent Field (SCF) method, which is almost the same as HF method except for the exchange-correlation potential.

When the SCF calculation is achieved, the total electron energy $E_{KS}[\rho]$ is obtained by the converged density, $\rho(\mathbf{r})$, as

$$E_{KS}[\rho] = \sum_{i=1}^{N} \varepsilon_i - \frac{1}{2}\int \rho(\mathbf{r}) v_{\mathrm{Hartree}}(\mathbf{r}) d\mathbf{r} + E_{\mathrm{xc}}[\rho] - \int \rho(\mathbf{r})\, v_{\mathrm{xc}}(\mathbf{r}) d\mathbf{r} \qquad (2\text{-}16)$$

where $E_{\mathrm{xc}}[\rho]$ is the exchange-correlation energy functional, which is related to the exchange-correlation potential as $\frac{\delta E_{\mathrm{xc}}[\rho]}{\delta \rho(\mathbf{r})} = v_{\mathrm{xc}}(\mathbf{r})$. Since the Kohn-Sham method treats a hypothetical system without interaction, the potential differs from that of a real system with interaction. For the exchange-correlation potential $v_{\mathrm{xc}}(\mathbf{r})$, there are many approximate models. Note here that the results depend strongly on the choice of $v_{\mathrm{xc}}(\mathbf{r})$. The local density approximation (LDA) and generalized gradient approximation (GGA), which are often used in the calculation of bulk systems, might not be suitable for the molecular system. Next, I explain the hybrid functional method.

2-6-1-2-3 Hybrid functional method (B3LYP method) (Ref. 2-16)

The Hybrid functional method is a class of approximations to the exchange- correlation energy functional in Kohn-Sham DFT. It linearly couples (hybrids) the exchange and correlation energies obtained by non-empirical or empirical methods with the exact exchange energy from Hartree-Fock theory. This exchange energy functional is called an implicit functional because it is expressed in terms of Kohn-Sham orbitals rather than densities. One of the most commonly used hybrid function is B3LYP (Becke,3-parameter, Lee-Yang-Parr). The main portion of the exchange correlation functional is Hartree-Fock's exact exchange functional defined as

$$E_{\mathrm{x}}^{\mathrm{HF}} = -\frac{1}{2}\sum_{i,j}\iint \psi_i^*(\mathbf{r}_1)\psi_j^*(\mathbf{r}_1)\frac{1}{r_{12}}\psi_i(\mathbf{r}_2)\psi_j(\mathbf{r}_2)d\mathbf{r}_1 d\mathbf{r}_2 \qquad (2\text{-}17)$$

and the other portion comes from the LDA and GGA types of exchange functional. The B3LYP (Becke, 3-parameter, Lee-Yang-Parr)-exchange correlation functional is expressed by the following equation

$$E_{\mathrm{XC}}^{\mathrm{B3LYP}} = E_{\mathrm{X}}^{\mathrm{LDA}} + a_0\left(E_{\mathrm{X}}^{\mathrm{HF}} - E_{\mathrm{X}}^{\mathrm{LDA}}\right) + a_x\left(E_{\mathrm{X}}^{\mathrm{GGA}} - E_{\mathrm{X}}^{\mathrm{LDA}}\right) + E_{\mathrm{c}}^{\mathrm{LDA}} + a_{\mathrm{c}}(E_{\mathrm{c}}^{\mathrm{GGA}} - E_{\mathrm{c}}^{\mathrm{LDA}}) \qquad (2\text{-}18)$$
$$(a_0 = 0.20, a_x = 0.72, a_c = 0.81)$$

where $E_{\mathrm{x}}^{\mathrm{GGA}}$ and $E_{\mathrm{c}}^{\mathrm{GGA}}$ are the GGA exchange and the correlation functions and $E_{\mathrm{x}}^{\mathrm{LDA}}$ and $E_{\mathrm{c}}^{\mathrm{LDA}}$ are the LDA exchange and correlation functions, respectively. The parameters, $\{a_0, a_x, a_c\}$, determining the weighting of the individual functions are usually determined by fitting the predictions of the functions to experimental or precisely calculated thermochemical data. These three parameters defining B3LYP are incorporated without modification from Becke's original fitting of the analogous B3PW91 functional to a series of atomization energies, ionization potentials, proton affinities, and total atomic energies.

2-6-2 Polarization continuum model (Ref. 2-17, 2-18, 2-19)

When one wants to model the solvent mediated reaction explicitly, the reactant solute molecules and surrounding solvent molecules should be treated. In this situation, the computational cost becomes too large to handle it owing to vast solvent molecules. If one treats the solvent as a polarizable continuous medium, the first-principles calculation of the solute becomes possible. There are two types of polarizable models. One is Dielectric Polarization Continuum Model (D-PCM) which treats the polarizing continuum. The other is Conductor-like Polarization Continuum Model (C-PCM) which treats the continuum as dielectric medium, similar to the one used in conductor-like screening model (**Ref. 2-17**).

The solvation Gibbs energy of a molecule G_{sol} is expressed as the sum of three terms as follows.

$$G_{\mathrm{sol}} = G_{\mathrm{es}} + G_{\mathrm{dr}} + G_{\mathrm{cav}} \qquad (2\text{-}19)$$

where G_{es} is electrostatic term, G_{dr} is Dispersion force - repulsion term, and G_{cav} is cavitation term. As a variant of PCM, the solvation model based on density (SMD) method was developed by Truhlar *et al.* for the

prediction of the solvation free energy with specific parameters **(Ref. 2-19)**. SMD is often superior to the conventional PCM methods.

2-6-3 Gaussian16 (Refs. 2-20 and Ref. 2-21) as a program used for first-principles calculations

Gaussian is a software for quantum chemical calculation designed by John Pople in 1970. The name of this software comes from the Gaussian orbitals. He introduced them instead of Slater orbitals to speed up the calculations. The quantum chemical calculation package Gaussian16 software was used for first-principles calculations in Chapter 4 and 5, and the initial structures were created using GaussView, which is a graphical user interface for Gaussian.

2-7 Molecular Dynamics (MD) Simulations (Ref. 2-22)

2-7-1 Basics of MD simulations

Molecular dynamics (MD) simulation is a type of molecular simulation method that uses a computer to numerically integrate the equations of motion for each of the particles (atoms and molecules) that make up a material system. MD simulation keeps track with changes in position, velocity, energy, and other parameters over time. In general, the time evolution of the state of an *N*-body system is simulated by solving the Newtonian equations of motion.

$$m_i \frac{d^2}{dt^2}\mathbf{r}_i = -\nabla_i U(\mathbf{r}_1 \cdots \mathbf{r}_N) \qquad i = 1 \sim N \tag{2-20}$$

where m_i and r_i are the mass and the position of particle i, t is time, and U is the potential energy of the whole system. The potential energy is the sum of four energy components: (2-1) bond stretching energy in the molecule, (2-2) angular deformation energy of the bond angle, (2-3) torsion energy of the torsion angle, and (2-4) nonbonding interaction energy (Coulomb interaction and van der Waals (vdW) interaction).

The function that evaluates these energy components, together with the parameters that define the function, is called the molecular force field (FF). More simply, in the classical MD simulation of proteins, the molecules are described by a model in which atoms are represented as point charge and are connected to each other by springs whose strength is determined by the type (and combinations) of atoms connected. The molecular FF corresponds to the spring. The molecular force field differs depending on the application and is defined as a

group of empirical parameters according to the application. Typical types of molecular force fields are as follows:

AMBER (Assisted Model Building and Energy Refinement) **(Ref. 2-23)**

CHARMM (Chemistry at HARvard Macromolecular Mechanics) **(Ref. 2-24)**

MM2, MM3, and MM4 (Molecular Mechanics program 2, 3, or 4).

For example, the Amber force field is defined by

$$U(\{\mathbf{r}_i\}) = \sum_{\substack{\text{bonds}\\(ij)}} k_{\mathrm{r}}(r_{ij} - r_{\mathrm{eq}})^2 + \sum_{\substack{\text{angle}\\(ijk)}} \mathrm{k}_{\theta}(\theta_{ijk} - \theta_{\mathrm{eq}})^2 + \sum_{\substack{\text{dihedrals}\\(ijkl)}} \frac{\mathrm{V_n}}{2}\left[1 - \cos\left(n\phi_{ijkl} - \gamma\right)\right] + \sum_{\mathrm{i<j}} \left[\frac{A_{\mathrm{ij}}}{r_{ij}^{12}} - \frac{B_{\mathrm{ij}}}{r_{ij}^{6}}\right] + \sum_{i<j} \frac{q_i q_{\mathrm{j}}}{\varepsilon r_{ij}} \quad \cdot\cdot\cdot\cdot (2-21)$$

The first through third terms are related to the intermolecular potentials (bond, angle, and torsion), while the fourth and fifth terms represent non-bonding potentials. In the first term, r_{eq}, indicates the equilibrium distance between atoms associated with chemical bonding and k_r is the corresponding spring constant for this bond. In the second term, θ is the angle among the three atoms i, j, and k, connected by two chemical bonds. Also, θ_{ijk} and k_θ are the equilibrium angle and the corresponding spring constant for this angle. In the third term, ϕ is defined as the dihedral angle for four connected atoms with the three bonds. Also, ϕ_{ijkk} and V_n are the equilibrium torsion angle and the strength for this torsion angle. Since this term is represented by triangular function, n and g are parameters for the number of periodically identical positions and the phase shift, respectively. The fourth and fifth terms are vdW and Coulombic interactions acting between atoms i and j. A_{ij} and B_{ij} are the parameters for vdW interaction. q_i is the charge of i -th particle and e is the dielectric constant.

2-7-2 GROMACS (Ref. 2-25) as a program used for MD simulations

Gromacs is a software package for MD simulations developed at the University of Groningen. It is an open-source software and is programmed for parallel computing. The majority of the programs are written in C and are based on GROMOS, which was previously developed by the same group. GROMACS is very fast due to algorithmic and processor-specific optimizations and General-purpose Graphical Processing Unit (GPGPU) acceleration. GROMACS is very flexible due to its support for different force fields such as Amber, Charmm, and so on. In this work, I used GROMACS for entire MD calculations in Chapter 3 and 5.

2-8 Empirical Methods

It is a method to derive results statistically using past data rather than physical calculations in reality. Some methods are derived using equations, but multiple regression analysis and neural networks are also utilized.

2-8-1 PerMM (Refs. 2-26 and 2-27)

PerMM is a web-based calculation tool managed and operated by the University of Michigan. It calculates the molecular structure change during the membrane permeation and the permeability index of various membranes empirically by using the PDB file which records the atomic positions in the space. This system first calculates the membrane permeability of the drug in DOPC and then calculates the permeability in other membranes. In order to evaluate, the membrane resistance *R*, which is the inverse of the permeability coefficient *P*, is modelled by

$$R = \frac{1}{P} = \int_{-d/2}^{d/2} \frac{dz}{K(z)D(z)} \tag{2-22}$$

In the above equation, $K(z)$ is the distribution coefficient $K(z) = e^{-\Delta G} transf^{(z)}/RT$ and $D(z)$ is the diffusion coefficient $D(z) = k\frac{D_0}{ASA}$, where k, D_o, and ASA are a constant value, diffusion constants depending on the different types of membranes, and accessible surface area, respectively. Now, one defines P_Σ for different membrane types by fitting between experimental and calculated value by (2-22) and using the linear regression curves such as $LogP = a + bLogP_\Sigma$ $(a = LogkD_o)$ among experimental values for different membranes as

$$\log P_{\Sigma}^{\mathrm{BLM}} = -\log\left(ASA\int_{-d/2}^{d/2}\frac{dz}{K(z)}\right) \tag{2-22}$$

$$\log P_{\mathrm{calc}}^{\mathrm{BLM}} = 1.063\log P_{\Sigma}^{\mathrm{BLM}} + 3.669 \tag{2-23}$$

$$\log P_{0\mathrm{calc}}^{\mathrm{BBB}} = 0.375\log P_{\Sigma}^{\mathrm{BLM}} - 1.600 \tag{2-24}$$

$$\log P_{0\mathrm{calc}}^{\mathrm{Caco-2/MDCK}} = 0.272\log P_{\Sigma}^{\mathrm{BLM}} - 2.541 \tag{2-25}$$

Even for large molecules, the calculation results are output in a few hours. I used PerMM to evaluate the membrane permeability of Bottromycin in Chapter 5.

2-9 References

1. B. Alberts, A. D. Johnson, J. Lewis, D. Morgan, M. Raff, K. Roberts, P. Walter, *Molecular Biology of the Cell* (book).2015
2. K. Shirahama, G. Sugihara, *et al.*, "Fundamentals of Biophysical Chemistry - For Understanding Biological Phenomena" (in Japanese), Sankyo Shuppan Co. Ltd. (2003)
3. N. Ise, I. Sogami, "Structure formation in polymer physics giant ion systems" (in Japanese), Asakura Shoten Co., Ltd. (2004)
4. T. Rezai, B. Yu, G. L. Millhauser, M. P. Jacobson, R. S. Lokey, "Testing the Conformational Hypothesis of Passive Membrane Permeability Using Synthetic Cyclic Peptide Diastereomers", *J. Am. Chem. Soc.* **128**(8), 2510-2511 (2006).
5. S. Riniker, "Toward the elucidation of the mechanism for passive membrane permeability of cyclic peptides", Future Med. Sci., **11**(7), 637-639 (2019).
6. K. Shimizu, "Research on middle molecule drug discovery for the realization of next generation treatment and diagnosis that combines computational chemistry and experiment", Master thesis, Rikkyo University (2019).
7. K. Masuya, "New trends in drug discovery and development by constrained peptides", *Folia Pharmacologica Jpn.*, **148**, 322-328 (2016)
8. N.Parquet, E. Reignequ, "New oral formulation of cyclosporin A (Neoral) pharmacokinetics in allogeneic bone marrow transplant recipients", *Bone Marrow Transplant*, **25**, 965-968 (2000).
9. M. Ohuchi, H. Murakami, H. Suga, "The flexizyme system: a highly flexible tRNA aminoacylation tool for the translation apparatus", *Curr. Opin. Chem. Biol.*, **11**(5), 537-554 (2007).
10. K. Manfred, S. Frank, G. Klaus, "Physicochemical High Throughput Screening: Parallel Artificial Membrane Permeation Assay in the Description of Passive Absorption Processes", *Med. Chem.* **41**(7), 1007-1010 (1998).
11. 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, P.A. Kollman, "A Second-Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules", *J. Am. Chem. Soc.*, **117**(19), 5179-5197 (1995).
12. A. Szabo, N.S. Ostrund, "Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory" Chapter 2, Ed. Ostlund, Macmillan Publishing Co., Inc., U.S.A (1982).

13. H. Shirai, “First-principles methods and density functional theory” Chapter 3, Ed. Sanken Osaka University (2005). http://www.cmp.sanken.osaka-u.ac.jp/~koun/Lecs/dft.pdf (accessed 2020-1-20).
14. P. Hohenberg, W. Kohn, “Inhomogeneous electron gas”. *Phys. Rev.*, **136** (3B), B864–B871 (1964).
15. W. Kohn, L. J. Sham “Self-Consistent Equations Including Exchange and Correlation Effects”. *Phys. Rev.*, **140** (4A), A1133-1138 (1965).
16. A.D. Becke. “A new mixing of Hartree-Fock and local density-functional theories”. *J. Chem. Phys.* **98** (2), 1372-1377 (1993).
17. J. Tomasi, B. Mennucci, R.Cammi, "Quantum Mechanical Continuum Solvation Models." *Chem. Rev.* **105**(8), 2999-3094 (2005).
18. M. Cossi, N. Rega, G. Scalmani, V. Barone, "Energies, structures, and electronic properties of molecules in solution with the C-PCM solvation model." *J. Comput. Chem.* **24**(6), 669-681 (2003).
19. A. V. Marenich, C. J. Cramer, D. G. Truhlar, “Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions”, *J. Phys. Chem. B*, **113**(18), 6378-6396 (2009).
20. Gaussian 16, Revision A.03, M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, G. Scalmani, V. Barone, G.A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A.V. Marenich, J. Bloino, B.G. Janesko, R. Gomperts, B. Mennucci, H.P. Hratchian, J.V. Ortiz, A.F. Izmaylov, J.L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V.G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J.A. Montgomery, Jr., J.E. Peralta, F. Ogliaro, M.J. Bearpark, J.J. Heyd, E.N. Brothers, K.N. Kudin, V.N. Staroverov, T.A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A.P. Rendell, J.C. Burant, S.S. Iyengar, J. Tomasi, M. Cossi, J.M. Millam, M. Klene, C. Adamo, R. Cammi, J.W. Ochterski, R.L. Martin, K. Morokuma, O. Farkas, J.B. Foresman, D.J. Fox, Gaussian, Inc., Wallingford CT (2016).
21. M.J. Abraham, T. Murtola, R. Schulz, PállS, J.C. Smith, B. Hess, I.E. Lin-dah, “GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers”, *SoftwareX*, **1**(2), 19-25 (2015).

22. S. Tamar, "Pursuing Laplace's Vision on Modern Computers", Mathematical Approaches to Biomolecular Structure and Dynamics". *The IMA Volumes in Mathematics and its Applications*. **82**., 219-247. (1996).
23. Y. Duan, C. Wu, S. Chowdhury, M. C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, J. Caldwell, J. Wang, P. Kollman, "A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations". *J. Comput. Chem.* **24** (16), 1999-2012 (2003).
24. B.R. Brooks, R.E. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan, M. Karplus, "CHARMM: A program for macromolecular energy, minimization, and dynamics calculations". *J. Comput. Chem.* **4** (2), 187-217(1983).
25. H.J.C. Berendsen, D. van der Spoel, R. van Drunen. "GROMACS: A message-passing parallel molecular dynamics implementation". *Comput. Phys. Commun.* **91** (1-3): 43-56 (1995).
26. A.L. Lomize, Irina D. Pogozheva, "Physics-based method for modeling passive membrane permeability and translocation pathways of bioactive molecules", *J. Chem. Inf. Model.* **59**(7), 3198-3213 (2019).
27. A.L. Lomize, J.M. Hage, K. Schnitzer, K. Golobokov, M.B. LaFaive, A.C. Forsyth, I.D. Pogozheva, "PerMM: A web tool and database for analysis of passive membrane permeability and translocation pathways of bioactive molecules", *J. Chem. Inf. Model.* **59**(7), 3094-3099 (2019).

Chapter 3: Distribution of Counterions at the Negatively Charged Lipid/Water/Air Interface: A Molecular Dynamics Study

3-1. Introduction

Biological membranes are indispensable in the field of life science and have been extensively studied both experimentally and by theoretical analysis. In particular, biological membranes have a unique structure consisting of a lipid bilayer **(Ref.3-1,3-2)** whose composition varies depending on the cell type and function, and which separates the inner and outer regions of the cell. The mechanism by which biological membranes**(Ref.3-1,3-2)** selectively pass substances through membrane proteins and receive signals such as hormones is also a scientifically fascinating subject of research. Thus, understanding biological membranes leads not only to understanding the activities of cells, but also to understanding the individual as a whole. In fact, some insecticides take advantage of the differences between the cell membranes of insect and mammalian cells to produce drugs that are effective only on insect cells **(Ref.3-2)**. An important aspect of understanding biological membranes are the behavior of molecules at the interface between the biological membranes and the water layer. Since some phospholipids are charged, and biomolecules such as proteins and nucleic acids often have charged sites on their surfaces, the electrostatic potential should affect the static and dynamic behavior of biomolecules. Such environmental changes can affect biomolecules in several ways. For example, it has recently been shown that biological membranes serve as a site for αβ-amyloid formation **(Ref.3-1,3-2)**. Therefore, it is necessary to clarify the physical properties at the boundary. Theoretically, the charged surface (electrode) in solution and the ion distribution around it has been modeled by the static continuum theory based on the Poisson-Boltzmann equation **(Ref.3-3,3-4)**, but these models have been extended from the Debye-Hückel (DH) model **(Ref.3-5)** around charged particles in solution, such as the Gouy-Chapman model and the Stern model **(Ref.3-6)** as mentioned in Chapter 2.

Some of these models have also been extended to deal with biological membranes, but this has proved difficult to attempt. Compared to solid electrodes, biological membranes are very soft and the charge distribution of lipid molecules is not scattered evenly but sparse and point charge. On the other hand, atomistic

models, such as molecular dynamics (MD) simulations **(Ref.3-7)** using classical force fields, are easily applied not only to biological membranes, but also to more complex systems, such as membrane proteins in both membrane and water environments, to elucidate fundamental molecular mechanisms. Recent advances in experimental measurements, particularly sum-frequency generation spectroscopy (SFG) **(Ref. 3-8, 3-9, 3-10)**, have made it possible to measure changes in physical properties and dynamics at interfaces (see also some reviews of water interfaces). However, detailed analysis using atomic models has not yet been fully explored. In this chapter, I present MD simulations with sufficiently large cell sizes to understand the behavior of ions and water near and away from the interface and compare them with classical analytical methods. In particular, I revisit classical concepts of ion distribution, such as the Debye model, and discuss the orientation of ions to the P atoms of lipids at different solution concentrations and types of cations (Na^+ and Mg^{2+}).

3-1-1 Gouy-Chapman model

As mentioned earlier, an analytic solution for the distribution of ions in water in contact with an electrically charged polar plate has been obtained in previous studies by Gouy-Chapman and others. The formula for the original equation is Poisson-Boltzmann's formula shown as

$$\nabla \circ \left(\varepsilon(\mathbf{r})\nabla\psi(\mathbf{r})\right) = -4\pi\left(\rho^{\mathrm{f}}(\mathbf{r}) + \rho^{\mathrm{ion}}(\mathbf{r})\right) \tag{3-1}$$

In the above equation, ρ is the density of each ion, ψ is the electrostatic potential, and ε is the dielectric constant of the solvent. Since it cannot be solved exactly as it is, they made the following approximation:

- The charge spreads uniformly across the polar plate.
- ψ and ρ are the functions of the direction perpendicular to the polar plate, *x*, only.
- The dielectric constant $\varepsilon(\mathbf{r})$ of the solvent is constant regardless of the location.
- $\rho^{\mathrm{f}}(\mathbf{r})$ will be zero except at the origin.

If one introduces these assumptions, then ρ and ψ can be related to as

$$\rho^{\mathrm{ion}} = \rho^{+} - \rho^{-} = \rho^{+}_{(\infty)}\exp\{-\beta z^{+} e\psi\} - \rho^{-}_{(\infty)}\exp\{-\beta z^{-} e\psi\} \tag{3-2}$$

Using the above equation, Eq. (3-1) can be solved as

$$\varepsilon \frac{d^2\psi}{dx^2} = 8\pi z e n_0 \sinh(\beta z e \psi) \quad \Longrightarrow \quad \psi(x) = \frac{2}{ze\beta} \ln \frac{1+\gamma \exp(-\kappa x)}{1-\gamma \exp(-\kappa x)} \tag{3-3}$$

By substituting this ψ into equation (3-2), we can obtain ρ as illustrated in Fig. 3-1.

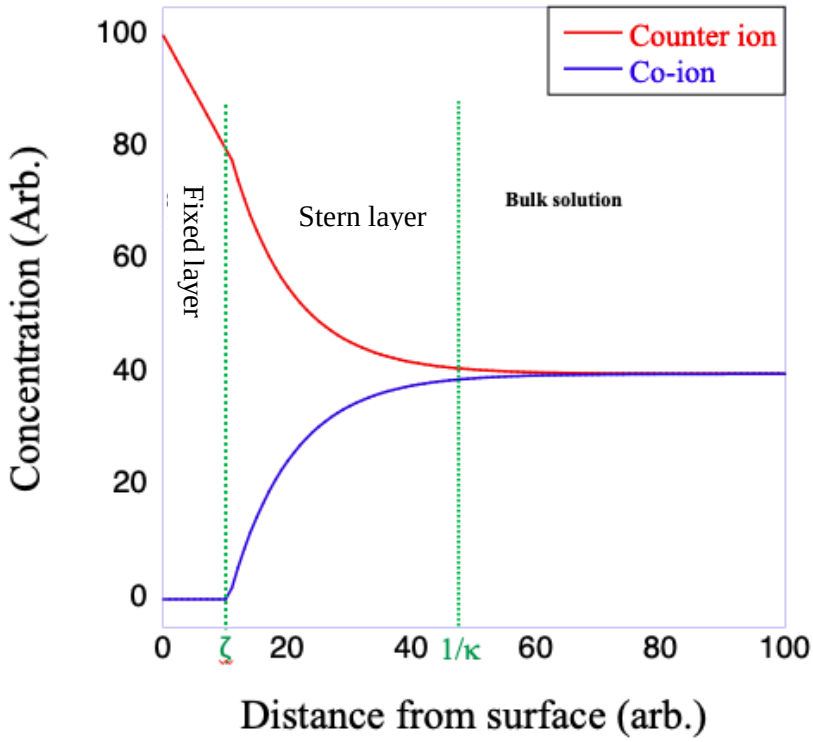


Figure 3-1. Electric double layer model, where the thickness of the Stern layer is 10Å at 0.1 M NaCl, and 3Å at 1 M NaCl evaluated from κ^{-1}.

In Figure 3-1, the membrane charges are located on the left side. The horizontal axis of the graph is the distance from the biological membrane. Ions in water that have the opposite charge to the membrane charge are called counter ions, and those that have the same charge as the membrane charge are called co-ions. The vertical axis of the graph represents the concentration of ions in the solution. The layer where the counter ions are strongly attracted to the membrane charge is called the Stern layer, and its approximate length is calculated as the Debye length, 1/κ, as mentioned before.

3-2 Methods of MD simulations

To prepare the model structure of the monolayer in solution as shown in Fig. 3-2(a), I first created a membrane model consisting of 80 DPPGs [1,2-dipalmitoyl-sn-glycero-3-phospho-(1-rac-glycerol)]. (80 Na^+ or 40 Mg^{2+} ions) and some additional ions (NaCl or $MgCl_2$), and prepared aqueous solutions of different concentrations using the software package Packmol **(Ref.3-11,3-12)**. Next, MD simulations were performed using Gromacs-5.0.4 **(Ref.3-13,3-14,3-15)** in order to create a stable equilibrium membrane model. The TIP3P model **(Ref.3-16)** was used for the water and the gromos54a7.ff force field was used for the membrane, with a time step of 2 fs. The volume of the box used in the simulations is V = 70.0Å× 70.0Å× 530.0Å. The *z*-direction was made sufficiently long compared to the previous study by Dreier *et al* **(Ref. 3-17)**. The temperature of the system was then gradually increased from 0 K to 300 K using the NVT ensemble to maintain the three phases. The temperature was controlled by the Nosé-Hoover thermostat **(Ref.3-18,3-19)**. The electrostatic interaction was processed using the PME (Particle Mesh Ewald) method **(Ref.3-20)**. I use a monolayer instead of a bilayer to model the lipid/water/air interface, but the effect of the hydrophobic part is negligible because the water layer is sufficiently large. And I conducted 10ns simulation.

(a) (b)

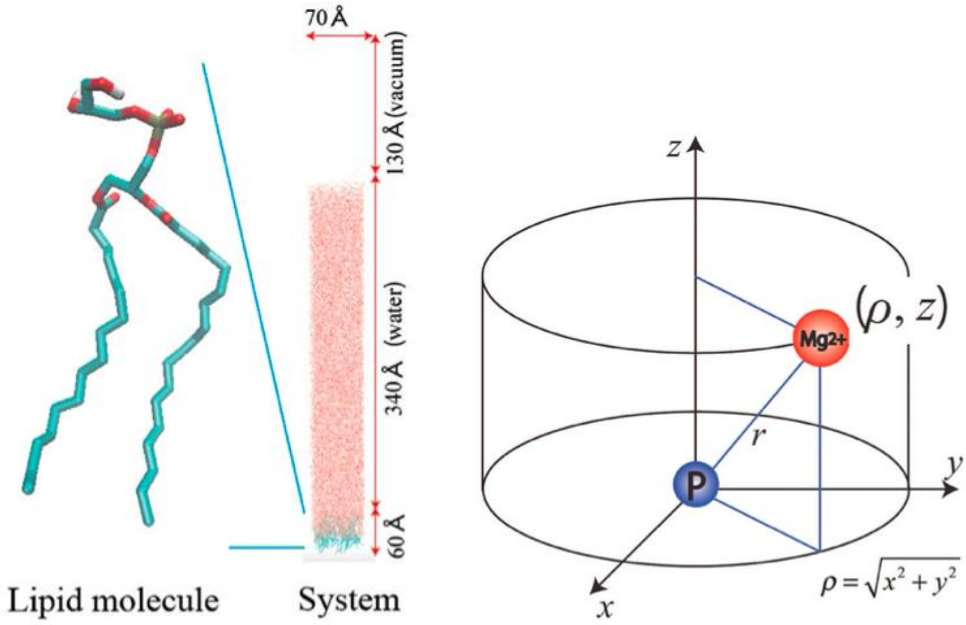


Figure 3-2. (a) Lipid structure and the model monolayer's membrane-water-air interface system. (b) Orientation of the cation closest to the P atom in the lipid.

3-3 Results and Discussion

3-3-1 Distribution of each ion in aqueous NaCl solution

In Figure 3-3(a), the negatively charged membrane is at the left end, and the horizontal axis is the vertical distance (Å) from the membrane. The vertical axis shows the concentration (mol/L) of each ion at that location; Na^+ (blue), Cl^- (red), and Gouy-Chapman's theoretical value (yellow), using NaCl as a sample. The figure can be divided into four main parts as

(i) Z = 0 to 7 Å: high and low concentration areas appear alternately, and maximum values are seen around Z = 3, 5 and 7.

(ii) Z = 8 to 30 Å: Na^+ decreases exponentially.

(iii) Z = 31 to 145 Å: Na^+ is sparsely present.

(iv) $Z \geqq$ 145 Å: Both Cl^- and Na^+ are present.

These behaviors result in graphs of approximately the same shape regardless of the concentration of NaCl, which is quite different from the theoretical values in yellow, especially the concentration of Na^+ decreases faster than the theoretical values. Figure 3-3 (b) shows the distribution of Na^+ near the charged membrane. We can see maximum values around Z = 3 and 5 Å clearly.

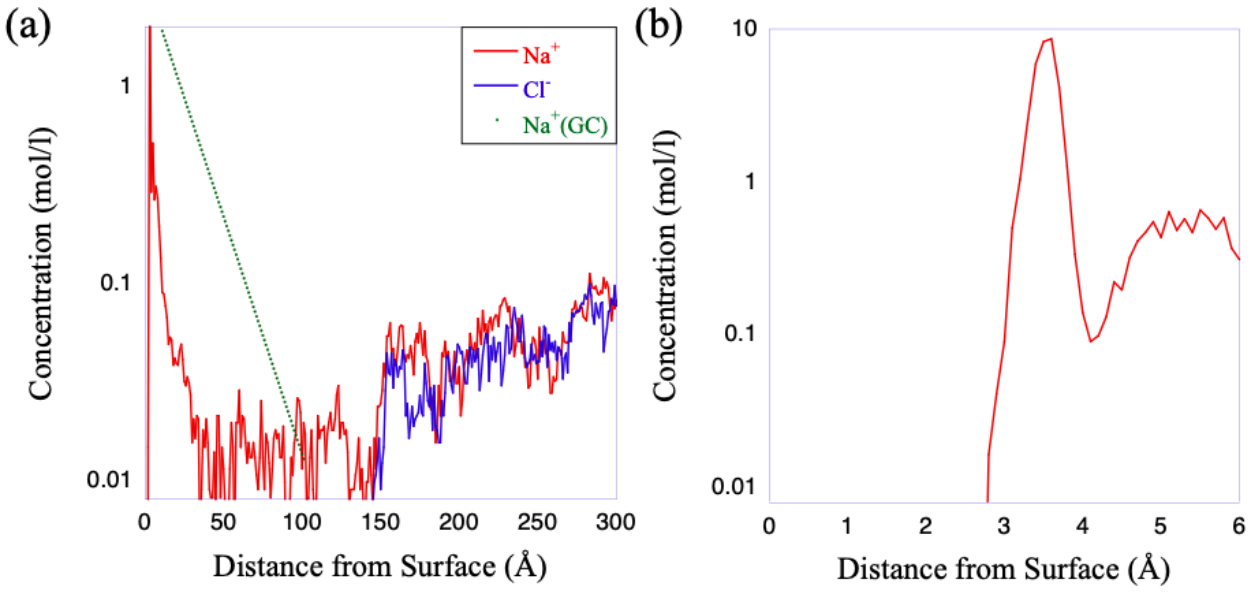


Figure 3-3. (a) Distribution of each ion in NaCl (0.03M) solution and (b) Na^+ ion nearby the membrane.

3-3-2 Difference in Salt of the Distribution of Cations

Figure 3-4 (a) shows the concentration distribution of Na^+ ions around the membrane in different concentrations of NaCl solution. There are maxima at specific locations Z = 3, 5, and 7 Å which are not predicted by Gouy-Chapman's theory.

Figure 3-4 (b) shows the concentration distribution of Mg^{2+} ions in different concentrations of $MgCl_2$ solution. The shape of the graph is more disordered than that of the Na^+ experiment, but again, regardless of the concentration, there is a fixed location where the concentration tends to gather, and the maximum value is found around Z = 5 and 7 Å. It is also important to note that the maximum concentration of Na^+ appears at Z = 3Å, but that of Mg^{2+} does not appear at Z = 3 Å to any great extent.

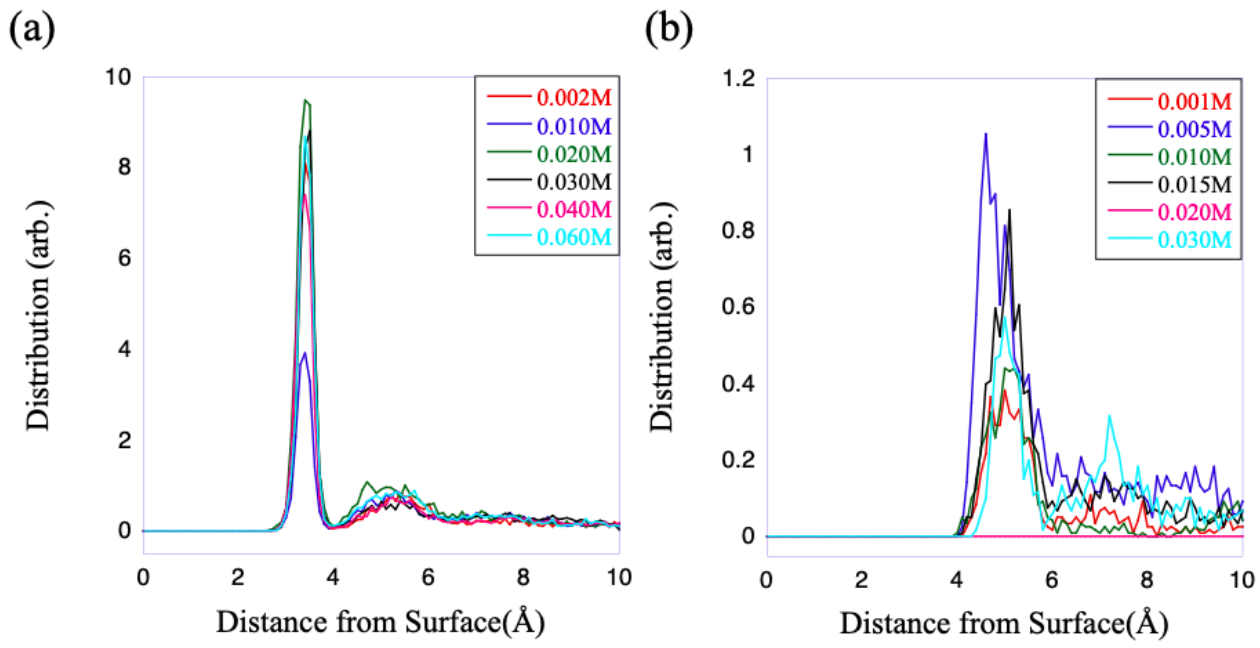


Figure 3-4. (a) Distribution of Na^+ and (b) Mg^{2+} around the membrane in NaCl and $MgCl_2$ solutions with different concentrations, respectively.

3-3-3 Three-dimensional graph of the distribution of counter ions around the membrane charge

The membrane charge in the cell membrane does not exist evenly in the membrane, but rather it exists sparsely as the point charges. Therefore, I investigated the distribution of the counter ions are distributed around the membrane charge in three dimensions. As shown in Fig. 3-5, *Z* represents the vertical distance from the membrane surface and ρ represents the distance parallel to the membrane. The concentration of the counter ion at the coordinate (r, *Z*) is shown in Fig. 3-5. The distribution of Na^+(0.03M) in Fig. 3-5 (a) shows that the first peak appears at around 3Å, which surrounds the membrane charge, followed by a small maximum at around 5Å. There is a slight variation in the distribution depending on the angle, but it is clear that the Na^+ ions are concentrated in an approximately isotropic manner.

The distribution of Mg^{2+} (0.03M) in Fig. 3-5 (b) shows that the first peak is at around 5Å, and that Mg^{2+} does not appear at around 3Å, where the first maximum of Na^+ appears. There is also a large bias in the direction of the surface, with less Mg^{2+} in the direction vertical to the membrane and more in the direction parallel to the membrane. This is because Mg^{2+} is a divalent ion, and a single Mg^{2+} ion is attracted by two membrane charges, so it tends to exist between the two charges, that is, parallel to the membrane.

Looking at the distribution of the higher concentration of Mg^{2+} (0.40 M) in Fig. 3-5 (c), it is surprising that the first maximum peak of Mg^{2+} appears at 3Å, where Mg^{2+} could not enter at the lower concentration (0.03 M). At first, Mg^{2+} could not enter the vicinity of 3Å because of its large radius. When there was no place for Mg^{2+} to escape, such as at high concentrations, Mg^{2+} entered the vicinity of 3Å, which is near the membrane charge.

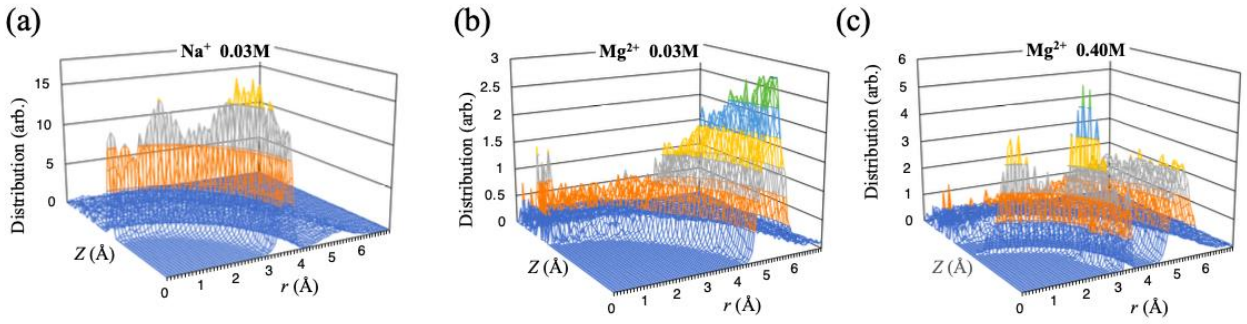


Figure 3-5. Counterion distribution around the membrane ion (a) Na^+ (0.03 M), (b) Mg^{2+} (0.03 M) and (c) Mg^{2+} (0.04 M).

3-3-4 Orientation of Water Molecules around an Ion

In order to investigate the phenomenon that the distance to the membrane charge changes with Mg^{2+} concentration as seen in section 3-3-3, I focused here on the water molecules around each ion. Since water molecules are polar molecules with a large dipole moment, they should be affected in the vicinity of ions. It is known that water molecules are connected by electrostatic forces to form a strong hydrogen bonding network structure. Therefore, I here consider that investigation on the distribution of water molecules would be a clue to solving the question of how the distance between the membrane charge and the Mg^{2+} changes depending on the Mg^{2+} concentration.

Specifically, the direction away from each ion was determined to be the positive direction, and the value was measured to be positive if the + polarity of the water (2 H atoms) was facing that direction (Fig 3-6).

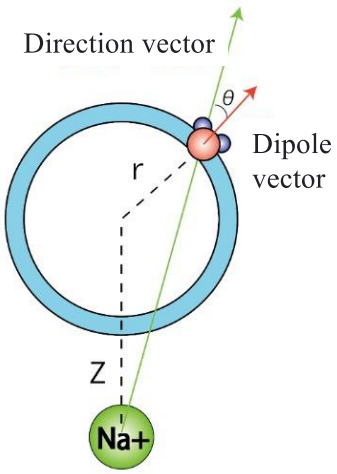


Figure 3-6. The way to calculate the orientation of water molecules. Ion as the origin, *z* and *r* represent coordinates, and the angle between the direction to the coordinate where the water molecule exists and the direction of polarity of the water molecule is θ.

First, the polarity direction of water molecules around Na+ and Mg^{2+} was investigated (Fig. 3-7). Figure 3-7(a) shows the polarity of water molecules around the Na^+ ion. There exist maxima around Z = 2.5, 4.5, and 6.5. It is found that water molecules cover the Na^+ ion in many layers with the minus polarity (O atom) on the Na^+ side and the plus polarity (2 H atoms) on the outside, first layer, second layer, and so on.

Figure 3-7(b) shows the polarity of water molecules around the Mg^{2+} ion. The maxima appear around Z = 2, 4, and 6 Å which are almost the same as those for Na^+. The major difference from the Na^+ case is that a strong layer of water molecules is formed not only in the first layer but also in the second layer. This is because the charge in the center is larger (2+) than Na^+ ion. Even if it is covered by the water molecules in the first layer, the charge inside cannot be fully covered, and the effect leaks strongly out of the first layer. Interestingly, the outside of the water layer is around Z = 3Å and 5Å. These lengths are the distances where the Na^+ and Mg^{2+} ions are distributed when they are attracted around the membrane ions.

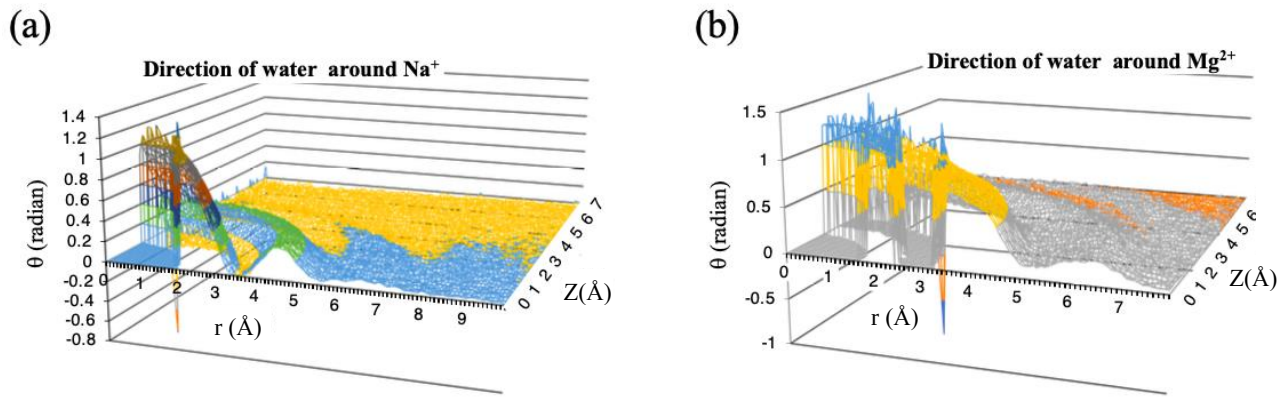


Figure 3-7. Polarity of water around (a) Na^+ and (b) Mg^{2+}

From this result, I consider that the result of the previous subsection, i.e., that Mg^{2+} could approach only around 5 Å from the membrane when the concentration was low, but could approach around 3 Å when the concentration was high, can be explained as follows. The $Mg2^+$ ion in water is surrounded in two layers of water and behaves as an ion with a radius two layers larger than its original ionic radius. However, as the concentration increases and there is no more space, they break through the outer water layer and come closer to each other. In this way, the results can be understood in a unified manner by considering the surrounding water environment. I visualized the layers of water molecules formed around the cations in the water (Figure 3-8).

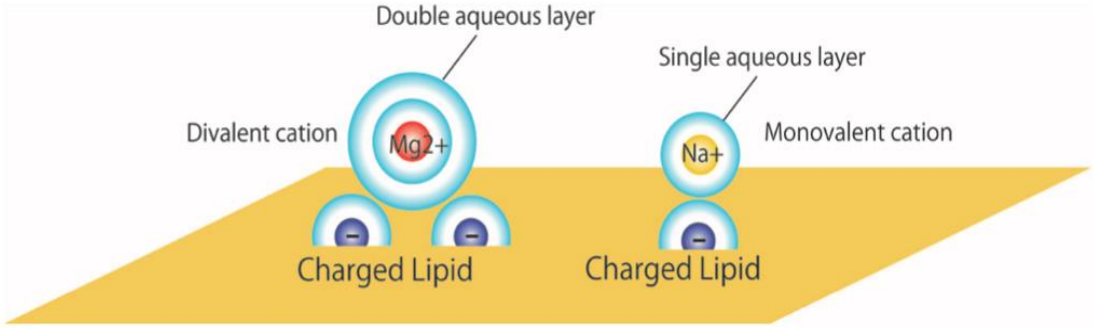


Figure 3-8. Schematic graph of layers of water molecules

Next, I studied the polarity of water molecules around the membrane charges. Figure 3-9 shows the orientation of the water molecules around the membrane ion (PO_4^-), which is negatively charged and thus oriented in the direction of the membrane charge. There is a negative extremum near Z = 4Å. Unlike the case of Na^+ and Mg^{2+}, the extremum does not appear clearly outside the first layer. It is thought that the reason for this is that the water molecules cannot arrange well because they are partly buried in the membrane and the surrounding membrane molecules interact with them. In fact, the second layer can be confirmed in the Z direction, not clearly in the ρ direction, and water molecules can align in the positive direction in close proximity to the membrane.

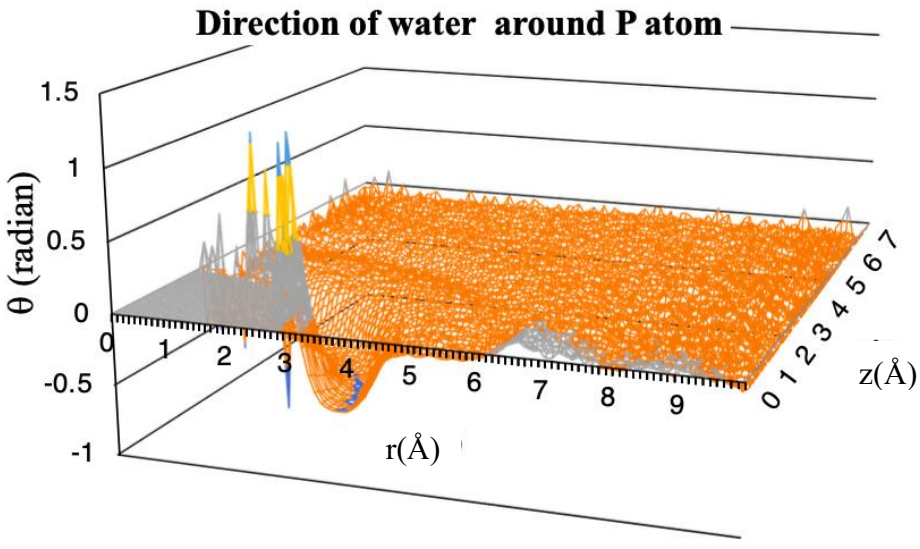


Figure 3-9. Polarity of water molecules around membrane ions. (*Z* is the distance perpendicular to the biological membrane and *r* is the distance horizontal to the biological membrane)

3-3-5 Distance from Membrane Charge and Polarity of Water Molecules

In order to investigate the effect of membrane charge on water molecules, the following analyses were performed. Figure 3-10 shows the change in the polarity of water molecules with distance from the membrane charge. Because the membrane charge is negative, the polarity of water molecules is negative in the vicinity, and the polarity of water becomes closer to 0 as it moves away from the membrane. This means that the effect weakens as water moves away from the membrane charge. However, the effect is still strong enough to reach around 300Å owing to the long-range Coulombic interaction. Now I examine the change in polarity when Na^+ and Mg^{2+} ions are added to the solution at different concentrations.

Comparing Mg^{2+} 0.001M (purple) and 0.015M (blue), the effect weakens rapidly at a distance of about Z = 5Å from the membrane. After that, the change becomes more gradual. It means that the charge polarity at the higher Mg^{2+} concentration 0.015M (blue) disappears as fast as that at lower Mg^{2+} concentration 0.001M (purple). Next, I compared Mg^{2+} 0.015M (blue) and Na^+ 0.030M (orange). Although the total amount of charge on both is the same, the former cancels the effect of the membrane charge more rapidly by around 5Å. It can be seen that the former cancels the effect of the membrane charge more effectively than the latter even at a distance. This behavior is quite natural since Mg^{2+} nearby membrane modulates the order of the water nearby the membrane.

Next, I compared two cases of Mg^{2+} 0.005 M (purple) and Na^+ 0.030 M (orange). Although both cases have the same amount of charges, behaviors are different from each other. This is also explained by the difference in the distribution of water molecules around Mg^{2+} and Na^+.

The strength of the charge of the ion often affects the physical properties of solution. For example, it is known that a small amount of divalent (2+) counterions can precipitate colloids more effectively than a high concentration of monovalent (+) counterions. As in this calculation, the colloids are surrounded by a layer of water molecules, and if counterions with high valent ions are added to the solution, the effect of polarity of the water molecules is effectively screened, and the colloids are brought closer together, leading to precipitation. The results of this calculation for the membrane system may be well related to this theory of colloids.

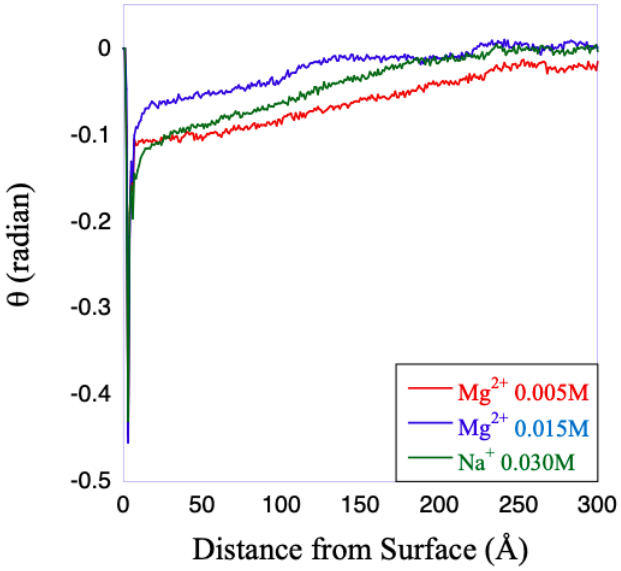


Figure 3-10. Change in distance from membrane charge and water polarity.

3-4 Summary

In the present study, the distribution of ions in the aqueous solution in contact with the membrane charge was investigated using MD simulations. It was found that the computed distribution was different from the theoretical value predicted by Gouy-Chapman model. Next, it was found that polar water molecules surrounded the ions in the aqueous solution, where one layer of water exists around monovalent Na^+ and two layers of water around divalent Mg^{2+}. Since PO_4^-ions partly buried in the membrane are hindered by the surrounding membrane molecules, the surrounding water layer cannot be formed properly. On the other hand, there is no hindrance structure around Na^+ and Mg^{2+} ions that are isolated in water, so that the water molecules can form a regular water layer. I think knowledge obtained from these results is also important for considering the structure of macromolecules in water, for example, middle-sized drug molecules in water considered in Chapter 5. The ions that exist nearby the drug molecules interact with them and can form a layer of water molecules different from free ion, which also affect the nature of the drug molecule itself.

It was found that the influence of the membrane charge can reach quite far in the water layer and be modulated by adding more counter ions to the aqueous solution. Furthermore, it was found that adding a small amount of divalent counterions was more effective than adding a large amount of monovalent counterions in screening out the effect of the membrane charge. This might be related to the fact that the addition of counterions with large valence is effective in precipitating hydrophilic colloids.

3-5 References

1. B. Alberts, A.D. Johnson, J. Lewis, D. Morgan, M. Raff, K. Roberts, P. Walter, "*Molecular Biology of the Cell*" (book).(2015)
2. T. Suzuki, T. Hongawa, I. Washitani, "*Series of New Biology*" (in Japanese), Suken Shuppan (2021).
3. D. A. McQuarrie, "*Statistical Mechanics*", University Science Books (2000).
4. W. J. Moore, "*Physical Chemistry*" (4th ed.), Longmans Green & Co. Ltd. (1963).
5. P. Debye and E. Hückel, "Zur Theorie der Elektrolyte. I. Gefrierpunktserniedrigung und verwandte Erscheinungen" *Phys. Z.*, **24**, 185-206. (1923).
6. Helmholtz H, "Ueber einige Gesetze der Vertheilung elektrischer Ströme in körperlichen Leitern mit Anwendung auf die thierisch-elektrischen Versuche" (German), Ann. Phys. Chem. **165** (6), 211-233. (1853).
7. B.J. Alder, T.E. Wainwright, "Studies in Molecular Dynamics. I. General Method", *J. Chem. Phys.* **31** (2), 459. (1959).
8. J.H. Hunt, P. Guyot-Sionnest, Y.R. Shen, "Observation of C-H stretch vibrations of monolayers of molecules optical sum-frequency generation", *Chem. Phys. Lett.*, **133**(3), 189-192 (1987).
9. P. Guyot-Sionnest, J. H. Hunt, Y.R. Shen, "Sum-frequency vibrational spectroscopy of a Langmuir film: Study of molecular orientation of a two-dimensional system", *Phys. Rev. Lett.*, **59**, 1597-1600 (1987).
10. P. Guyot-Sionnest, P. Dumas, Y.J. Chabal, G.S. Higashi, "Lifetime of an adsorbate-substrate vibration: H on Si(111)", *Phys. Rev. Lett*., **64**, 2156-2159 (1990).
11. L. Martínez, R. Andrade, E. G. Birgin, J. M. Martínez, "Packmol: A package for building initial configurations for molecular dynamics simulations", *J. Comput. Chem*. **30**(13), 2157-2164 (2009).
12. J. M. Martínez, L. Martínez, "Packing optimization for automated generation of complex system's initial configurations for molecular dynamics and docking" *J. Comput. Chem*. **24**(7), 819-825 (2003).
13. H.J.C. Berendsen, D, van der Spoel, R, van Drunen," GROMACS: A message-passing parallel molecular dynamics implementation", *Comput. Phys*. Commun. **91**(1–3) 43-56 (1995).
14. E. Lindahl, B. Hess, D. van der Spoel, "GROMACS 3.0: a package for molecular simulation and trajectory analysis", *J. Mol. Model.* **7**, 306-317 (2001).
15. D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A.E. Mark, H.J.C. Berendsen, "GROMACS: Fast,

flexible, and free", *J. Comput. Chem.*, **26**(16), 1701-1718 (2005).

16. W.L. Jorgensen, J. Chandrasekhar, J.D. Madura, R.W. Impey, M. Klein, "Comparison of simple potential functions for simulating liquid water", *J. Chem. Phys*. **79**, 926-935. (1983).
17. L. B. Dreier, Y. Nagata, H. Lutz, G. Gonella, J. Hunger, E. H. G. Backus, M. Bonn, "Saturation of charge-induced water alignment at model membrane surfaces", *Sci. Adv.* **4**(3), 28 Mar 2018
18. S. Nosé, "A unified formulation of the constant temperature molecular-dynamics methods", *J. Chem. Phys.* 81 (1), 511-519 (1984).
19. W.G. Hoover, "Canonical dynamics: Equilibrium phase-space distributions", *Phys. Rev. A* **31**(3), 1695-1697 (1985).
20. M. Di Pierro, R. Elber, B. Leimkuhler," A Stochastic Algorithm for the Isobaric–Isothermal Ensemble with Ewald Summations for All Long Range Forces", *J. Chem. Theory Comput.* **11**(12), 5624-5637 (2015).

Chapter 4: Accurate prediction of Log$P_{o/w}$ by semi-empirical electronic structure calculation using dielectric continuum model

4-1 Introduction

Cell membrane permeability of a drug is very important for drug development, because the drug cannot work unless it can pass through the cell membrane toward the target even if it is effect against a target. Several methods have been developed for the verification of the cell membrane permeability, but this time I decided to investigate the cell membrane permeability using first-principles calculations and empirical methods. Since the membrane permeability is quite difficult to estimate, I here first tackle the water-octanol partition coefficient in Chapter 4.

The membrane permeability of a drug is a very important factor in drug development **(Ref.4-1～Ref.4-7)**, and it is useful if it can be predicted before in vivo experiments **(Ref. 4-8)**. Octanol/water partition coefficient (Log$P_{o/w}$). Log$P_{o/w}$ is a numerical expression of the ratio of drug solubility in water and octanol. And Log$P_{o/w}$ has long been used as an indicator of the membrane permeability of drugs **(Refs. 4-9 and 4-10)**. For example, Log$P_{o/w}$=0 indicates that the drug is equally soluble in water and octanol. Assuming that blood is water and cell membranes are octanol, a drug that is soluble in both water and octanol can easily pass through the cell membrane from the blood and reach its target in the cell **(Ref. 4-11)**. In reality, membranes are not composed of octanol, but of many kinds of lipids, cholesterol, and membrane proteins such as channels and transporters, and the relationship between blood and cell membranes cannot be easily replaced by that between water and octanol. Despite these problems (membranes are not octanol, etc.), the measurement of Log$P_{o/w}$ is simple and useful as an indicator of the permeability of cell membranes to common drug molecules.

The shake flask method **(Ref. 4-12)** is one of the simplest experimental methods to measure Log$P_{o/w}$. Although this method has been used for a long time and is reliable, it has some weaknesses such as complicated and time-consuming operation and difficulty in measurement when the product is heterogeneously dissolved in water or octanol. If high-performance liquid chromatography is used instead of the shake flask method, a wider range of molecules with different Log$P_{o/w}$ values can be measured rapidly. However, this method

requires prior information on the molecule and the Log$P_{o/w}$ values of similar compounds, which makes it difficult to apply to unknown substances or molecules with large structural changes **(Ref. 4-13～Ref. 4-22)**. Therefore, one of the solutions to these problems is to calculate the solvation free energy of the solute in water and octanol solvents using a computer. If the solvation free energy is obtained, it can provide the Log$P_{o/w}$ value. However, the computational cost is too high to evaluate the Log$P_{o/w}$ values of thousands of molecules for drug screening. Therefore, a method with low computational cost is required. In this chapter, I develop the low-cost Log$P_{o/w}$ method by using combined semi-empirical quantum chemical calculation (Semi-empirical quantum chemistry methods are based on the Hartree–Fock formalism, but make many approximations and obtain some parameters from empirical data. Since the reliability of semi-empirical methods strongly depends on the approximations and parameters used, the limitations of each calculation method must be fully understood when using them. In well-considered situations, semi-empirical methods can be effective in terms of balancing computational complexity and reliability. Currently, widely used methods include the PPP (Pariser-Parr-Pople) method **(Ref.4-23, Ref.4-24)**, which calculates only pi-electrons, the MNDO (Modified Neglect of Diatomic Differential Overlap) method **(Ref.4-25)**, the AM1 (Austin Model1) method **(Ref.4-26)**, the MNDO-PM3 (Parametric Method No. 3) **(Ref.4-27)**, which calculate only valence electrons. which calculates only valence electrons. PM6 and PM7 **(Ref.4-28)** are methods that have been devised to reproduce the experimental values by readjusting the parameters of the PM3 method) **(Ref.4-29)** and Quantitative Structure-Activity Relationship (QSAR) method (QSAR method is a statistical analysis method that clarifies the relationship between structure and function. the method of inferring property values from structural features using linear regression, multiple regression, etc. is equivalent to this.) **(Ref.4-30)**, where the former calculation data are corrected by the descriptors from the chemical structure of a given molecule. And I used the PCM method described in Chapter 2, which considers the solvent is considered as a series of dielectric materials, not as particles, and the free energy of the stabilized structure and the free energy in the cavity are calculated. **(Ref. 4-31～Ref. 4-37)**

Previous research on predicting Log$P_{o/w}$ using a computer was conducted by Prof. Chuman *et al.* in Tokushima University **(Ref.4-11)**. They classified small molecules into nine: three classes (almost unrelated to hydrogen bonds, acceptors of hydrogen bonds, and donors of hydrogen bonds) and three groups (benzene substituent compounds, non-aromatic compounds, and the whole class). They adopted B3LYP, one of the first principles calculations, to calculate the solvation free energy ΔG of each molecule in octanol and in water. They performed multiple regression analyses using ΔG values and the accessible surface area (ASA) of the

atoms in the structure to predict the Log$P_{o/w}$. The predicted results were very high accuracy with a high correlation coefficient R^2 =0.90 to 0.98 (depending on the class). **(Ref.4-11)** However, this previous study uses first-principles calculations, which take a long time to calculate. In order to perform comprehensive calculations for tens to hundreds of thousands of drug candidates, it is necessary to achieve both shorter calculation time and higher accuracy. Therefore, the purpose of this study was to find a quicker and more accurate method of predicting Log$P_{o/w}$ by using various calculation methods.

4-2 Methods

4-2-1 How to calculate Log$P_{o/w}$ from ΔG

At first, one evaluates the solvation free energy (difference of free energy) $\Delta G = G_{solv}$-G_{gas} calculated from the free energy of (G_{solv}) in solution and (G_{gas}) in gas phase for water and *n*-octanol. Using these values, the octanol/water partition coefficient (Log$P_{o/w}$) is defined as follows

$$\mathrm{Log}P_{o/w} = (\Delta G_{wat} - \Delta G_{oct})/\ 2.302RT \qquad (4\text{-}1)$$

where ΔG_{wat} and ΔG_{oct} are the calculated solvation free energies of water and n-octanol, respectively, R is the gas constant, and T is the absolute temperature. The free energies G_{gas}, G_{wat}, and G_{oct} were calculated using the semiempirical PM6/PM7 and CPCM, IEF-PCM, and SMD **(Ref. 4-38～Ref. 4-43)**, respectively. And Density functional theory (DFT) **(Ref. 4-44～Ref. 4-48)** B3LYP/6-31++g(d,p) and CPCM, SMD, respectively. The experimental results of the solvation energy were taken from the supporting literature. **(Ref. 4-48, Ref. 4-49)**.

4-2-2 R^2 value

I want to find a computational method to predict the Log$P_{o/w}$ value of a drug with high accuracy. To numerically test the accuracy of the prediction method, I will use the R^2 value. The definition of the R^2 value is as follows

$$R^2 \equiv 1 - \frac{\sum(y_i - f_i)^2}{\sum(y_i - \bar{y})^2} \qquad (4\text{-}2)$$

where y_i is the value of the i-th data, $\bar{y}$ is the average value of all data, and f_i is the value of the i-th data predicted by the theory. If practical predictions can be made, an R^2 value of 95% or higher is desirable, where 95% is judged as a statistically accurate accuracy.

4-2-3 Multiple Regression Analysis

Next, a multiple regression analysis was performed to refine the obtained calculated data with other variables. In this multiple regression analysis, the following equation was employed.

$$Y = \sum a_j \times X_j \quad (4\text{-}3)$$

where the objective variable is Y and the explanatory variable of the j th component is X_j. Using the mean square method, I obtained the coefficient a_j that had the smallest error with the experimental data. *P*-value and *t*-value represent the importance of explanatory variables in multiple regression analysis, so I examined these values in detail as well as comparing R^2 values. In the multiple regression analyses, the t-value indicates the magnitude of the influence of the explanatory variable on the target variable, and if it is less than 2, the explanatory variable is judged to have "no influence" on the target variable. Also the P-value indicates the significance of the coefficient, and if it is greater than 0.05, the explanatory variable is judged to be "not related" to the target variable.

4-2-4 Quantum Chemical Calculations

All quantum chemical calculations**)** using PM6 and PM7 were performed by the Gaussian 16 program site **(Ref.4-50)** was used, but for those using density functional tight binding (DFTB) method were performed by the GAMESS program site **(Ref.4-51)** because PCM with DFTB isn't implemented in the Gaussian 16 site. It should be noted that Gaussian and GAMESS program sites may not give the same results due to different SMD default settings and implementations, and GAMESS does not have PM6 or MP7 using CPCM or IEF-PCM. Therefore, only SMD/DFTB, which is a similar calculation but with a different set of parameters for DFTB, will be considered here.

4-2-5 Polarizable Continuum Model

In this study, I adopt the polarizable continuum model (PCM) **(Ref. 4-38～Ref. 4-43)** to treat the solvent effects implicitly, as described in Chapter 2. In the PCM method, the solvent is treated as a continuous homogeneous dielectric with dielectric constant ε, the solute is enclosed in a cavity in the dielectric, and the electrostatic interaction between the solute and the solvent is calculated self-consistently. The definition of the cavity and some of the variables of the electrostatic interaction are determined empirically so that the calculated and experimental values agree for many different substances. Due to differences in the values of these

empirically determined variables and in the way the cavity is treated, several different calculation methods exist for the PCM method.

Here, I investigate whether a combination of various PCM methods and semi-empirical methods can better predict the solvation free energy in water and in octanol, while keeping computational costs as low as possible. In the PCM methods, I used Conductor PCM (CPCM) **(Ref. 4-42)**, Integral Equation Form PCM (EEF-PCM) **(Ref. 4-42)**, and Density Based Solvation Model (SMD) **(Ref. 4-43)**. In the semi-empirical methods, I used PM6, PM7, and partially Density Functional Tight Binding (DFTB) methods **(Ref. 4-44～Ref. 4-47)**.

For this purpose, 205 different molecules (mainly organic compounds) were used as a benchmark set. In order to further improve the accuracy of the calculations, in addition to the purely calculated Log$P_{o/w}$ values, a multiple regression analysis **(Ref. 4-32～Ref. 4-34)** was performed using the number of multiple bonds in the molecule and the number of multiple elements as explanatory variables, with the experimental values of Log$P_{o/w}$ as the objective variables. Based on the results, I corrected the value of Log$P_{o/w}$ using the explanatory variables (descriptors) as weights and improved the accuracy.

4-3 Results

4-3-1 Overall Trend of LogPo/w and Solvation Energy

First, Table 4-1 shows the Gibbs energy ΔG calculated by the SMD/PM7 method and the Log$P_{o/w}$ values obtained from the calculated ΔG and experimental data.**(Ref.4-48, Ref.4-49)**

Table 4-1. Calculated and experimental Log$P_{o/w}$ values and calculated free energy differences (kcal/mol) obtained by SMD/PM7 method.

	SMD/PM7 Method	Calculation data			Experimental data **(Ref. 4-48, Ref.4-49)**
No	Name	ΔG_{wat}	ΔG_{oct}	Log $P_{o/w}$	Log $P_{o/w}$
1	methyl propanoate	-5.22	-6.03	0.59	0.83
2	2-chloropropane	-1.18	-4.42	2.37	1.90
3	triethylphosphate	-13.35	-12.41	-0.69	0.79
4	1,1,2-trichloroethane	-5.55	-8.95	2.49	1.89
5	m-xylene	-0.82	-3.89	2.25	3.23
6	4-methylpyridine	-5.81	-7.28	1.08	1.22
7	methanol	-5.48	-4.73	-0.55	-0.91
8	morpholine	-10.89	-9.00	-1.39	-0.87
9	methyl formate	-6.36	-4.36	-1.47	0.03
10	benzamide	-12.10	-10.78	-0.97	0.64
11	p-bromophenol	-5.52	-7.47	1.44	2.54
12	hexanoic acid	-9.18	-10.81	1.19	1.91
13	toluene	-0.40	-4.03	2.66	2.68
14	tetrachloroethene	1.60	-2.92	3.32	3.15
15	n-butane	0.94	-2.56	2.56	2.89
16	nitromethane	-8.61	-6.64	-1.44	-0.32
17	1-chloropropane	-0.34	-3.80	2.54	2.05
18	ethane	1.14	-1.05	1.61	1.81
19	phenol	-5.33	-6.42	0.80	1.52
20	n-propane	0.99	-1.86	2.09	2.36

21	formaldehyde	-5.10	-3.31	-1.31	
22	chloroethane	-0.58	-3.34	2.02	1.43
23	ethanol	-6.06	-5.69	-0.27	-0.48
24	methyl pentanoate	-8.84	-9.45	0.45	1.88
25	butyl acetate	-10.02	-10.61	0.43	1.77
26	benzonitrile	-1.49	-4.15	1.96	1.46
27	1,2-dichlorobenzene	-0.21	-4.56	3.19	3.41
28	cyclohexane	-0.06	-4.26	3.08	3.44
29	cyclopropane	0.21	-2.38	1.90	1.72
30	1-bromobutane	-0.31	-4.56	3.12	2.75
31	naphthalene	-0.92	-5.16	3.11	3.36
32	pyrrole	-3.33	-5.00	1.22	
33	p-cresol	-4.34	-7.12	2.04	1.98
34	pyridine	-4.06	-4.71	0.48	0.47
35	thiophene	0.37	-2.22	1.90	1.81
36	trichloromethane	-1.06	-4.45	2.49	2.01
37	2,2,2-trifluoroethanol	-8.33	-8.07	-0.19	0.37
38	acetic acid	-10.44	-8.25	-1.61	-0.26
39	benzene	0.02	-3.13	2.31	2.09
40	n-octane	2.07	-3.97	4.43	5.19
41	2-methylaniline	-4.73	-7.15	1.78	1.32
42	p-xylene	-0.71	-4.83	3.02	3.21
43	1,1-difluoroethane	-1.26	-3.31	1.50	0.75
44	tetrahydropyran	-5.72	-6.81	0.80	0.80
45	1,1,1-trichloroethane	-1.60	-5.49	2.85	2.52
46	dichloromethane	-1.55	-4.17	1.92	1.25
47	acetone	-8.28	-7.38	-0.67	-0.51
48	cyclopentanone	-7.76	-8.13	0.27	0.24
49	methyl benzoate	-4.83	-5.42	0.43	2.46
50	benzaldehyde	-5.40	-7.01	1.18	1.55
51	methyl acetate	-7.56	-5.88	-1.23	0.16

52	o-cresol	-5.73	-7.75	1.48	1.92
53	nitrobenzene	-5.83	-6.11	0.21	1.84
54	2-methylpyridine	-4.51	-5.73	0.90	1.11
55	fluorobenzene	0.57	-2.71	2.41	2.27
56	1,4-dichlorobenzene	0.42	-4.13	3.33	3.42
57	1,2-dimethoxyethane	-8.66	-6.66	-1.47	-0.21
58	pentanoic acid	-8.64	-9.79	0.84	1.51
59	aniline	-4.31	-6.24	1.42	0.89
60	s-trans-1,3-butadiene	2.30	-0.40	1.97	1.99
61	2-ethylpyrazine	-6.88	-6.48	-0.29	0.65
62	methyl butanoate	-8.43	-8.23	-0.15	1.29
63	n-pentane	0.99	-3.16	3.04	3.51
64	4-methylaniline	-4.57	-8.88	3.16	1.39
65	2-methoxyethanol	-10.91	-8.67	-1.65	-0.69
66	quinoline	-4.99	-6.84	1.36	
67	3-methylpyridine	-4.14	-5.18	0.76	1.20
68	1-nitrobutane	-7.33	-10.00	1.95	1.49
69	1,1,1-trifluoropropan-2-ol	-4.97	-6.00	0.75	0.70
70	4-ethylpyridine	-4.37	-6.03	1.22	2.24
71	piperazine	-6.76	-7.76	0.74	-0.83
72	anthracene	-1.73	-7.08	3.92	4.58
73	propanal	-4.96	-4.56	-0.30	0.51
74	1,4-dioxane	-9.28	-7.85	-1.04	-0.12
75	tetrafluoromethane	3.57	1.34	1.64	1.22
76	n-hexane	0.80	-4.04	3.54	4.03
77	ethyl acetate	-9.80	-9.81	0.01	0.70
78	p-dibromobenzene	-0.82	-5.50	3.43	3.79
79	hydrazine	-4.19	-3.88	-0.23	-2.07
80	anisole	-3.15	-5.04	1.39	2.21
81	acetonitrile	-3.16	-3.95	0.57	
82	propyl acetate	-9.74	-9.98	0.17	1.24

83	tribromomethane	-2.42	-6.16	2.74	2.67
84	bromobenzene	-0.46	-4.38	2.87	2.93
85	chlorobenzene	0.11	-3.72	2.81	2.85
86	m-hydroxybenzaldehyde	-10.40	-8.39	-1.48	1.38
87	1-bromopropane	-0.39	-4.00	2.65	2.10
88	N-methylaniline	-4.87	-6.90	1.49	1.66
89	p-hydroxybenzaldehyde	-12.52	-10.31	-1.63	1.38
90	cyclopentane	-1.12	-4.76	2.67	2.82
91	N,N-dimethylformamide	-11.26	-9.09	-1.59	
92	ethylbenzene	1.14	-3.15	3.14	3.14
93	nitroethane	-5.90	-6.47	0.42	0.16
94	butanal	-5.91	-6.05	0.10	1.06
95	n-heptane	0.74	-4.73	4.01	4.66
96	m-cresol	-5.64	-7.23	1.16	1.99
97	tetrahydrofuran	-6.50	-6.17	-0.25	0.34
98	2-methylpyrazine	-7.29	-6.21	-0.80	0.22
99	bromomethane	-0.20	-2.39	1.61	1.18
100	methane	2.00	0.63	1.01	1.09
101	methylcyclohexane	0.22	-4.46	3.44	3.61
102	dibromomethane	-2.14	-5.02	2.11	1.52
103	2-bromopropane	-1.11	-4.53	2.51	2.14
104	o-xylene	-0.90	-5.03	3.03	3.06
105	piperazine	-10.97	-9.02	-1.43	
106	2-nitropropane	-7.16	-6.49	-0.49	0.80
107	1-nitropropane	-8.80	-8.58	-0.16	0.81
108	bromoethane	-0.80	-2.91	1.55	1.61
109	1,1,1,3,3,3-hexafluoropropan-2-ol	-5.56	-6.75	0.87	1.46
110	1,1-dimethyl-3-phenylurea	-12.87	-13.16	0.21	2.56
111	1,2-ethanediol	-15.7	-11.09	-3.38	-1.36

112	1-bromo-1-chloro-2,2,2-trifluoroethane	-0.32	-4.05	2.73	2.30
113	1-bromopentane	-0.41	-5.30	3.58	3.37
114	1-butanol	-5.68	-6.77	0.80	0.73
115	1-butene	1.34	-1.83	2.33	2.40
116	1-heptanol	-5.82	-8.85	2.22	2.57
117	1-hexanol	-5.70	-8.07	1.75	1.98
118	1-hexene	1.43	-2.95	3.21	3.39
119	1-hexyne	1.61	-2.26	2.84	2.73
120	1-octanol	-6.90	-10.59	2.71	2.96
121	1-pentanol	-5.70	-7.43	1.27	1.42
122	1-pentyne	1.62	-1.63	2.38	2.05
123	1-propanethiol	-0.67	-4.05	2.47	1.81
124	1-propanol	-5.59	-6.06	0.34	0.14
125	2,2',3'-trichlorobiphenyl	-0.84	-7.64	4.99	5.23
126	2,2'-dichlorobiphenyl	-0.68	-6.89	4.55	4.90
127	2,2-dichloroethenyl dimethyl phosphate	-13.18	-13.43	0.18	1.45
128	2,2-dimethylpropane	0.81	-2.95	2.76	3.11
129	2,3-dichlorobiphenyl	-0.77	-7.01	4.57	4.97
130	2,6-dichlorobenzonitrile	-1.66	-5.39	2.73	2.90
131	2-butanone	-8.26	-7.62	-0.47	0.10
132	2-heptanone	-8.85	-9.41	0.41	1.91
133	2-hexanone	-8.10	-8.84	0.55	1.27
134	2-methyl-1-nitrobenzene	-4.97	-5.88	0.67	2.35
135	2-methylpropane	0.99	-2.37	2.47	2.76
136	2-methylpropene	0.64	-1.18	1.34	2.34
137	2-octanone	-8.93	-10.14	0.89	2.57
138	2-pentanone	-8.03	-8.19	0.11	0.60
139	3,3-dimethylbutanone	-7.04	-7.76	0.53	1.20
140	3-bromopropene	0.14	-3.02	2.32	1.79

141	3-methylaniline	-3.67	-6.02	1.73	1.39
142	3-pentanone	-7.09	-7.09	0.00	0.70
143	4-amino-3,5,6-trichloropyridine-2-carboxylic acid	-13.86	-12.12	-1.27	0.30
144	9-methyladenine	-18.30	-18.01	-0.21	-0.03
145	acetophenone	-7.56	-8.51	0.70	
146	allyl alcohol	-6.40	-6.51	0.08	0.14
147	bromotoluene	-0.40	-4.79	3.22	2.93
148	bromotrifluoromethane	1.95	-0.95	2.13	1.86
149	butanoic acid	-8.57	-7.72	-0.63	0.89
150	butanonitrile	-2.26	-4.46	1.61	0.45
151	butylamine	-4.62	-6.25	1.20	0.76
152	chlorodifluoromethane	-0.21	-2.45	1.64	1.08
153	diethyl 2,4-dichlorophenyl thiophosphate	-7.90	-13.47	4.08	5.14
154	diethyl 4-nitrophenyl thiophosphonate	-10.47	-12.53	1.51	3.70
155	diethyl ether	-5.58	-5.48	-0.08	0.83
156	diethyl sulfide	-0.97	-4.88	2.86	1.95
157	diethylamine	-4.66	-5.87	0.88	0.50
158	difluorodichloromethane	2.15	-1.23	2.48	2.16
159	dimethyl 2,4,5-trichlorophenyl thiophosphate	-6.31	-12.05	4.21	4.86
160	dimethyl 4-bromo-2,5-dichlorophenyl thiophosphate	-7.70	-12.83	3.76	4.84
161	dimethyl 4-nitrophenyl thiophosphate	-13.16	-13.28	0.09	2.99
162	dimethyl disulfide	-1.89	-4.63	2.01	1.77
163	dimethyl ether	-3.76	-3.44	-0.23	0.10
164	dipropylamine	-4.67	-6.90	1.64	1.73
165	1,2-dichloroethene	0.77	-2.37	2.30	2.09

166	ethene	1.76	0.05	1.25	1.13
167	ethyl 4-cyanophenyl phenylthiophosphonate	-5.62	-12.23	4.85	4.37
168	ethyl phenyl ether	-4.87	-7.33	1.80	2.52
169	ethylamine	-4.87	-5.10	0.16	-0.30
170	ethyne	2.35	0.82	1.12	
171	fluorotrichloromethane	1.42	-2.45	2.84	2.53
172	hydrogen	1.95	0.55	1.03	0.42
173	isopropanol	-7.75	-7.97	0.16	-0.10
174	methyl 3-methyl-4-thiomethoxyphenyl thiophosphate	-10.86	-14.31	2.53	4.13
175	methyl isopropyl ether	-5.13	-5.91	0.57	1.93
176	methyl propyl ether	-3.25	-5.98	2.00	1.44
177	methylamine	-5.15	-4.51	-0.47	-0.57
178	methylhydrazine	-6.01	-5.28	-0.54	-1.05
179	O-ethyl O'-4-bromo-2-chlorophenyl S-propyl phosphorothioate	-11.66	-16.18	3.31	4.69
180	propanoic acid	-9.73	-8.34	-1.02	1.02
181	propene	1.34	-1.07	1.77	1.77
182	propionitrile	-2.42	-3.96	1.12	-0.14
183	propylamine	-4.86	-5.67	0.59	0.28
184	propyne	1.16	-0.69	1.35	0.94
185	t-butyl methyl ether	-4.96	-5.44	0.35	0.94
186	thioanisole	-2.86	-6.61	2.75	2.74
187	thiophenol	-1.43	-4.94	2.58	2.52
188	trichloroethene	0.99	-2.80	2.78	2.46
189	trimethylamine	-5.05	-5.08	0.02	0.27
190	tripropylphosphate	-11.84	-16.04	3.08	1.87
191	water	-7.74	-6.06	-1.23	-1.38
192	N,N-dimethylacetamide	-12.31	-10.41	-1.39	-1.85

193	trimethylphosphate	-14.61	-12.11	-1.84	-0.65
194	N-methyl-2-pyrrolidinone	-12.14	-11.04	-0.81	
195	t-butanol	-6.61	-7.27	0.49	
196	urea	-16.56	-12.33	-3.10	-2.10
197	1,1,2-trichloro-1,2,2-trifluoroethane	2.52	-2.08	3.37	3.16
198	1,1-dichloro-2,2-difluoroethyl methyl ether	-3.50	-6.30	2.05	
199	1-decanol	-6.91	-11.32	3.24	
200	3-ethyl-2-methoxypyrazine	-9.37	-8.53	-0.61	
201	formamide	-12.43	-8.68	-2.75	
202	g-butyrolactone	-10.57	-8.99	-1.16	
203	N-methylformamide	-12.73	-9.10	-2.66	
204	2-pyrrolidinone	-12.72	-10.57	-1.58	
205	Z-1,2-dichloroethene	-0.27	-3.15	2.11	1.86

Using ΔG_{wat} and ΔG_{oct} calculated by different calculation methods, Log$P_{o/w}$ was calculated using equation (1). Figure 4-1 shows the correlation between the experimental Log$P_{o/w}$ values and calculated ones obtained by using (a) CPCM/PM7, (b) IEF-PCM/PM7, and (c) SMD/PM7 methods. To investigate the accuracy of the agreement between the calculated and experimental values, the R^2 value was examined; when the R^2 value (Eq. (4-2)) close to 1, the agreement is high and when the R^2 value closer to 0, the agreement is worse. The R^2 values of Log$P_{o/w}$ were found to be significantly lower in CPCM and IEF-PCM than in SMD.

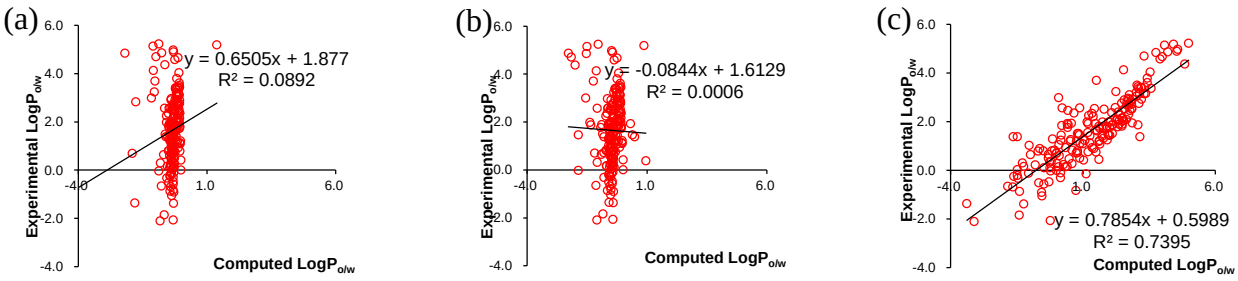


Figure 4-1. Correlation diagram between experimental and calculated Log$P_{o/w}$ values by (a) CPCM/PM7, (b) IEF-PCM/PM7, and (c) SMD/PM7 for 205 compounds

As shown in Table 4-2, I evaluated Log$P_{o/w}$, ΔG_{wat} and ΔG_{oct} for a total of 11 different combinations of calculation methods. Clearly CPCM and IEF-PCM had no correlation between experimental and calculated Log$P_{o/w}$ values. Table 4-2 summarizes the R^2 values for the prediction of Log$P_{o/w}$, ΔG_{wat} and ΔG_{oct} for each calculation method. The R^2 values of Log$P_{o/w}$ are extremely low not only for CPCM but also for IEF-PCM. The reason for the low accuracy of Log$P_{o/w}$ for CPCM and IEF-PCM is that the values of ΔG_{wat} and ΔG_{oct}, which are the basis for calculating Log$P_{o/w}$, are much lower than those calculated by SMD, especially for CPCM and IEF-PCM The accuracy of ΔG_{wat} is significantly lower using the semi-empirical methods. Among the semi-empirical methods, SMD/PM7 has the highest accuracy, with an R^2 value of about 73%, but as mentioned above, this result is far from practical. The R^2 value tends to be lower for ΔG_{wat} than ΔG_{oct} than for ΔG_{wat}, indicating that semi-empirical methods are generally not good at calculating ΔG_{oct}, regardless of any solvent effects. Interestingly, the individual R^2 values of CPCM/B3LYP for ΔG_{wat} and ΔG_{oct} are better than those by SMD/PM7, but do not correlate with the calculation of Log$P_{o/w}$. It is also noteworthy that the Log$P_{o/w}$ values are poorly correlated for both density functional tight binding method 2 (DFTB2) and DFTB3, despite the use of the SMD method. This means that the parameter set does not improve the inaccuracy of SMD/DFTB

as implemented in GAMESS. In particular, the accuracy of ΔG_{oct} is lower than that of CPCM and IEF-PCM/PM6,7. In the semi-empirical method, the initial values and variables are set based on the experimental values, and the difference in the accuracy of ΔG between calculation methods, is due to differences in the implementation and default settings of Gaussian16 and GAMESS program sites. Therefore, before predicting Log$P_{o/w}$ values, it is advisable to examine the methodology and program combination employed against a known set.

Table 4-2. R^2 values of Log$P_{o/w}$, ΔG_{wat} and ΔG_{oct} when the combination of electronic structure calculation levels with solvent models is adopted. (I compared the experimental **(Ref.4-48, Ref.4-49)** and calculated values for Log$P_{o/w}$, ΔG_{wat} and ΔG_{oct})

Methods	PCM	Log$P_{o/w}$	ΔG_{wat}	ΔG_{oct}
B3LYP/ 6-31++g(d,p)	CPCM	0.0361	0.8155	0.9447
	SMD	0.8340	0.9227	0.8948
PM6	CPCM	0.0359	0.6385	0.5596
	IEF-PCM	0.0021	0.6273	0.5248
	SMD	0.6997	0.7557	0.6706
PM7	CPCM	0.0892	0.5877	0.5418
	IEF-PCM	0.0006	0.5754	0.5265
	SMD	0.7395	0.7447	0.6765
DFTB3 3OB31/UFF	SMD	0.1356	0.5644	0.3761
DFTB2 MIO11/DFT	SMD	0.0922	0.4890	0.3384

4-3-2 Dependence of Log$P_{o/w}$ on Functional Group and Atom

Next, I investigated whether the calculated Log$P_{o/w}$ could be corrected by using the characteristic structures of the molecules in Table 4-1. Here, I classified the molecules according to their characteristics and examined how the slope, intercept, and R^2 values of the first-order approximation line of the correlation diagram changed when only molecules with such characteristics were used. (See Table 4-3). Hereafter, I explain the tendency for each group.

Table 4-3. R^2 values of Log$P_{o/w}$, ΔG_{wat} and ΔG_{oct} using SMD/PM7 when the data were classified by atoms and functional groups. (This table show the number of molecules with each structure, and the slope, intercept and R^2 value of the regression analysis.)

Functional group/atom	Number of molecules	slope	intercept	R^2
alkane	140	0.7703	0.6309	0.7307
alkene	13	0.7185	0.6171	0.7431
alkyne	4	0.9466	-0.1253	0.9780
aromatic ring	51	0.6119	1.4730	0.5905
heteroaromatic ring	13	0.6127	0.6880	0.6929
N atom	50	0.7079	0.3853	0.4914
O atom	86	0.8395	0.8870	0.7244
F atom	13	0.7747	0.2225	0.8491
Cl atom	30	0.9735	0.1463	0.7793
P atom	12	0.7516	1.6114	0.7539
S atom	14	0.6212	1.6540	0.3505
Br atom	17	1.1886	-0.5035	0.6631
phenolic group	8	0.2490	1.7272	0.6058
hydroxyl group	17	0.7618	0.3911	0.8758
carboxyl group	6	0.6271	1.1577	0.8497

nitro group	8	0.8329	1.0761	0.4927
ether linkage	9	0.7172	0.7407	0.7738
ester linkage	9	0.8335	1.2206	0.6144
aldehyde group	5	−0.0018	1.1727	3×10^{-5}
ketone group	10	1.7733	0.6739	0.9413
Overall	205	0.7854	0.5989	0.7395

4-3-2-1 Structures containing alkyl chains

Table 4-3 shows that 140 of the 205 molecules contain alkanes. The slope is 0.7703, the intercept is 0.6309, and the R^2 value is 0.7307, indicating that the R^2 value is in close agreement with the overall trend. The molecule with the highest error is No. 161 dimethyl 4-nitrophenyl thiophosphate (No. 161 in Table 4-1), which has a large molecular weight and a long chain length. In addition, the presence of P and S atoms, as described below, is also thought to be the cause of the large error. There are 13 molecules containing alkenes. The slope is 0.7185, the intercept is 0.6171, and the R^2 value is 0.7431, which is still similar to the overall trend. The molecule with the largest error is No.127 2,2-dichloroethenyl dimethyl phosphate (No.127), which also has a large molecular weight and a long alkyl chain; the presence of O, Cl, and P atoms is also thought to be a cause of the large error. There are four molecules containing alkynes. The slope is 0.9466, the intercept is -0.1253, and the R^2 value is 0.978, which is an extremely high R^2 value. But the sample size is too small to determine a trend. Despite the good correlations, the slopes and intercepts are quite different from the overall trend. It can be concluded that the molecules with alkyl chains show a good correlation between the experimental and calculated the Log$P_{o/w}$ values.

4-3-2-2 Structure containing an aromatic ring

Out of 205, there are 51 molecules containing aromatic rings. The slope is 0.6119, the intercept is 1.473, and the R^2 value is 0.5905, indicating that the R^2 value is rather low compared to the overall trend. The molecule with the largest error is No.89 p-hydroxybenzaldehyde (No. 89), which has a medium molecular weight, and the presence of O atoms is thought to be the cause of the large error. There are 13 compounds containing heteroaromatic rings. The slope is 0.6127, the intercept is 0.688, and the R^2 value is 0.6929, indicating that the R^2 values are comparable to the overall trend. The molecule with the largest error is No.143 4-amino-3,5,6-

trichloropyridine -2-carboxylic acid (No. 143), which has a large molecular weight, and the presence of N, O, and Cl atoms are thought to be the cause of the large error. Compared to the results for molecules containing aromatic and heteroaromatic rings, the slopes are similar, but the intercepts are different. This result suggests that the presence of a heteroatom in the ring results in a certain shift, which requires a correction term.

4-3-2-3 Dependence on included atoms

The analysis in sections 4-3-2-1 and 4-3-2-2 suggested that molecules containing heteroatoms and heavy elements behave differently from the overall trend, so I investigated whether classification could be performed by specific atoms.

First, I consider here the molecules with the elements of period 2 of the periodic table: N, O, F, and C atoms. There are 50 molecules out of 205 that contain N atoms. The slope is 0.7079, the intercept is 0.3853, and the R^2 value is 0.4914, indicating that the R^2 value is extremely small compared to the overall trend. 86 molecules contain O atoms. The slope is 0.8395, the intercept is 0.887, and the R^2 value is 0.7244, indicating that the R^2 value is moderate compared to the overall trend. 13 molecules contain F atoms. The slope is 0.7747, the intercept is 0.2225, and the R^2 value is 0.8491, indicating a very high R^2 value compared to the overall trend and other period 2 elements. The reason for the small errors for all the molecules is that the F atom behaves very similarly to the H atom. For example, the H atom has a van der Waals radius of 1.20Å and the F atom has a van der Waals radius of 1.35Å. Both form single bonds with the C atom. Despite this similarity, the intercept for the F atom is smaller than the overall trend. These atoms, along with the C atom, belong to period 2 of the periodic table, but the regression curves for molecules containing these atoms show no similarity.

Now let's consider the third period elements: 12 molecules containing P atoms. The slope is 0.7516, the intercept is 1.6114, and the R^2 value is 0.7539, showing a good R^2 value compared to the overall trend. The intercept shows the second highest value among the atoms studied here; there are 14 molecules containing S atoms with a slope of 0.6212, an intercept of 1.654, and an R^2 value of 0.3505. The intercept is very similar to the result for the P atom, but the R^2 value is much lower. Additionally, there are 30 molecules containing Cl atoms. The slope is 0.9735, the intercept is 0.1463, and the R^2 value is 0.7793, indicating that the R^2 value is high compared to the overall trend. Again, no general trend was found for the third period elements.

There are 17 molecules containing Br atoms. The slope is 1.1886, the intercept is -0.5035, and the R^2 value is 0.6631, indicating that the R^2 value is low compared to the overall trend. Among all the atoms, only the

intercept of Br atom is negative. When comparing the results obtained for the halogen atoms (group 17 elements, i.e., F, Cl, and Br atoms), the slope and intercept do not match. This behavior is also true for pnictogens (group 15 elements) and chalcogens (group 16 elements).

Based on the above, I conclude that there is no specific trend in the linear regression curves associated with the columns (families of elements) or rows (periods of elements) of the periodic table. Because several molecules with large errors contain different types of atoms at the same time, a simple linear regression analysis cannot successfully show a general trend, indicating that multiple regressions may be necessary. This point is discussed later in the paper.

4-3-2-4 Containing a specific chemical group

When one calculated the acid dissociation constant (pK_a) and redox potential (E_{redox}) of the compounds using PCM, one found that the linear regression curves were strongly dependent on the chemical group. The difference in the solvent environment used to evaluate the Log$P_{o/w}$ is not the same as the change in pK_a or E_{redox} calculations do not cause a change in the total charge of the compound, but here I also investigate the chemical group dependence of the Log$P_{o/w}$ values in order to better understand the trends in Log$P_{o/w}$. Here I have limited to investigating only chemical groups containing O atoms in order to further resolve the good correlation behavior.

There are eight molecules out of 205 that contain a phenolic group. The slope is 0.249, the intercept is 1.7272, and the R^2 value is 0.6058, indicating that the R^2 value is rather low compared to the overall trend. The molecule with the largest error was p-hydroxybenzaldehyde (No. 89). The large molecular weight and the simultaneous inclusion of an aldehyde group are thought to be the causes of the large error. There are 17 molecules containing hydroxyl groups. The slope is 0.7618, the intercept is 0.3911, and the R^2 value is 0.8758, indicating that the R^2 value is higher than the overall trend. The molecule with the highest error is No.111, 1,2-ethanediol (No. 111), which has a small molecular weight and a short chain length of alkyl chain. There are six molecules that containing carboxyl groups. The slope is 0.6271, the intercept is 1.1577, and the R^2 value is 0.8497, indicating that the R^2 value is very high compared to the overall trend. The molecule with the highest error is No.143 4-amino-3,5,6-trichloropyridine-2-carboxylic acid (No. 143), which has a large molecular weight and a heteroaromatic ring. The presence of N, O, and Cl atoms are also thought to be a cause of the large error. There are eight molecules containing nitro groups. The slope is 0.8329, the intercept is 1.0761, and the R^2 factor is 0.4927, indicating that the R^2 value is extremely small compared to the overall trend. The

molecule with the highest error is No. 154 diethyl 4-nitrophenyl thiophosphate, which has a large molecular weight and an aromatic ring. The presence of N, O, P, and S atoms is also thought to be the cause of the large error. There are 11 molecules that containing amino groups, the slope is 0.5813, the intercept is -0.0305, and the R^2 value is 0.8045, which is extremely high R^2 value compared to the overall trend. The molecule showing the largest error is 4-methylaniline (No. 64), which is a large molecular weight molecule with an aromatic ring. The presence of N atoms is also thought to be a cause of the large error. There are nine molecules containing ether bonds. The slope is 0.7172, the intercept is 0.7407, and the R^2 value is 0.7738, indicating that the R^2 values are average compared to the overall trend. The molecule showing the largest error is methyl isopropyl ether (No. 175). It is characterized by a small molecular weight and a short chain length of alkyl chains. There are nine molecules containing ester bonds. The slope is 0.8335, the intercept is 1.2206, and the R^2 value is 06144, indicating that the R^2 value is extremely small compared to the overall trend. The molecule with the largest error is No.49 Methyl Benzoate (No. 49), which has a large molecular weight and an aromatic ring. There are five molecules with aldehyde groups. The slope is -0.0018, the intercept is 1.1727, and the R^2 value is 3×10^{-5}, indicating that the R^2 value is extremely small compared to the overall trend. The molecule showing the largest error is p-hydroxybenzaldehyde (No. 89). The large molecular weight and the presence of an aromatic ring are thought to be the cause of the large error. There are 10 molecules with ketone group. The slope is 1.7733, the intercept is 0.6739, and the R^2 value is 0.9413, indicating an extremely high R^2 value compared to the overall trend. The molecule showing the largest error is 2-octanone (No. 137). It is characterized by a large molecular weight and a long alkyl chain length.

4-3-2-5 Summary of functional groups and atoms

The intercepts and slopes of the linear regressions for each characteristic structure are different for each skeletal structure, atom, and chemical group, indicating that it is not appropriate to represent the correlations for all molecules containing each structure by a simple linear regression, even though pK_a or E_{redox} values can be calculated in a similar way. Thus, it can be seen that different corrections need to be made for each structure.

4-3-3 Multiple Regression Analysis of Log$P_{o/w}$

Based on the results in section 4-3-2, multiple regression analysis was conducted to correct the value of Log$P_{o/w}$ using multiple features. In this analysis, I used the experimental value of Log$P_{o/w}$ as objective variable and 12 characters including multiple bonds and heavy atoms (see below) as explanatory variables.

The calculated value of LogPo/w was corrected by using the coefficients obtained from the multiple regression analysis and calculated by (Eq. 4-3). The following 12 types of explanatory variables (#var) were used.

① Log$P_{o/w}$ value obtained by calculation (value obtained by calculating ΔG_{wat} and ΔG_{oct} by each calculation method)

② Number of multiple bonds in the cyclic structure (example; benzene has three)

③ Number of multiple bonds contained in compounds other than cyclic structures (example; ethine has two)

④ 1 for compounds without C and for compounds containing C

⑤ 1 for cycloalkanes and 0 for non-cycloalkanes

⑥ - ⑫ Number of each type of atom in the structure (N, O, F, Cl, P, S, Br)

First, I conducted a multiple regression analysis using above 12 explanatory variables for each calculation method and examined the resulting *t*-values and *P*-values. Of these, I considered that explanatory variables with *t*-values less than 2, and *P*-values more than 5% did not affect the analysis, so I omitted these explanatory variables and conducted the multiple regression analysis again (see Table 4-4)

In Table 4-4, the best R^2 value is obtained by using the solvation free energy difference of SMD/PM7 and the other 11 explanatory variables for the correction, and $R^2 \sim 0.95$ is achieved. It is noteworthy that this result is similar to that obtained with SMD/B3LYP, indicating that the computational constants can be significantly reduced. We can complete the calculation of 205 molecules in half a day with SMD/PM7, compared to a week with SMD/B3LYP on a standard computer. Furthermore, this method has no error values greater than 4 in the correction values, which means that it can predict Log$P_{o/w}$ values without molecules whose predictions are far off, used in this study. In each calculation, I focused on *t*-values and *P*-values and found that I could calculate the values with almost the same accuracy, except for N_{err}, even if I focused only on the explanatory variables relevant to the analysis.

Table 4-5 shows the contribution of the 12 explanatory variables in multiple regression analysis using the results of SMD/PM7. If the *t*-value is less than 2 or the *P*-value is 5% or more, the explanatory variable can be generally excluded from the analysis as being of low importance. The correction can be done by multiplying the "coefficients" discharged by the multiple regression analysis by the "numerical values of explanatory variables" and adding the numerical values.

Table 4-4. The results obtained by multiple regression method. The notations #var, N_{ex}, and N_{err} are the type of explanatory variables, the number of explanatory variables, and the number of numerators with errors greater than 4, respectively.

Methods	PCM	#var	N_{ex}	R^2	N_{err}
B3LYP	CPCM	all	12	0.6947	3
		②③⑤⑥⑨⑪⑫	7	0.6912	3
	SMD	all	12	0.9461	0
		①②⑥	3	0.9425	0
PM6	CPCM	all	12	0.7137	3
		①②③⑤⑥⑦⑨⑩⑪⑫	10	0.7070	3
	IEF-PCM	all	12	0.7200	2
		①②③⑤⑥⑨⑪⑫	8	0.7113	2
	SMD	all	12	0.9491	0
		①②④⑥⑦	5	0.9476	0
PM7	CPCM	all	12	0.7116	3
		①②③⑤⑥⑦⑨⑩⑪⑫	10	0.7045	3
	IEF-PCM	all	12	0.7234	2
		①②③⑤⑥⑧⑨⑪⑫	9	0.7198	2
	SMD	all	12	0.9502	0
		①②④⑥⑦⑩	6	0.9490	1

Table 4-5. Contribution of explanatory variables in multiple regression analysis to the calculation results of SMD/PM7.

#var	coefficient	standard error	*t*- value	*P*- value
①	0.8605	0.0296	29.0562	0.0000
②	0.1894	0.0294	6.4385	0.0000
③	0.0771	0.0581	1.3283	0.1858
④	−0.9582	0.3620	−2.6466	0.0089
⑤	0.2870	0.2348	1.2227	0.2231
⑥	−0.2110	0.0653	−3.2299	0.0015
⑦	0.5441	0.0482	11.2849	0.0000
⑧	−0.0346	0.0517	−0.6702	0.5036
⑨	−0.0003	0.0526	−0.0066	0.9948
⑩	−0.6912	0.2592	−2.6666	0.0084
⑪	−0.0554	0.1536	−0.3604	0.7190
⑫	0.0348	0.0999	0.3483	0.7281

Table 4-6 shows the R^2 values after correction for the case where all 12 explanatory variables are used and the case where only the necessary explanatory variables are used. Even when the number of explanatory variables is reduced to only the important ones, the values are comparable to those obtained when all explanatory variables are used.

In Table 4-4, the values of Log$P_{o/w}$ calculated from ΔG were compared with the experimental values, while in Table 4-6, ΔG_{oct} and ΔG_{wat} before calculating Log$P_{o/w}$ were corrected using multiple regression analysis, and the correlation with the experimental values was examined. In addition, Log$P_{o/w}$ was calculated using the corrected ΔG_{oct} and ΔG_{wat}, and compared with the experimental Log$P_{o/w}$. As a result, the best value of R^2 between calculated Log$P_{o/w}$ and experimental Log$P_{o/w}$ is slightly smaller than that in Table 4-4. In other words, it was found that the accuracy of the correction was higher if Log$P_{o/w}$ was first calculated using ΔG_{oct} and ΔG_{wat} and then the correction by multiple regression analysis was applied.

Table 4-6. R^2 values when the values of ΔG_{oct} and ΔG_{wat} are corrected for the explanatory variables obtained from the multiple regression analysis.

Methods	PCM	Log$P_{o/w}$	ΔG_{wat}	ΔG_{oct}
B3LYP/ 6-31++g(d,p)	CPCM	0.6963	0.8898	0.9447
	SMD	0.9480	0.9611	0.9769
PM6	CPCM	0.7244	0.8694	0.9507
	IEF-PCM	0.6852	0.8699	0.9459
	SMD	0.9383	0.9367	0.9726
PM7	CPCM	0.7528	0.8623	0.9498
	IEF-PCM	0.7130	0.8621	0.9464
	SMD	0.9400	0.9319	0.9715
DFTB3 3OB31/UFF	SMD	0.6769	0.8832	0.9159
DFTB2 MIO11/DFT	SMD	0.6602	0.8100	0.8090

4-4 Summary

When one simply adopted the semi-empirical method was adopted to estimation of ΔG_{oct} and ΔG_{wat} using the dielectric solvent model as a method to predict Log$P_{o/w}$, the best R^2 value of the calculated Log$P_{o/w}$ was at most 0.70, which was not practical. However, when corrections were applied to Log$P_{o/w}$ by multiple regression analysis using the number of heavy bonds, the number of aromatics, and the type and number of heavy elements as explanatory variables, the R^2 value increased to 95%. This suggests that Log$P_{o/w}$ can be predicted with considerable accuracy by using a semi-empirical method instead of the time-consuming DFT calculation. In addition, the difference between the calculated and experimental values is less than 4. (If the log$P_{o/w}$ value differs by more than 4, then the $P_{o/w}$ value differs by more than 1000 times, and the error in the prediction can be considered to be beyond the acceptable range.) Indicating that the Log$P_{o/w}$ values can be predicted with some reliability. Since the explanatory variables used in the correction were limited to the number of heavy bonds and the number of heavy elements, the variables can be easily set based on the molecular structure. Before calculating Log$P_{o/w}$, I also tried to correct ΔG_{oct} and ΔG_{wat}, which are the basis for calculating Log$P_{o/w}$, by multiple regression analysis. However, the accuracy of the method to directly correct Log$P_{o/w}$ by multiple regression analysis could not be exceeded. In the present study, I compared the calculated values with the experimental values using only small molecules with simple structures, and I was able to obtain the calculated values with some accuracy.

4-5 References

1. G. G. Briggs, “Theoretical and experimental relationships between soil adsorption, octanol-water partition coefficients, water solubilities, bioconcentration factors, and the parachor”, *J. Agric. Food Chem.*, **29**, 1050-1059 (1981).
2. C. Hannesschlaeger, A. Horner, P. Pohl, “Intrinsic Membrane Permeability to Small Molecules”, *Chem. Rev.*, **119**, 5922-5953 (2019).
3. G. Yun, W. Youn, H. Lee, S. Y. Han, M. B. Oliveira, H. Cho, F. Caruso, J. F. Mano, I. S. Choi. “Dynamic Electrophoretic Assembly of Metal–Phenolic Films: Accelerated Formation and Cytocompatible Detachment”, *Chem. Mater.*, **32**, 7746-7753 (2020).
4. C. J. Porter, J. R. Werber, M. Zhong, C. J. Wilson, M. Elimelech. “Pathways and Challenges for Biomimetic Desalination Membranes with Sub-Nanometer Channels”, *ACS Nano*, **14**, 10894-10916 (2020).
5. A. S. Kamenik, J. Kraml, F. Hofer, F. Waibl, P. K. Quoika, U. Kahler, M. Schauperl, K. R. Liedl. “Macrocycle Cell Permeability Measured by Solvation Free Energies in Polar and Apolar Environments”, *J. Chem. Inf.*, **60**, 3508-3517 (2020).
6. A. Ebert, K.-U. Goss. “Predicting Uncoupling Toxicity of Organic Acids Based on Their Molecular Structure Using a Biophysical Model”, *Chem. Res. Toxicol.*, **33**, 1835-1844 (2020).
7. J. A. H. Schwöbel, A. Ebert, K. Bittermann, U. Huniar, Kai-Uwe Goss, Andreas Klamt. “COSMOperm: Mechanistic Prediction of Passive Membrane Permeability for Neutral Compounds and Ions and Its pH Dependence”, *J. Phys. Chem. B*, **124**, 3343-3354 (2020).
8. A. Kovalenko, F. Hirata, “Self-consistent description of a metal–water interface by the Kohn–Sham density functional theory and the three-dimensional reference interaction site model”, *J. Chem. Phys.*, **110**, 10095 (1999).
9. D. Roy, N. Blinov, A. Kovalenko, “Predicting Accurate Solvation Free Energy in n-Octanol Using 3D-RISM-KH Molecular Theory of Solvation: Making Right Choices”, *J. Phys. Chem. B*, **121**, 9268-9273 (2017).
10. K. Ogata, M. Hatakeyama, S. Nakamura, “Effect of Atomic Charges on Octanol–Water Partition Coefficient Using Alchemical Free Energy Calculation”, *Molecules*, **23**, 425 (2018).

11. H. Chuman, A. Mori, H. Tanaka, “Prediction of the 1-Octanol/H_2O Partition Coefficient, Log P, by Ab Initio MO Calculations: Hydrogen-Bonding Effect of Organic Solutes on Log*P*”, *Anal. Sci.*, **18**, 1015 (2002).
12. J. P. Hansen, I. McDonald, “Theory of Simple Liquids”, 3rd ed.; *Elsevier Academic Press, London Burlington*, 2006.
13. J. Johnson, D. A. Case, T. Yamazaki, S. Gusarov, A. Kovalenko, T. Luchko, “Small molecule hydration energy and entropy from 3D-RISM”, *J. Phys.*, **28**, 344002 (2016).
14. A. Kovalenko, Condens. “Molecular theory of solvation: Methodology summary and illustrations”, *Matter. Phys.*, **18**, 32601 (2015).
15. E. L. Ratkova, D. S. Palmer, M. V. Fedorov, “Solvation Thermodynamics of Organic Molecules by the Molecular Integral Equation Theory: Approaching Chemical Accuracy”, *Chem. Rev.*, **115**, 6312-6356 (2015).
16. A. Kovalenko, “Multiscale modeling of solvation in chemical and biological nanosystems and in nanoporous materials”, *Pure Appl. Chem.*, **85**, 159-199 (2013).
17. A. Kovalenko, “Multiscale Modeling of Solvation“, *in Springer Handbook of Electrochemical Energy,* Springer-Verlag, Berlin Heidelberg, 2017.
18. J. F. Truchon, B. M. Pettitt, P. A. Labute, “A Cavity Corrected 3D-RISM Functional for Accurate Solvation Free Energies”, *J. Chem. Theory Comput.*, **10**, 934-941 (2014).
19. Y. Maruyama, N. Yoshida, H. Tadano, D. Takahashi, M. Sato, F. Hirata, “Massively parallel implementation of 3D-RISM calculation with volumetric 3D-FFT”, *J. Comput. Chem.*, **35**, 1347-1355 (2014).
20. T. Luchko, S. Gusarov, D. R. Roe, C. Simmerling, D. A. Case, J. Tuszynski, A. Kovalenko, “Three-Dimensional Molecular Theory of Solvation Coupled with Molecular Dynamics in Amber”, *J. Chem. Theory Comput.*, **6**, 607-624 (2010).
21. R. E. Skyner, J. L. McDonagh, C. R. Groom, T. van Mourik, J. B. O. Mitchell, “A review of methods for the calculation of solution free energies and the modelling of systems in solution”, *Phys. Chem. Chem. Phys.*, **17**, 6174-6191 (2015).
22. M. J. S. Dewar, E. G. Zoebisch, E. F. Healy, J. J. P. Stewart, “Development and use of quantum mechanical molecular models. 76. AM1: A New General Purpose Quantum Mechanical Molecular Model”. *J. Am.*

Chem. Soc. **107**, 3902-3909 (1985).

23. Pariser, R. and Parr, R. G. "A Semi-Empirical Theory of the Electronic Spectra and Electronic Structure of Complex Unsaturated Molecules. I.", *J. Chem. Phys.*, **21**(1953), 466-471, *ibid.*, **21**(1953), 767
24. Pople, J. A. "Electron interaction in unsaturated hydrocarbons", *Trans. Farad. Soc.*, **49**(1953), 1375-1385
25. Dewar, M. J. S. and Thiel, W. "Ground states of molecules. 38. The MNDO method. Approximations and parameters", *J. Am. Chem.* Soc., **99**(1977), 4899-4907
26. Dewar, M. J. S., Zoebisch, E. G., Healy, E. F. and Stewart, J. J. P. "Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model", *J. Am. Chem. Soc.*, **107**(1985), 3902-3909
27. Stewart, J. J. P. : "Optimization of parameters for semiempirical methods II. Applications", *J. Comp. Chem.*, **10**(1989), 221-264
28. Jiri Hostas, Jan Rezac, Pavel Hobza, "On the performance of the semiempirical quantum mechanical PM6 and PM7 methods for noncovalent interactions", *Chemical Physics Letters* **568–569**, 161-166 (2013)
29. R. Todeschini, V. Consonni, "Molecular Descriptors for Chemoinformatics", *Methods and Principles in Medicinal Chemistry* **41**. Wiley, (2009).
30. S. Basavaraj, G. V. Betageri, "Can formulation and drug delivery reduce attrition during drug discovery and development—review of feasibility, benefits and challenges", *Acta Pharm. Sin. B.*, **4**, 3-17 (2014).
31. 31.H. D. Williams, N. L. Trevaskis, S. A. Charman, R. M. Shanker, W. N. Charman, C. W. Pouton, C. J. H. Porter, " Strategies to Address Low Drug Solubility in Discovery and Development", *Pharmacol. Rev.*, **65**, 315-499 (2013).
32. C. J. Cramer, "Essentials of Computational Chemistry: Theories and Models", *John Wiley & Sons, Chichester* (2013).
33. J. C. Kromann, C. Steinmann, J. H. Jensen, "Improving solvation energy predictions using the SMD solvation method and semiempirical electronic structure methods", *J. Chem. Phys.*, **149**, 104-102 (2018).
34. J. Tomasi, B. Mennucci, R. Cammi, "Quantum Mechanical Continuum Solvation Models", *Chem. Rev.*, **105**, 2999-3094 (2005).
35. A. Klamt, G. Schuurmann, "COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient", *J. Chem. Soc.*, **2**, 799-805 (1993).
36. C. J. Cramer, D. G. Truhlar, "A Universal Approach to Solvation Modeling", *Acc. Chem. Res.*, **41**, 760-

768 (2008).

37. C. Steinmann, K. L. Blædel, A. S. Christensen, J. H. Jensen, “Interface of the Polarizable Continuum Model of Solvation with Semi-Empirical Methods in the GAMESS Program”, *PLoS One*, **8**, e67725 (2013).
38. J. Tomasi, B. Mennucci, R. Cammi, “Quantum Mechanical Continuum Solvation Models”, *Chem. Rev.*, 105, 2999-3094 (2005).
39. B. Mennucci, Comput. “Polarizable continuum model”, *Mol. Sci.*, **2**, 386-404 (2012).
40. V. Barone, M. Cossi, “Quantum Calculation of Molecular Energies and Energy Gradients in Solution by a Conductor Solvent Model”, *J. Phys. Chem A*, **102**, 1995-2001 (1998).
41. M. Cossi, N. Rega, G. Scalmani, V. Barone. “Energies, structures, and electronic properties of molecules in solution with the C-PCM solvation model”, *J. Comput. Chem.*, **24**, 669-681 (2013).
42. B. Mennucci, E. Cancès, J. Tomasi, “Evaluation of Solvent Effects in Isotropic and Anisotropic Dielectrics and in Ionic Solutions with a Unified Integral Equation Method: Theoretical Bases, Computational Implementation, and Numerical Applications”, *J. Phys. Chem. B*, **101**, 10506-10517 (1997).
43. A. V. Marenich, C. J. Cramer, D. G. Truhlar, “Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions”, *J. Phys. Chem. B*, **113**, 6378-6396 (2009).
44. J. J. P. Stewart, “Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements”, *J. Mol. Model*., **13**, 1173-1213 (2007).
45. J. J. P. Stewart, “Molecular Modeling and Process Simulation: Real Possibilities and Challenges”, *J. Mol. Model.*, **17** 19-29 (2003).
46. M. Elstner, D. Porezag, G. Jungnickel, J. Elsner, M. Haugk, T. Frauenheim, S. Suhai, G. Seifert, “Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties”, *Phys. Rev. B*, **58**, 7260 (1998).
47. M. Gaus, Q. Cui, M. Elstner, “DFTB3: Extension of the Self-Consistent-Charge Density-Functional Tight-Binding Method (SCC-DFTB)”, *J. Chem. Theory Comput.*, **7**, 931-948 (2011).
48. R. Dipankar, B. Nikolay, K. Andriy, “Predicting Accurate Solvation Free Energy in n-Octanol Using 3D-RISM-KH Molecular Theory of Solvation: Making Right Choices”, *J. Phys. Chem. B*, **121**, 9268–9273 (2017).

49. V. Aleksandar, J. Christopher, G. Donald, “Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions”, *J. Phys. Chem. B.*, **113**, 6378–6396 (2009)
50. M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria,M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, D. J. Fox, in Gaussian 16, (Revision A.02), Gaussian, Inc., Wallingford CT, 2016.
51. M.W. Schmidt, K.K. Baldridge, J.A. Boatz, S.T. Elbert, M.S. Gordon, J.H. Jensen, S. Koseki, N. Matsunaga, K.A. Nguyen, S. Su, T.L. Windus, M. Dupuis, J.A. Montgomery,” General atomic and molecular electronic structure system”, *J. Comput. Chem.*, **14**, 1347-1363 (1993).

Chapter 5: Evaluation of Membrane Permeability of Cyclic Peptides of Middle-sized Molecules

5-1 Introduction

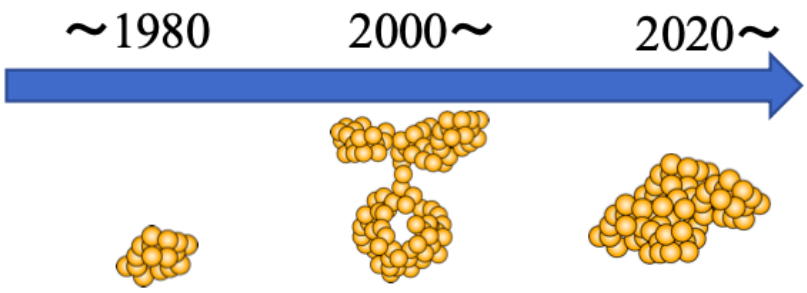
In the field of drug discovery, it is very important to measure the membrane permeability of middle-sized molecular drugs. Although the membrane permeability has been measured by using artificial membranes such as in PAMPA experiment as explained later, I explored an efficient method for numerically evaluating physical properties concerning membrane permeability using computer simulations and analyze the correlation between the physical property and the experimental membrane permeability.

5-1-1 Molecular Weight and Characteristics of Drugs

In drug discovery, the development of small molecule drugs has been actively pursued for a long time. The small molecule drugs are drugs composed of molecules with a molecular weight of 0.5 kDa or less and have been studied for a long time because they are extracted from familiar places and are easy to chemically synthesize. Since the 2000s, the approach to drug discovery research has shifted from screening drug discovery of small molecules to genomic drug discovery, and thus high-molecular drugs such as antibodies, which consist of a kind of protein with a molecular weight of 150 kDa or more, emerged as a result. While small molecule drugs have the advantage of high cell membrane permeability, they have the disadvantage of low specificity and severe side effects caused by working on non-targets. On the other hand, the antibodies **(Ref. 5-1)** have the advantage of the high specificity and low the side effects but have the disadvantage of low/no permeability to cell membranes, and then they are not suitable for oral administration.

In recent years, with the progress of cell engineering represented by nucleic acid engineering and iPS cells, middle-sized molecular drugs(0.5〜5 kDa) **(Ref. 5-2)** have emerged as new drug targets such as nucleic acid drugs and cell therapies, which are in line with the era of genome drug discovery. The middle-sized molecular drugs mainly include peptides, nucleic acids, and natural products which have various structures. Particularly, the middle-sized molecular drugs with high lipid bilayer membrane permeability can target on intracellular protein complexes, which have been considered difficult for the antibody-based drugs. Owing to the advantage,

the middle-sized molecular drugs are expected to be used in a wide range of applications. However, there are still many unknowns about this middle-sized molecular drug, and it is expected to be a drug that combines the good points of both small molecules and antibodies, such as the excellent membrane permeability of small molecules and the specificity to the target of antibodies and is expected to be the subject of active research in the future. The characteristics of these drugs are summarized in Fig. 5-1.



	Small Molecule	Antibody	Middle Molecules
Molecular Weight	～0.5 kDa	150 kDa～	0.5～5kDa
Specificity	△	○	○
Side Effect	○	△	△
Protein-Protein Interaction	×	○	○
Membrane-Permeable	○	×	?

Figure 5-1. Properties of small molecules, antibodies, and middle molecules.

5-1-2 Cell Membrane Permeability of Middle-sized Molecular Drugs

It is hard to believe that middle-sized molecule drugs with large molecular weight have excellent membrane permeability to biological membranes as well as small molecules. This is because, in general, as shown in Fig. 5-2.

(1) The diffusion coefficient is inversely proportional to the molecular radius. **(Ref. 5-3)** Therefore, it is difficult for middle-sized molecules with a large radius to move in the aqueous phase and in passing through the membrane.

(2) The drug proceeds to rearrange all the interactions between the solvent and the drug surface. If one thinks about it, the middle-sized molecule with a large surface area has many interactions with solvent molecules, and it is difficult to move while cutting all of them at the same time. Therefore, it is considered that the membrane permeation of middle-sized molecular drugs is overwhelmingly more difficult than that of small molecules.

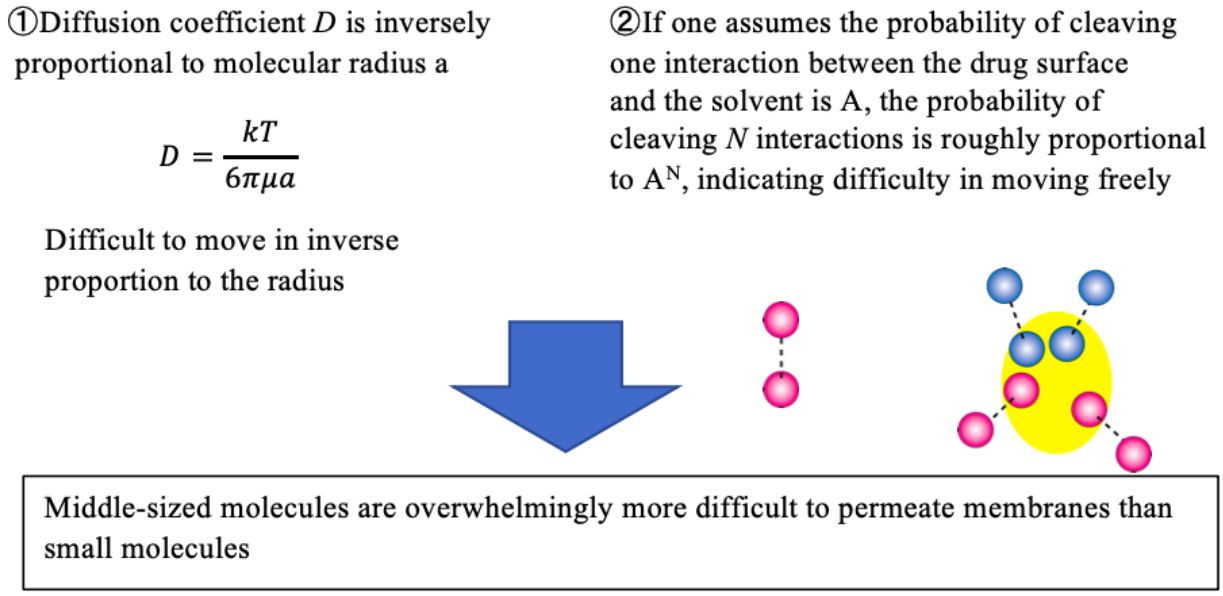


Figure 5-2. Relationship between molecular size and membrane permeation.

5-1-3 About cyclic peptides

In recent years, middle-sized drugs with a special structure called cyclic peptides **(Ref. 5-3)** have been attracting much attention from both experimental and theoretical points of view. These cyclic peptides are thought to have a structural property that enhances membrane permeability. The following is a discussion of the membrane permeability of cyclic peptides compared to that of linear peptides.

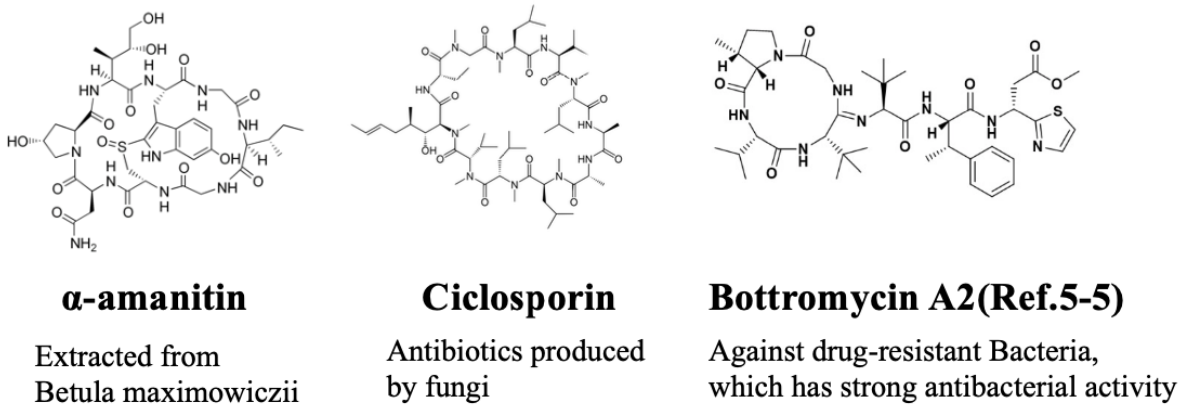


Figure 5-3. Examples of cyclic peptides.

In general, molecules with a linear structure are less likely to pass through the cell membrane than those with a cyclic structure. When a molecule with a linear structure enters the cell membrane from water, if there are hydrophilic and hydrophobic moieties at both ends of a linear chain, the hydrophilic moiety is stabilized outside the membrane, while the hydrophobic moiety is stabilized inside the membrane. When the hydrophilic moiety should penetrate the membrane, it becomes remarkably unstable, resulting in a high energy barrier during the membrane permeation process. The larger the molecule, the more stable it becomes, and the membrane permeability of middle-sized molecules is clearly lower than that of small molecules.

In contrast, in the case of a cyclic structure, if there are hydrophilic and hydrophobic groups on the inside and outside, the hydrophilic and hydrophobic properties of the entire molecular surface can be changed at once by turning the front and back sides over as illustrated in Fig. 5-4. This is because the cyclic structure is similar to that of micelles and is considered to be a suitable structure for membrane permeation *in vivo*.

Recently, a system to synthesize special cyclic peptides with infinite chain length by mRNA display has been established by using an artificial synthetic system, and it has become easy to obtain structural diversity.

However, not only cyclic peptides but also middle-sized molecular drugs have a vast number of possible structures (configurations) compared to small molecules, and the experimental evaluation of physical properties has not been progressed, and there has not been many comparisons between the experimental results and corresponding theoretical calculations.

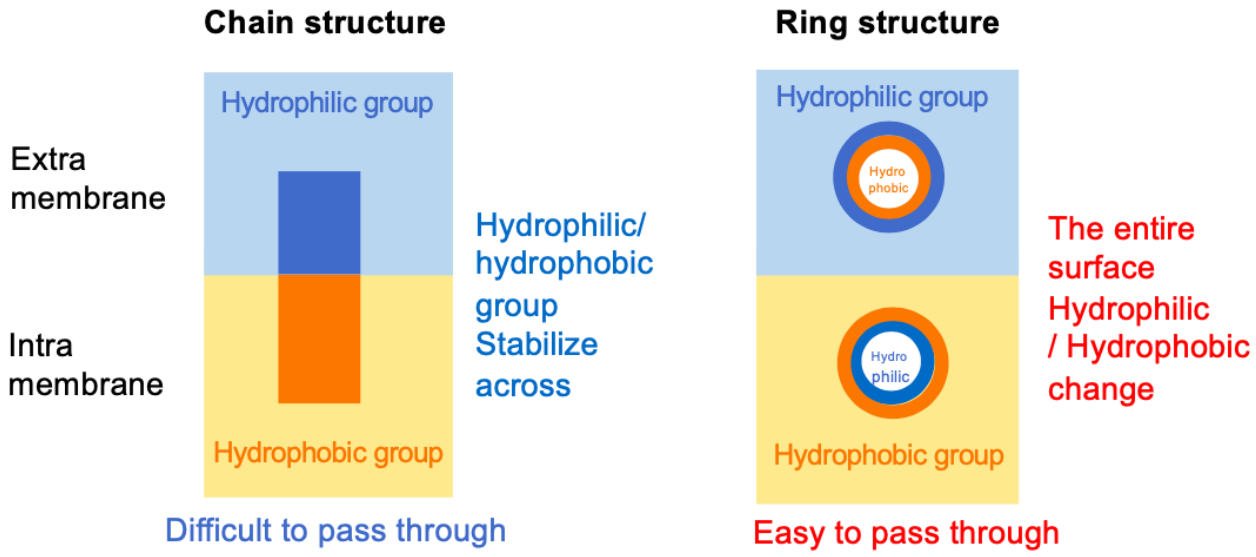


Figure 5-4. Relationship between molecular structure and membrane permeation.

5-1-4 Artificial Membrane Permeability Test

As mentioned earlier, one of the difficulties in the research of this middle-sized molecular drug is that it takes an order of magnitude more time to calculate the molecular backbone structures than in the case of small molecules, even when experiments using computers are performed because of the high molecular weight. In order to shorten the computation time, several methods have been devised, but these methods are still under development and in the research stage. Here, I describe the experimental method to evaluate the membrane permeability before the computer simulation methods.

Artificial Membrane Permeability Assay (PAMPA) **(Ref. 5-6)** is used to investigate the membrane permeability of existing intermediates. It is a method to evaluate the membrane permeability from acceptor to donor by using a phospholipid membrane coated on a membrane filter as a substitute for a biological membrane (see Fig. 5-5). The membrane permeation time is in the range of 2 to 20 hours, and nowadays, automated robots are commercially available to analyze the data. One can obtain reliable results from the direct membrane permeation process is observed even though it is an artificial membrane rather than a cell membrane. However, this method is not perfect. Due to the nature of the experiment, (1) one cannot be confident that the artificial membrane is the same as the biological membrane, and (2) the newly synthesized drug for the experiment must have been synthesized using chemical methods. In this work, I will examine the accuracy of the calculation by comparing the experimental values to those measured by PAMPA. At present, PAMPA has been developed as a device that automatically performs experiments and numerical analyses.

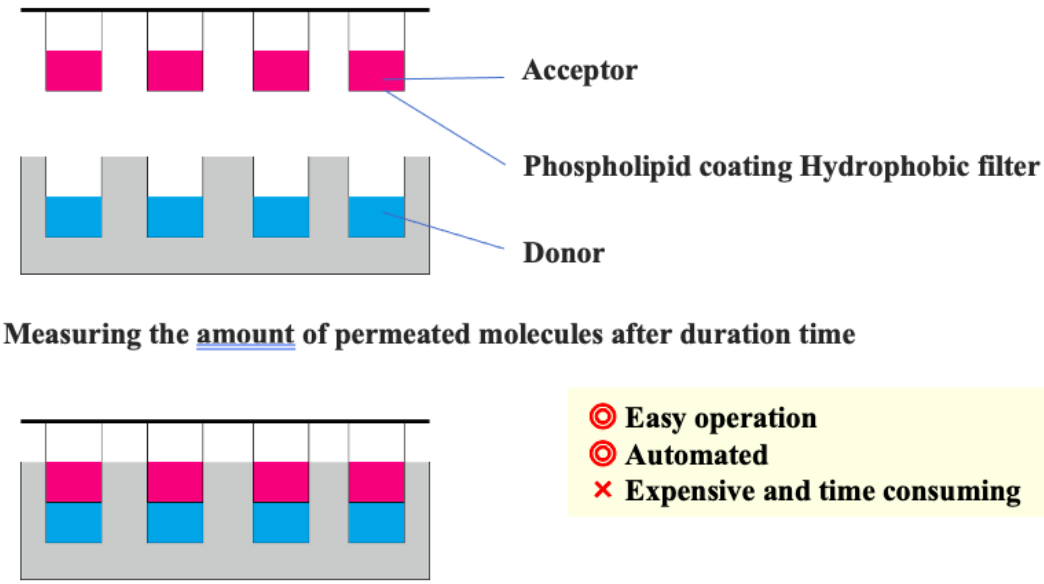


Figure 5-5. Artificial Membrane Permeability Suitability Test (PAMPA) experiment.

5-2 Calculation Methods

5-2-1 Calculation of Log$P_{o/w}$ using quantum chemical methods

As described in Chapter 4, the water-octanol partition coefficient, Log$P_{o/w}$, is a numerical value that indicates the ratio of water and octanol dissolved in a flask, and has been used as an indicator of the membrane permeability. According to Chapter 4, I calculated the Gibbs free energies ΔG_{wat} and ΔG_{oct} in water and octanol using a first-principle methods, **(Ref.5-7～Ref.5-9)** and used these values to calculate Log$P_{o/w}$.

$$\mathrm{Log}P_{\mathrm{o/w}} = \frac{\Delta G_{\mathrm{wat}} - \Delta G_{\mathrm{oct}}}{2.302\ RT} \tag{5-1}$$

where R and T are again the gas constant and the temperature, respectively.

The same method as used in the second study, SMD DFT and SMD PM7 are used to calculate the Gibbs free energy of the optimized structure of the drug molecules in water and octanol, and the value is used to estimate Log$P_{o/w}$. The calculation time for the SMD PM7 is faster than that for SMD DFT, though the accuracy is said to be lower. In the case of the small molecules, the accuracy of the calculation was comparable to that of DFT. However, it is not clear whether the calculation with high accuracy can be performed for middle-sized molecules with much larger molecular weights than small ones. Here I examine the accuracy of SMD PM7 compared to DFT.

5-2-2 Sampling membrane permeation processes by PaCS-MD

It is expected that molecular dynamics (MD) simulation can be used to investigate the membrane permeability of drugs. However, the calculation time is limited to about μ seconds in MD simulation, while phenomena such as structural change and membrane permeation of middle-sized molecule drugs are so-called rare events, which take much longer time than MD simulation can reach. Therefore, conventional MD calculations cannot sample the membrane permeation process with reasonable computation time, and it is necessary to devise a new method. In this study, I used PaCS-MD (Parallel Cascade Selection Molecular Dynamics) **(Ref.5-10)** as one of the methods.

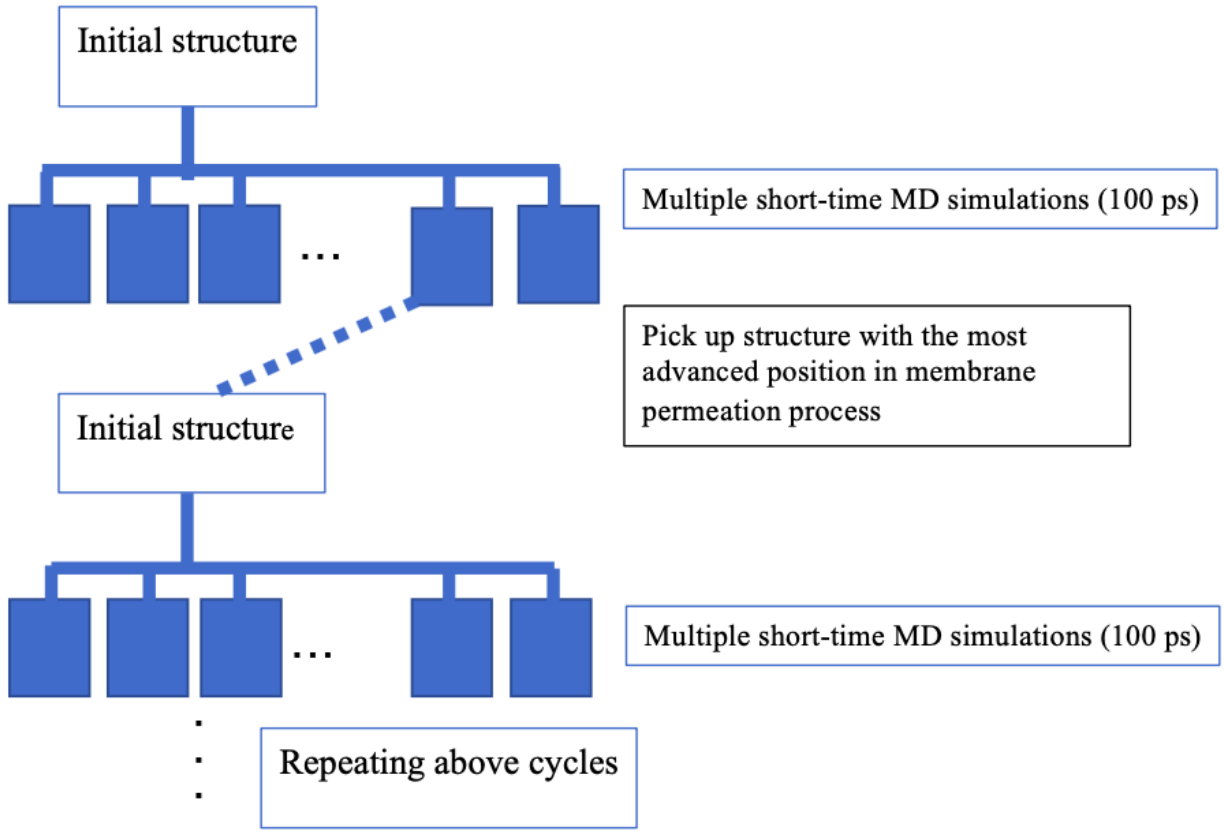


Figure 5-6. Flowchart of PaCS-MD simulation.

To omit the backward motion, I first simulated 100 ps from the initial structure, selected the structure that moves forward the most in the 100 ps, and set the structure as the new initial structure for the next 100 ps simulations. (Fig. 5-6) By using this method, the membrane permeation of the drug can be realized about 100 times faster than the usual MD calculation. However, it is noted that there is a risk that the structural change of the drug during the permeation will not be sufficiently kept up. Especially in the case of drugs with a large

molecular structure such as the Bottromycin and its derivatives used here it is necessary to be careful, since the structural change takes a long time.

5-2-3 PerMM

PerMM **(Ref. 5-11～Ref. 5-12)** is a computational method to investigate the membrane permeability of drugs to biological membranes using empirical methods based on molecular backbone information. In about 10 seconds, it calculates the structural change of drug molecule and approximate free energy in 1,2-Dioleoyl-sn-glycero-3-phosphocholine (DOPC) which is one of the membranes, and then it calculates the membrane permeability index for the black lipid membrane (BLM), blood-brain barrier (BBB) **(Ref. 5-13)**, CACO2**(Ref. 5-14)** (immortalized cell line of human colorectal adenocarcinoma cells), PAMPA-DS, and Plasma Membrane by applying correction.

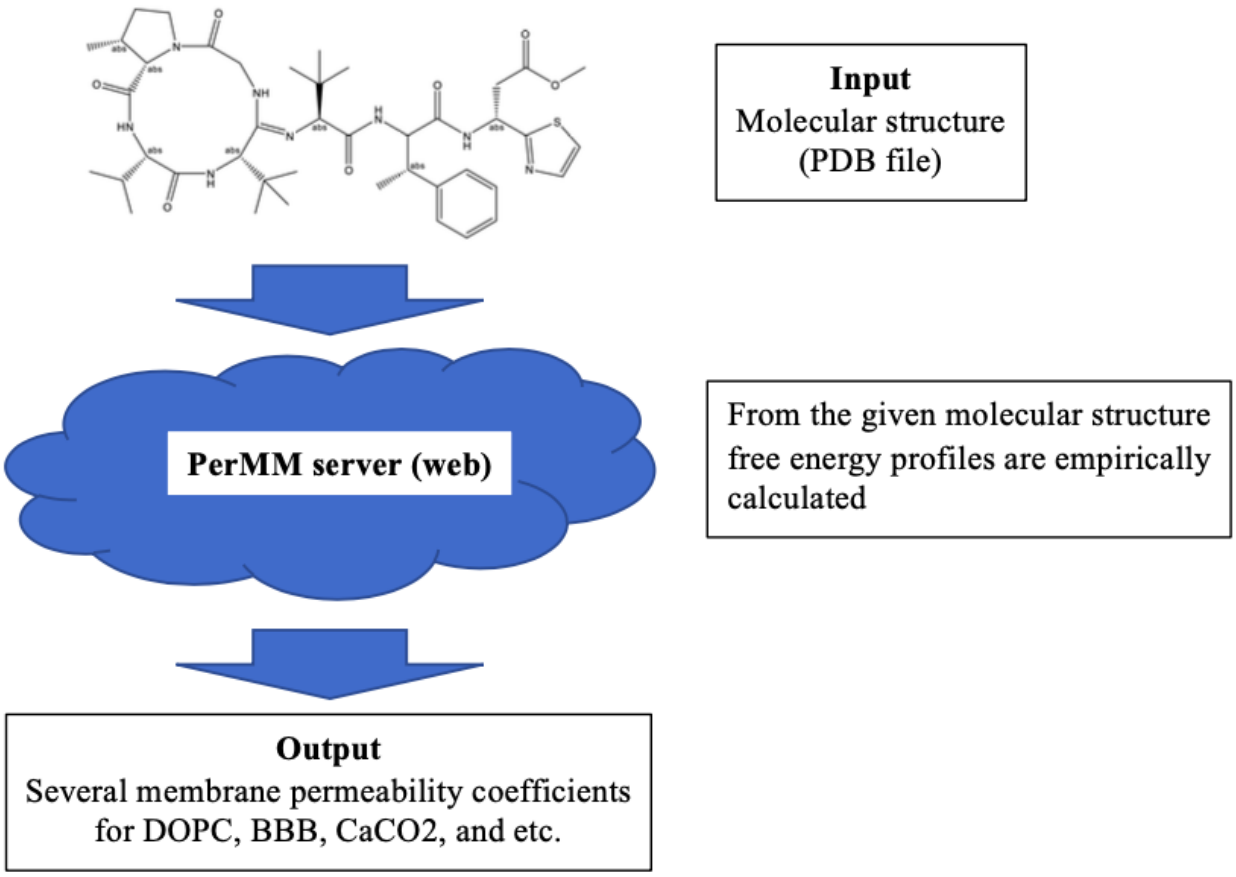


Figure 5-7. Schematic diagram of PerMM calculation.

Originally, it was developed to investigate the membrane permeability of drugs with small molecular weight, but I investigated here, whether it is possible to calculate correctly even for drugs with large molecular weight. Even if it is a middle-sized molecular drug, it discharges the calculation result in a few seconds just like a small molecule. When I saw the calculation result, I felt uncomfortable. Originally, I can see the structural change of the molecule when the drug molecule moves through the water layer, membrane layer, and water layer, but because the molecule is too large for the middle-sized molecular, the molecule does not fit in the water layer from the starting stage, and the calculation starts from the state where the molecule is partially trapped in the cell membrane. Therefore, PerMM may not be good at handling large molecules. In addition, the difference in the initial structure has a big influence on the calculation result because the PDB file is input.

5-3 Results and Discussion

5-3-1 Bottromycin A2 and its derivatives used in the experiments

BottorocycinA2 is a cyclic peptide with potent antibacterial activity against drug-resistant bacteria. In this calculation, I illustrated BottromycinA2 and its derivatives with substituted side chains, where the experimental values of PAMPA are written next to the names of the derivatives (Fig. 5-8). As a general tendency, the membrane permeability (P_e, PAMPA experimental value) is lower when the substituted side chain has a negative charge (for example B3) or a long side chain (for example B5). Although there is only one data set, it can be seen that when the substituted side chain is positively charged (B1), P_e value is remarkably large.

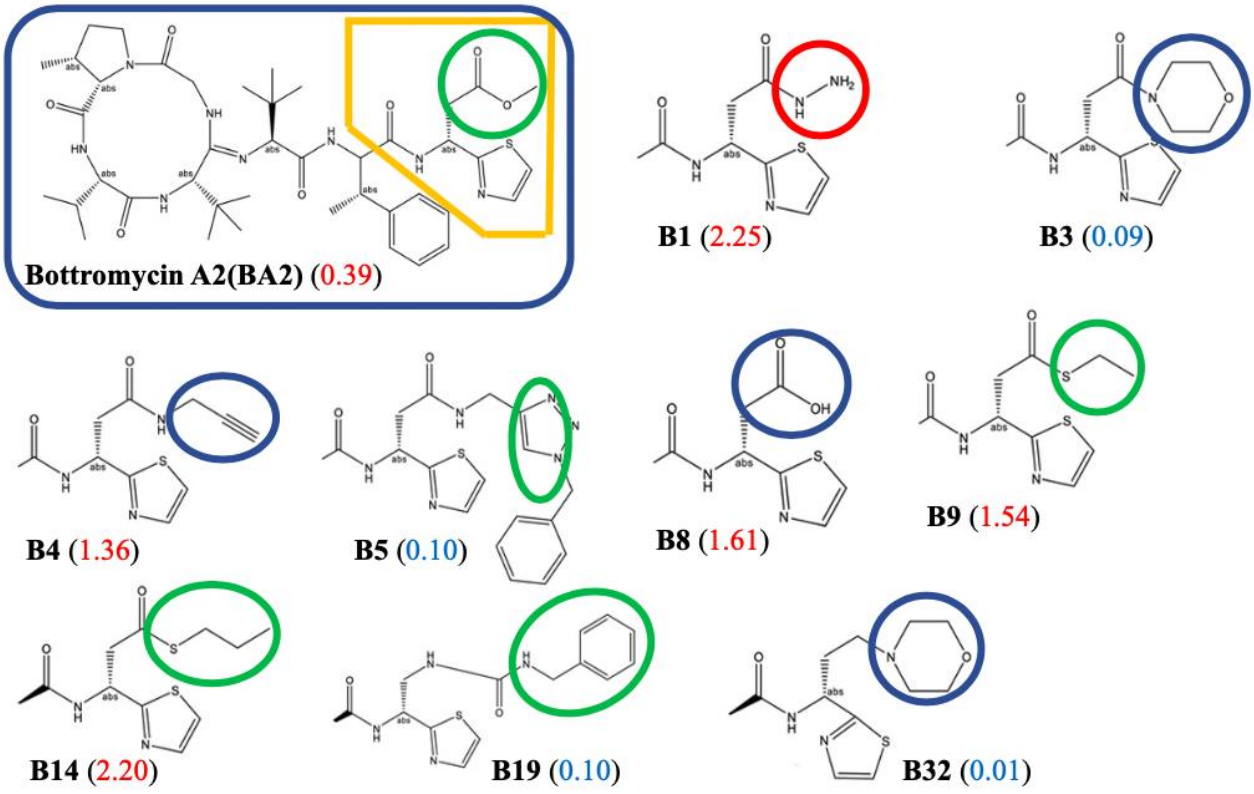


Figure 5-8. BottromycinA2 and its derivatives. The numbers at the right side of the compound name are the experimental PAMPA value, P_e, which is represented by the velocity ($\times 10^{-6}$cm/s), where the red and blue show the high and low membrane permeability. Blue, read, and green circles mean negative, positive, and no charge in the derivatives.

5-3-2 Analyses with Log$P_{o/w}$ by quantum chemical methods

The correlation between Log$P_{o/w}$ calculated by SMD/DFT(B3LYP) and the experimental value of PAMPA, P_e, was plotted in Fig. 5-9 (a). The R^2 value for the entire molecule is around 0.19, indicating that the correlation is not so good. Therefore, the correction was made by the multiple regression analysis used in Chapter 4. As a result, I was able to greatly improve the R^2 value to around 0.84 (Fig.5-9(b)). I also performed a multiple regression analysis of the Log$P_{o/w}$ and LogP_e, resulting in the R^2 value of about 0.99 (Figure 5-9(c)). Although the R^2 value after the multiple regression analysis seems to be high, I cannot say that this method is a good one because the calculation time is very long, taking about one month per molecule. Therefore, I think this method is inappropriate.

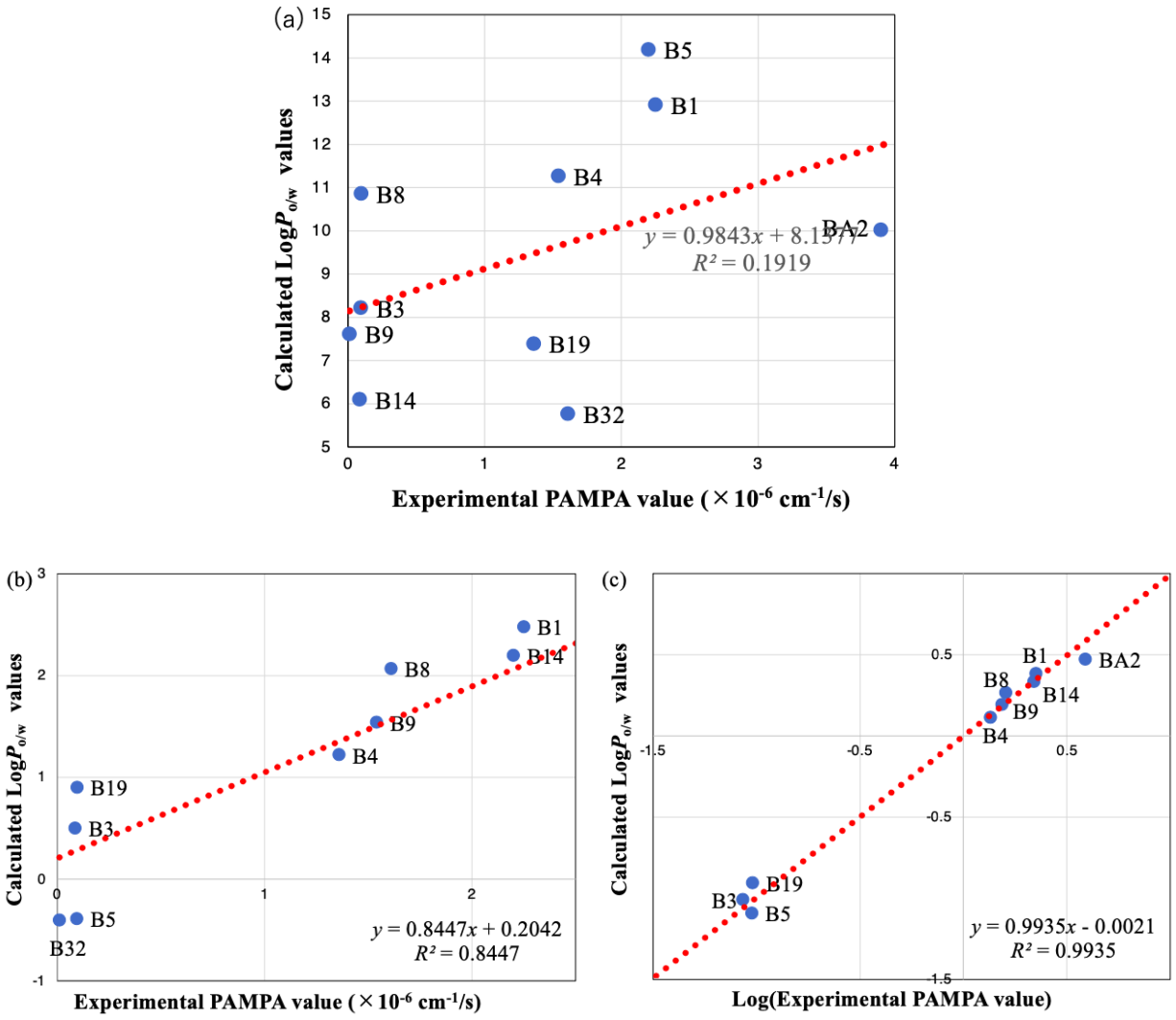


Figure 5-9. Correlations (a) between Log$P_{o/w}$ values by SMD B3LYP and the experimental P_e values, (b) between the corrected Log$P_{o/w}$ values and the experimental P_e values and (c) between the corrected Log$P_{o/w}$ values and the experimental LogP_e values.

As seen in the second study, SMD/PM7, which requires less computation time than the SMD/DFT calculation, can make the predictions as accurate as SMD/DFT for the molecules with small molecular weights. Since this result was applicable only for the molecules with small molecular weights, and it is still undecided whether it can predict middle-sized molecules with sufficient accuracy due to their overwhelmingly large molecular weights. Here I also adopt SMD/PM7 in evaluating Log$P_{o/w}$ values of Bottromycin and its derivatives. As can be seen in Fig.5-10, comparing the calculated Log$P_{o/w}$ by SMD/PM7 with the experimental PAMPA values, the R^2 value is 0.11, which is the same tendency for SMD/DFT and is worth for the prediction. Thus, I next consider the correction by conducting the multiple regression analysis using the explanatory variables.

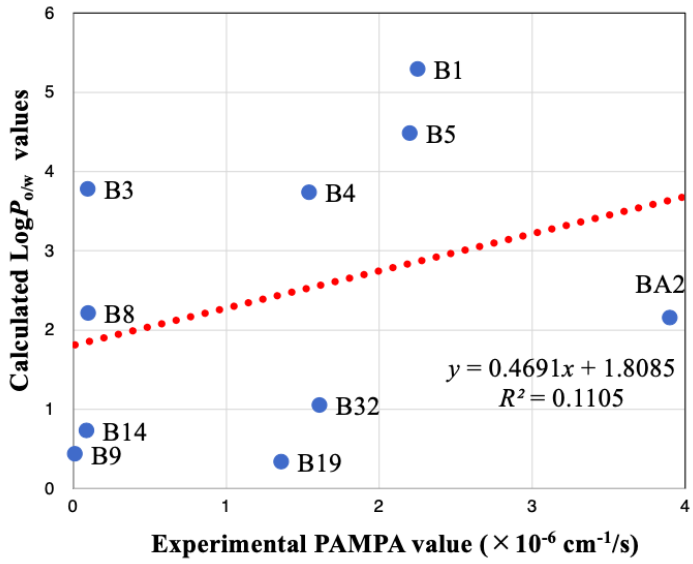


Figure 5-10. Correlation between the calculated Log$P_{o/w}$ by SMD/PM7 and experimental P_e values.

Table 5-1. The summary of the multiple regression analysis for small 205 compounds using SMD/PM7.

	Coefficient
Intercept	0.0000
Log$P_{o/w}$ value	0.8605
Number of multiple bonds in the cyclic structure	0.1894
Number of multiple bonds contained in compounds other than cyclic structures	0.0771
1 for cycloalkanes and 0 for non-cycloalkanes	0.2870
Number of N atoms	-0.2110
Number of O atoms	0.5441
Number of S atoms	-0.0554

Since I have established the correction scheme for Log$P_{o/w}$ values of the small molecules I first examine here the applicability of the correction scheme to the PAMPA values for the Bottromycin and its derivatives. The coefficients are summarized in Table 5-1. When the same method was used for 205 small molecules in the second study, the R^2 value did not improve much (Fig.5-11), from 0.11 to 0.14. I thought that the reason for this was that if I tried to use coefficients obtained from only small molecules to correct toward the middle-sized molecules (Table.5-2) with large molecular weights, the error would be too large, and it would not be able to adequately correct for large middle-sized molecules.

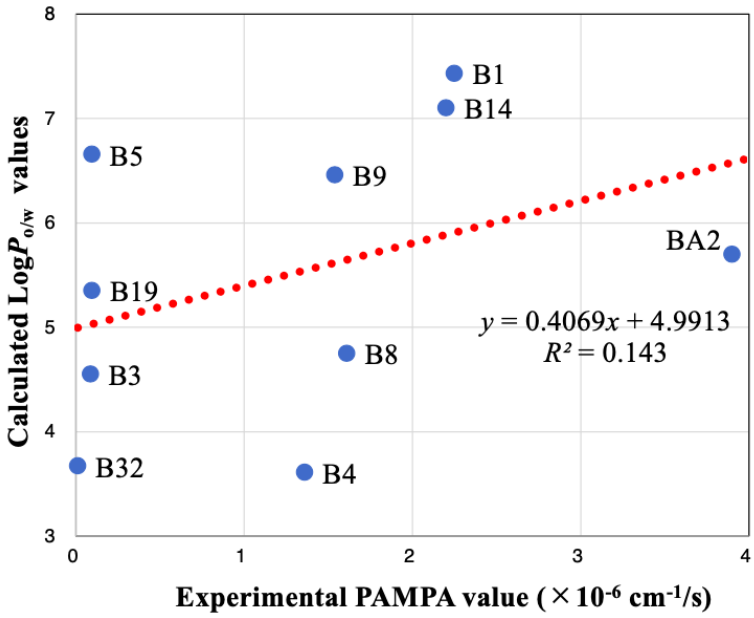


Figure 5-11. Correlation between the corrected Log$P_{o/w}$ values with the coefficient obtained for 205 small compounds and the experimental PAMPA values, P_e.

Next, I adopted the SMD PM7 calculation results for the Bottromycin and the derivatives for the multiple regression analysis by using the following explanatory variables: ① Log$P_{o/w}$ value obtained by calculation (value obtained by calculating ΔG_{wat} and ΔG_{oct} by each calculation method), ② Number of multiple bonds in the cyclic structure, ③ Number of multiple bonds contained in compounds other than cyclic structures, ④ Number of cycloalkane, ⑤Number of N atoms, ⑥ Number of O atoms, and ⑦ Number of S atoms.

Multiple regression analysis was also performed with the experimental values of PAMPA as the objective variables, and the correlation with the experimental values of PAMPA was examined by correcting the values of ① Log$P_{o/w}$. The R^2 value is considerably improved to 0.86. The *t*-value in the multiple regression analysis also shows that the value of Log$P_{o/w}$ is the largest and has the greatest influence as expected. However, when the number of target derivatives is small, having many explanatory variables will give arbitrary result.

Since multiple regression analysis cannot be performed correctly if the number of explanatory variables is increased when the data is small. Judging from the *t*- and *P*-values in Table 5-2, only three, ① Log$P_{o/w}$, ⑤ the number of N atoms, and ⑦ the number of S atoms are relatively of great significance among them. To reduce the arbitrariness, the multiple regression analysis was again performed by adopting these three

explanatory variables. Even after reducing the number of the explanatory variables, I'm able to maintain a relatively high R^2 value (0.74).

Looking at the last result only, it can be seen that the accuracy of the SMD PM7 calculations itself is not high enough. I found that the coefficients used in the second study were not sufficient for the correction, and a new multiple regression analysis with seven explanatory variables gave the same high accuracy as DFT. R^2 value is 0.74, which is a promising result. However, in this study, the number of target derivatives was small, and the multiple regression analysis could not be fully used. If I have a larger number of Bottromycin derivatives, I could increase the number of explanatory variables and obtain predictive values comparable to those of small molecules. The correlation between the calculated Log$P_{o/w}$ and LogP_e is shown in (Fig. 5-11(b)).

Table 5-2. Multiple regression analysis using 7 explanatory variables.

	coefficient	standard error	*t*-value	*P*-value
intercept	12.6130	15.8624	0.7952	0.5099
① Log$P_{O/W}$ value	1.0480	0.9381	1.1172	0.3801
② Number of multiple bonds in the cyclic structure	0.5466	1.1292	0.4841	0.6762
③ Number of multiple bonds contained in compounds other than cyclic structures	0.6462	0.9102	0.7100	0.5513
④ Number of cycloalkane	0.8600	3.2554	0.2642	0.8164
⑤ Number of N atoms	-2.0031	2.1613	-0.9268	0.4519
⑥ Number of O atoms	-0.0667	1.3305	-0.0501	0.9646
⑦ Number of S atoms	-3.3718	3.3193	-1.0158	0.4166

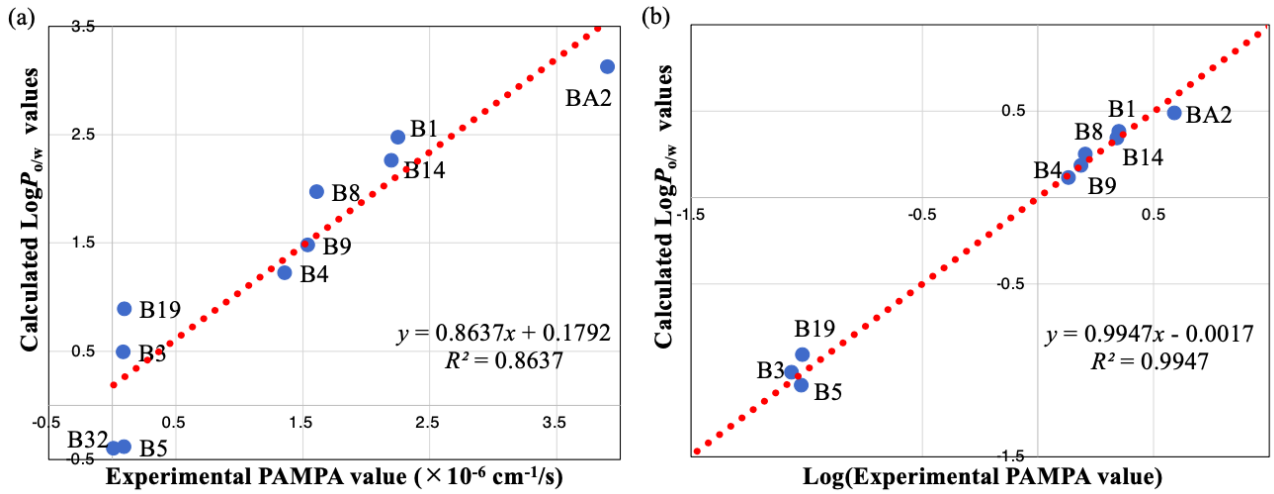


Figure 5-12. Correlations (a) between the corrected Log$P_{o/w}$ values and P_e values and (b) between the corrected Log$P_{o/w}$ values and LogP_e values by using the multiple regression analysis with 7 explanatory variables.

Table 5-3. Multiple regression analysis using Log$P_{o/w}$ and the numbers of N and S atoms.

	Coefficient	Standard error	*t*-value	*P*-value
intercept	12.3137	3.1973	3.8513	0.0084
Log$P_{o/w}$	0.6830	0.2104	3.2468	0.0175
N atoms	-1.0940	0.2869	-3.8135	0.0088
S atoms	-2.2504	0.9772	-2.3028	0.0609

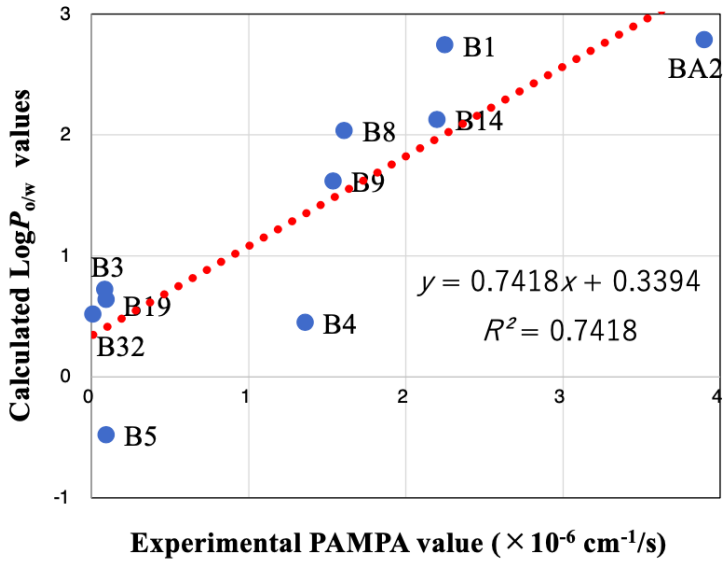


Figure 5-13. Correlations (a) between the corrected Log$P_{o/w}$ values and P_e values by using the multiple regression analysis with 3 explanatory variables (Log$P_{o/w}$ and the numbers of N and S atoms).

5-3-4 PaCS-MD

In previous study, Ms. Shimizu in Rikkyo University (Ref.5-15) reported that the Steered MD simulation using syringolin showed a good correlation between the results of the computer experiment and the laboratory experiment. The steered MD applies an artificial force to the drug molecules to the direction of the membrane to shorten the membrane permeation time compared to normal MD simulation. Sometimes the artificial force causes unphysical behavior when we treat the membrane permeation of a large molecule. Here, I wanted to investigate the membrane permeability of Bottromycin, which has a larger molecular weight than syringolin, using PaCS-MD. PaCS MD is expected to keep track the natural structural changes during the membrane permeation process, rather than the steered MD.

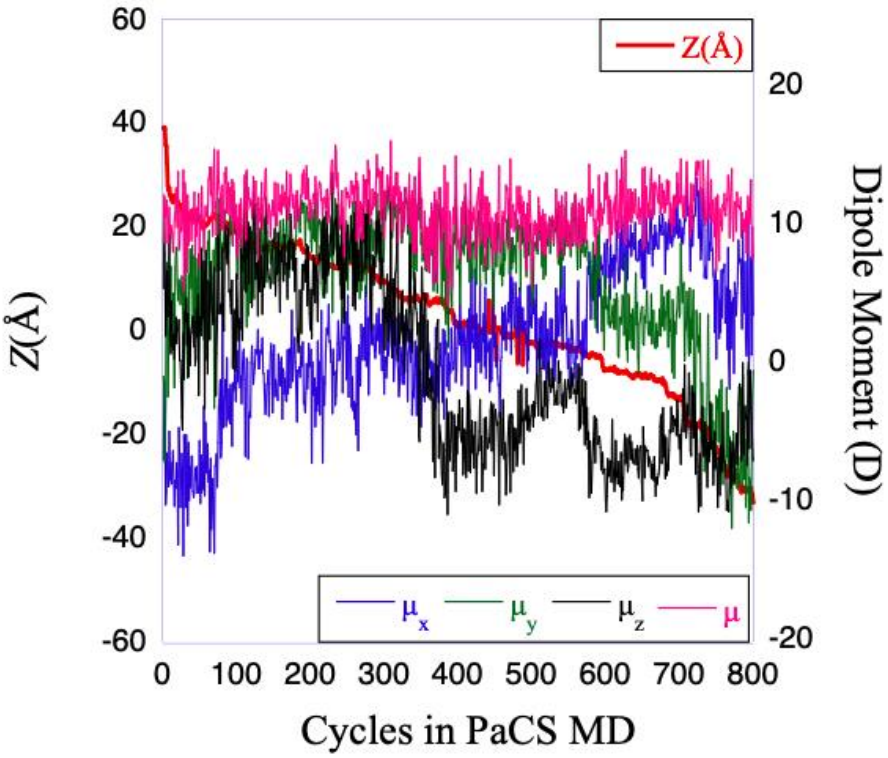


Figure 5-14. Z-coordinates and dipole moments, μ, of the ligand molecule during the cycles of PaCS-MD simulation. μ_x, μ_y, and μ_z are the magnitudes of the dipole moments for the *x*, *y*, and *z* directions, respectively.

Figure 5-14 shows the change in the membrane permeation of a drug using PaCS-MD. The right vertical axis is the *Z*-coordinate of the center of gravity of the drug, the left vertical axis is the Dipole-Moment, and the horizontal axis is the number of cycles in PaCS-MD. It is noteworthy that the Dipole-Moment (yellow) in

the *Z*-direction is reversed at the center of the membrane, indicating that when the drug enters the membrane, it starts from the cyclic part and finally enters the linear structure, reverses in the membrane, and exits from the linear structure and finally exits the cyclic structure. If the linear structure is left behind in the water layer outside the membrane and becomes stabilized, or if the inversion is prevented by binding to the membrane surface, it is expected to have difficulty in permeating the membrane.

Table 5-4. Number of cycles in PaCS-MD required for the drug molecule to pass through the membrane when each initial structure is made with water, aniline and octanol calculated by SMD/B3LYP calculations.

Passing through cycle	Water	Aniline	Octanol
B1	1720	1171	1190
B3	1233	**2917**	1593
B4	1149	**2293**	806
B5	2206	1642	2174
B8	1985	1420	1108
B9	1906	1354	1745
B14	2965	1078	1226
B19	995	1446	1591
B32	1201	1921	960
BA2	1043	**2718**	1503

Table 5-4 shows the number of cycles required for the membrane permeations of the drug molecules stabilized in water, aniline and octanol using the density functional theory. The results show that the number of cycles required for the same molecule to pass through the membrane varies greatly depending on the difference among the initial structures, which suggests that the drug molecule has not undergone sufficient structural change during permeation. In addition, I focused on B3, B4, and BA2, which take much longer to permeate the membrane when the initial structure is made in aniline solvent.

I made the initial structure with this aniline and sampled the membrane permeation with PaCS-MD and correlated the number of cycles required for the membrane permeation with the experimental value of PAMPA (Fig. 5-15). Since the edge of the derivative shown in Fig. 5-8 can be divided into three types: (i) negatively charged, (ii) nonpolar, and (iii) positively charged derivatives, I divided the derivative into three groups and correlated each group with the experimental value, but the R^2 value did not improve to around 0.3. Figure 5-

15 shows that B3, B4, and BA2 are far apart from the other groups, and it seems that a good correlation can be made by classifying them. B3, B4, and BA2 are derivatives with a significantly larger number of cycles required for membrane permeation than the others, and I thought there might be a cause for this.

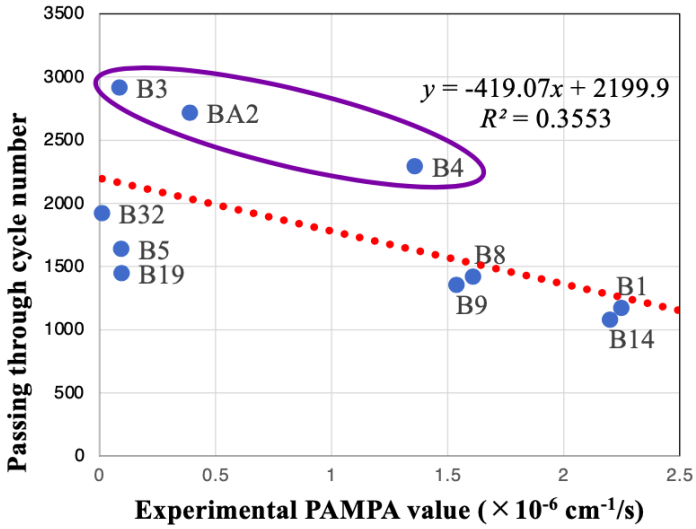


Figure 5-15. Number of cycles in PaCS-MD required to pass through the membrane when each initial structure is stabilized in aniline solvent calculated by SMD/B3LYP calculations.

To consider the reasons for this, the molecular structure was again examined. B3, B4, and BA2 contain amide and ester bonds in their substituent groups, and it is thought that these bonds with atoms on the membrane surface prevented the membrane permeation. B1, B5, and B8 also contain amide and ester bonds, but they were not exposed to the surface during the membrane permeation. Other researchers have used PaCS-MD to investigate how botromycinA2 enters the membrane layer from the water layer by holding water inside the molecule. Therefore, for this membrane permeation, I must consider the surrounding water molecules as an integral part of the membrane, as I investigated in the experiments in Chapter 3.

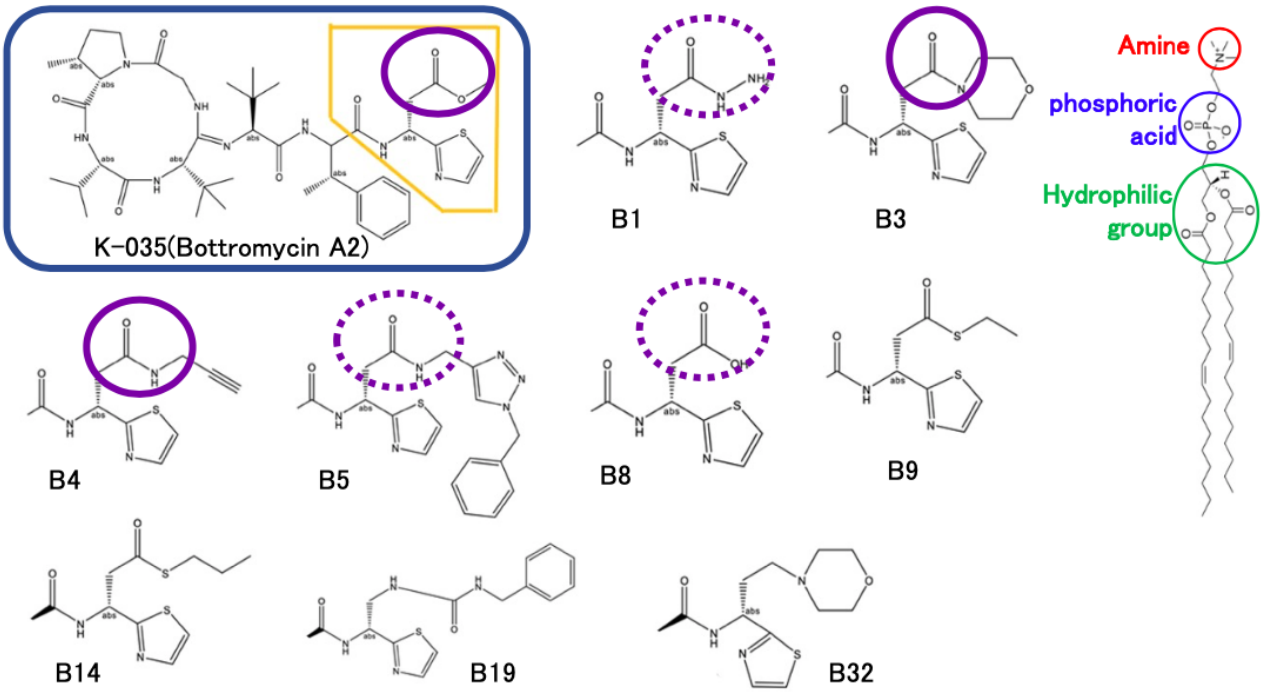


Figure 5-16. BotromycinA2 and its derivatives used in the experiment. Amide bond, carboxylic acid. Purple circles are amide bond, carboxylic acid. I believe that solid lines tend to expose functional groups on the surface, as opposed to dotted lines.

5-3-5 Analysis by PerMM

I have used DFT to calculate the structure of bottromycinA2 derivatives stabilized in aniline as an initial structure and submitted it to PerMM. PerMM gives a pdb file of 15 structures of the drug on the way to the membrane. Figure 5-17 shows the structures at the moment of membrane penetration among the 15 structures. It can be seen that the cyclic structure enters the membrane first and the linear part is still outside the membrane. Figure 17 (right) shows behavior of structural change upon permeation out of the 15 structures that can be easily seen during membrane permeation. The upper blue layer and the lower red layer corresponds to the inner and outer cell membrane, respectively.

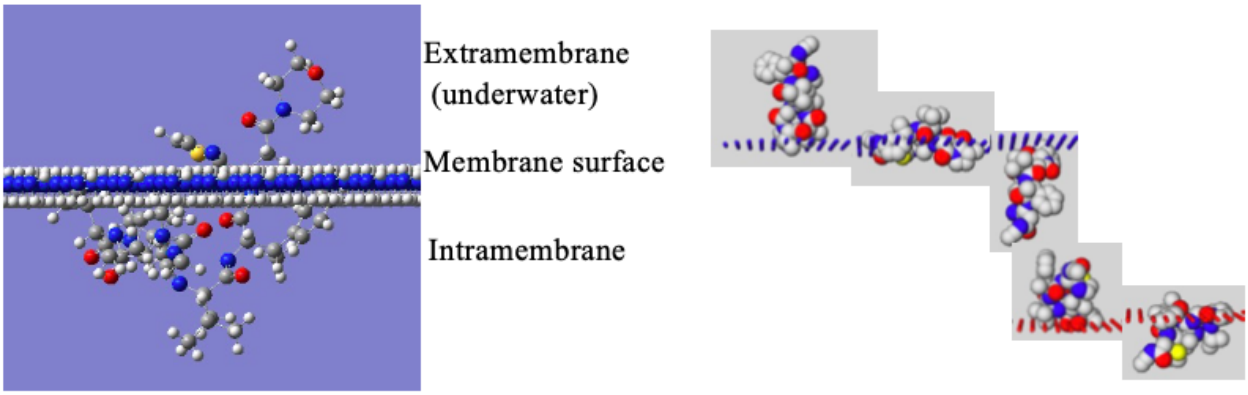


Figure 5-17. Changes in the molecular structure of Bottromycin during the membrane permeation sampled by PerMM.

Figure 5-18 (a) or (b) shows the scatter plot between the membrane permeabilities for DOPC or the PAMPA-DS values, which are ones of the membrane permeation indices calculated by PerMM and the experimental PAMPA values. Table 5-5 shows that the R^2 value of DOPC is 0.015, which is quite poor, while the R^2 values of the other values are slightly better, around 0.17. Also, according to the correlation of PAMPA-DS as illustrated in Fig. 5-18(b), the derivatives can be classified into three groups, i.e., negative charge, non-polar, and positive charge ones. In fact, as shown in this table, the values of all indices except for DOPC can be increased to around 0.7. As seen in Fig. 5-18 (a), it is difficult to classify them by the charge of derivatives, but three data, B3, B4, and BA2, seem to be different from the others. As will be shown later, this group is a group of derivatives that have amide and ester bonds in the substituents.

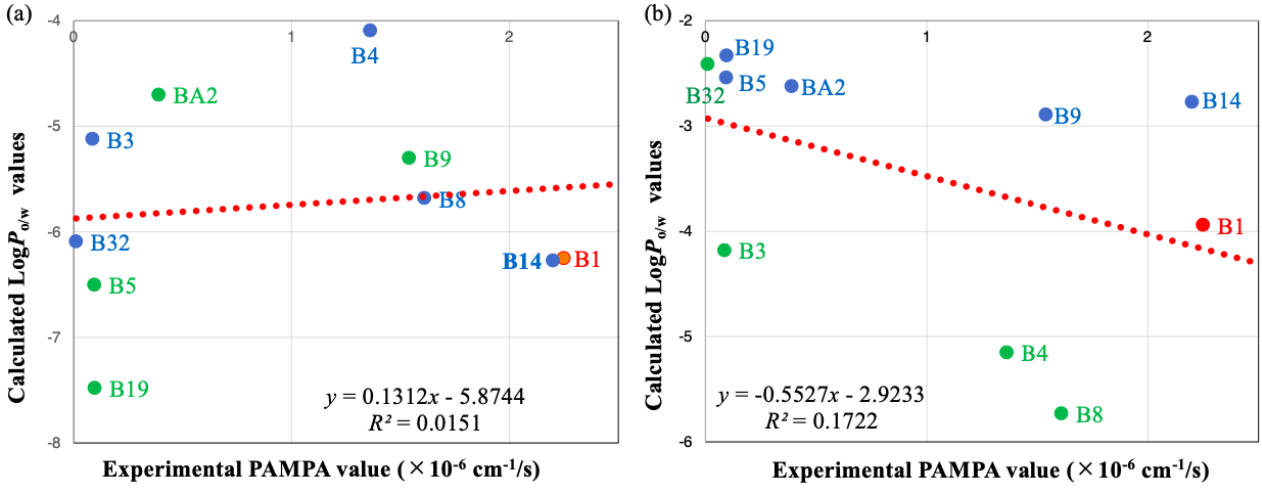


Figure 5-18. Correlation between PerMM calculated values (a) DOPC or (b) PAMPA-DS and experimental PAMPA values. Green, Red, and Blue points mean non-polar, positive charged, and negative charged chemical group in the Bottromycin derivative, respectively.

Table 5-5. Correlation between PerMM-calculated values and PAMPA experimental values and those with classification by charge.

Additional molecule	**Correlation with experimental values (R^2 value)**					
	DOPC	**BLM**	**BBB**	**CACO2**	**PAMPA-DS**	**plasma membrane**
All molecules	0.0151	0.1717	0.1712	0.1719	0.1722	0.1723
-Charge	0.1609	0.7793	0.7786	0.781	0.7803	0.7792
Non-polar	0.0606	0.6601	0.6298	0.7131	0.6581	0.6561

However, since PerMM uses 3-dimensional coordinates, it is possible that differences in the initial stabilized structure may affect the results. In order to confirm the initial structure dependence, I used PerMM to calculate

the structure stabilized by water, aniline, and octanol obtained by the DFT calculations (Table 5-6). As seen the results in Table 5-6, it is found that the results depend greatly on the initial stabilized structure, even for drugs with the same molecular structure, which is different from our initial prediction. BottromycinA2, which is used in these calculations, has a large structure and several conformations, so that I thought it would be very difficult to discuss the results of PerMM, which is greatly affected by the initial structure.

Table 5-6. Relationship between initial stabilized structure and PerMM calculated values.

		DOPC	BLM	BBB	CACO2	PAMPA-DS
B1	water	-7.03	-0.33	-3.01	-3.57	-1.54
	aniline	-6.86	-0.24	-2.98	-3.54	-1.45
	octanol	-7.81	-0.04	-2.91	-3.49	-1.26
B3	water	-5.93	-1.78	-3.52	-3.93	-2.87
	aniline	-4.69	-3.11	-3.99	-4.28	-4.10
	octanol	-8.07	0.08	-2.87	-3.46	-1.15
B4	water	-7.54	2.32	-2.08	-2.89	0.91
	aniline	-5.60	-1.66	-3.48	-3.90	-2.75
	octanol	-10. 0	0.92	-2.57	-3.24	-0.38
B5	water	-6.90	0.16	-2.84	-3.44	-1.08
	aniline	-7.38	0.66	-2.66	-3.31	-0.62
	octanol	-7.97	-0.07	-2.92	-3.50	-1.29
B8	water	-5.85	-2.16	-3.66	-4.03	-3.22
	aniline	-5.34	-2.83	-3.89	-4.20	-3.84
	octanol	-6.45	0.69	-2.65	-3.30	-0.59
B9	water	-7.42	0.30	-2.79	-3.40	-0.95
	aniline	-6.60	0.59	-2.69	-3.33	-0.68
	octanol	-6.39	-0.51	-3.07	-3.61	-1.69
B14	water	-7.70	-0.26	-2.99	-3.55	-1.45
	aniline	-6.42	-0.93	-3.22	-3.72	-2.08
	octanol	-7.41	-0.42	-3.04	-3.59	-1.61
B19	water	-8.23	0.89	-2.58	-3.25	-0.41
	aniline	-8.00	2.05	-2.17	-2.96	0.66
	octanol	-7.72	0.68	-2.65	-3.31	-0.60
BA2	water	-5.73	-0.60	-3.11	-3.63	-1.78
	aniline	-6.98	-1.61	-3.46	-3.89	-2.72
	octanol	-8.62	1.79	-2.26	-3.02	0.43

5-4 Summary

I have attempted to use three calculation-based methods for evaluation of membrane permeability of cyclic peptides, each of which has its own advantages and disadvantages as follows.

The method using density functional theory (SMD/DFT) and multiple regression analysis, had a high accuracy in R^2 value of around 0.84 for Bottromycin and derivatives. The disadvantage is that the calculation time takes a month for some cases. The reason for this is that it is difficult to explore completely stable structures because of the conformations. Unlike small molecules, the method was still not practical. Using the semiempirical method, the accuracy of SMD/PM7 is comparable with SMD/DFT calculations. By conducting the multiple regression analyses, I could predict with high accuracy comparable to the SMD/DFT calculations. In addition, the calculation of SMD/PM7 can be performed in a shorter time than SMD/DFT, which makes it suitable for practical use in the prediction.

The second method focused on PaCS-MD based on MD calculations. Although the membrane permeation time is very fast, and there is a risk that there is not enough time to change the structures of the middle-sized molecular drug. This risk can be seen that the different initial structure gives different steps in PaCS-MD simulations. In addition, it was reported by other researchers that when the drug moves from the water layer to the membrane layer, it takes on a structure that holds water molecules in its molecular structure. I think it is necessary to consider the structure of water molecules that cover the surface of the drug, mainly ions, as I considered in the research of Chapter 3. It is necessary to prepare the initial structure carefully because differences in the initial structure can have a significant effect on the calculation results.

As the third method, the PerMM approach is easy to use because it can calculate using only the molecular structure, and the calculation time is very short, and it can get the result in a few seconds even if it is a huge BottromycinA2. However, there was a problem that the results differed depending on the initial stabilized structure because the 3-dimensional structure was inputted. Moreover, the accuracy was not so good, and a correction was necessary to obtain the reliable result.

Considering the three calculation methods, I conclude here that the first method using SMD PM7 gives the most practical and reliable results. Although it is not very accurate enough ($R^2 = 0.15$), by performing multiple regression analysis, I have been able to increase the R^2 value to 0.86. This result is comparable to the first SMD DFT calculation, which has a higher calculation accuracy. Nevertheless, the computation time required for

SMD PM7 is much less than that for SMD DFT. It is expected that SMD PM7 can be used for practical prediction in a short calculation time. In the present work, I'm not able to perform multiple regression analysis sufficiently because the number of target drug molecules was only about 10. By increasing the number of target drug molecules, I expect to be able to further determine the explanatory variables in detail and make highly accurate predictions.

5-5 References

1. C. Hamers-Casterman, T. Atarhouch, S. Muyldermans, G. Robinson, C. Hammers, E. Bajyana Songa, N. Bendahman, R. Hammers, "Naturally occurring antibodies devoid of light chains", *Nature* **363**, 446-448 (1993)
2. "*The Potential of Middle Molecule Drug Discovery: Developing Peptides for Drug Discovery*" (in Japanese) in Chemistry Today / Modern Chemistry Editorial Group, ed. (543): 47-51 (2016)
3. S. Hayashi, "*An introduction to the transport theory*" (in Japanese) Toyo Shoten, (2007).
4. D. J. Craik. "Seamless Proteins Tie Up Their Loose Ends". *Science* **311**, 1563-1567(2006).
5. T. Yamada, M. Yagita, Y. Kobayashi, G. Sennari, H. Shimamura, H. Matsui, Y. Horimatsu, H. Hanaki, T. Hirose, S. Omura, T. Sunazuka, "Synthesis and Evaluation of Antibacterial Activity of Bottromycins", *J. Org. Chem.*, **83**(13), 7135-7149(2018).
6. C. Masungi, J. Mensch, A.Van Dijck, C. Borremans, B. Willems, C. Mackie, M. Noppe, E. Brewster, "Parallel Artificial Membrane Permeability Assay (PAMPA) combined with a 10-day multiscreen Caco-2 cell culture as a tool for assessing new drug candidates", *Pharmazie* **63**, 194-199 (2008).
7. W. Kohn; L. J. Sham, "Self-Consistent Equations Including Exchange and Correlation Effects". *Phys. Rev.* 140 (4A), A1133-1138(1965).
8. R. G. Parr; W. Yang, "Density-Functional Theory of Atoms and Molecules" (International Series of Monographs on Chemistry Book 16) Oxford University Press, (1989).
9. T. Tsuneda, "Basics of Density Functional Thoery" (in Japanese), Kodansha Scientific books (2013).
10. R. Harada, A. Kitao, *J. Chem. Phys.*, **139**(3), 035103 (2013).
11. A.L. Lomize, Irina D. Pogozheva, "Physics-based method for modeling passive membrane permeability and translocation pathways of bioactive molecules", *J. Chem. Inf. Model.* **59**(7), 3198-3213 (2019).
12. A.L. Lomize, J.M. Hage, K. Schnitzer, K. Golobokov, M.B. LaFaive, A.C. Forsyth, I.D. Pogozheva, "PerMM: A web tool and database for analysis of passive membrane permeability and translocation pathways of bioactive molecules", *J. Chem. Inf. Model.* **59**(7), 3094-3099 (2019).
13. H. Ridley*, "The Anatomy of the Brain"*, London, Printers to the Royal Society (1965).
14. I.J. Hidalgo, T.J. Raub, R.T. Borchardt, "Characterization of the human colon carcinoma cell line (Caco-2) as a model system for intestinal epithelial permeability", *Gastroenterology,* **96**(3), 736-749 (1989).

15. K. Shimizu, “Research on middle molecule drug discovery for the realization of next generation treatment and diagnosis that combines computational chemistry and experiment”, Master thesis, Rikkyo University (2019).

Chapter 6 General Conclusion

In these three studies, I have been able to elucidate the permeability of drugs to the cell membrane.

In the first study, I was able to investigate the behavior of ions and water molecules around the cell membrane using the MD simulations. I obtained the different ion distribution of Na^+ and Mg^{2+} from that predicted theoretically by the Guoy-Chapman theory. As mentioned earlier, for the many-body problem, I believe that the results of the computer experiments such as MD simulations describe the events more accurately than algebraic theory such as Gouy-Chapman theory. I also investigated the behavior of water molecules around ions and found that water molecules formed various layers around the ions depending on their valence. This is a new finding that could not be seen even with a microscope. This finding is important not only for the evaluation of membrane permeability of drugs but also for the investigation of the cause of protein folding and motor protein movement.

In the second study, the Log$P_{o/w}$ values of 205 compounds were evaluated using the combinations of the semiempirical quantum chemical method and the polarizable continuum model, Log$P_{o/w}$ values of 205 compounds were evaluated. Since the R^2 value between experimental and calculated Log$P_{o/w}$ values was up to 0.7, I used the calculation data and several explanatory variables to predict the Log$P_{o/w}$ values. The multiple regression analysis gave a better correlation between the experimental and computed Log$P_{o/w}$ values, and the R^2 value was improved to around 0.95. In particular, the results SMD-PM7 gave the best the water-octanol partition coefficient Log$P_{o/w}$ values, which enabled us to predict the membrane permeability index with practical accuracy, and the calculation time was as short as half a day for 205 compounds, while DFT requires one week for them.

In the third study, I explored the best way to evaluate the cell membrane permeability of drugs using computer simulations. For BottromycinA2, a typical middle-sized molecular drug, I estimated Log$P_{o/w}$ calculations using SMD DFT and SMD PM7 (quantum mechanical calculations), PaCS-MD (molecular dynamics simulations), and PerMM (empirical model), to find the correlation between their calculated values and experimental membrane permeability by PAMPA experiments. As a result, Log$P_{o/w}$ values using SMD PM7 were found to be the best fit to the experimental membrane permeability. The calculation time was short and could be completed within half a day to a day for one Bottromycin derivative, and the accuracy of the

prediction was high with R^2 values around 0.86 when using the multiple regression analysis using 7 explanatory variables. In this study, I'm only able to use about 10 kinds of molecules, but if I can increase the number of molecules used in the experiments, I will be able to make more accurate prediction model for the Bottromycin and its derivatives.

In this thesis, I used several computational biophysical methods at the atomic level, such as first-principles calculations, molecular dynamics calculations, and data science methods, to clarify the physical properties of membranes through the interaction of water and molecules with them. In particular, I found that the ion distribution around biological membranes differ from the classical macroscopic theory and clarified the physical mechanism of the difference. This knowledge is quite general in the atomic-level simulation, but worth mentioning how this simulation is useful to understand the physics of the membrane. In addition, I developed a highly accurate prediction method of $\mathrm{Log}P_{\mathrm{o/w}}$ value, one of the indexes of membrane permeability, using first-principles calculations, and examined the applicability of this method to drug screening by correlating it with experimental data of membrane permeability tests of middle-sized molecules. The study of biological membranes is a fast-growing frontier in the life sciences, and technological advances in this field have enabled significant progress and discoveries not only in the field of biophysics, but also in drug discovery, medicine, and agriculture. Therefore, this doctoral dissertation is of extremely high academic value, as it pioneers research in an interdisciplinary field that crosses material science and life science based on physical theory.

Acknowledgements

I would like to take this opportunity to express my best gratitude to Professor Yasuteru Shigeta, Ryuhei Harada, and Toru Matsui for their assistance throughout my doctoral studies. I also acknowledge the collaborators, Dr. Yuki Nagata for the MD simulation protocol in the first work and Prof. Hiroyuki Miyachi, Prof. Katsumi Maenaka, and Prof. Toshiaki Sunazuka for providing the experimental data of PAMPA.

List of Publications

1. K. Yanagi, Y. Mitsuta, K. Takaoka, T. Takahashi, K. Hengphasatporn, R. Harada, Y. Shigeta, “Solubility and Membrane Permeability of Cyclic Dipeptides Approximately Estimated by Quantum Chemistry and Molecular Dynamics Simulations”, *Chemistry Letters*, 50(12), 1964-1967 (2021). **DOI**: 10.1246/cl.210488
2. T. Takahashi, T. Matsui, K. Hengphasatporn, Y. Shigeta, “A Practical Prediction of Log$P_{o/w}$ through Semiempirical Electronic Structure Calculations with Dielectric Continuum Model”, *Bulletin of Chemical Society of Japan*, **94**(7), 1807-1814 (2021). **DOI**: 10.1246/bcsj.20210035
3. T. Takahashi, R. Harada, Y. Shigeta, "Distribution of Counter Ions in Negatively-charged Lipid/Water/Air Interface: Molecular Dynamics Study", *Chemistry Letters* **49** (4), 361-363(2020). **DOI**:10.1246/cl.200043