

Kyoto university



DOCTORAL THESIS

**Molecular dynamics study
focusing on
flexible molecular assemblies**

By

Yuki Yamamoto

Theoretical Chemistry Group

Department of Chemistry

Graduate School of Science

Supervisor: Prof. Shigehiko Hayashi

December, 2024

Contents

1	General introduction	3
2	Liquid Structures and Diffusion Dynamics of Alkyl-Pyrene Liquids	6
2.1	INTRODUCTION	7
2.2	COMPUTATIONAL METHODS	10
2.2.1	Force field	10
2.2.2	MD simulation	13
2.2.3	Body-fixed coordinates	16
2.3	RESULTS	18
2.3.1	Translational diffusion	18
2.3.2	Rotational diffusion	22
2.3.3	The van Hove distribution function	26
2.3.4	Static structures	26
2.3.5	Dynamic structures	30
2.3.6	Spatially resolved two-body time correlation function	32
2.4	DISCUSSIONS AND CONCLUDING REMARKS	38
2.5	APPENDIX	42
2.5.1	Force field parameters	42
2.5.2	Parameter fitting for the mixed KWW function	44
2.5.3	Confidence interval estimations	45
3	Self-assembly of V-shaped Polyaromatic Amphiphiles	46
3.1	INTRODUCTION	47
3.2	COMPUTATIONAL DETAIL	50
3.2.1	Force field	50
3.2.2	MD simulation	54
3.2.3	The VPAA assembly definition	55
3.3	RESULTS	56
3.3.1	Time evolution of the VPAA assemblies	56
3.3.2	Lifetime of molecular assembly	58
3.3.3	Molecular symmetry corrected RMSD	62
3.3.4	Branch-and-prune method	64
3.3.5	Similarity matrix	68
3.3.6	Distribution of the symRMSD	70
3.3.7	Structural clustering	71

3.3.8	Representative structure	75
3.3.9	Assemblies of two VPAAAs	76
3.3.10	Assemblies of more than two VPAAAs	79
3.4	DISCUSSION	84
3.5	CONCLUSION	87
3.6	APPENDIX.....	88
3.6.1	Sample size	88
3.6.2	Confidence interval of MSD	89
3.6.3	Confidence interval of mean lifetime	91
4	Molecular-Oriented RMSD with Branch-and-Bound	103
4.1	INTRODUCTION	104
4.2	METHOD	106
4.2.1	Molecular-oriented RMSD.....	106
4.2.2	Branch-and-bound.....	107
4.2.3	Pruning rule.....	107
4.2.4	Branching strategy.....	110
4.2.5	Search strategy.....	112
4.2.6	Multi-component system.....	113
4.2.7	mobbRMSD	118
4.3	RESULTS	119
4.3.1	Implementation	119
4.3.2	Random coordinates	119
4.3.3	Computational cost.....	120
4.3.4	Test with random coordinates.....	122
4.3.5	Comparison with previous studies	125
4.3.6	Test with aromatic micelles	129
4.4	DISCUSSION	132
4.4.1	Methodological characteristics	132
4.4.2	Correlation between RMSD and computational cost	132
4.4.3	Early termination	137
4.4.4	Minimum spanning tree	139
4.5	CONCLUSION	140
4.6	DATA AND SOFTWARE AVAILABILITY.....	140
4.7	APPENDIX.....	141
4.7.1	Proof of Eq. 4.7.....	141
4.7.2	Proof of Eq. 4.12	142
4.7.3	Proof of Eq. 4.13	143
4.7.4	DSTS algorithm for multi-component system	144
4.7.5	Memoizing.....	147
4.7.6	Thinning out of p	147
4.7.7	Coordinate ordering	148
5	Conclusion and the future prospects	149

Chapter 1

General introduction

In recent years, flexible molecular assemblies have attracted much attention in the context of novel functional materials, [1–3] including functional molecular liquids (FMLs), [1] aromatic micelles, [2] and soft crystals. [3] FMLs are a new generation of fluid materials which is a solvent-free and non-volatile at room temperature, combining the excellent optoelectronic properties produced by the incorporated π -conjugated assemblies with the ease of handling and stability of liquid materials. [1] Aromatic micelles are a series of amphiphiles with aromatic moieties and polar alkyl chains as hydrophobic and hydrophilic groups, respectively, which spontaneously form micellar molecular assemblies consisting of several molecules in water, encapsulating various extended aromatic molecules and solubilizing them in water. [2, 4, 5] Soft crystals are a group of solid-state materials that have highly ordered structures but can transition between different phases with low activation free energy; [3] they exhibit remarkable stimulus-response properties such as mechanochromism [6–9] and vapochromism. [10, 11] Excellent concepts utilizing flexible molecular assemblies have been proposed in rapid succession, and they are paving the way for new fields of chemistry. However, their molecular design is not easy; the relationship between molecular interactions and macroscopic flexibility is complex, and it is difficult to synthesize them predictably. Currently, it can only be assessed by exhaustive synthesis and measurement of macroscopic physical properties. Therefore, it is urgent to predict *in-silico* the macroscopic properties resulting from intermolecular interactions.

There are two major trends in *in-silico* chemistry: data chemistry and simulations based on theoretical chemistry. Data chemistry predicts physical properties based on the classification of data, [12–14] but it does not provide causal explanations. Theoretical simulations can reproduce chemical properties from basic physics equations and chemical structures. [15, 16] Simulations of the flexible molecular assemblies given by intermolecular interactions require atomistic precision and long-time dynamics. In particular, molecular dynamics (MD) simulations based on the Newton equation are good way to achieve atomically accurate long-time dynamics, but reproducing the flexibility of molecular assemblies is a difficult task due to limitations in computational resources.

MD simulations originated with Aneesur Rahman's work on liquid argon, [17] and continue to develop and play an increasingly role in physical chemistry today. [16] In general, MD simulations are based on three assumptions: to represent particle interactions as a func-

tion of position, to adequately sample the statistical configuration of the molecular system, and to compute the desired physical properties from the sample. MD simulations are often successful models that link atomic resolution and macroscopic properties with statistical mechanics. The MD simulation is a well-successful model despite its bold simplifications. [15, 16, 18, 19] The statistical dynamic ensemble of sampled atomic resolution and several macroscopic physical quantities are linked by statistical mechanics. [15] However, the purpose of MD simulations is not only to compute physical quantities; that is, causal modeling based on domain knowledge is needed to derive essential insights. [16, 18, 19] Even if improved computational power replaces the three assumptions of MD simulations with better ones, causal modeling is still necessary and is an important task for researchers. [20]

In this thesis, we present two MD simulation studies that reveal causal explanations for the flexibility of molecular assemblies: viscosity changes of FMLs and self assemblies of aromatic micelles.

In Chapter 2, we present the study of the rheological changes of FMLs caused by alkyl chains. We focus on four types of FMLs with pyrene as the functional π -core, [21] which are non-volatile Newtonian liquids at room temperature. The four pyrene-FMLs differ only in the number and substitution position of branched alkyl chains. The rheological properties of these molecules deviate significantly from empirical theories of the viscosity of molecular liquids, *i.e.* the Stokes-Einstein relationship, which assumes a hydrodynamic sphere with Brownian motion in a fluid. [22] Elucidating this mechanism at atomic resolution is important for the FML strategy of tuning physical properties by modulating alkyl chains. We performed long MD simulations of tens of microseconds each, totaling submilliseconds, which are sufficiently descriptive of the fluid properties of the pyrene FMLs. Furthermore, through analyses using several spatio-temporal correlation functions reveal their characteristic local liquid structures and provide a consistent causal explanation for the origin of their anomalous rheology. This issue is reprinted with permission from *J. Phys. Chem. B* **2023**, 127, 21, pp.4870-4885. Copyright **2023** American Chemical Society.

In Chapter 3, we present the study of aromatic micelles, focusing on their self-assemblies and ability to encapsulate a wide range of extended aromatic molecules. [2, 4, 5] We report MD simulations of two types of aromatic micelles called V-shaped polyaromatic amphiphiles (VPAAAs) which form micelle-like non-bonded self-assemblies in aqueous solution. [23] The broad encapsulation capability indicates the high flexibility of self-assemblies such as micelles. However, this flexibility masks experimental atomic resolution details, making it difficult to gain insights that lead to molecular design. To sufficiently sample the equilibrium structure of the VPAA assemblies in water, 10 microseconds MD simulations and 1 microsecond replica exchange MD simulations were performed. Then, to reveal the detailed molecular structures of the VPAA assemblies, we developed a novel similarity metric that can be used for structural analysis of molecular assemblies and performed structural clustering. Our simulations and new analytical methods provide a complete picture of the ensemble of molecular structures of the VPAA assemblies and an explanation for their extensive encapsulation ability. This issue is reprinted with permission from *J. Phys. Chem. B* **2024** (<https://doi.org/10.1021/acs.jpcb.4c06937>). Copyright **2024** American Chemical Society.

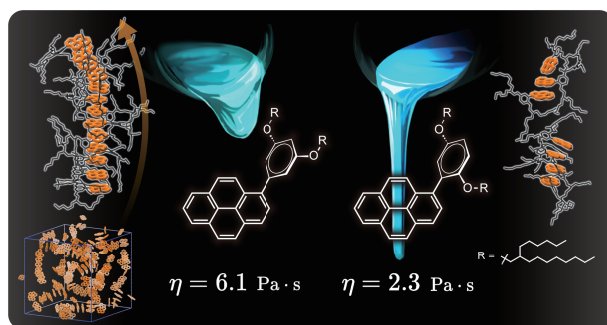
In Chapter 4, we introduce a new methodology that improves on the method developed

in the previous chapter. Root mean squared deviation (RMSD) is one of the most common metrics for comparing the similarity of three-dimensional chemical structures. However, calculating the RMSD of a molecular assembly requires establishing optimal atom mapping, which creates difficulties due to combinatorial explosion. The novel method, called molecular-oriented RMSD with branch-and-bound (mobbRMSD), uses the branch-and-bound method to obtain an exact solution for RMSD. mobbRMSD is formulated in molecular-oriented coordinates and can handle large, complex chemical systems that are difficult to handle using conventional methods, such as molecular liquids, solute solvation, and self-assemblies of large molecules. Our methodology integrates 3D structures and data chemistry of soft molecular assemblies, paving the way to explaining the causal relationships between the dynamics and properties of soft molecular assemblies that have remained a black box.

Chapter 5 is a conclusion and provides a summary of this thesis and future research prospects.

Chapter 2

Liquid Structures and Diffusion Dynamics of Alkyl-Pyrene Liquids



Functional molecular liquids (FML) based on alkylated π -conjugated molecules have attracted attention as solvent-free and nonvolatile liquid materials with prominent optoelectronic features. Recently, novel FML compounds containing pyrene as the functional core were synthesized, and their rheological and photochemical properties were investigated. Although the molecules differ only in the number of alkyl chain substituents and their substitution positions, their viscosity coefficients are largely different beyond Stokes-Einstein relation on the assumption of the identical microscopic friction, indicating that local microscopic molecular interactions are crucial for the macroscopic rheological properties. Here we report a theoretical study on rheological properties of the alkyl-pyrene liquids by means of atomistic molecular dynamics (MD) simulations. We performed long-time MD simulations for tens of microseconds to obtain ample statistical samples of the alkyl-pyrene liquids, and analyzed their liquid structures and diffusion dynamics based on spatio-temporal correlation functions. We found formation of characteristic local liquid structures of π - π stacking of the pyrene moieties and locally anisotropic and anomalous diffusion dynamics which remarkably vary depending on the alkyl substituent patterns. The present results provide an atomistic insight into macroscopic rheological properties of alkyl- π FMLs and molecular design strategy for them.

2.1 INTRODUCTION

Functional molecular liquids (FMLs) have attracted attention as a new generation of solvent-free, nonvolatile room temperature fluid materials [1, 24] for various applications such as organic light-emitting diodes, [25–27] luminescence, [28–35] photoconduction, [36–39] molecular recognition, [40, 41] and energy storage. [42–45] Recently, a synthetic strategy called alkyl- π engineering has been developed, allowing the synthesis of FMLs with prominent properties by combinations of bulky yet flexible branched-alkyl chains and optoelectronically active π -conjugated moieties. [24, 46, 47] The combinations lead to the formation of entropy-driven fluid states because of the suppression of intermolecular interactions between π -conjugated moieties by the alkyl chains. This synthetic approach can therefore be easily applied to a variety of π -conjugated molecules and thus is expected to greatly expand the FML library. [1, 45, 47] Furthermore, aggregation structures of the π -conjugated moieties in FMLs are controlled by the alkyl chain substituents and their substitution positions, occasionally generating unexpected photochemical and rheological properties. [48]

Recently alkyl- π liquid compounds **1-4** containing pyrene (**Figure 2.1a**) as the functional π -moiety were synthesized, and their rheological and photochemical properties were investigated. [21] One phenyl group with two 2-hexyldecoxy substituents (-OC₆C₁₀) is attached to pyrene for the compounds **1** and **2**, and four of them are attached for the compounds **3** and **4**. Positions of the two 2-hexyldecoxy substituents are also different. The two 2-hexyldecoxy substituents are attached at the 3,5-positions of the phenyl group(s) for the compounds **1** and **3**, while they are attached at the 2,5-positions for the compounds **2** and **4**. All these solvent-free alkyl-pyrene liquids behave as highly viscous Newtonian liquids at room temperature, and any long-range order is absent in the neat liquid state.

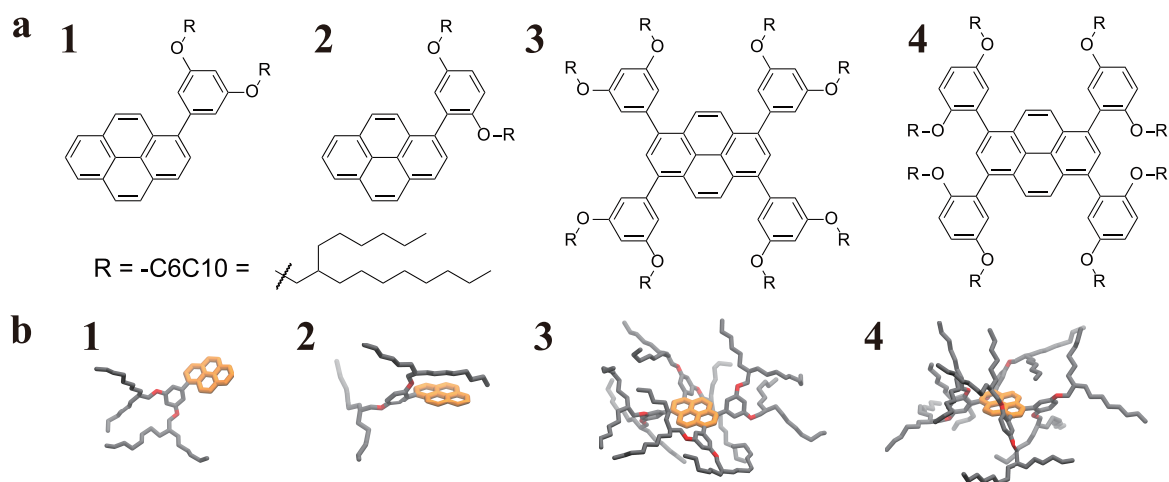


Figure 2.1: Molecular structures of alkyl-pyrene liquids. (a) Chemical structures of the compounds **1-4**. (b) Representative snapshots of molecular structures of the compounds **1-4** in MD simulations of the alkyl-pyrene liquids. [48]

Although molecular structures of the alkyl-pyrene liquids **1-4** differ only in the number of the alkyl chain substituents and their substitution positions, their rheological and photochemical properties differ remarkably. [48] For the rheological properties, an experimentally

measured complex viscosity coefficient of the compound **1** is five-fold larger than that of **2**, although the number of the substituents of the alkyl chains is identical and only the positions of the two 2-hexyldecoxy substituents are different (**Table 2.2**). Similarly, a complex viscosity coefficient of the compound **4** is ten-fold larger than that of the compound **3**. In a comparison between the compounds with the same substitution positions, the viscosity coefficients of the compounds **1** and **2** are larger than those of the compounds **3** and **4**, respectively. The viscosity coefficient of the compound **1** is then close to that of the compound **4**. [48]

The variations of the viscosity coefficients among the compounds cannot be explained by Stokes-Einstein relation on the assumption of the identical microscopic friction, [22] indicating that local microscopic molecular interactions govern the macroscopic rheological properties. Rotational relaxation measured by solid-state NMR experiments also showed that rotational diffusion of the compound **3** is the slowest, those of the compounds **4** and **1** are comparable, and that of the compound **2** is the fastest. [48] These experimental results indicate that the number of the alkyl chain substituents and their positions in the alkyl- π liquids determine local liquid structures and intermolecular interactions.

However, the underlying molecular mechanism of the peculiar dynamics depending on the number of the substituents and their position in the alkyl- π liquids remains obscure. Especially, how the small difference in the positions of the alkyl substituents even for the compounds with the same alkyl- π composition ratios, *i.e.*, the differences between the compounds **1** and **2**, and between the compounds **3** and **4**, strongly influence the dynamics are unclear. Microscopic understanding of molecular interactions in the dynamic processes of the liquids are therefore necessary to be obtained to explain the rheological properties controlled by the alkyl- π engineering.

To regulate key properties of alkyl- π liquids such as liquid viscosity and separation of functional π -conjugated moieties, empirical molecular design has been used. [24, 46] However, further molecular design of alkyl- π liquids through straightforward molecular screening suffers from combinatorial explosion, making exhaustive synthetic exploration impractical. In addition, it is not easy to experimentally determine at the atomic level microscopic molecular structures of disordered materials, limiting understanding of their physical properties and rational control of the weak π - π interactions of the functional moieties.

Here we present a theoretical study on the rheological properties of the alkyl-pyrene liquids, the compounds **1-4**, by means of atomistic molecular dynamics (MD) simulations. MD simulations have been utilized to investigate rheological properties of liquids such as viscosity coefficients through analyses of microscopic dynamics and diffusion processes of molecules in liquids underlying the rheological properties for various molecular fluids including hydrocarbons, [49–51] ionic liquids, [52, 53] and deep eutectic solvents. [54–58] For the bulk alkyl- π systems, the initial processes of order-disorder phase transition were investigated by MD simulations for 1 μ s in total, and their static liquid structures and optoelectronic properties were analyzed. [59, 60]

However, because of highly viscous nature of the alkyl-pyrene liquids, much more long-time MD simulations and more elaborated analyses are required. We therefore performed in the present study long-time MD simulations for tens of microseconds to obtain ample statistical samples of liquid conformations as well as their dynamics on the long-time scale

in the viscous alkyl-pyrene liquids, and examined liquid structures and diffusion dynamics simultaneously through extensive analyses based on spatio-temporal correlation functions. We uncovered in atomic detail characteristic local liquid structures and locally anisotropic and anomalous diffusion dynamics which remarkably vary depending on the number of the alkyl substituents and their substitution positions. The macroscopic rheological behavior of the alkyl-pyrene liquids is successfully explained by the microscopic liquid structures and dynamics elucidated, providing insight into the molecular mechanism that governs the rheology of alkyl- π liquids.

We first describe the MD simulation method and force field parameters of the compounds **1-4** along with coordinate systems for theoretical analysis in COMPUTATIONAL METHODS section. In the following RESULTS section, characteristics of diffusion of the alkyl-pyrene liquids such as anomalous diffusion behavior and anisotropy are examined, and viscosity coefficients are evaluated. Moreover, microscopic liquid structures and dynamics are analyzed based on various spatio-temporal correlation functions. Finally, we discuss microscopic molecular interactions and dynamics determining the rheological properties of the alkyl-pyrene liquids along with implication for their photochemical properties in DISCUSSIONS AND CONCLUDING REMARKS section.

2.2 COMPUTATIONAL METHODS

2.2.1 Force field

For computational efficiency, the molecules were modeled with united atoms. A CH_x ($x = 1-3$) group is treated as a single interaction site located at the carbon center. We employed the TraPPE-UA force field, [61, 62] which is known to reproduce physical properties of hydrocarbon compounds, [49, 63] with supplemental parameters taken from the OPLS-UA force field [64] and those determined by quantum chemistry calculations. Partial charges of most of the hydrocarbon united atoms are zero, while those of the united atoms around the alkoxy groups are nonzero and were estimated using the resp method [65, 66] for coordinates on the heavy atoms of model molecule. All quantum chemistry calculations were performed using GAMESS (US) (2012 version). [67, 68]

Atom type definition

For atoms in the aromatic rings, sp^2 carbon-hydrogen united atoms are represented by the CH atom type of TraPPE-UA force field, and sp^2 carbon atoms are represented by the C atom type. In addition, four atom types related to biphenyl-like bonds between the pyrenyl and phenyl groups were newly defined (**Figure 2.2a-c**), as potential energy surface around the biphenyl-like bonds are complex and thus potential energy functions related to the bonds are newly determined by quantum chemistry calculations (see next section). The atom types of the carbon atom in the pyrenyl ring attached to the phenyl group are defined as CP, and that in the phenyl group is defined as CB. In addition, CX and CY were defined to distinguish the bond angles of the phenyl group. Force field parameters for non-bonding interactions of those newly defined atom types are set to be the same as those of the CH and C ones. sp^3 carbons of the alkyl chains with 1, 2, and 3 C-H bonds are treated as united atoms represented by the C1, C2, and C3 atom types of the TraPPE-UA force field, respectively. The force field parameters are summarized in **Tables 2.8, 2.9**.

Intramolecular interaction

In the use of the TraPPE-UA force fields, bond lengths are supposed to be fixed. Thus, potential energy parameters of bond stretching force constant are not provided. However, the PMEMD of Amber program packages, which was computationally very efficient and was therefore utilized in the present study, is not able to fix the bond lengths of heavy atoms. The bond stretching motions were therefore allowed in the present MD simulations by the PMEMD program with the force constant parameters taken from OPLS-UA.

According to the Amber-type force field, the total potential energy, $U^{\text{TOT}}(\mathbf{q})$, which is a function of the coordinates of the united atoms, $\mathbf{q}$, is divided into a non-binding component $U^{\text{NB}}(\mathbf{q})$ and an intramolecular bonded terms as given by

$$U^{\text{TOT}}(\mathbf{q}) = U^{\text{NB}}(\mathbf{q}) + \sum_{i \in \text{bonds}} U_i^{\text{BOND}}(\mathbf{q}) + \sum_{j \in \text{bonds}} U_j^{\text{BEND}}(\mathbf{q}) + \sum_{k \in \text{dihedrals}} U_k^{\text{DIHED}}(\mathbf{q}). \quad (2.1)$$

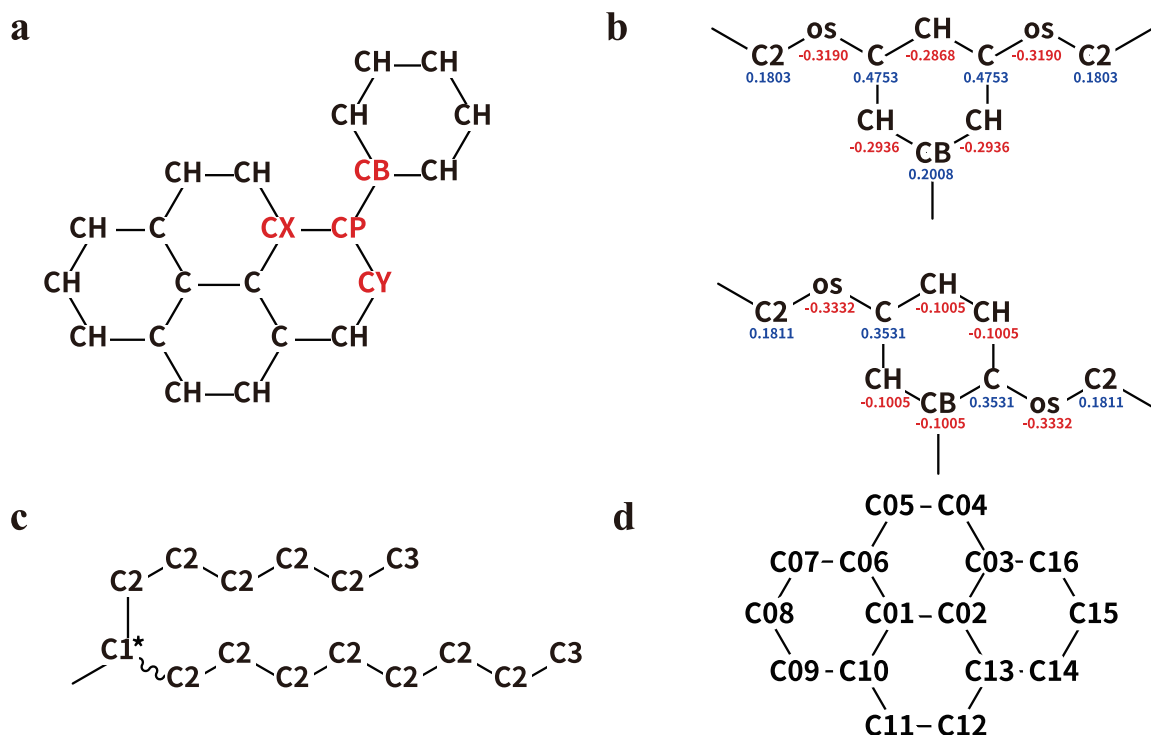


Figure 2.2: Atom type definition. **(a)** Atom types in a 1-phenylpyrene part. Newly defined atom types are depicted in red. **(b)** Atom types and partial charges of 2,5- and 3,5-dialkoxyphenyl groups. All partial charges not shown here are set to zero. **(c)** Carbon atom types in a 2-C6C10 group. C1 atom marked with an asterisk is an asymmetric carbon atom. **(d)** Atom labels for defining the molecular axes. Here, C16 is located at the 1-substituent position.

Here, these sums include all bonds, bends, dihedral angles, and out-of-plane bends (improper torsions). Those potential energy contributions are expressed as

$$U_i^{\text{BOND}}(\mathbf{q}) = \frac{1}{2} k_{a(i)b(j)}^{\text{BOND}} (r_{ij} - r_{a(i)b(j)}^0)^2, \quad (2.2)$$

$$U_i^{\text{BEND}}(\mathbf{q}) = \frac{1}{2} k_{a(i)b(j)c(k)}^{\text{BEND}} (\theta_{ijk} - \theta_{a(i)b(j)c(k)}^0)^2 \quad (2.3)$$

and

$$U_i^{\text{DIHED}}(\mathbf{q}) = \frac{1}{2} k_{a(i)b(j)c(k)d(l),n}^{\text{DIHED}} [1 + \cos(n\varphi_{ijkl} - \omega_{a(i)b(j)c(k)d(l),n}^0)]^2 \quad (2.4)$$

where r_{ij} is the bonding length between atoms i and j , and $r_{a(i)b(j)}^0$ is the neutral bonding length of the corresponding atom types. The subscripts $a(i)$ and $b(j)$ denote the atom types for atoms i and j . Similarly, θ_{ijk} and $\theta_{a(i)b(j)c(k)}^0$ are the bending angles and the neutral ones. The dihedral angle potential is represented by a linear combination of n out-of-phase Fourier series. φ_{ijkl} is the dihedral angle and $\omega_{a(i)b(j)c(k)d(l)}^0$ is a phase shift parameter. These potential functions are characterized by the corresponding force constants k_{ab}^{BOND} , k_{abc}^{BEND} and $k_{abcd,n}^{\text{DIHEDRAL}}$, respectively.

Nonbonded interaction

The nonbonded energy consists of a sum of a two-body repulsive and dispersion energy term, $U^{\text{LJ}}(\mathbf{q})$, and a Coulomb interaction energy, $U^{\text{COUL}}(\mathbf{q})$,

$$U^{\text{NB}}(\mathbf{q}) = U^{\text{LJ}}(\mathbf{q}) + U^{\text{COUL}}(\mathbf{q}), \quad (2.5)$$

where

$$U^{\text{LJ}}(\mathbf{q}) = \sum_{i>j} 4\epsilon_{a(i)b(j)} \left(\left(\frac{\sigma_{a(i)b(j)}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{a(i)b(j)}}{r_{ij}} \right)^6 \right), \quad (2.6)$$

$$U^{\text{COUL}}(\mathbf{q}) = \sum_{i>j} \frac{q_i q_j}{4r_{ij}\pi\epsilon_0}, \quad (2.7)$$

r_{ij} is an interatomic distance, $\sigma_{a(i)b(j)}$ and $\epsilon_{a(i)b(i)}$ are the Lennar-Jones parameters. We employed Lorentz-Berthelot combination rules for the nonbonded parameter mixing, which are also used in TraPPE-UA and AMBER. ϵ_0 is the electric constant, and q_i is the partial charge. According to standard heterocyclic and hydrocarbon chain models of the TraPPE-UA force field, the partial charges of the pyrene rings and the alkyl chains are set to zero. Regarding the phenyl rings with methoxy-groups, partial charges at heavy atoms, which represent the Coulombic interaction sites of the united atoms, were determined by a restrained electrostatic potential (RESP) fitting on the heavy atoms at the M06-2X/6-31G(d) level of theory (**Figure 2.2b**). In addition, 1-2 and 1-3 non-bonded interactions are excluded, and 1-4 Lennard Jones and Coulombic interactions were reduced to 0 and 1/2, respectively.

Refinement of intramolecular force field parameters

Potential energy functions around the biphenyl-type bonds are very complex due to hydrogen atom collisions and stabilization by electronic delocalization between conjugated π orbitals. We, therefore, refined the intramolecular force field parameters around the biphenyl-type bonds. Quantum chemistry calculations at the M06-2X/6-31G(d) level of theory were carried out for 1-phenylpyrene and potential energy curves along six degrees of freedom around the biphenyl-type bond were computed. Through the least square fitting to the potential energy curves obtained by the quantum chemistry calculations, force field parameters of $k_{\text{CP-CB}}^{\text{BOND}}$, $k_{\text{CX-CP-CB}}^{\text{BEND}}$, $k_{\text{CY-CP-CB}}^{\text{BEND}}$, $k_{\text{CP-CB-C*}}^{\text{BEND}}$, $k_{\text{***-CP-CB-***}}^{\text{DIHEDRAL}}$, improper, $r_{\text{CP-CB}}^0$, $\theta_{\text{CX-CP-CB}}^0$, and $\theta_{\text{CY-CP-CB}}^0$ were determined (Tables 2.8, 2.9). The same parameters were used for all the compounds 1-4 studied.

2.2.2 MD simulation

Non-bonded interactions were treated with a 14 Å spherical cutoff and analytic tail corrections for the Lennard-Jones interactions. The particle mesh Ewald method was used for Coulomb interaction. Time evolutions of the particles were computed by the velocity Verlet algorithm, and the Langevin thermostat was utilized for temperature control. All intramolecular degrees of freedom were not constrained. Molecular dynamics simulations were performed using AMBER 14 and 16 packages. [69]

To determine an appropriate time step, preliminary calculations were carried out for the compound 4 in NVE ensemble for 10 ns with different time steps, *i.e.*, 2, 4, and 5 fs, respectively. For the time steps of 2 fs and 4 fs, the total energy, a conserved quantity, was generally conserved with fluctuations around the mean value; the magnitudes of the fluctuation were 0.01% at 2 fs (the mean is 35,318 kcal/mol, and RMS fluctuation is 4.5 kcal/mol) and about 0.05% at 4 fs (the mean is 35,928 kcal/mol, and RMS fluctuation is 18 kcal/mol) (Figure 2.3a). However, numerical integration immediately diverged for the time step of 5 fs. We therefore used the time step of 2 fs for the production calculations of long-time MD simulations, considering a stability required for analyses of the long-time dynamic behavior. We also confirmed that the total energy and thermodynamic quantities (temperature, pressure, and density) remained stable for the system in NPT ensemble for 2 μ s with the time step of 2 fs (Figure 2.3b).

Simulation cells in a three-dimensional periodic boundary condition consisted of 200 molecules. The 2,5-substituents of the liquid alkyl-pyrene compounds 2 and 4 possess possible conformational isomers because of rotation of the phenyl groups. Hence, the compound 4 has eight atropisomers in the given molecular symmetry. Furthermore, a chiral carbon exists at the second carbon atom from the ether oxygen in the 2-hexyldecoxy (OC6C10) group. Two chiral chains are therefore present in the 2,5- or 3,5-OC6C10 groups. Experimentally, those groups were shown to be racemic. [21, 70] In the present simulation systems, the atropisomers were evenly mixed. Preliminary simulations at 300 K confirmed that the atropisomers of the phenyl group were exchangeable by thermal fluctuation. For the enantiomers, the two OC6C10 groups were modeled to be one (S)-type and one (R)-type substituents, respectively, for computational simplicity. Bulk properties are expected

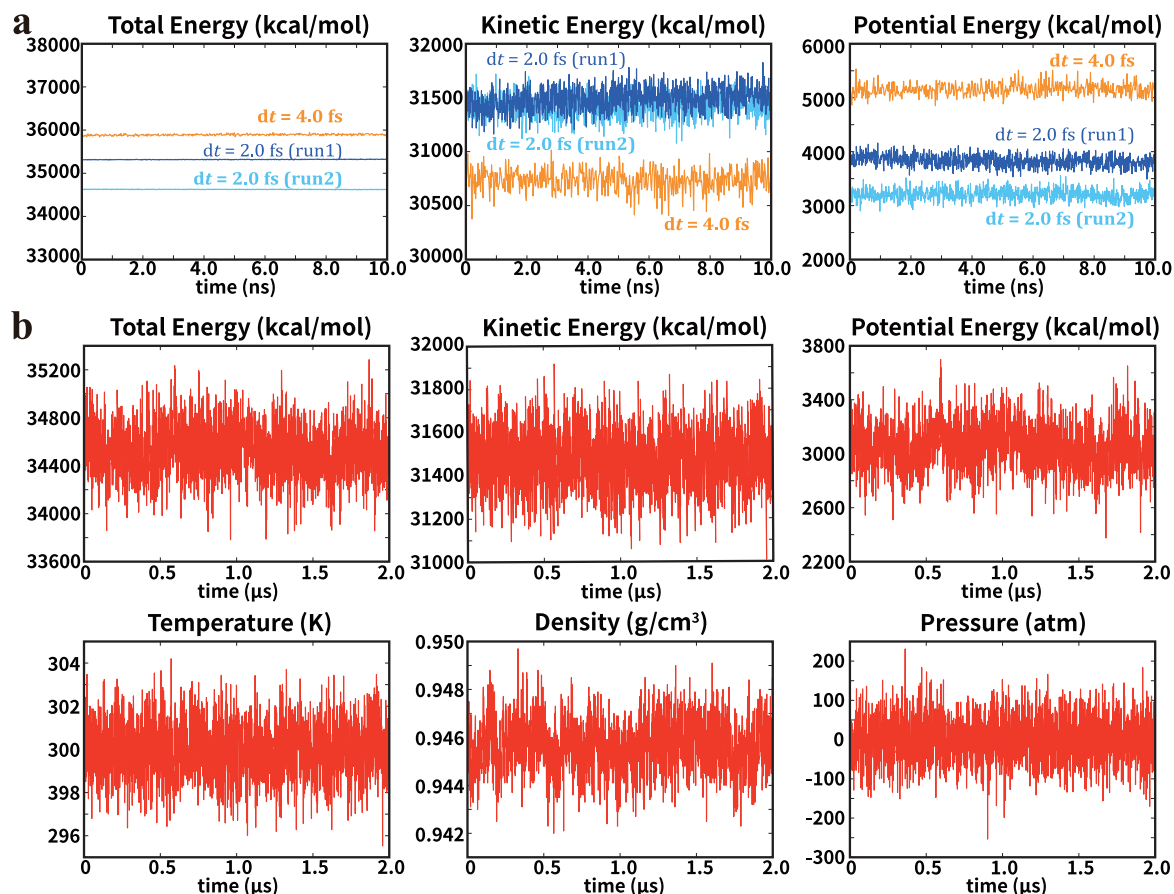


Figure 2.3: Time evolutions of quantities along MD trajectories. **(a)** Energy components along trajectories in NVE ensemble calculated for 10 ns with time steps of 2 fs and 4 fs, respectively. **(b)** Energy components and thermodynamic properties along a trajectory in NVT ensemble calculated for 2 μ s with time steps of 2 fs.

to be less affected by enantiomers, because the two substituents on the chiral carbon are the *n*-hexyl group and the *n*-octyl group, respectively, which are approximately the same length.

An initial configuration of molecules in the simulation box was set by packing 200 molecules which were translated and rotated as a rigid body from a single molecular structure optimized in the gas phase. The initial simulation boxes are cubic ones with a side length of 85 Å for **1-2** and 105 Å for **3-4**, respectively. The set up was performed by Packmol. [71] After the initial coordinates were energy minimized, MD simulations for equilibration were performed in NPT ensemble at 1 atm. The system was heated to 1,000 K in 100 ps, randomized at 1,000 K for 100 ns, cooled to 400 K in 5 ns, and equilibrated at 400 K for 10 μ s. The system was then cooled down to 300 K in 5 ns and equilibrated at 300 K for 5 ns (**Figure 2.4a**). The production runs were performed at 300 K from the equilibrated systems.

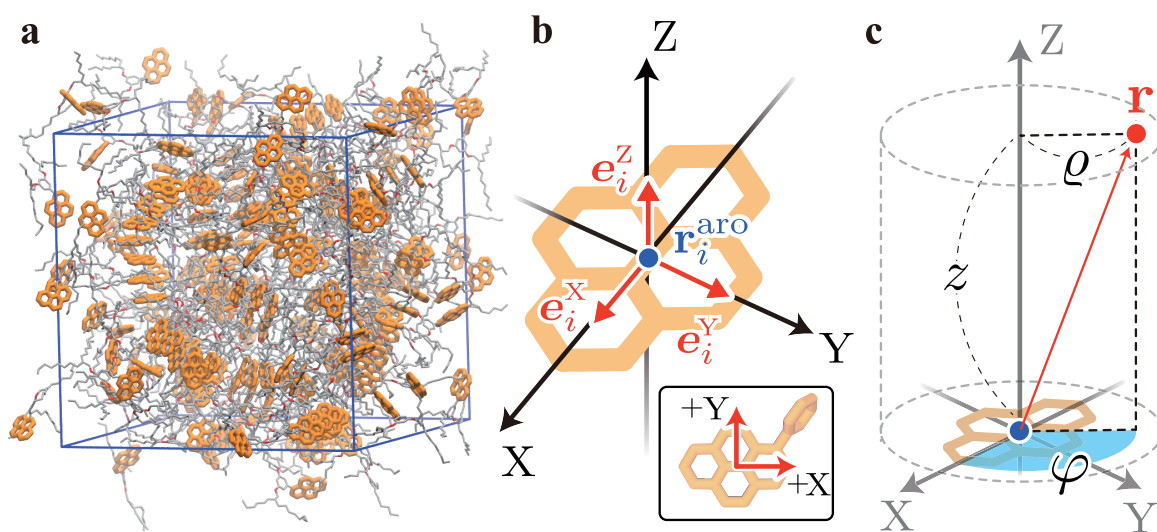


Figure 2.4: A simulation system and definition of molecular coordinates. **(a)** A snapshot of an equilibrated simulation system of **1**. **(b)** Definition of the body-fixed coordinates. The origin was set at the center of gravity of the pyrene ring. X-, Y-, and Z-axes were defined by the long axis of the pyrene ring, the short axis of it, and the axis perpendicular to the pyrene plane, respectively. The unit vectors associated with these axes are labeled e_i^x , e_i^y , and e_i^z , respectively. The positive and negative directions of X- and Y-axes for the compounds **1** and **2** are distinguished as they are in C_s symmetry. The 1-position of pyrene is located in the first quadrant. **(c)** Definition of the cylindrical coordinate system in the body-fixed coordinates. Three coordinates (ρ , φ , z) are defined by an axial distance ρ from Z-axis, the azimuth φ from X-axis on XY-plane, and the axial coordinate z , respectively.

We performed MD simulations for 50 μ s for compounds **1** and **2**, and for 25 μ s for **3** and **4**, respectively. The coordinates were sampled at every 10 ps. In addition, as the alkyl-pyrene compounds are viscous liquids and thus their molecular dynamics on a wide time scale are correlated, structural samplings at smaller time steps are required to examine the time correlation. To reduce conformational storage, the correlation time domain was divided into three domains, *sim* 10 ps, *sim* 10 ns, and *sim* 10 μ s. For the shortest time domain of *sim* 10 ps, we performed 5,000 and 2,500 short MD simulations for 100 ps

starting at snapshots of every 10 ns in the aforementioned MD simulations for 50 μ s and 25 μ s, respectively, and sampled the coordinates at every 10 fs. Displacement of the coordinates of the center of gravity of the entire system was removed for analysis of time correlations.

2.2.3 Body-fixed coordinates

Unlike simple systems composed of isotropic monoatomic molecules, structures and dynamics of molecular liquids are in general complex. Especially in alkyl- π liquids, rigid aromatic π -rings and flexible alkyl chains coexist in a molecule, producing anisotropic aggregation and diffusion. To analyze the complex intermolecular structure and dynamics, the molecular position and orientation are expressed in the molecular frame defined by the long axis of pyrene as the X-axis, the short axis as the Y-axis, and the normal to the aromatic plane as the Z-axis (**Figure 2.4b**).

Given rotational degrees of freedom of the biphenyl-like bonds of the alkyl-pyrene liquids, the compounds **1** and **2** in the ensemble average were treated with the C_s symmetry with a mirror plane of the pyrene aromatic ring lying in the plane of the X- and Y-axes. Hence, no distinction between the positive and negative values in the Z-axis was made. Here, the X-axis is defined such that the phenyl group at 1-position is located in the first quadrant. The compound of **3** and **4** were treated with the D_{2h} symmetry as well, and no distinctions between the positive and negative values in the X-axis and Y-axis, respectively, were made in addition.

To measure intermolecular correlations, a local molecular coordinate system of a target molecule is introduced. We consider the body-fixed coordinate system, which is a local molecular coordinate system of the pyrene aromatic moiety, to monitor time evolutions of environments around the target molecule. In the following, unless otherwise specified, the indices $i, j, \dots$ are labels of molecules, and the number of molecules contained in the system is represented by N . The centroid of the pyrene ring of the i -th molecule is expressed as

$$\mathbf{r}_i(t) = \frac{1}{n_p} \sum_{l=1}^{n_p} \mathbf{q}_{il}(t), \quad (2.8)$$

where $\mathbf{q}_{il}$ is the coordinate of the l -th united atom in the pyrene ring of the i -th molecule, and n_p is the number of the united atoms constituting the pyrene ring.

A set of three orthogonal unit vectors $\mathbf{e}_i^X$, $\mathbf{e}_i^Y$, and $\mathbf{e}_i^Z$ that represent the molecular frame of the i -th molecule at time t is also defined with snapshot coordinates in the trajectory as follows: First, unit vectors $\{\tilde{\mathbf{e}}_i^X, \tilde{\mathbf{e}}_i^Y\}$ connecting the geometric centers of specific united atoms are defined as

$$\tilde{\mathbf{e}}_i^X(t) = \frac{\mathbf{r}_i^X(t) - \mathbf{r}_i^{X'}(t)}{|\mathbf{r}_i^X(t) - \mathbf{r}_i^{X'}(t)|} \quad (2.9)$$

and

$$\tilde{\mathbf{e}}_i^Y(t) = \frac{\mathbf{r}_i^Y(t) - \mathbf{r}_i^{Y'}(t)}{|\mathbf{r}_i^Y(t) - \mathbf{r}_i^{Y'}(t)|}, \quad (2.10)$$

where $\mathbf{r}_i^X$, $\mathbf{r}_i^{X'}$, $\mathbf{r}_i^Y$, and $\mathbf{r}_i^{Y'}$ are geometrical centroids of the atomic sets chosen to define long and short axes of a pyrene ring. For the long axis, the atoms C01, C06, C07, C08, C09 and

C10 are selected for $\mathbf{r}_i^X$, and the atoms C02, C03, C13, C14, C15 and C16 are selected for $\mathbf{r}_i^{X'}$. For the short axis, the atoms C01, C02, C03, C04, C05 and C06 are selected for $\mathbf{r}_i^Y$, and the atoms C01, C02, C10, C11, C12 and C13 are selected for $\mathbf{r}_i^{Y'}$. $\mathbf{e}_i^Z$ is first defined by the cross product of $\mathbf{e}_i^Z = \tilde{\mathbf{e}}_i^X \times \tilde{\mathbf{e}}_i^Y$. Since $\{\tilde{\mathbf{e}}_i^X, \tilde{\mathbf{e}}_i^Y\}$ of snapshot coordinates are not orthogonal in general due to intramolecular deformation, we defined $\mathbf{e}_i^X$ and $\mathbf{e}_i^Y$ by $\mathbf{e}_i^X = \tilde{\mathbf{e}}_i^X$ and $\mathbf{e}_i^Y = \mathbf{e}_i^X \times \mathbf{e}_i^Z$.

Furthermore, a time-dependent affine transformation function P_i is defined as

$$\mathbf{P}_i(\mathbf{r}, t) = \mathbf{R}_i(t) [\mathbf{r} - \mathbf{r}_i(t)], \quad (2.11)$$

where $\mathbf{r}$ is a three-dimensional position vector and $\mathbf{R}_i(t)$ is a rotation matrix associated with the orthonormal basis vectors $\{\mathbf{e}_i^\alpha(t)\}$ given by

$$\mathbf{R}_i(t) = \left(\mathbf{e}_i^X(t), \mathbf{e}_i^Y(t), \mathbf{e}_i^{Z(t)} \right)^\top. \quad (2.12)$$

The function $\mathbf{P}_i(\mathbf{r}, t)$ moves the coordinate to the body-fixed coordinates of the i -th molecule at time t , and provides a vector with three scalar components

$$\mathbf{P}_i(\mathbf{r}, t) = \left(P_i^X(\mathbf{r}, t), P_i^Y(\mathbf{r}, t), P_i^Z(\mathbf{r}, t) \right)^\top \quad (2.13)$$

which represents coordinates in the body-fixed system. This transformation is generally non-commutative because it is an affine transformation consisting of translation and linear transformation.

2.3 RESULTS

2.3.1 Translational diffusion

Since the alkyl-pyrene liquid system has no macroscopic anisotropy, translational diffusion is isotropic in a long-time domain regardless of initial orientations of the molecules. However, molecular liquids generally undergo anisotropic translational diffusion in a short time domain because of anisotropic shape of the molecule. To quantify the anisotropy of the diffusion process in a short-time domain, the alkyl-pyrene liquid trajectories were analyzed by anisotropic mean square displacements (a-MSDs) expressed as

$$\sigma_{\alpha}^2(t, t_0) = \frac{1}{N} \sum_{i=1}^N (P_i^{\alpha}(\mathbf{r}_i(t + t_0), t_0))^2, \quad (2.14)$$

where α is associated with the X-, Y- and Z-axis, respectively. The XYZ components of $\sigma^{\alpha}(t, t_0)$ are defined in the molecular local frame at $t = t_0$ expressed by the basis vectors $\{\mathbf{e}^{\alpha}(t_0)\}$. The sum of the three a-MSDs is the standard isotropic mean square displacement (MSD),

$$\sigma^2(t, t_0) = \sigma_X^2(t, t_0) + \sigma_Y^2(t, t_0) + \sigma_Z^2(t, t_0). \quad (2.15)$$

The diffusion anisotropy ratio is defined by normalizing the anisotropic values as

$$\varsigma(t, t_0) = \frac{\sigma_{\alpha}^2(t, t_0)}{\sigma^2(t, t_0)}. \quad (2.16)$$

Given that the alkyl-pyrene liquids do not exhibit macroscopic bulk anisotropy experimentally, the three components converge to 1/3 in the long-time limit.

The translational diffusion processes of the compounds **1** to **4** are presented in **Figure 2.5**. A common feature of the alkyl-pyrene liquids is an anomalous diffusion [72, 73] due to a cage effect appearing from sub-picoseconds and lasting up to microseconds. After simple diffusions on the sub-picosecond time scale, the diffusions drastically slow down in sub-picoseconds to tens of nanoseconds, and then gradually increases on the microsecond time scale (**Figure 2.5a, b**). In the time region of sub-picoseconds to tens of nanoseconds, the exponents of time for MSDs decrease to ~ 0.2 , indicating a strong subdiffusion (**Table 2.1**). After 5 microseconds, then, the diffusion processes become normal again; the exponents of time are ~ 1.0 . This transition from subdiffusion to normal diffusion is typically seen for dynamics of caged particles in a crowded environment. [74, 75] The temporal change in the diffusion characteristics is closely correlated with behavior of the diffusion anisotropy (**Figure 2.5d, e**). In the short time less than sub-picoseconds, the diffusion anisotropy ratios are almost the same, expressing a simple diffusion before the molecules experience a cage effect. The diffusion anisotropy ratios then remarkably deviate from the average value in sub-picoseconds to tens of nanoseconds. In the compounds **1** and **2**, the anisotropic diffusion components along the X-axis representing the long axis of the pyrene ring become the fastest ones. The components along the Y-axis representing the short axis of the ring are also faster than the average values. The anisotropic diffusion components along the Z-axis perpendicular to the pyrene ring then become very slow. In

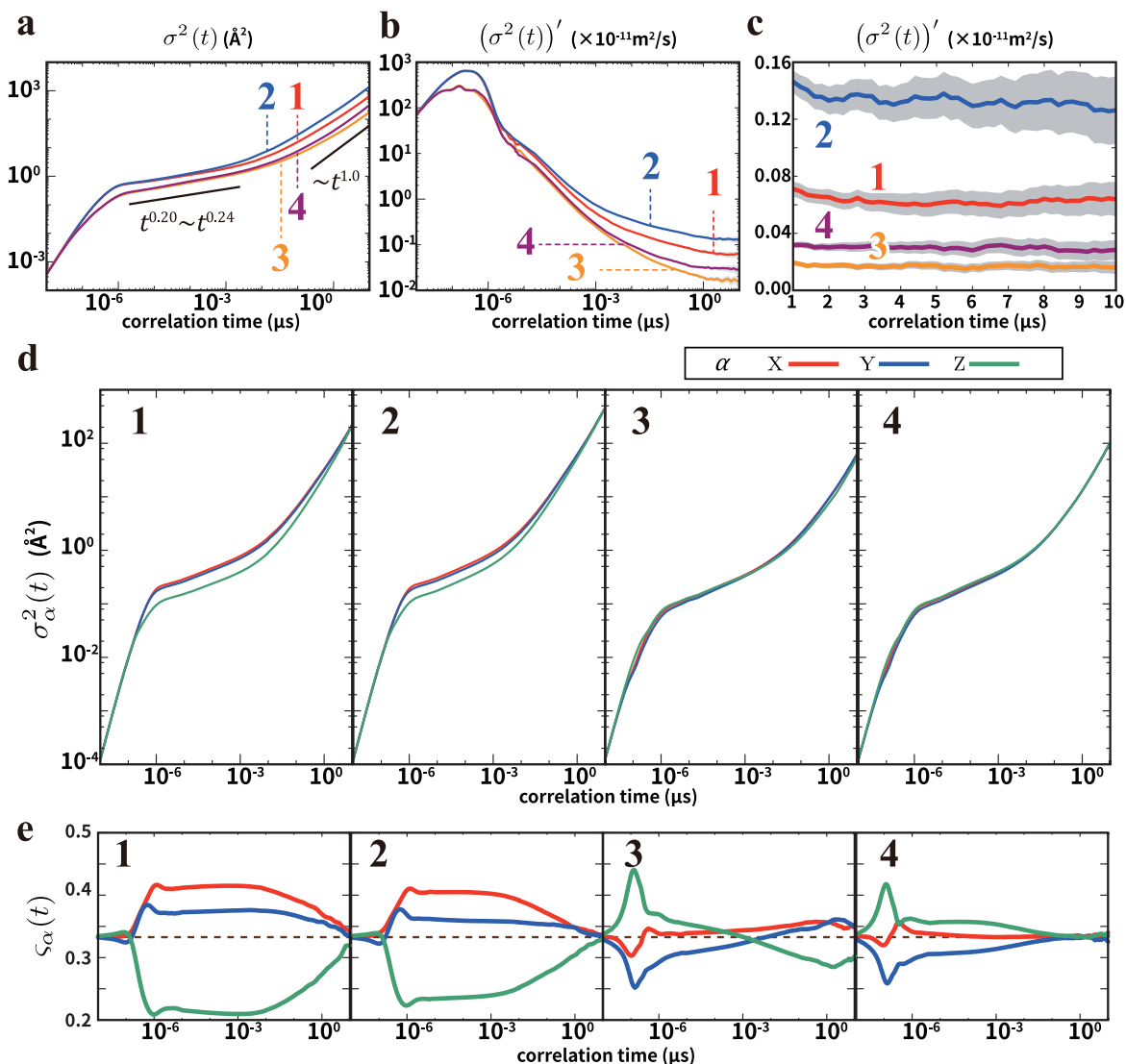


Figure 2.5: Translational diffusion properties of the compounds **1-4**. **(a)** Isotropic MSDs of the compound **1-4**. The slopes of the lines were calculated by least squares for the specified intervals. The detailed values are summarized in **Table 2.1**. **(b),(c)** Time derivatives of the isotropic MSDs. Gray areas indicate the 99.9% confidence intervals for the Student's t distribution. **(d)** a-MSDs along X-, Y-, and Z-axes (red, blue, and green), respectively. **(e)** Diffusion anisotropy ratios, ς_α , defined by Eq. 2.16. Dashed line indicates 1/3 for the isotropic convergence.

the case of the compounds **3** and **4**, in contrast, the anisotropy components along the Z-axis and the Y-axis become the fastest and slowest ones, respectively. The components along the X-axis do not largely deviate from the average values. The deviations of the anisotropic diffusion components are attributed to characteristic local intermolecular structures in the alkyl-pyrene liquids as described below. The anisotropy of diffusions then decays in the long-time domain after tens of nanoseconds, except for the compound **3** which presumably exhibits a conformational relaxation slower than the simulation time.

Table 2.1: Exponents of anomalous diffusion in time domains of 1 ps to 1 ns and 5 to 10 μ s.

		1	2	3	4
1 ps to 1 ns	σ_X^2	0.20	0.22	0.23	0.23
	σ_Y^2	0.20	0.22	0.24	0.24
	σ_Z^2	0.19	0.23	0.21	0.24
	σ^2	0.20	0.22	0.23	0.24
5 to 10 μ s	σ_X^2	1.07	1.08	1.18	1.11
	σ_Y^2	1.09	1.09	1.10	1.09
	σ_Z^2	1.09	0.97	1.09	1.05
	σ^2	1.07	1.09	1.07	1.08

To characterize the rheologic properties of the FMLs, we compare viscosities of the liquids obtained from the computed translational diffusion coefficients above with the experimental ones. [21] Note that the complex viscosity coefficients were experimentally measured with a rheometer and no frequency dependence was found for all compounds. [21] The complex viscosity coefficients were therefore regarded as the viscosities of the liquids which are defined with the complex viscosity coefficients in the zero frequency limit. [76]

A straightforward computational evaluation of a viscosity coefficient would be based on the Green-Kubo theory with time correlations of stress tensor components. [15] However, it is a formidable task to compute a well-converged accurate time correlation for viscosity for the highly viscous alkyl-pyrene liquid systems. Note that the alkyl-pyrene liquids become to exhibit the normal diffusion after *sim*5 μ s as seen above (**Figure 2.5**), and thus the calculations of the time-correlations require impractically long time MD trajectory calculations and huge data storages. We therefore introduced an approximation based on the Stokes-Einstein relation to qualitatively evaluate a viscosity from the computed diffusion coefficient. The viscosity, η , is thus obtained with a diffusion coefficient, D , as

$$\eta \approx \frac{k_B T}{A \pi a D}, \quad (2.17)$$

where k_B is the Boltzmann constant, T is the absolute temperature, and a is the radius of the hydrodynamic equivalent sphere of the molecule. The quantity A depends on whether the boundary conditions are assumed to be "sticky" ($A = 6$) or "slippery" ($A = 4$). We employed the latter value, which is known to be more realistic for neat molecular liquid. [74, 77]

Using the slope of the isotropic MSD of the diffusion region in the long time limit, the diffusion coefficients calculated by MDs for the finite periodic boundary systems [78, 79] are given by

$$D^{\text{MD}} = \lim_{t \rightarrow \infty} \frac{1}{6} \frac{d\sigma^2(t, t_0)}{dt}. \quad (2.18)$$

The diffusion coefficients in the thermodynamic limit were then evaluated by an analytic correction formula derived by Yeh and Hummer [80] as

$$D \approx D^{\text{MD}} \left(1 - \frac{A\xi a}{6L} \right)^{-1} \quad (2.19)$$

where, $\xi \approx 2.837297$, and L is length of the cubic simulation cell.

The radius of the equivalent sphere of the molecule, a , was estimated with the averaged radius of gyration as

$$a = \frac{1}{t_{\text{sim}}} \int_0^{t_{\text{sim}}} dt \sqrt{\frac{1}{n_{\text{apm}}} \sum_{l=1}^{n_{\text{apm}}} \|\mathbf{q}_{il}(t) - \mathbf{r}_i(t)\|^2}. \quad (2.20)$$

where n_{apm} is the number of united atoms in the molecule, $\mathbf{q}_{il}$ is the coordinate of the l -th united atom of the i -th molecule, and t_{sim} is the simulation time.

The diffusion coefficients and the radii of the equivalent spheres of the molecules were evaluated with the averages of the time derivatives of the isotropic MSDs (**Figure 2.5b, c**) and the radius of gyration in 9-10 μs where the time derivatives of the isotropic MSDs are temporally almost constant. The statistical errors during the correlation time are sufficiently small (**Figure 2.5c**), indicating that clear statistical differences in the diffusion processes between the compounds exist. It should be noted, however, that the diffusion process of the compound **3** did not completely reach the equilibrium as their anisotropy components were not fully converged (**Figure 2.5e**). The diffusion coefficient derived from Eq. 2.18 with the isotropic MSD evaluated by the present simulations even for tens of microseconds is therefore still an approximated estimate.

Table 2.2 summarizes the viscosity and diffusion coefficients along with the averaged radii of gyration and the simulation cell sizes. The computed viscosity coefficients of the compounds **1**, **2**, and **4** (3.5, 1.8, and 5.6 Pa·s, respectively) are in qualitative agreement with the experimental ones (6.1, 2.3, and 6.2 Pa·s, respectively). [21] The viscosity coefficient of the compound **2** is the smallest one, and those of the compounds **1** and **4** are 2-3 times larger than that of the compound **2**. Although the absolute magnitudes of the computed viscosity coefficients of the compounds **1** and **4** are slightly deviated from the experimental ones, respectively, presumably due to lack of sufficient accuracy of the force field and the simulation time, the relative order of the coefficients is successfully reproduced by the present simulations.

The computed viscosity coefficient of the compound **3** (10.6 Pa·s) is, in contrast, largely underestimated compared with the experimental one (58.0 Pa·s), although it is the largest one among the compounds considered and thus is in line with the experimental evidence. The large discrepancy of the computed coefficient may originate from lack of the sufficient simulation time despite the present long-time MD simulation calculation for tens of

Table 2.2: Molecular weights, averaged radii of gyration, a , length of the cubic simulation cell, L , computed translational diffusion coefficients, D^{MD} , thermodynamic limit translational diffusion coefficients, D , and viscosity coefficients, η , of alkyl-pyrene liquid compounds **1-4**. Numbers in parentheses for a and L indicate the standard deviations, and numbers with $\pm$ for D^{MD} and D indicate the 99.9% confidence intervals for the Student’s t distribution. See Appendix for methods for the calculation of confidence intervals.

	1	2	3	4
Molecular weight	759.19	759.19	2430.23	2430.23
a (Å)	6.93(0.28)	6.81(0.30)	10.12(0.31)	9.26(0.29)
L (Å)	63.29(0.04)	63.35(0.04)	94.72(0.04)	94.86(0.04)
D^{MD} ($\times 10^{-12} \text{m}^2 \text{s}^{-1}$)	0.11 $\pm$ 0.02	0.21 $\pm$ 0.04	0.027 $\pm$ 0.006	0.046 $\pm$ 0.011
D ($\times 10^{-12} \text{m}^2 \text{s}^{-1}$)	0.14 $\pm$ 0.02	0.27 $\pm$ 0.05	0.034 $\pm$ 0.008	0.058 $\pm$ 0.014
η (Pa·s)	3.5 $^{+0.8}_{-0.5}$	1.8 $^{+0.4}_{-0.3}$	10.6 $^{+3.3}_{-2.0}$	5.6 $^{+1.7}_{-1.1}$
η (Pa·s) (Exp.) ^a	6.1	2.3	58.0	6.2

^a Experimental values. [21]

microseconds. In fact, the anisotropic MSDs for the compound **3** (**Figure 2.5e**) are not yet converged in the simulation time as described above. A longer MD simulation would therefore be required to obtain the converged value of the larger viscosity coefficient.

2.3.2 Rotational diffusion

We analyzed rotational self-diffusion of the molecules which also exhibits anisotropy. [81–83] Given the anisotropic shape of the pyrene molecule, the time evolutions of the rotational dynamics of $\mathbf{e}_i^X(t)$, $\mathbf{e}_i^Y(t)$, and $\mathbf{e}_i^Z(t)$, which represent the body fixed axes of the molecular frame of the pyrene moiety of the i -th molecule were monitored individually. The time correlation function of the rotational dynamics is given by

$$\omega_\alpha(t, t_0) = \frac{1}{N} \sum_{i=1}^N \omega_i^\alpha(t, t_0), \quad (2.21)$$

where ω_i^α is a rotational time correlation function of the i -th molecule expressed as

$$\omega_i^\alpha(t, t_0) = P^{(2)}(\mathbf{e}_i^\alpha(t + t_0) \cdot \mathbf{e}_i^\alpha(t_0)), \quad (2.22)$$

and $P^{(2)}$ is a normalized quadratic Legendre polynomial defined by

$$P^{(2)}(x) = \frac{3}{2}x^2 - \frac{1}{2}. \quad (2.23)$$

The time correlation function which exhibited complex behavior involving decays on multiple time scales was then fitted by an analytic function to estimate decay time constants of the correlation.

For the analytic function, we employed a linear combination of the Kohlrausch-Williams-Watts (KWW) extended exponential functions [84–86] given by

$$\tilde{\omega}_\alpha(t) = \sum_{k=1}^K c_k^\alpha \exp \left[- \left(\frac{t}{t_k^\alpha} \right)^{\kappa_k^\alpha} \right]. \quad (2.24)$$

Here, t_k^α is the KWW correlation time, and κ_k^α is a distribution parameter which stretches the time decay curve when the value lies between zero and one and compresses the decay curve when the value is greater than one. c_k^α is a coefficient of the linear combination and is constrained so that the sum of the coefficients is one. The average lifetime of the rotational correlation function was then obtained by

$$\tau_\alpha^\omega = \sum_{k=1}^K \frac{c_k^\alpha t_k^\alpha \Gamma[(\kappa_k^\alpha \tau_k^\alpha)^{-1}]}{\kappa_k^\alpha}, \quad (2.25)$$

where $\Gamma[x]$ is the gamma function. The fitting to the mixed KWW function was performed with $K = 4$, and the optimized parameters are shown in **Table 2.3**. See Appendix for details of the parameter fitting.

The rotational correlation functions of the compounds **1-4** are shown in **Figure 2.6**. For all the rotational correlations around the X-, Y-, and Z-molecular axes, those of **2** decay most rapidly, followed in order by **1**, **4**, and **3**. The order of the relaxation times is consistent with experimental observations of viscosity and solid-state NMR measurements. [21] The initial decreases within ~ 1 ps can be interpreted as a simple rotational diffusion of the molecule before the molecule experiences a cage. [77] In the initial time domain, **1** and **2** decrease by $\sim 4\%$ and **3** and **4** by $\sim 1\%$. The decays are then considerably slowed until nanoseconds due to restriction of the cage, followed by faster exponential relaxation on a longer time scale.

Table 2.4 summarizes the average rotational lifetime τ_α^ω . The average decay times of **1** are more than two-fold larger than those of **2** for all the molecular axes, and similarly the decay times of **3** are about 1.5 times larger than those of **4**, despite that the numbers of the substituent are the same for the compounds **1** and **2**, and for **3** and **4**, respectively. The relative order of the rotational decay times in terms of the compounds is the same as that of the diffusion coefficients of translational diffusion seen above. For the compounds **1** and **2**, clear rotational anisotropy was seen; τ_Z^ω are ~ 1.4 times and ~ 1.8 times larger than τ_X^ω and τ_Y^ω , which nearly coincide, for the compounds **1** and **2**, respectively, attributed to remarkably slow decay of ω_Z of the compound **1** in the time region of microseconds or longer (**Figure 2.6c**). In contrast, no difference in the decay time among the molecular axes was found for **3** and **4**, indicating that those compounds exhibit isotropic rotational relaxation.

Table 2.3: Fitting parameters for a parametric analytical rotational correlation function $\tilde{\omega}_\alpha(t)$.

α	k	1			2		
		c_k^α	t_k^α (ns)	κ_k^α	c_k^α	t_k^α (ns)	κ_k^α
X	1	0.945	432	0.617	0.902	172	0.670
	2	0.025	8.40×10^{-2}	0.515	0.072	4.29×10^{-1}	0.333
	3	0.021	2.66×10^{-4}	1.638	0.021	3.27×10^{-4}	2.009
	4	0.009	6.34×10^{-4}	1.171	0.005	6.30×10^{-4}	1.731
Y	1	0.940	425	0.609	0.905	167	0.664
	2	0.023	5.93×10^{-2}	0.561	0.058	2.59×10^{-1}	0.381
	3	0.022	1.76×10^{-4}	1.846	0.021	4.30×10^{-4}	1.501
	4	0.015	4.72×10^{-4}	1.286	0.016	1.73×10^{-4}	2.286
Z	1	0.928	792	0.616	0.862	267	0.725
	2	0.039	2.90×10^{-2}	0.312	0.084	3.05	0.461
	3	0.022	1.67×10^{-4}	2.310	0.032	2.22×10^{-4}	1.741
	4	0.011	4.49×10^{-4}	1.392	0.022	7.37×10^{-3}	0.449

α	k	3			4		
		c_k^α	t_k^α (ns)	κ_k^α	c_k^α	t_k^α (ns)	κ_k^α
X	1	0.747	3329	1.219	0.945	1420	0.885
	2	0.231	451	0.773	0.040	11.6	0.508
	3	0.014	2.09×10^{-1}	0.428	0.008	5.47×10^{-3}	0.446
	4	0.008	2.74×10^{-4}	1.037	0.007	3.50×10^{-4}	1.278
Y	1	0.947	2423	0.897	0.928	1342	0.919
	2	0.037	18.8	0.492	0.053	23.7	0.505
	3	0.009	1.17×10^{-4}	1.755	0.010	1.68×10^{-4}	1.139
	4	0.007	2.39×10^{-3}	0.409	0.009	4.18×10^{-3}	0.380
Z	1	0.944	2483	0.843	0.925	1245	0.877
	2	0.037	14.1	0.468	0.050	16.7	0.516
	3	0.010	1.32×10^{-3}	0.366	0.014	5.48×10^{-3}	0.352
	4	0.009	1.27×10^{-4}	2.007	0.011	1.97×10^{-4}	1.205

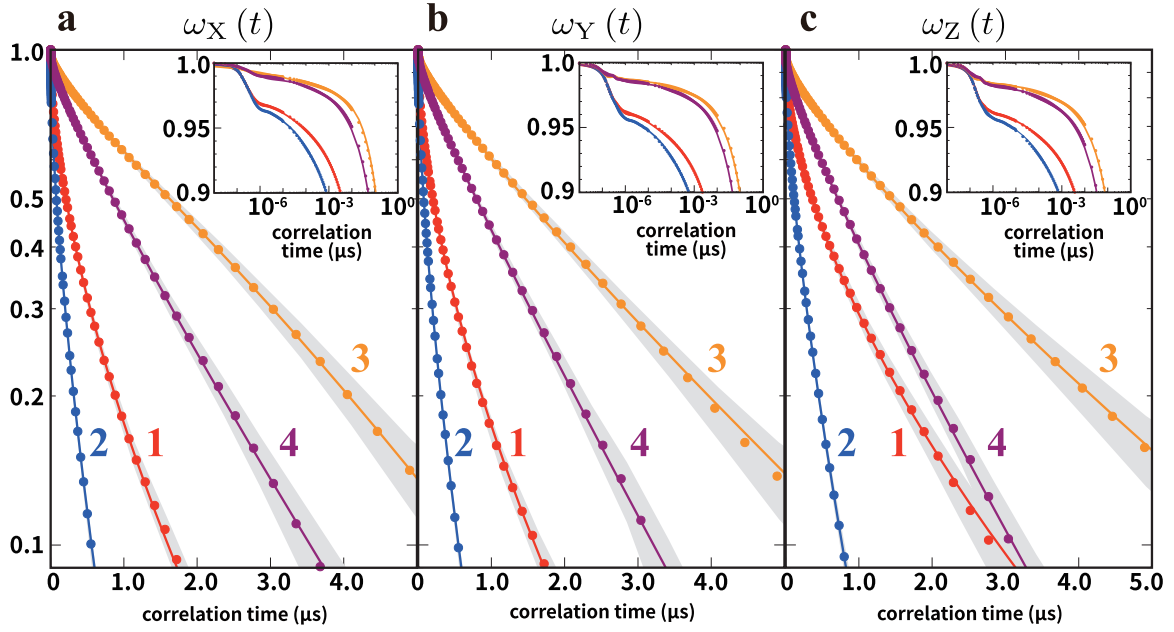


Figure 2.6: Rotational relaxation functions. **(a)** $\omega_X(t)$. **(b)** $\omega_Y(t)$. **(c)** $\omega_Z(t)$. Dots indicate calculated values of $\omega_\alpha(t)$ at sample times, the solid lines are analytical stretched exponential curves, $\tilde{\omega}_\alpha(t)$, fitted to the calculated values, and gray areas are the 99.9% confidence intervals for the Student's t distribution. Insets show initial relaxations ($0.9 \leq \omega_\alpha \leq 1.0$) in a short time region.

Table 2.4: Lifetimes τ_α^ω of rotational correlations. τ_α^ω is defined by Eq. 2.25 and was calculated by approximating the analytical function defined by Eq. 2.24. The numbers in parentheses are the RMSE for the data of the fitted curves.

	1	2	3	4
τ_X^ω (ns)	593 (0.0556)	205 (0.0524)	2382 (0.1135)	1425 (0.0813)
τ_Y^ω (ns)	589 (0.0556)	202 (0.0498)	2419 (0.0739)	1298 (0.0701)
τ_Z^ω (ns)	1071 (0.0612)	283 (0.0601)	2563 (0.0997)	1231 (0.0670)

2.3.3 The van Hove distribution function

Static and dynamic nature of alkyl-pyrene liquids in liquid equilibrium state is characterized through the van Hove distribution function which is a spatio-temporal correlation function of two-body density fluctuation. [87, 88] The van Hove distribution function in the body-fixed coordinates of the molecules, $G(\mathbf{r}, t)$, is given by

$$G(\mathbf{r}, t) = \frac{1}{N\rho(t)} \sum_{i=1}^N \sum_{j=1}^N \delta(\mathbf{r} - \mathbf{P}_i(\mathbf{r}_j(t), 0)) \quad (2.26)$$

where $\rho(t)$ is the number density of the system given by $N/V(t)$, and $V(t)$ is the volume of the system at time t . The initial time of the time correlation, at which the body-fixed frame of the molecule i is determined, is set to 0 in Eq. 2.26. The van Hove distribution function can be divided into a self-component and a distinct-component

$$G(\mathbf{r}, t) = G^{\text{SELF}}(\mathbf{r}, t) + G^{\text{DIST}}(\mathbf{r}, t). \quad (2.27)$$

The self-component is given by

$$G^{\text{SELF}}(\mathbf{r}, t) = \frac{1}{N\rho(t)} \sum_{i=1}^N \delta(\mathbf{r} - \mathbf{P}_i(\mathbf{r}_i(t), 0)) \quad (2.28)$$

and the distinct-component is

$$G^{\text{DIST}}(\mathbf{r}, t) = \frac{1}{N\rho(t)} \sum_{i=1}^N \sum_{j \neq i}^N \delta(\mathbf{r} - \mathbf{P}_i(\mathbf{r}_j(t), 0)). \quad (2.29)$$

The static pair correlation function called the spatial distribution function (SDF), [89, 90] $g(\mathbf{r})$, is given by the distinct-component at $t = 0$ as

$$g(\mathbf{r}) = G^{\text{SELF}}(\mathbf{r}, 0). \quad (2.30)$$

2.3.4 Static structures

The SDFs given by Eq. 2.30 microscopically characterize static molecular structures of the alkyl-pyrene liquids in equilibrium. **Figure 2.7** depicts SDFs of **1-4** around the pyrene moieties. One can discern sharp and periodic peaks of the densities of **1** and **2** which develop from the pyrene moiety along the Z-axis at every 4 Å. As the aromatic plane of the pyrene moiety lies in the XY-plane, the periodic peaks represent columnar multiple π - π stacking of the pyrene rings. Note that the FMLs do not exhibit any macroscopic system anisotropy. Thus, the columnar structures do not create a persistent one-dimensional order, but repeatedly form and collapse as described later.

The local coordination number was calculated by integrating the SDF with respect to a subspace. [89] To analyze the anisotropic columnar structure developing along the Z-axis more quantitatively, a spatial distribution function in a cylindrical coordinate (**Figure 2.4c**)

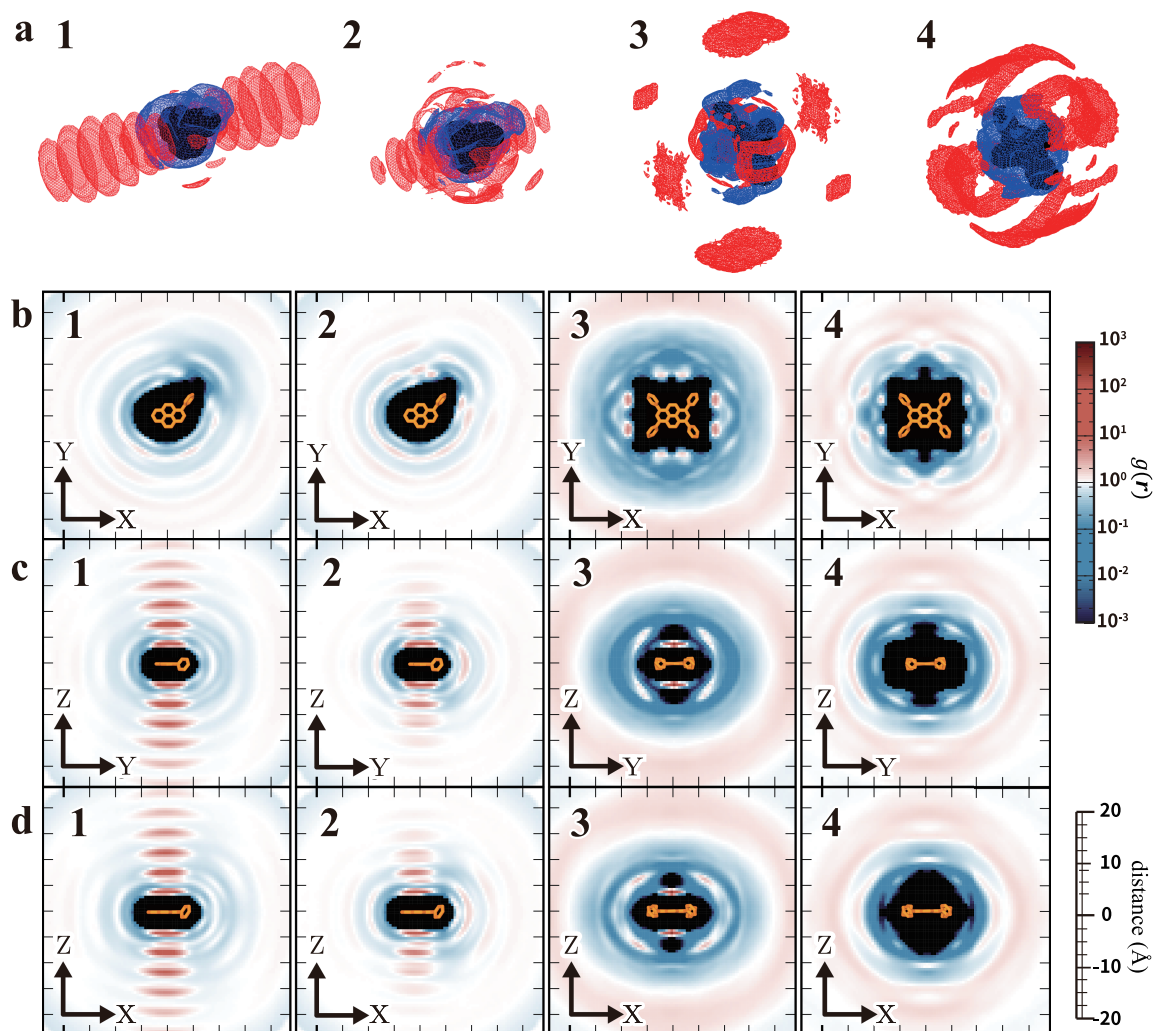


Figure 2.7: SDFs of pyrene rings around the reference molecule of the compounds 1-4. (a) Isosurfaces of the SDFs at values of 0.50 (blue mesh surfaces) and 2.50 (red mesh surfaces). (b)-(d) Two dimensional maps of the SDFs on planes spanned by two of the three axes of the body-fixed coordinates of the reference molecule, XY-plane (b), YZ-plane (c), and XZ-plane (d). In each row, the compounds 1-4 are aligned from left to right. All figures are depicted to the same scale, and images of the reference molecules without alkyl chains were included.

is introduced. Similarly, the number density of the columnar structure along the Z-axis, $n^G(z, t)$, is defined by

$$n^G(z, t) = \bar{\rho} \int_0^{\varrho_a} \varrho d\varrho \int_0^{2\pi} d\varphi \int_{-\inf}^{\inf} dz' G^{\text{DIST}}(\mathbf{r}, t) \delta(z' - z), \quad (2.31)$$

where $\bar{\rho}$ is the time averaged number density of molecules given by $\bar{\rho} = N/\bar{V}$, $\bar{V}$ is the time averaged volume of the system, ϱ is the radial distance from the Z-axis, and φ is azimuth. ϱ_a is the cut-off distance and was set to 5.0 Å. In particular, at $t = 0$, n^G is the cylindrical integral of the SDF, denoted as $n^g(z) = n^G(z, 0)$. The accumulated number of the SDF in the cylinder along the Z-axis, $N^g(z)$, is given by

$$N^G(z) = \int_0^z n^g(z') dz'. \quad (2.32)$$

The number distribution in the cylinder along the X-axis is compared with the homogenous number distribution

$$N^H(z) = \bar{\rho} \int_0^{\varrho_a} \varrho d\varrho \int_0^{2\pi} d\varphi \int_{-\inf}^{\inf} dz' = \pi \varrho_a^2 z \bar{\rho} \quad (2.33)$$

which represents the number distribution of liquid without structure. As the number distributions are symmetric along the Z-axis, those of the positive and negative z are averaged.

The periodic peaks of the SDF of **1** are bolder than those of **2** (**Figure 2.7**). In fact, $n^g(z)$ of the compound **1** persists progression of the periodic peaks up to 20 Å, while the periodic peaks of **2** disappear in 8 Å (**Figure 2.8a**). The observation indicates that the columnar multiple π - π stacking of **1** is stronger than that of **2**, despite that the numbers of the phenyl substituents in **1** and **2** are identical (one). In the case of **1**, the two hexyldecyl alkyl chains substituted at the 3,5-positions of the phenyl group extend outward from the pyrene ring and thus the π - π stacking of the pyrene rings are less disturbed by the alkyl chain substituents (**Figure 2.1b**). In contrast, one of the two OC6C10 alkyl chains substituted at the 2-position of the phenyl group in **2** points inward and covers the pyrene ring, which hinders stronger π - π stacking interaction of the pyrene rings (**Figure 2.1b**). Because of the strong columnar multiple π - π stacking of **1**, the accumulated number N^g of **1** along the Z-axis is considerably larger than that of the homogenous number distribution, N^H , and differs by about 1 in the limit of $z \rightarrow \infty$ (**Figure 2.8b**).

The SDFs of **3** and **4** do not exhibit periodic peaks of the columnar structures (**Figures 2.7** and **2.8**). The numbers of the phenyl substituents in **3** and **4** are increased to four, and thus influence of the substituents become substantial. Nevertheless, the compound **3** forms a π - π stacking pair which is represented by strong peaks of the SDF above and below the pyrene ring (**Figure 2.7c, d**) and a single strong peak of $n^g(z)$ (**Figure 2.8a**). As described above, the two OC6C10 alkyl chains substituted at the 3,5-positions of the phenyl groups extend outside from the pyrene ring in **3** (**Figure 2.1b**), allowing the compound **3** to establish one strong dimeric π - π stacking interaction between two molecules. However, the increased volume of the substituents compared with **1** and **2** prevents elongation to the columnar structures of more than trimer with multiple π - π stacking of the pyrene rings (**Figure 2.8a**). In the case of **4**, the pyrene ring is completely isolated from those of other molecules (**Figure 2.7**)

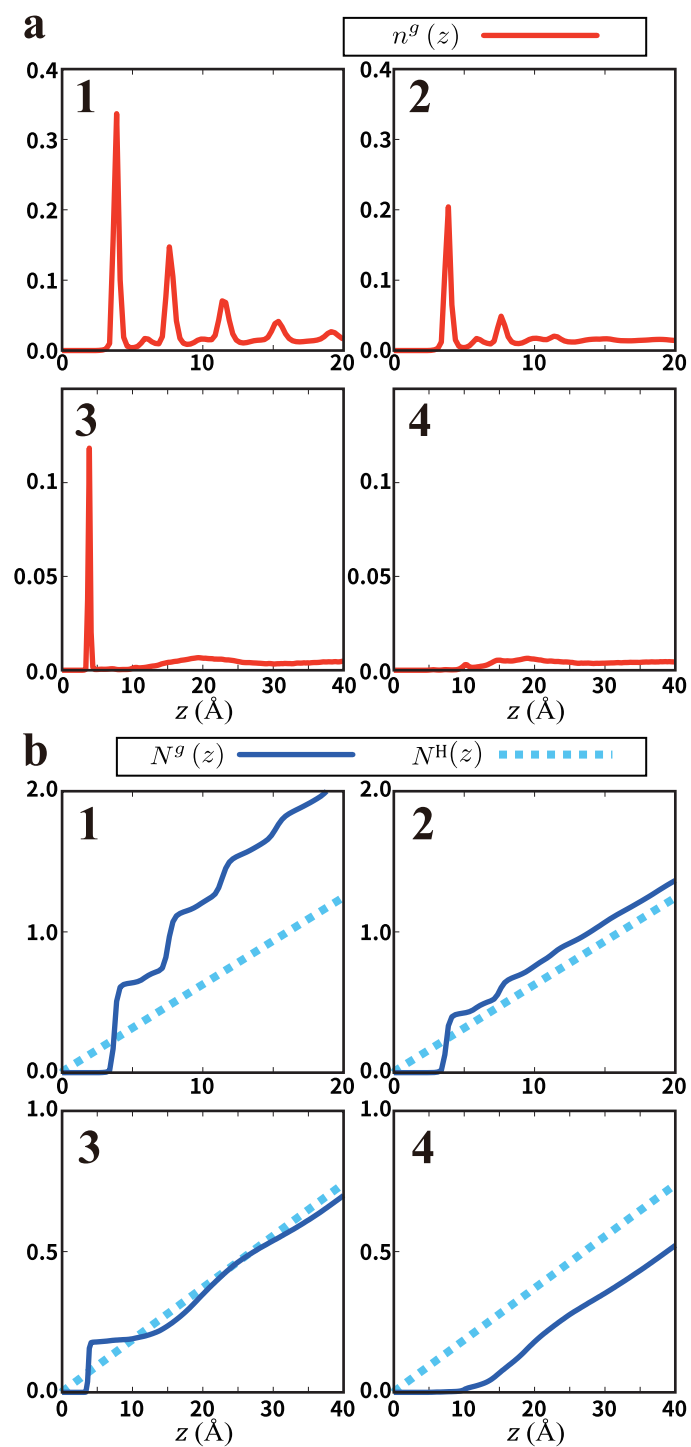


Figure 2.8: Average number densities and number distributions of the SDFs in the body-fixed cylindrical coordinates of the compounds **1-4**. **(a)** Average number densities $n^g(z)$. **(b)** Number distributions $N^g(z)$ and number distributions for the homogeneous distribution, $N^H(z)$.

because the pyrene ring is fully covered by the OC6C10 alkyl chains substituted at the 2-position of the four phenyl groups (**Figure 2.1b**) as described above. The closest molecules are located in diagonal positions in the XZ- and YZ-planes (**Figure 2.7c, d**) where the molecules can maximally avoid steric hindrance of the OC6C10 alkyl chains at the 2- and 5- positions of the phenyl groups. Accordingly, molecules cannot approach to the pyrene ring along the Z-axis, and thus the accumulated number of **4** along the Z-axis is significantly smaller by 0.3 than that of the homogeneous distribution (**Figure 2.8b**).

2.3.5 Dynamic structures

Spatio-temporal behavior of the van Hove distribution functions is examined to gain a microscopic insight into the rheological properties of the alkyl-pyrene liquids based on the liquid structures analyzed above. **Figure 2.9a-c** depicts temporal changes of the self-diffusion components of the correlation functions $G^{\text{SELF}}(\mathbf{r}, t)$. The order of the decay times of the positional distribution peaks is $2 < 1 < 4 < 3$, which is in line with the viscosity coefficients (**Table 2.2**). The compounds **1** and **2** exhibit temporally anisotropic decays of the positional distribution peaks appearing as elliptical distributions in the cross-sections of the XZ- and YZ-planes. The retarded decays along the Z-axis indicate that the diffusion along the Z-axis which is perpendicular to the plane of the pyrene ring is slowed in those compounds. In contrast, an isotropic decay of the positional distribution peak with a nearly complete round distribution is seen for the compound **4**.

The observed temporal anisotropy of the microscopic positional distributions underlies the anisotropies of the a-MSD components shown in **Figure 2.5d, e**. The anisotropy of the a-MSD components of **1** and **2** are significantly larger than that of **4**. The a-MSD components of **1** and **2** along the Z-axis is considerably smaller than those along the X- and Y-axes. The observation of the anisotropy of a-MSD is consistent with the temporal behavior of $G^{\text{SELF}}(\mathbf{r}, t)$ described above. Given that the pyrene aromatic rings form the columnar structures through the multiple π - π stacking interactions in **1** and **2**, which is absent in **4**, the anisotropy of the diffusion is governed by the characteristic liquid structures of the pyrene rings which are controlled by the alkyl chains of the substituent groups.

In the case of **3**, temporally anisotropic decay of the positional distribution peaks of $G^{\text{SELF}}(\mathbf{r}, t)$ is also observed (**Figure 2.9a-c**). The distribution peak is distorted into an elliptical shape on 10 nanoseconds or longer time scale, and thus shows a slower translational diffusion along the Z-axis. The anisotropic decay is attributed to the strong pairing of the pyrene rings through π - π stacking interaction described above. However, unlike **1** and **2**, the π - π stacking does not grow to form a columnar structure and is terminated at the formation of the strongly bound dimeric pair as described above. The smaller a-MSD components (**Figure 2.5e**) therefore represent the translational diffusion of the overall dimeric form of which the shape is not significantly anisotropic, while the translational diffusion of each monomer molecule in the dimer is strongly affected by the strong anisotropic π - π stacking interaction, which underlies the anisotropy of the positional distribution peaks of $G^{\text{SELF}}(\mathbf{r}, t)$.

Difference in anisotropy among the compounds **1-4** described above is also well explained by their structural characteristics. The columnar aggregates of **1** and **2** give rise to

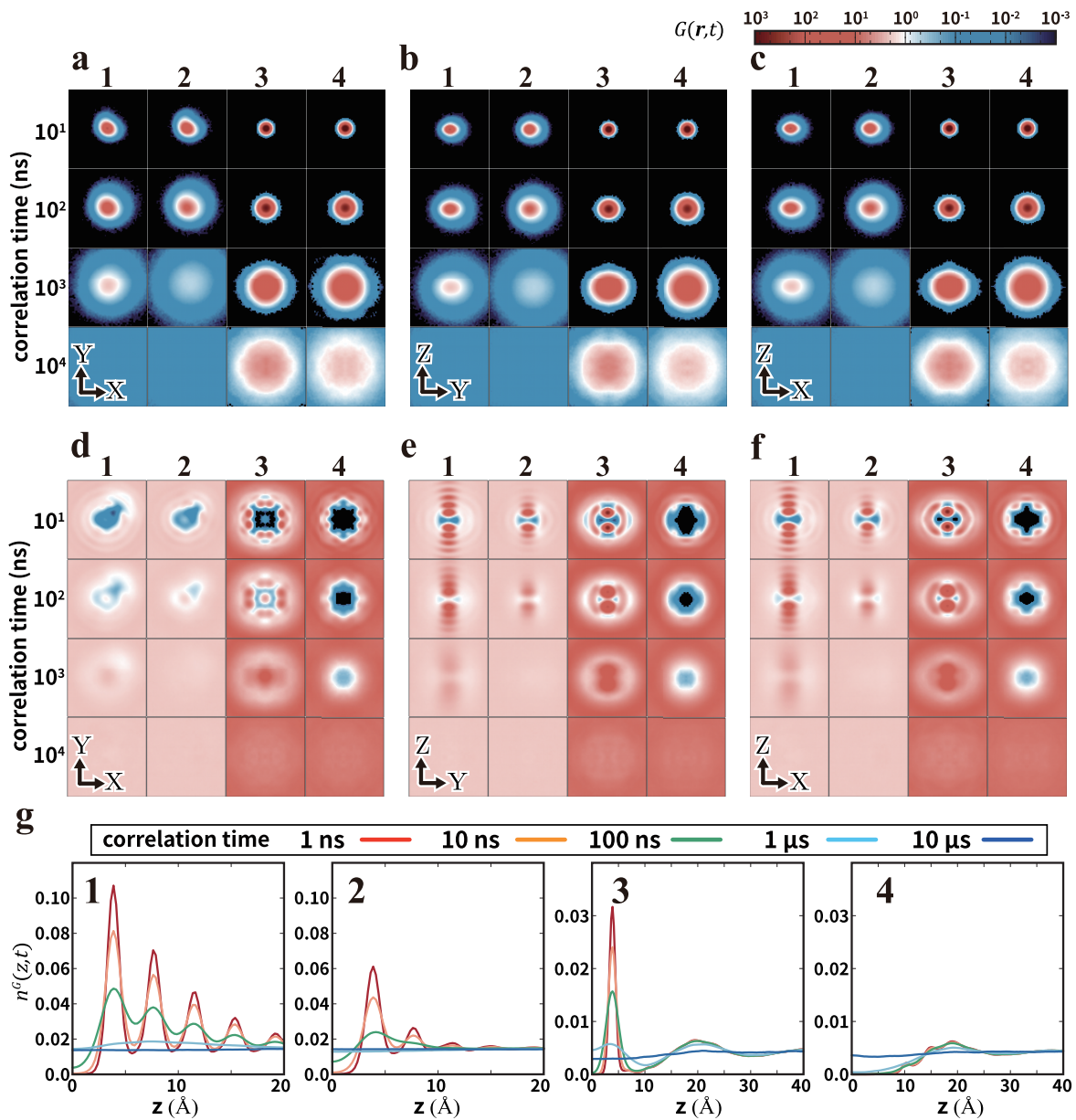


Figure 2.9: The van Hove distribution functions around the reference molecules of the compounds **1-4**. **(a)-(c)** Two dimensional maps of self-parts, $G^{\text{SELF}}(\mathbf{r}, t)$, on planes spanned by two of the three axes of the body fixed coordinates of the reference molecule, XY-plane **(a)**, YZ-plane **(b)**, XZ-plane **(c)**. In each panel, the compounds **1-4** are shown from the left, and correlation time of 10, 100, 1,000 and 10,000 nanoseconds are shown from the top. **(d)-(f)** Two dimensional maps of distinct parts, $G^{\text{DIST}}(\mathbf{r}, t)$. **(g)** The number density distributions along the Z-axis, $n^G(z, t)$, defined by Eq. 2.31.

the rotational anisotropy of τ_α^ω where τ_Z^ω , *i.e.*, the out-of-plane rotational decay of the stacking axis is remarkably larger than the in-plane decays of τ_X^ω and τ_Y^ω while the isolated pyrene structure of **4** results in the anisotropic behavior of τ_α^ω (**Figure 2.4** and **Table 2.4**). Similar to the a-MSD components, the isotropic behavior of **3** is attributed to the strong π - π stacking pair where the overall pyrene rings rotate together.

Spatio-temporal behavior of the liquid structures are also examined by the distinct component of the van Hove correlation function, $G^{\text{DIST}}(\mathbf{r}, t)$, shown in **Figure 2.9d-g**. Progression of peaks along the Z-axis expressing the columnar structure for **1** is maintained on the sub-microsecond time scale, while that for **2** decays faster and disappears in tens of nanoseconds. A pair of peaks of the strong π - π stacking pair of **3** is also kept on the sub-microsecond time scale. Thus, marked difference in the dynamic behavior of the liquid structures among those compounds was observed. It should, however, be noted that the temporal decay of $G^{\text{DIST}}(\mathbf{r}, t)$ involves both internal collapse of the structured aggregates and diffusion of the entire structured aggregates without collapse, and thus $G^{\text{DIST}}(\mathbf{r}, t)$ is not capable of distinguishing between them to identify the structural dynamic origin of the temporal character of the liquids.

2.3.6 Spatially resolved two-body time correlation function

To clearly describe collapse of the π - π stacking structure, we introduce a spatially resolved two-body time correlation function of the mutual orientation of the pyrene planes as

$$\Psi^\alpha(\mathbf{r}, t) = \frac{1}{N\rho(t)} \sum_{i=1}^N \sum_{j \neq i}^N \delta(\mathbf{r} - \mathbf{P}_i(\mathbf{r}_j(0), 0)) \xi_{ij}^\alpha(t, 0), \quad (2.34)$$

where ξ_{ij}^α is the two-body rotational correlation function given as

$$\xi_{ij}^\alpha(t, t_0) = P^{(2)}(\mathbf{R}_i(t + t_0)\mathbf{e}_j^\alpha(t + t_0) \cdot \mathbf{R}_i(t_0)\mathbf{e}_j^\alpha(t_0)). \quad (2.35)$$

Eq. 2.34 is therefore a two-body rotational spatio-temporal correlation function of the molecule j in the body-fixed coordinate system of the molecule i which changes in time. Since the mutual orientation of the pyrene planes is a good index of the π - π stacking structure, the decay of the correlation function represents the collapse of the stacking structure. Note that, unlike the van Hove distribution function which approaches to a uniform distribution of $1/\rho$ in the long-time limit, the correlation function given by Eq. 2.34 decays to zero if the π - π stacking structure completely collapses because it includes the rotational correlation.

Similar to the analysis of SDFs described above, we compute the number density distribution along the Z-axis as

$$n_\alpha^\Psi(z, t) = \bar{\rho} \int_0^{\varrho_a} \varrho d\varrho \int_0^{2\pi} d\varphi \int_{-\infty}^{\infty} dz' \Psi^\alpha(\mathbf{r}, t) \delta(z' - z) \quad (2.36)$$

with $\varrho_a = 5$ Å, which well represents distribution of the columnar structure (**Figure 2.10**). To quantify the lifetime of the stacking structure, the average lifetime along the Z-axis is calculated as

$$\tau_\alpha^\Psi(z, t) = \int_0^\infty \frac{n_\alpha^\Psi(z, t)}{n_\alpha^\Psi(z, 0)} dt. \quad (2.37)$$

The temporal changes of peaks of $n_\alpha^\Psi(z, t)$ were fitted with the mixed KWW equation as used for the fitting of the rotational correlation function described above (Eq. 2.24), and the average lifetimes of the peaks were evaluated. The fitting to the mixed KWW function was performed with $K = 4$ and the sum of c_k^α was modified to match $n_\alpha^\Psi(z, 0)$. The optimized parameters are summarized in **Table 2.5** and **2.6**. The peaks of which temporal behaviors were examined are the first, second, and third peaks at $z=4.0, 8.0$, and 12.0 Å for **1** and **2**, the first sharp peak at $z=4.0$ Å for **3**, and broad peaks at $z=20$ Å for **3** and **4**, respectively (**Figure 2.11a-c**). Anisotropy of the correlation time was also examined by the ratio as

$$p^\Psi(z) = \frac{2\tau_Z^\Psi(z)}{\tau_X^\Psi(z) + \tau_Y^\Psi(z)}. \quad (2.38)$$

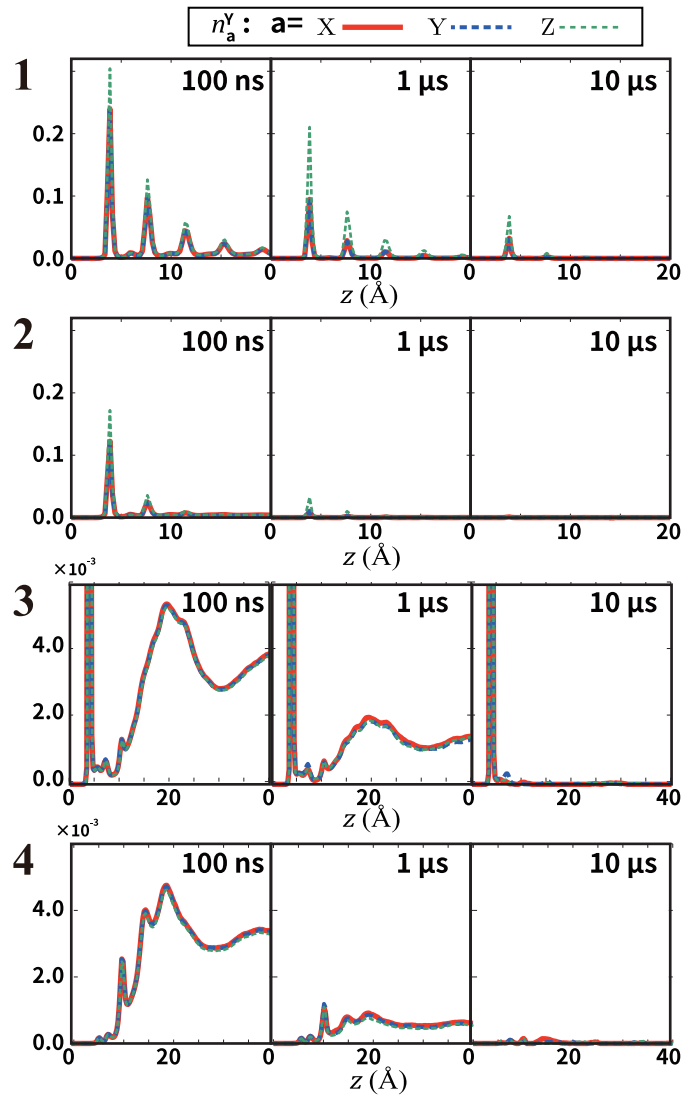


Figure 2.10: Number density distributions along the Z-axis for spatial decompositions of rotation relaxation correlation functions, $n_\alpha^\Psi(z, t)$, for the compounds **1-4**.

Table 2.5: Fitting parameters of the KWW function for $n_\alpha^\Psi(z, t)$ for compounds **1** and **2** for $z=4.0$ and 8.0 Å. Terms with c_k^α of less than 1% are omitted.

$z=4.0$ Å							
		1			2		
α	k	c_k^α	t_k^α (ns)	κ_k^α	c_k^α	t_k^α (ns)	κ_k^α
X	1	0.296	999	0.491	0.193	377	0.509
	2	0.023	260	1.24	-	-	-
	3	0.014	2.56×10^{-4}	1.28	-	-	-
	4	0.003	2.00×10^{-2}	5.84×10^2	-	-	-
Y	1	0.319	863	0.497	0.180	403	0.497
	2	0.008	6.40×10^{-4}	4.52×10^2	0.011	99.9	1.41
	3	-	-	-	0.010	2.47×10^{-4}	1.25
	4	-	-	-	0.003	1.00×10^{-2}	3.17×10^3
Z	1	0.309	3769	0.776	0.194	1161	0.768
	2	0.010	5120	5.72×10^{10}	0.005	8.13×10^{-5}	5.00×10^1
	3	0.009	8.00×10^{-5}	5.85×10^3	0.004	1.81×10^{-2}	0.342
	4	0.007	1.77	0.241	-	-	-
$z=8.0$ Å							
		1			2		
α	k	c_k^α	t_k^α (ns)	κ_k^α	c_k^α	t_k^α (ns)	κ_k^α
X	1	0.134	565	0.507	0.043	182	0.603
	2	0.005	160	7.26×10^{13}	0.003	2.69×10^{-4}	1.44
	3	-	-	-	0.002	6.92×10^{-2}	0.528
	4	-	-	-	0.001	2.84×10^3	2.00
Y	1	0.106	607	0.447	0.044	187	0.591
	2	0.033	354	1.00	0.003	354	0.350
	3	0.005	1.60×10^{-4}	1.21×10^8	0.002	1.60×10^{-4}	1.05×10^5
	4	0.003	6.40×10^{-4}	1.21×10^8	0.001	5.17×10^3	7.57
Z	1	0.138	2653	0.692	0.045	617	0.698
	2	0.006	1.35×10^{-4}	2.07	0.004	1.54×10^{-4}	1.94
	3	0.002	6.27×10^{-1}	0.505	0.001	0.118	0.464

Table 2.6: Fitting parameters of the KWW function for $n_\alpha^\Psi(z, t)$ for compounds **1** and **2** ($z=12.0$ Å) and compounds **3** and **4** ($z=20.0$ Å). Terms with c_k^α of less than 1% are omitted.

$z=12.0$ Å							
α	k	1			2		
		c_k^α	t_k^α (ns)	κ_k^α	c_k^α	t_k^α (ns)	κ_k^α
X	1	0.055	415	0.510	0.0178	94.2	0.552
	2	0.008	248	1.180	0.0013	2.89×10^{-4}	1.439
	3	0.003	2.03×10^{-4}	1.807	0.0004	7.73×10^{-2}	1.051
	4	0.001	9.81×10^5	0.179	-	-	-
Y	1	0.055	392	0.521	0.0173	86.0	0.518
	2	0.006	230	1.254	0.0014	2.72×10^{-4}	1.40
	3	0.003	391	0.521	0.0006	138	2.32
	4	0.002	4.52×10^3	0.304	-	-	-
Z	1	0.061	1528	0.612	0.0167	212	0.574
	2	0.003	1.43×10^{-4}	2.083	0.0014	2.17×10^{-4}	1.53
	3	-	-	-	0.0012	4.72	0.567
	4	-	-	-	0.0002	5.30×10^{-2}	1.80

$z=20.0$ Å							
α	k	3			4		
		c_k^α	t_k^α (ns)	κ_k^α	c_k^α	t_k^α (ns)	κ_k^α
X	1	1.00×10^{-3}	2921	0.225	2.39×10^{-3}	1153	0.985
	2	0.34×10^{-3}	432	1.86	0.28×10^{-3}	26.6	0.409
	3	0.08×10^{-3}	2.09	0.225	0.26×10^{-3}	1.16×10^4	4.32
	4	0.07×10^{-3}	2.75	1.40	0.08×10^{-3}	4.84×10^{-4}	0.749
Y	1	1.06×10^{-3}	2172	1.50	1.26×10^{-3}	3729	1.73
	2	0.27×10^{-3}	277	2.12	1.24×10^{-3}	629	1.23
	3	0.09×10^{-3}	2.17	0.152	0.40×10^{-3}	62.8	0.471
	4	0.08×10^{-3}	2.83	1.45	0.09×10^{-3}	3.69×10^{-4}	0.734
Z	1	1.06×10^{-3}	2437	1.30	1.91×10^{-3}	1082	0.568
	2	0.27×10^{-3}	268	1.95	0.95×10^{-3}	1511	1.37
	3	0.09×10^{-3}	1.83	0.177	0.08×10^{-3}	2.77×10^{-4}	0.900
	4	0.08×10^{-3}	2.58	1.27	0.05×10^{-3}	1.62×10^{-1}	0.504

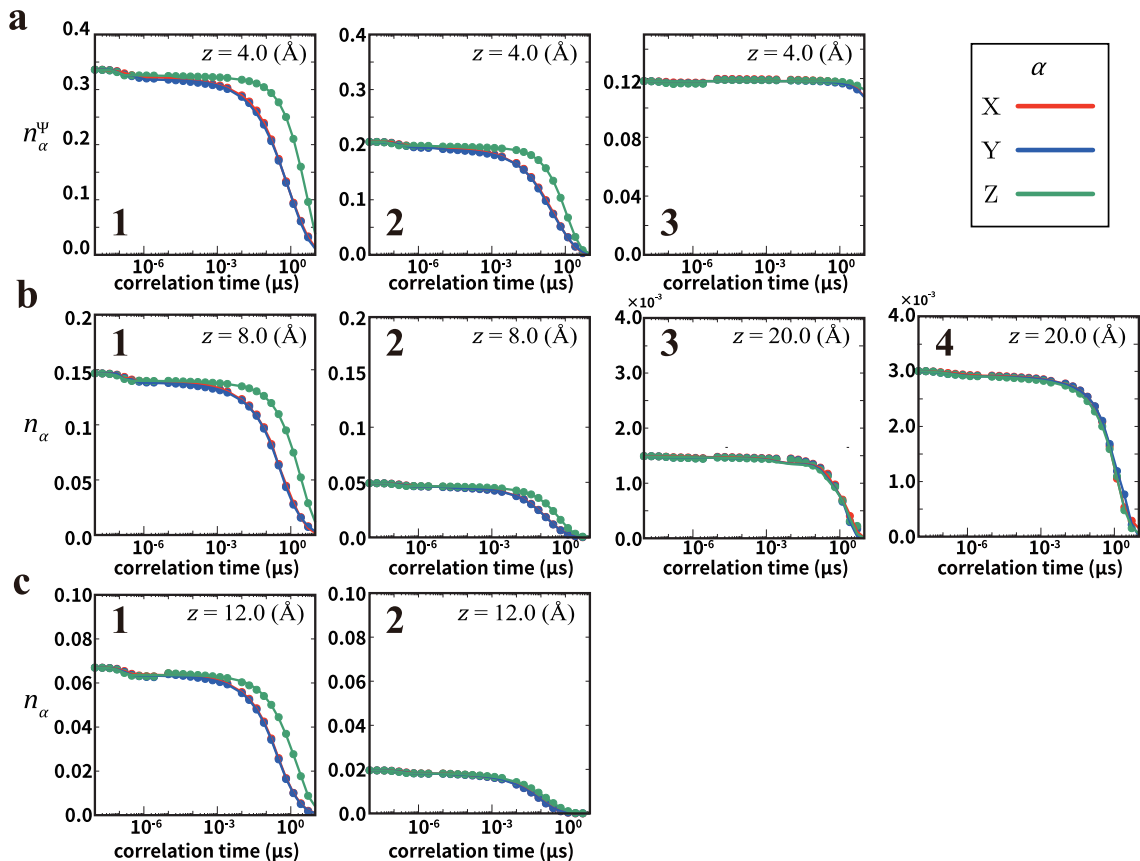


Figure 2.11: Temporal decays of peaks of the number density distributions along the Z-axis for spatial decompositions of rotation relaxation correlation functions, $n_{\alpha}^{\Psi}(z, t)$, for the compounds **1-4**. (a) Decay curves of the first peaks at $z=4.0$ Å for **1**, **2** and **3**. (b) Decay curves of the second peaks at $z=8.0$ Å for **1**, **2**, and $z=20.0$ Å for **3** and **4**. (c) Decay curves of the third peaks at $z=12.0$ Å for **1** and **2**.

Figure 2.10 depicts temporal changes of $n_{\alpha}^{\Psi}(z, t)$ and their peaks. The sharp peaks of the compounds **1** and **2** decay in microseconds, and their lifetimes, $\tau_{\alpha}^{\Psi}(z)$, were remarkably different from one another as listed in **Table 2.7**. First, large anisotropies, $p^{\Psi}(z)$, of about 2 to 3 were observed for all the peaks of **1** and **2**. The large anisotropy indicates that the decay of the Z-axis of the body-fixed coordinate is significantly slower than the other axes. The different rotational decays are explained by the columnar structure where the π - π stacking of the pyrene ring elongates along the Z-axis. The mutual rotations of the X- and Y- axes can proceed in the π plane while keeping the π - π stacking, whereas that of the Z-axis needs to destroy the stacking, giving rise to the large anisotropies. As described above, the decay of the Z-axis therefore is a good index of the collapse of the columnar structure.

The decay of the peak for the stacking pair of the compounds **1** and **2** is distance dependent. The decay of the peak at a larger distance in the columnar structure is faster (**Table 2.5**). $\tau_Z^{\Psi}(z)$ of the peak at 12.0 Å for the compound **1**, 2,471 ns, is significantly smaller than that at 4.0 Å, 4,175 ns. The decay time constant at 12.0 Å for the compound **2**, 290 ns, is four times smaller than that at 4.0 Å, 1,391 ns. Note that the pair at a larger distance in the columnar structure more includes the pyrene rings at the edges of the col-

Table 2.7: Average lifetimes of spatial decompositions of rotation relaxation correlation functions at peak positions z , $\tau_\alpha^\Psi(z)$, and correlation time anisotropy ratios, $p^\Psi(z)$.

	1			2			3		4
z (Å)	4.0	8.0	12.0	4.0	8.0	12.0	4.0	20.0	20.0
τ_X^Ψ (ns)	1837	1009	825	687	311	144	-	1896	1859
τ_Y^Ψ (ns)	1656	1178	767	722	325	147	-	1708	1661
τ_Z^Ψ (ns)	4175	3219	2471	1391	718	290	-	1682	1554
p^Ψ	2.4	2.9	3.1	2.0	2.3	2.0	-	0.9	0.9

umn. The distance dependency of the decay of the peak therefore indicates that the columnar structure collapses faster at its edges.

The decay of the peaks of the compound **1** is slower than that of the compound **2** (**Table 2.7**), although those compounds form the columnar structures. The decay time constant of **1** at 4.0 Å is three times larger than that of **2**, and the ratio of the time constants increases as the pair distance is larger. The difference in the decay time constant between the compounds **1** and **2** indicates that the π - π stacking of the former is stronger than that of the latter. The longer columnar structure of **1** than **2** shown in **Figures 2.7** and **2.8** is consistent with the observation. Despite the strong π - π stacking of **1**, however, the columnar structure undergoes collapse in the microsecond timescale, maintaining liquidity by preventing macroscopic grow of a solid columnar structure.

Unlike the compounds **1** and **2**, the compounds **3** and **4** do not form a columnar structure as described above (**Figures 2.7** and **2.8**). A broad peak of $\tau_\alpha^\Psi(z)$ appears at 20.0 Å for **4** (**Figure 2.10**), and its anisotropy is small ($p^\Psi(z)$ is 0.9) (**Table 2.7**), indicating that the pyrene ring is isolated. The decay time constant is 1,554 ns, which is remarkably smaller than that of **1** and larger than that of **1**. In the case of **3**, a broad peak of $\tau_\alpha^\Psi(z)$ at 20.0 Å similar to that of **4** is also observed. However, in addition, a strong peak of $\tau_\alpha^\Psi(z)$ at 4.0 Å for the pairing by the direct π - π stacking described above (**Figures 2.7** and **2.8**) is found. The pairing peak at 4.0 Å lasts during the simulation time, indicating that the π - π stacking dimeric structure is considerably stable. Because the small anisotropy and the decay time constant of the peak at 20.0 Å for **3** resemble those for **4**, $\tau_\alpha^\Psi(z)$ of the peak at 20.0 Å for **3** expresses overall dynamics of the stable dimeric form.

2.4 DISCUSSIONS AND CONCLUDING REMARKS

We performed long-time MD simulations of the alkyl-pyrene liquids of the compounds **1-4** and comprehensive analyses of their molecular structures and dynamics. The alkyl-pyrene liquids consist of the alkyl- π molecules in which a planar aromatic π conjugated system of pyrene is linked with modifying phenyl groups with long alkyl chains, furnishing the compounds with the nature of liquid by preventing the pyrene from forming a crystal. Rheological properties of the alkyl- π systems are controlled by changing composition ratio of the alkyl chains to the pyrene π -core. Furthermore, changing substituent positions of the alkyl chains also remarkably alters the rheological properties even though the composition ratio is kept identical.

To examine the rheological properties of the alkyl- π liquids, long-time dynamics simulations are required because of their high viscosities. However, atomistic models of the alkyl- π liquids without coarse graining are also necessary to understand the molecular mechanism of the rheology control by changing the substituent positions. Previous theoretical studies based on MD simulations for 1 μ s in total showed that chemical stimulated phase transition of alkyl- π liquids to an ordered phase where experimentally measured viscosity significantly increases is attributed to formation of multilayer π - π stacking columnar structures. [59, 60] However, the MD analyses were limited to an initial process of the ordered structure formation and static properties of the liquid structures due to short trajectory lengths. We therefore carried out long-time MD simulations for tens of microseconds in total with atomic molecular models, and succeeded in semi-quantitatively reproducing the rheological properties experimentally observed (**Table 2.2**). [21] The comprehensive analyses for the long-time trajectories shed light on tight correlation of structural and dynamic characteristics of the alkyl-pyrene liquids in an atomic detail underlying the rheological properties. The spatio-temporal analysis unveiled that the slight differences in the substituent position significantly affect solvent structures and dynamics, and in turn alter the rheological properties, showing a critical importance of the atomistic models for elucidation of the rheological properties of the alkyl- π liquids.

The MD simulations and the spatiotemporal analysis based on the van Hove functions uncovered that the alkyl-pyrene liquids are underlay by the characteristic local liquid structures depending on the substituents (**Figures 2.7-2.11, Table 2.7**). The compounds with one substituent of the phenyl group with the long alkyl chains, *i.e.*, the compounds **1** and **2**, form the columnar structures established by the π - π stacking between the pyrene rings. Because of the smaller ratios of the substituents to the pyrene core, the substituent effect is moderate, invoking the π - π stacking between the pyrene rings in the local structures. Moreover, the positions of the substituents of the alkyl chains influence the π - π stacking character. The π - π stacking columnar structures of the compound **1** with the 3,5-substitutions of the alkyl chains to the phenyl ring develop longer in length and lasted longer in time for several microseconds. In contrast, the compound **2** with the 2,5-substitutions of the alkyl chains form shorter columnar structures with a shorter lifetime. The more fragile nature of the local columnar structure of the compound **2** is attributed to the alkyl chain substituted at the 2-position. Because the alkyl chain at the 2-position points toward the pyrene ring (**Figure 2.1**), it can cover the pyrene ring and thus facilitate dissociation of their π - π stacking.

In the case of the compounds **3** and **4** which are chemically modified with four phenyl substituents with the alkyl chains, no elongated columnar structure was observed. Because of the larger composition ratio of the alkyl chains, the progressive formation of the π - π stacking of the pyrene rings is arrested. Instead, the compound **3** forms the π - π stacking structure of a pair of the pyrene rings without elongation. The pairing stacking structure lasted for more than the simulation time, showing its high stability. Despite the large composition ratio of the alkyl chains, the high flexibility of the alkyl chains permits the pyrene rings to form the strong π - π stacking pair interaction. However, as the bulky alkyl chains are in turn excluded from the π - π stacking side of the pyrene rings and thus are rendered to move to the other sides, the progression of the π - π stacking leading to formation of a columnar structure is prohibited. In the compound **4**, the pyrene ring is surrounded by the alkyl chains and isolated from other pyrene rings. As described above, the alkyl chains substituted at the 2-positions of the phenyl rings covering the pyrene ring prevent the pyrene rings from establishing the π - π stacking interaction.

The local structures of the alkyl-pyrene liquids also underlie the anisotropic characters of the molecular diffusion and the rotational relaxation (**Figures 2.5, 2.6, 2.7-2.11** and **Table 2.4**). The formation of the columnar structures in the alkyl-pyrene liquids of the compounds **1** and **2**, considerably slows translational relaxation along the Z-axis in the molecular frame which is the axis perpendicular to the pyrene π plane as well as rotational relaxation of the Z-axis. The strong π - π stacking interactions in the columnar structures arrest those motions destroying the π - π stacking. In contrast, any anisotropy of the relaxations is absent for the compound **4** because the pyrene ring is isolated from other pyrene rings and thus lacks the π - π stacking interaction. In the case of the compound **3**, the strong pairing π - π stacking structure moves together, and the π - π stacking pair exhibits an isotropic overall movement.

The rheological properties of the alkyl-pyrene liquids were strongly influenced by their local liquid structures. First, dynamics of the alkyl-pyrene liquids clearly exhibit anomalous diffusions. Given that the anomalous diffusions appear for the alkyl-pyrene liquid of the compound **4** where the π - π stacking is absent as well as for the other alkyl-pyrene liquids which forms the stacking local structures, the basic nature of the anomalous diffusion is not related to the π - π stacking but originates from the long alkyl chains of the substituents which are expected to give a viscoelastic caging effect as seen in polymers [91, 92] and liquid crystals. [93] In addition, as the remarkable variation of the rheological properties of the alkyl-pyrene liquids are created by the differences in the composition ratios of the alkyl chains to the pyrene ring and the substituent positions of the formers, the local contact interaction between the pyrene rings and the alkyl chains in the liquids determines the rheological properties. It is suggested that more extensive microscopic contacts between the π plane of the pyrene rings and the alkyl chains where the entangled interaction is absent lead to faster diffusion, which well explains that the alkyl-pyrene liquids of the compounds **1** and **2** possess a higher composition ratio of the pyrene ring to the alkyl chain providing more contacts between the π plane of the pyrene ring and the alkyl chains than those of the compounds **3** and **4** (**Figure 2.1a**), and the mean diffusion coefficients of the former are larger than the latter (**Table 2.2**).

Furthermore, the local liquid structures which remarkably vary depending on the sub-

stituent positions of the alkyl chains also affect the diffusion processes. As the compound **3** forms the pairing π - π stacking structure, the contact interaction between the pyrene plane and the alkyl chain in the liquid of the compound **3** is less than that of the compound **4**, well explaining the smaller diffusion coefficient of the former despite the identical composition ratio of the pyrene ring to the alkyl chains (**Table 2.2**). Although the slower diffusion of the compound **3** observed in the present MD simulations is in qualitative agreement with that observed in the experiment, [21] however, the decrease of the diffusion coefficient is quantitatively underestimated. It is noteworthy that long-chain alkanes composed of 15 or 16 carbon atoms comparable with the alkyl chain substituent of the alkyl-pyrene liquids possesses the melting point around room temperature. [94] Given the dense alkyl chains in the liquid of the compound **3** which prevent the π - π pairing structure from elongation to the columnar one described above, therefore, the liquid may approach to a less fluidic phase where the diffusion is largely arrested. Unfortunately, modeling of such a significantly less fluidic phase would require a MD simulation on a much longer time scale with more elaborated molecular models, which is impracticable with the current computational resources. In fact, the anisotropic components of the mean square displacement are not yet converged in the simulation time even for several microseconds (**Figure 2.5e**).

The lack of temporal convergence for the compound **3** in the long-time region directly affects evaluation of the diffusion coefficient discussed in relation to the results in **Figure 2.5** and **Table 2.2**, as the normal diffusion was only observed in the long-time region. However, the analyses other than that of the diffusion coefficient in the present study, *i.e.*, the analyses of the rotational relaxation and the spatio-temporal correlations mainly showed temporal-behaviors of time-correlations on the shorter time scale, and thus the lack of temporal convergence in the long-time region does not directly influence the analyses. Nevertheless, as thermal equilibrium of the system of the compound **3** is not ensured because of the lack of temporal convergence in the long-time region in **Figure 2.5e** and the pairing stacking lasted for more than the simulation time in **Figure 2.11a**, a possibility that the initial conformations of the time-correlation functions are not in thermal equilibrium remains for the compound **3**.

The compounds **1** and **2** form the characteristic π - π stacking columnar local structures in their liquids. Their diffusions are therefore retarded by decrease of the contact interaction between the π plane of the pyrene rings and the alkyl chains due to the formation of the π - π stacking columnar structure. The columnar structure of the compound **2** is short because of interference of the alkyl chain at the 2-position of the substituent phenyl ring with the π - π stacking of the pyrene rings. Together with the high composition ratio of the pyrene rings to the alkyl chains, therefore, the diffusion coefficient of the compound **2** is large (**Table 2.2**). In contrast, the longer columnar structure of the compound **1** due to absence of the interference of the alkyl chain significantly reduces the contact interaction between the π plane of the pyrene rings and the alkyl chains, and thus slows the diffusion (**Table 2.2**). The collapse dynamics of the columnar local structures examined by the spatiotemporal analysis little affect the diffusion processes as the lifetimes of the columnar structures are sub-micro to microseconds (**Table 2.7** and **Figure 2.11**), while the difference in the MSD between the compounds **1** and **2** already appears in a much shorter time, *i.e.*, sub-nano to nano seconds (**Figure 2.5b**).

As described above, a diffusion coefficient computed for a finite-size system in a periodic boundary condition is affected by a hydrodynamic interaction. As seen in the analytic correction formula for the finite-size effect derived by Yeh and Hummer [80] (Eq. 2.19), the correction is comparable to the diffusion coefficient computed for the finite-system if the size of the molecule is comparable to the system size.

Although the long columnar local structures of the compound **1** which extended to lengths comparable to half of the system size were formed, however, the lifetime of the columnar local structure was shown to be shorter than the time scale of the normal diffusion (**Figures 2.5b, c and 2.11, and Table 2.7**). Thus, it is more appropriate to treat the diffusion as diffusion of the individual small molecules in the inhomogeneous micro-environment than diffusion of the large aggregates of which the structures are kept during the diffusion. The finite-size effect is therefore less significant for the present systems (**Table 2.2**).

In summary, the contact area between the aromatic rings, pyrenes, and the alkyl chains determines the magnitude of the viscosity coefficient of alkyl-pyrene liquids. The compounds **1** and **2** have fewer substituted alkyl chains than **3** and **4**, resulting in a larger area of pyrene-alkyl chain interface, smaller intermolecular friction, and faster diffusion rate. In addition, the π - π stacking columnar structure reduces the number of pyrene rings exposed to the alkyl chain environment. Thus the compound **1**, which forms a longer stacked structure, has a slower diffusion rate than the compound **2**, because of greater intermolecular friction. Similarly, the compound **3**, which forms a pairing π - π stacking structure, diffuses slower than the compound **4**, whose π -core is in an isolated environment. In previous studies of other alkyl- π liquids, the phase transition to an ordered phase in which multilayer π - π stacking grows in chemical stimuli response, and a viscosity increase was observed. [59, 60] Our mechanism can provide a consistent explanation for this as well.

Finally, it is noteworthy that the alkyl-pyrene liquids of the compounds **1-4** feature photoluminescence as the pyrene core exhibits strong absorption and fluorescence of visible light. [21] Interestingly, photochemical properties of absorption and fluorescence spectra and photochemical dynamics remarkably differ among the compounds **1-4**. The local liquid structures characterized by the interaction between the pyrene rings unveiled in the present study is therefore expected to control the photochemical properties of the alkyl-pyrene liquids. A theoretical analysis of the photochemical properties of the alkyl-pyrene liquids is now on-going in our group and will be reported elsewhere.

2.5 APPENDIX

2.5.1 Force field parameters

Table 2.8: Force field parameters of liquid-pyrenes for non-bonding and bonding.

LJ parameters.				
Atom type	σ (Å)	ϵ (kcal/(mol) $^{-1}$)	Mass	
CH,CY	3.695	0.100	13.02	Aromatic
C,CP,CB,CX	3.700	0.060	12.01	Aromatic
C1	4.650	0.020	13.02	sp3
C2	3.950	0.091	14.03	sp3
C3	3.750	0.195	15.04	sp3
O	2.800	0.109	16.00	ether

Bonding parameters.		
Bond type	k_{ab}^{BOND} (kcal/(molÅ 2) $^{-1}$)	r_{ab}^0 (Å)
CP-CB	340.0	1.483 ^a
C*-C*	469.0	1.400
C1-C2	310.0	1.540
C2-C2	310.0	1.540
C2-C3	310.0	1.540
C -O	372.4	1.373
O -C2	308.6	1.416

^a Parameters newly determined by quantum chemistry calculations.

Table 2.9: Force field parameters of liquid-pyrenes for bending, dihedral, and improper torsion.

Bending parameters.				
Bond type		k_{abc}^{BEND} (kcal/(mol) ⁻¹)	θ_{abc}^0 (deg.)	
CB-CP-CX		52.755 ^a	118.275	
CB-CP-CY		52.755 ^a	121.725	
CP-CB-C*		53.389 ^a	121.030	
C*-C*-C*		85.000	120.000	
C*-C -O		87.289	119.290	
C -O -C2		67.843	118.200	
O -C2-C1		88.706	121.870	
* -C1-*		124.25	112.000	
* -C2-*		124.25	114.000	

Dihedral parameters.				
Bond type	paths No.	$k_{abcd,n}^{\text{DIHEDRAL}}$ (kcal/(mol) ⁻¹)	$\varphi_{abcd,n}^0$ (deg.)	n
-CP-CB--	4	1.657 ^a	180.0	2
		1.380 ^a	0.0	4
		-0.274 ^a	180.0	6
-C2-C2--	1	0.706	0.0	1
		-1.136	180.0	2
		1.573	0.0	3
-C1-C2--	1	0.700	180.0	1
		0.900	180.0	2
		0.550	0.0	3
** -C -C -**-	1	3.625	180.0	2
** -C -CH-**-	1	2.650	180.0	2
** -CH-CH-**-	1	5.300	180.0	2
** -O -C2-**-	1	0.383	0.0	3
** -C -O -**-	1	1.610	180.0	2

Improper torsion parameters.			
Bond type	$k_{abcd,n}^{\text{DIHEDRAL}}$ (kcal/(mol) ⁻¹)	$\varphi_{abcd,n}^0$ (deg.)	n
CP-CB-**-**-	4.31 ^a	180.0	2
CB-CP-**-**-	6.63 ^a	180.0	2
C*-C*-**-**-	1.1	180.0	2

^a Parameters newly determined by quantum chemistry calculations.

2.5.2 Parameter fitting for the mixed KWW function

Parameters of the mixed KWW function Eq. 2.24, *i.e.*, $\{c_k\}$, $\{t_k\}$, and $\{\kappa_k\}$ with $k = 1, 2, \dots, K$, were numerically determined by the following fitting procedure. Since the parameter fitting is non-linear, a regularized least squares method was employed. The cost function is given by a function of a sampling time vector $\mathbf{s} \equiv \{s_i\}$ with $i = 1, 2, \dots, M$, the parameter vectors $\mathbf{c}$, $\mathbf{t}$, $\mathbf{\kappa}$, and a regularization coefficient λ as

$$\chi(\mathbf{s}, \mathbf{c}, \mathbf{t}, \mathbf{\kappa}, \lambda) = \frac{1}{2} \sum_{i=1}^M \{\tilde{\omega}(\mathbf{s}, \mathbf{c}, \mathbf{t}, \mathbf{\kappa}) - \omega_i\}^2 + \lambda L(\mathbf{c}, \mathbf{t}, \mathbf{\kappa}), \quad (2.39)$$

where ω_i is the estimated correlation function at time s_i , and $L(\mathbf{s}, \mathbf{c}, \mathbf{t}, \mathbf{\kappa})$ is the L2-norm regularization term

$$L(\mathbf{c}, \mathbf{t}, \mathbf{\kappa}) = \frac{1}{2} \sum_{k=1}^K (c_k^2 + t_k^2 + \kappa_k^2). \quad (2.40)$$

To avoid being trapped in a local solution, an annealing optimization with an inverse temperature parameter β was performed. The procedure is summarized as follows:

Set initial values of the KWW coefficients $\mathbf{c} = \{c_{k,0}\}$, the inverse temperature $\beta = \beta_0$, and the cost value $E = E_0 \gg 1$. Generate candidate coefficients $\mathbf{c}' = \{c'_k\}$ with a random number sequence $\epsilon = \{\epsilon_k\}$, $k = 1, 2, \dots, K$ given from a multi-normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$, where $\mathbf{0}$ is a K -dimensional zero vector and $\mathbf{I}$ is a $K \times K$ identity matrix. To ensure the condition that the coefficient is a positive value, scaling with an exponential function is performed as

$$c'_k = \frac{c_k \exp(\epsilon_k)}{\sum_{k=1}^K c_k \exp(\epsilon_k)}. \quad (2.41)$$

Find $\arg \min_{\mathbf{\tau}, \mathbf{\kappa}} \chi(\mathbf{t}, \mathbf{c}', \mathbf{\tau}, \mathbf{\kappa}, \lambda)$ to obtain candidate KWW function parameters $\mathbf{t}'$ and $\mathbf{\kappa}'$ by using the BFGS method. Accept the candidate parameters with probability

$$P(\mathbf{c}', \mathbf{t}', \mathbf{\kappa}'; \mathbf{c}, \mathbf{t}, \mathbf{\kappa}) = \min(1, \exp[-\beta(\chi(\mathbf{c}', \mathbf{t}', \mathbf{\kappa}') - \chi(\mathbf{c}, \mathbf{t}, \mathbf{\kappa}))]) \quad (2.42)$$

with the Metropolis-Hastings criteria, or return to Eq. 2.41 if rejected.

If the cost value $E' = \chi(\mathbf{c}', \mathbf{t}', \mathbf{\kappa}')$ is lower than the minimum value in the previous iteration cycles, update E with E' , and the inverse temperature β is replaced by β' multiplied by the scaling factor $\xi > 1$. The calculation ends when the number of iterations exceeds the specified number N^{ANIL} . If not, go back to Eq. 2.41.

The calculation above was performed with $K = 1$ with $\lambda = 1.0 \times 10^{-5}$, $\beta_0 = 1.0$, $\xi = 1.2$, and $N^{\text{ANIL}} = 1000$. The parameters determined for $K = 1$ were used as the initial values of the corresponding parameters of the calculation for $K = 2$. An initial value of 0.01 is assigned to the newly added coefficient c_2 . The components of the mixed KWW function were increased up to K in the same manner. The expansion of linear combination in the KWW function of which the number of terms is K is terminated when the coefficient of the newly added $K + 1$ -th term is less than 1.0×10^{-5} .

2.5.3 Confidence interval estimations

Accuracy evaluations of computational estimation of physical quantities have been carried out through the following procedures. For N independent random variable samples $\{X_1, X_2, \dots, X_N\}$, the unbiased variance (UV) is given by

$$S^2 = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2 \quad (2.43)$$

Here, $\bar{X}$ is the sample mean which is the estimator to be found.

The confidence intervals were evaluated for the long-time domain of the time correlation function to assess statistical convergence for physical quantities representing diffusion behavior. A time-average for a time correlation function $C_i(t, t')$ over the simulation time t_{sim} for the molecule i ,

$$X_i(t) = \int_0^{t_{\text{sim}}-t} C_i(t', t' + t) dt' \quad (2.44)$$

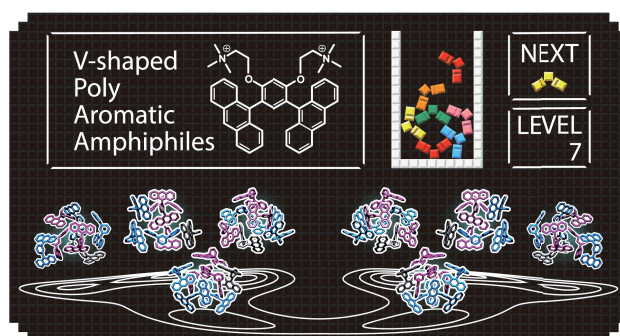
was treated as a single sample which was assumed to be independent when estimating the confidence interval of the autocorrelation function for the delay time t . Thus, the number of independent samples is 200, which is the number of molecules in the simulation system. Let μ_X be the population mean and σ_X^2 be the population variance of the probability distribution that X_i follows. We also assume from the central limit theorem that the distribution of the sample mean $\bar{X}$ converges to a normal distribution with the mean μ_X and the variance σ_X^2/N as the number of samples $N \rightarrow \infty$. In this case, the 99.9% confidence interval of the Student's t distribution is determined by

$$X \in \left[\bar{X} - s_{N-1} \sqrt{\frac{S^2}{N}}, \bar{X} + s_{N-1} \sqrt{\frac{S^2}{N}}, \right] \quad (2.45)$$

where s_{N-1} is the two-sided 99.995% interval of the Student's t distribution with $N - 1$ degrees of freedom, and $s_{N-1} \approx 3.340$ when $N=200$.

Chapter 3

Self-assembly of V-shaped Polyaromatic Amphiphiles



V-shaped polyaromatic amphiphiles (VPAAAs) form micelle-like non-bonded self-assemblies in aqueous solution and feature prominent properties of encapsulation and solubilization for various types of hydrophobic molecules. To understand microscopic molecular characteristics underlying the wide capability of solubilization, the atomic-level molecular structures of the self-assemblies of VPAAAs were investigated by microseconds molecular dynamics (MD) simulations. The MD simulations showed that VPAAAs spontaneously formed quasi-stable self-assemblies in close agreement with experimental observations. To characterize the microscopic structures of the assemblies and their conformational changes, we developed an RMSD-based structural similarity metric applicable to molecular assemblies. Through conformational classification by an unsupervised clustering, highly stable conformations depending on the number of constituent molecules in the assemblies were identified. The analysis revealed that the stable conformations of the VPAA assemblies flexibly reorganized depending on the number of constituent molecules, well explaining the wide capability of solubilization of hydrophobic molecules of various sizes and shapes.

3.1 INTRODUCTION

Molecular capsules create nanometer-sized space inside, which can provide a microenvironment isolated from bulk solution and thus potentially utilized for containers through host-guest encapsulation [95–102] and reactors. [103–107] Self-assembly fragments forming molecular capsules through various interactions such as non-covalent hydrogen-bonding, [95, 96, 103, 104] coordinative metal-ligand, [97, 98] hydrophobic, [99] and van der Waals/C-H ones [100] have been developed. One can control capsules' molecular properties of host-guest encapsulation and reactors by altering the chemical structures of assembly fragments and molecular interactions among them. For example, a well-shaped molecular capsule constituted by rigid polyaromatic panels and assembled through strong metal-ligand coordination encapsulates C60 fullerene of which the size and the shape match those of the cavity of the capsule, and changes its photo-properties. [101] The well-formed cage also furnishes the capsule with the selectivity of encapsulation of substrates. [102–105] For the well-shaped self-assembly capsules with metal-ligand coordination, their formation processes were microscopically investigated by means of molecular simulations. [108–110]

Recently, Yoshizawa and co-workers have continued to synthesize a series of bent aromatic amphiphiles with aromatic moieties as hydrophobic groups and flexible alkyl chains with charged polar groups as the hydrophilic head. [2, 4, 5] The V-shaped polyaromatic amphiphiles (VPAAs) consist of two anthracene panels aligned in a V-shape as the hydrophobic groups which form micelle-like non-bonded self-assemblies in aqueous solution (**Figure 3.1a**). [23] VPAAs are characterized by prominent properties of encapsulation and solubilization for various types of hydrophobic molecules. [23, 111–116] Especially, unlike well-shaped molecular capsules, the self-assemblies can encapsulate a wide variety of hydrophobic carbons of different sizes and shapes, including hydrophobic dyes, [23, 111, 112] aromatic catalysts, [113] stacked planer aromatic compounds, [114–116] and curved nanocarbons. [114] The prominent feature of the wide capacity of encapsulation indicates the high flexibility of molecular conformations of the self-assemblies like micelles. Several experiments also revealed interesting properties of the interior of the hydrophobic cavity as a molecular reaction field, [113] and showed that the modulation of the V-shaped panel affects the encapsulation properties. [23, 111, 112, 116–119]

The self-assembly of VPAAs was examined by various experiments, such as ^1H NMR spectroscopies between poor and good solvents, [23] the dynamic light scattering analysis (DLS), [23] measurement of the average particle size by atomic force microscopy (AFM) on mica surfaces, [23] and the ultrafast spectroscopy using solvatochromic fluorescent probe molecule. [112] Based on the DLS and AFM results, the number of constituent molecules of VPAA assemblies in an aqueous solution was estimated to be 4–6. [23] However, the microscopic conformations of VPAA assemblies are experimentally unclear. The difficulty in microscopic characterization is due to the molecular flexibility of VPAA assemblies, which prevents the formation of certain dominant conformers. Thus, despite that numerous VPAA analogues have been synthesized, their design principle remains in the dark. [2, 5]

In the present study, we theoretically investigated microscopic structures of **VPAA1** and **VPAA2** assemblies in aqueous solution (**Figure 3.1c**). **VPAA1** and **VPAA2** differ only in their hydrophobic side chains, *i.e.*, **VPAA1** possesses two ammonium cations, while **VPAA2**

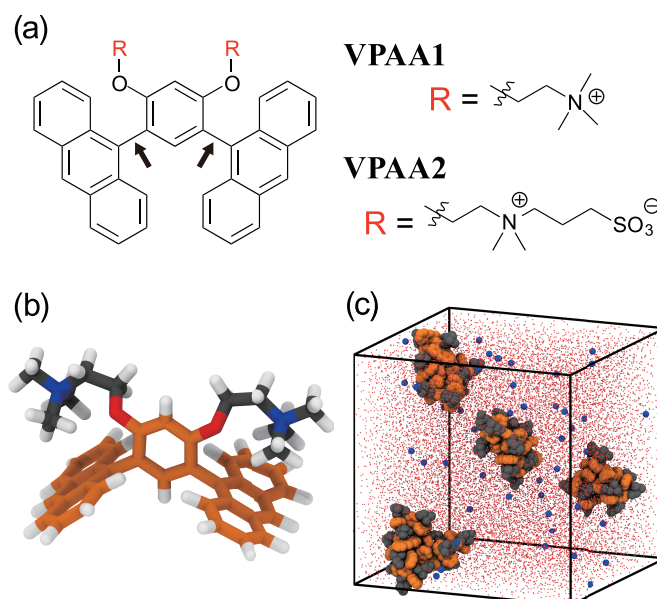


Figure 3.1: Structure of VPAA molecules and an MD simulation system. (a) Chemical structure of VPAA molecules with two different polar chains, **VPAA1** and **VPAA2**. Arrows indicate the bonds of which force field parameters were determined. (b) Molecular structure of **VPAA1**. (c) A snapshot of an MD simulation system of **VPAA1** in aqueous solution. The red and blue points represent water and chloride ions, respectively.

does two zwitterionic sulfobetaine groups (**Figure 3.1a**). By molecular dynamics (MD) simulations, sizes distributions and lifetimes of the spontaneously formed assemblies were examined. The lifetimes obtained by the MD simulations were found to be much longer than predictions of collision frequency by a simple coaggregation theory, [120] and indeed frequent collisions without molecular exchange were observed in time-series of molecular assembly distributions. This demonstrates that microscopic structures consisting of finite conformers are important for the stability of the assemblies.

We therefore characterized the microscopic structures and dynamics of the assemblies by unsupervised clustering of the trajectory conformational data using distance matrices with a root mean square displacement (RMSD) between the conformers, which has been a widely utilized metric to compare their structural differences. [18, 19] For the purpose, we first developed a methodology, symRMSD, which allows one to efficiently evaluate an RMSD between the flexible molecular assemblies in which the constituent molecules freely change their positions and orientations in the trajectories and thus straightforward atom numbering prerequisite for a standard RMSD calculation is precluded.

We then performed unsupervised clustering of the conformers by a hierarchical density-based clustering method, HDBSCAN, [121] using distance matrices based on symRMSDs. The conformational clustering revealed in clarifying microscopic molecular details of transient structures of the assemblies dynamically appearing in the trajectory data. The analyses revealed characteristic local structures in the assemblies which differ depending on the size of the assemblies and the side-chains, providing a molecular insight into the prominent encapsulation and solubilization properties of the aromatic amphiphiles VPAA.

This chapter is organized as follows. The COMPUTATIONAL DETAIL section provides details of MD simulations and the definition of molecular assemblies. In the RESULTS section, we first describe the computational results of the mean lifetimes of VPAA assemblies obtained by MD simulations and comparison of them with a simple coaggregation theory. Then, we introduce the RMSD-based distance function for quantitative structural comparison of VPAA assemblies. We also describe the algorithm of symRMSD to solve the combinatorial problem caused by the internal exchange and symmetry of molecular assemblies. The clustering analysis with the distance function to classify VPAA assembly conformers are then described. Differences in stability depending on the number of constituent molecules and the side-chains are discussed based on atomic coordinates of the conformers in the DISCUSSION section. Finally, a summary of the present study is given in the CONCLUSION section.

3.2 COMPUTATIONAL DETAIL

3.2.1 Force field

We treated two different VPAA molecules, **VPAA1** and **VPAA2**, which possess different polar alkyl chains, *i.e.*, mono-cationic trimethylamine groups and zwitterionic sulfobetaine groups, respectively (**Figure 3.1a**). All MD simulations were performed with a force field based on Amber. [69] We used the general amber force field [122] and TIP3P force field [123] in AmberTools 2015, [124] except for effective atomic charges, and parameters for intramolecular terms including bonds connecting the benzene ring and the anthracene panels, indicated by arrows in **Figure 3.1a**. The force field parameters of the connecting bonds between benzene and anthracenes were determined by quantum chemistry calculations.

The Amber energy function is described as

$$\begin{aligned}
 E = & \sum_k a_{\text{bond},k} (r_k - r_{\text{eq},k})^2 + \sum_l a_{\text{angle},l} (\theta_l - \theta_{\text{eq},l})^2 \\
 & + \sum_m \sum_n \frac{a_{\text{dihedral},mn}}{2} (1 + \cos(n\omega_{mn} - \gamma_{mn})) \\
 & + \sum_j \sum_{i>j} \left\{ \kappa_{ij}^{\text{Cl}} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} + \kappa_{ij}^{\text{LJ}} \epsilon_{ij} \left(\left(\frac{r_{0,ij}}{r_{ij}} \right)^6 - 2 \left(\frac{r_{0,ij}}{r_{ij}} \right)^{12} \right) \right\},
 \end{aligned} \tag{3.1}$$

where bond distances r_k , bond angles θ_l , dihedral angles ω_{mn} , and inter-atomic distances r_{ij} , are the variables of molecular geometry, ϵ_0 is the electric constant, and the others are the force field parameters. We employed the standard excluding/scaling parameters of κ_{ij}^{Cl} and κ_{ij}^{LJ} .

We first performed quantum chemistry calculations to obtain potential energy curves along degrees of freedom including the connecting bonds of a 1,3-di-anthracenyl-benzene part of VPAA. The effective charges were determined with the restrained electrostatic potential fitting for electronic density [65] obtained by quantum chemistry calculations at the M06-2X/6-31G(d) level of theory with the polarizable continuum model of water solvent. [125] For **VPAA2**, we used a basis set 6-31G+(d) for anionic sulphate groups. All the quantum chemistry calculations were carried out using GAMESS (US) (2012 and 2014 versions). [67] The atomic charges determined are summarized **Figure 3.2**.

The force field parameters of the connecting bonds were then determined by fitting the energy function, Eq. 3.1, to the potential energy curves of quantum chemistry calculations. The new atom types for the carbon atoms forming the connecting bonds in the anthracene rings and the benzene ones were defined as xa and xb, respectively. The standard atom types of ca and ha were also set to the other aromatic carbon atoms and the hydrogen ones in those rings. The bonding length and Lennard-Jones parameters of cp were employed for those of xa and xb. We determined the force field parameters of the energy terms of the following five angles: bending angles of ca-xa-xb and ca-xb-xa, improper dihedral angles of ca-ca-xa-xb and ca-ca-xb-xa, and a dihedral angle of ca-xa-xb-ca. We determined parameters of the energy terms of those angles, $a_{\text{angle},l}$ and $a_{\text{dihedral},mn}$, in Eq. 3.1 by least square fitting. The equilibrium angles of the bending angles of ca-xa-xb and ca-xb-xa, the improper dihedral

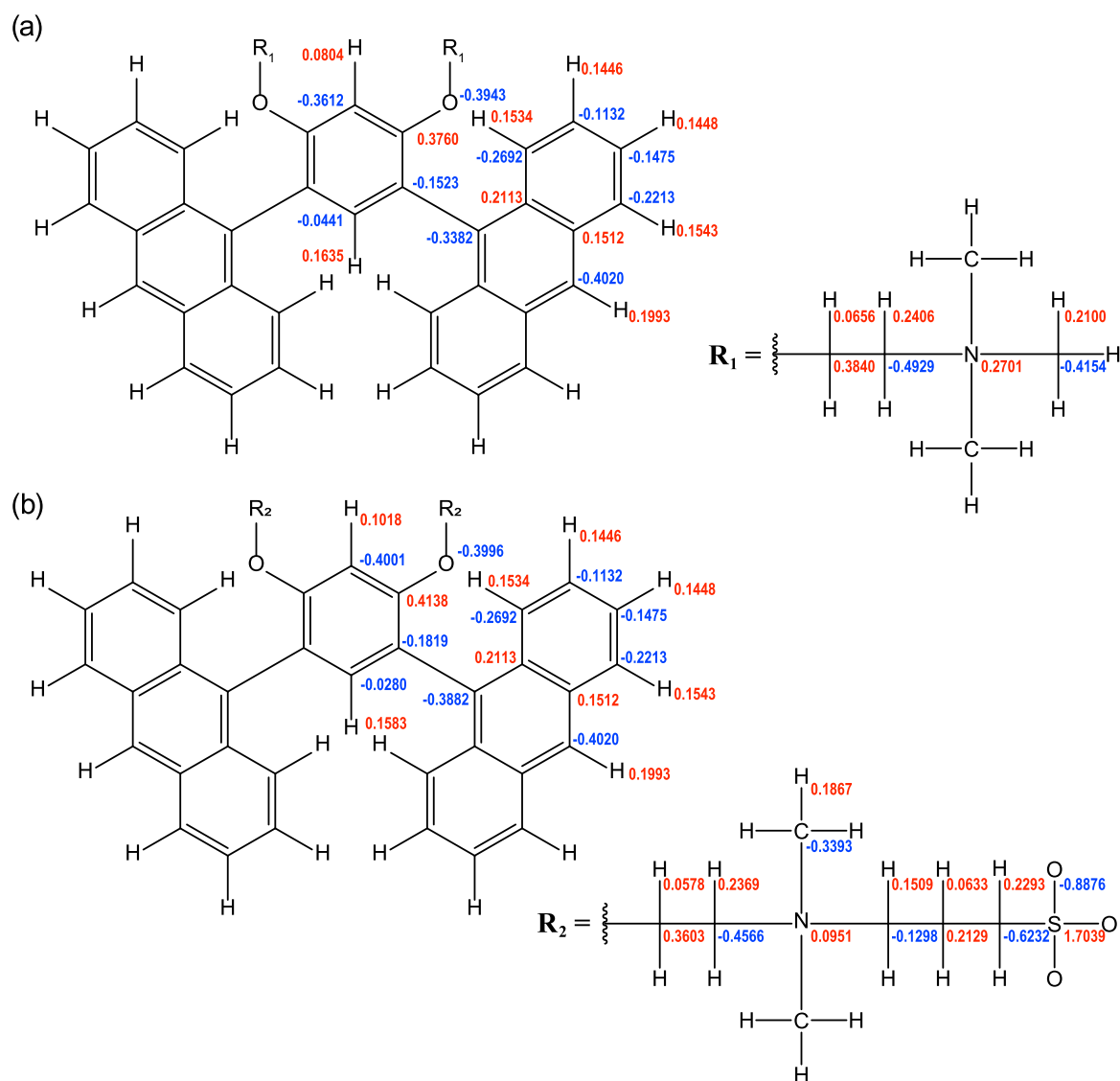


Figure 3.2: Effective atomic charges of VPAA1 (a) and VPAA2 (b) in the elementary charge unit. The effective charges of the symmetrically unique atoms are shown.

angles of ca-ca-xa-xb and ca-ca-xb-xa, and the dihedral angle of ca-xa-xb-ca were set to be 120, 180, and 0 degrees, respectively. The periodicities of $n = 2$ and $n = 2, 4, 6$, and 8 were chosen for the improper dihedral angles and the dihedral ones, respectively. The potential energies were evaluated only along each of the angles, since evaluation of the potential energies at a great number of geometry points in the correlated space of the five angles is computationally too demanding. We therefore performed a least square fitting of each of the amplitude parameters with respect to the corresponding potential energy curve, and iterated the fitting one by one until convergence. We employed 67 samples in total: 13 samples for ca-xa-xb, 14 samples for ca-xb-xa, 10 samples for ca-ca-xa-xb, 17 samples for ca-ca-xb-xa, and 13 samples for ca-xa-xb-ca. Despite that the procedure does not guarantee convergence of the iteration, the parameters were well converged in 50 iterations within negligible deviations of 0.2 kcal/mol per one sample conformation because of small geometric correlations of the angles found in the quantum chemical geometry optimizations for the evaluation of the potential energy curves.

Figure 3.3 shows potential energy curves with the newly determined forcefields around one of the two connecting bonds of 1,3-di-anthracenyl-benzene evaluated by quantum chemistry calculations at the M06-2X/6-31G(d) level of theory. In the calculation of the potential energy curves, geometry of all degrees of freedom except for each of the angles were optimized. One can discern that the potential energy profile of the dihedral angle around ca-xa-xb-ca is rather flat around the perpendicular angle due to the cancellation of the conjugation of the two π systems and the steric hinderance.

Table 3.1: Force field parameters of the connecting bond between a benzene ring and an anthracene panel of VPAA.

Bonding parameters			
atom type	a_{angle} (kcal/mol)	θ_{eq} (degree)	
ca-xa-xb	29.888	120	
ca-xb-xa	40.244	120	
Improper dihedral parameters			
atom type	a_{dihedral} (kcal/mol)	γ (degree)	n
ca-ca-xa-xb	2.140	180	2
ca-ca-xb-xa	2.085	180	2
Dihedral parameters			
atom type	a_{dihedral} (kcal/mol)	γ (degree)	n
ca-xa-xb-ca	-1.449	0	2
	0.405	0	4
	0.515	0	6
	0.222	0	8

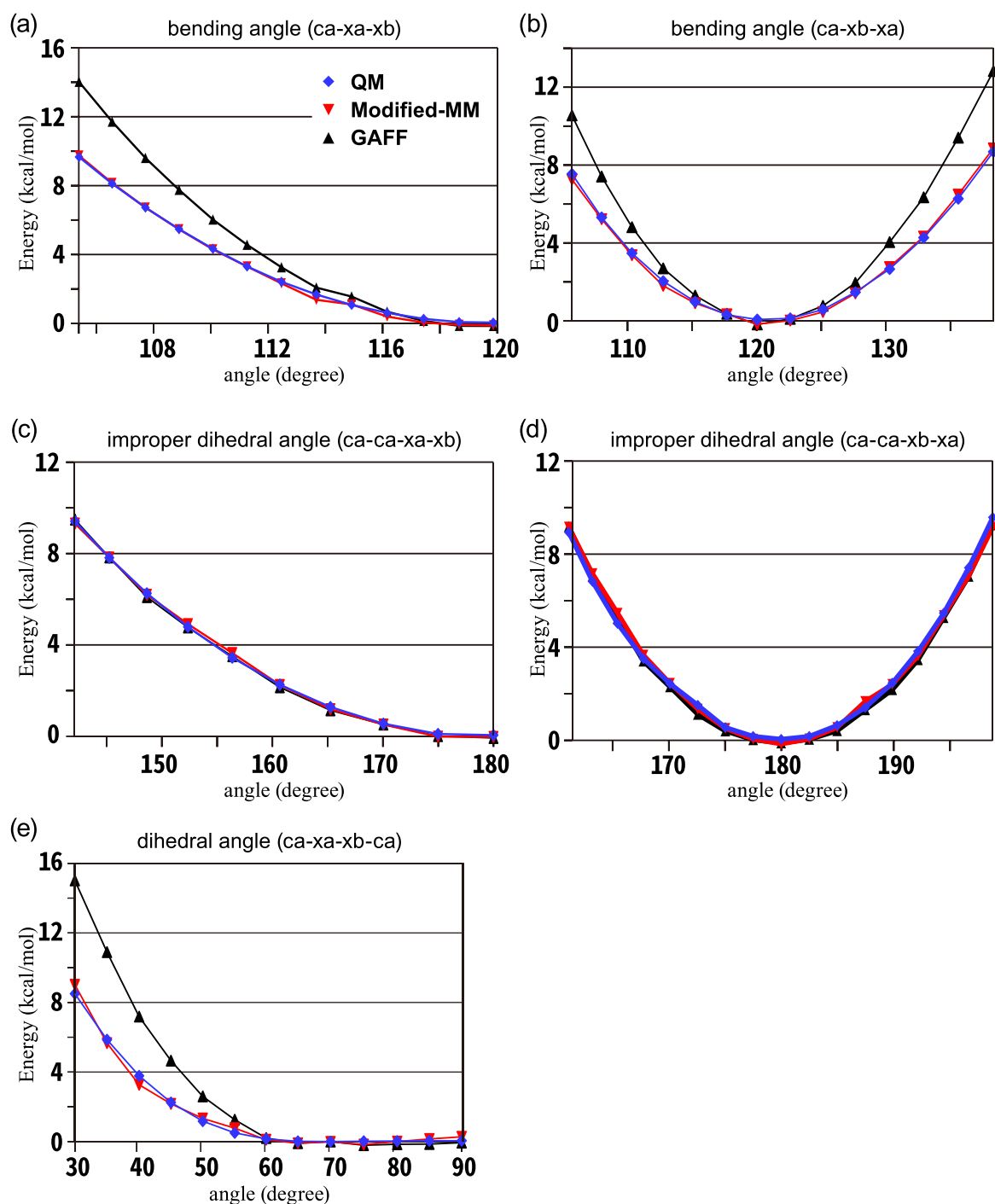


Figure 3.3: Potential curves for the newly determined parameters. (a), (b) Potential energy curves along bending angles, ca-xa-xb (a), ca-xb-xa (b). Lines with diamond symbols (blue), inverse triangle ones (red), triangle ones (black) represent the potential energy curves obtained by the quantum chemistry calculation, the modified molecular force field, and GAFF, respectively. (c), (d) Improper dihedral angles (ca-ca-xa-xb (c), ca-ca-xb-xa (d)). (e) A dihedral angles, ca-xa-xb-ca.

3.2.2 MD simulation

We performed conventional MD (cMD) simulations of 27 molecules of **VPAA1** and **VPAA2**, respectively, in aqueous solvents to examine the structures and dynamics of amphiphilic micelle-like assemblies of VPAAAs. Both simulation systems, **VPAA1** and **VPAA2**, contains 17,351 water molecules. The **VPAA1** system contains an additional 54 chloride ions. The **VPAA1** system and the **VPAA2** one contain 54,645 atoms and contains 55,077 atoms, respectively. We constructed the initial simulation systems randomly because experimental structural data for the VPAA assemblies were not available. A priori modeling of self-assemblies is extremely difficult because it is impossible to distinguish whether the instability is an inherent property of the self-assembly or due to modeling failure. The arrangement of solute molecules was constructed using the software, packmol. [71]

The procedure for randomizing the initial structure is as follows: First, energy minimizations of the systems were performed in 1,000 step. Then, the systems were heated up from 0 K to 300 K in 2 ns, and equilibrated for 2 ns at 300 K in NVT ensemble. NPT simulations for 10 ns at 300 K then followed to determine the simulation box sizes. After equilibrations, the box sizes were nearly constants and cubic, and they were determined to be 82.01 Å per side for **VPAA1** and 82.07 Å per side for **VPAA2**, respectively. Finally, we heated up the systems to 1,000 K and then cooled down them to 300 K in NVT ensemble (equilibration at 300 K for 2 ns, heating up from 0 K to 300 K in 2 ns, equilibration at 1,000 K for 10 ns, cooling down from 1,000 K to 300 K in 2 ns, and equilibration at 300K for 2 ns) to randomize the solute positions.

After equilibration to randomize initial conformations, the conformational sampling at 300 K for 10 μ s in NPT was performed. The box sizes of the simulation systems quickly converged to cubes of about 82.01 and 82.07 Å on a side for **VPAA1** and **VPAA2**, respectively. The concentration of VPAA in the simulation systems, 81 mM, was set to be high compared to experimental ones (~ 2.0 mM) [23] to stimulate the exchange of VPAAAs among assemblies in the systems and thus to obtain sufficient equilibrium distributions of the assembly size. Despite the concentrations of VPAAAs being high, the clusters were observed to be well isolated from one another, allowing analysis of the assembly size.

To obtain equilibrium ensembles, we also performed replica exchange molecular dynamics (REMD) simulations, [126] since the relaxation times of VPAA assemblies are relatively slow. First, we determined the highest temperature (380 K) of REMD at which sufficient exchanges between assemblies were obtained. The initial structures for replicas were prepared in the same manner as those of the cMD simulations. The initial structures of each replicas were prepared in the same manner as for the conventional MD simulations. The number of the replicas was 40 each for **VPAA1** and **VPAA2**, and the target temperatures were set to realize enough exchange between each replica as 300.00, 301.70, 303.41, 305.12, 306.85, 308.60, 310.37, 312.16, 313.97, 315.80, 317.65, 319.52, 321.41, 323.32, 325.25, 327.20, 329.17, 331.16, 333.17, 335.20, 337.25, 339.32, 341.41, 343.52, 345.65, 347.80, 349.97, 352.16, 354.37, 356.60, 358.85, 361.12, 363.41, 365.72, 368.05, 370.40, 372.77, 375.16, 377.57, and 380.00 K. The exchange attempts were performed at every 2 ps. Equilibrium simulations were performed for 1 μ s in NPT ensemble. The average exchange probability was 20%. Snapshots were recorded every 10 ps and used for analysis.

3.2.3 The VPAA assembly definition

We analyzed the temporal behavior of the assembly size defined geometrically. One of the simplest definitions of irregularly shaped molecular aggregates is to set a threshold of the intermolecular nearest neighbor distance. The nearest neighbor distance is defined as the shortest distance between two atoms belonging to different molecules. [127] The aromatic carbon atoms were used to calculate the interatomic distances, and the threshold was set to 5 Å. Any pair of molecules whose nearest neighbor distance is below a threshold is defined as being in the same assembly. In the following, molecular assemblies constituted by M molecules of **VPAA1** and **VPAA2** are denoted as $(\text{VPAA1})_M$ and $(\text{VPAA2})_M$, respectively.

3.3 RESULTS

3.3.1 Time evolution of the VPAA assemblies

In general, multiple assemblies with the same number of constituent molecules appear in the simulation systems during the trajectory calculation. The number of the VPAA assemblies of M molecules at time t is denoted by $N(M, t)$. The total number of all the VPAA assemblies is expressed as

$$N_0(t) = \sum_{M=1}^{\infty} N(M, t), \quad (3.2)$$

and the fraction of the VPAA assemblies of M molecules is $X(M, t) = N(M, t)/N_0(t)$. Using the average of M as $\bar{M}(t) = \sum_{M=1}^{\infty} M X(M, t)$, the product $N_0(t)\bar{M}(t)$ is a conservative quantity and is equal to the number of the VPAA molecules, N_{mol} , in the simulation systems.

The time evolutions of the fraction and their statistics are summarized in **Figure 3.4**. Changes of $X(M)$ indicating exchanges of VPAA among the assemblies are observed frequently and intermittently during the simulation (**Figure 3.4a**). These exchanges occur on a sub-microsecond timescale and that of **VPAA1** is particularly slow compared to **VPAA2**. Formation of more than eight assemblies in either system is rare, and they decay in sub-microseconds to microseconds.

As shown in (**Figure 3.4b**), the average sizes show moderate growth in cMD, while they quickly converge to constants in one microsecond REMD trajectories; these of **VPAA1** and **VPAA2** approach $\bar{M} \sim 6.0 \pm 0.8$ and $\bar{M} \sim 5.2 \pm 0.8$, respectively. Here, the first 200 ns samples of REMD are discarded, and statistics are computed from the last 800 ns samples. The numbers indicated by $\pm$ are the standard deviations. Equilibrium distributions of $X(M)$ in cMD and REMD are shown in **Figures 3.4d** and **e**, respectively. While the equilibrium distributions computed from cMD still contain large confidence intervals (**Figure 3.4d**) because of the long autocorrelations in the cMD trajectories, REMD shows adequate convergence (**Figure 3.4e**). Thus, we analyze the equilibrium distributions of REMD hereafter.

In the REMD size distributions, $(\text{VPAA1})_6$, $(\text{VPAA1})_7$, $(\text{VPAA2})_5$, and $(\text{VPAA2})_6$ are dominant (**Figure 3.4e**). These numbers of molecules in the dominant assemblies are equal to or slightly larger than the experimental estimates, $M \sim 4$ to 6. [23] The MD simulations were performed in high concentrations of ~ 81 mM compared to experimental one of ~ 2.0 mM to obtain sufficient samples within the limited simulation time as described above. Given that the critical micelle concentrations, which is the concentration that balances the growth due to intermolecular collisions and decay rate of aggregation, are ~ 1.0 mM for **VPAA1**, and ~ 0.03 mM for **VPAA2**, [23] the overestimation of the average number sizes in the simulation are therefore reasonable. Despite the high concentration conditions, a certain extent of free species ($M=1$) is present. This suggests that free species are involved in molecular exchange between the VPAA assemblies.

At high temperatures (340 K and 380 K), free molecules increase and the distribution peaks shift to smaller values (**Figure 3.4e**). However, the stability of the assemblies is not compromised, and even at 380 K the distribution still peaks at $M=5$ for both **VPAA1** and

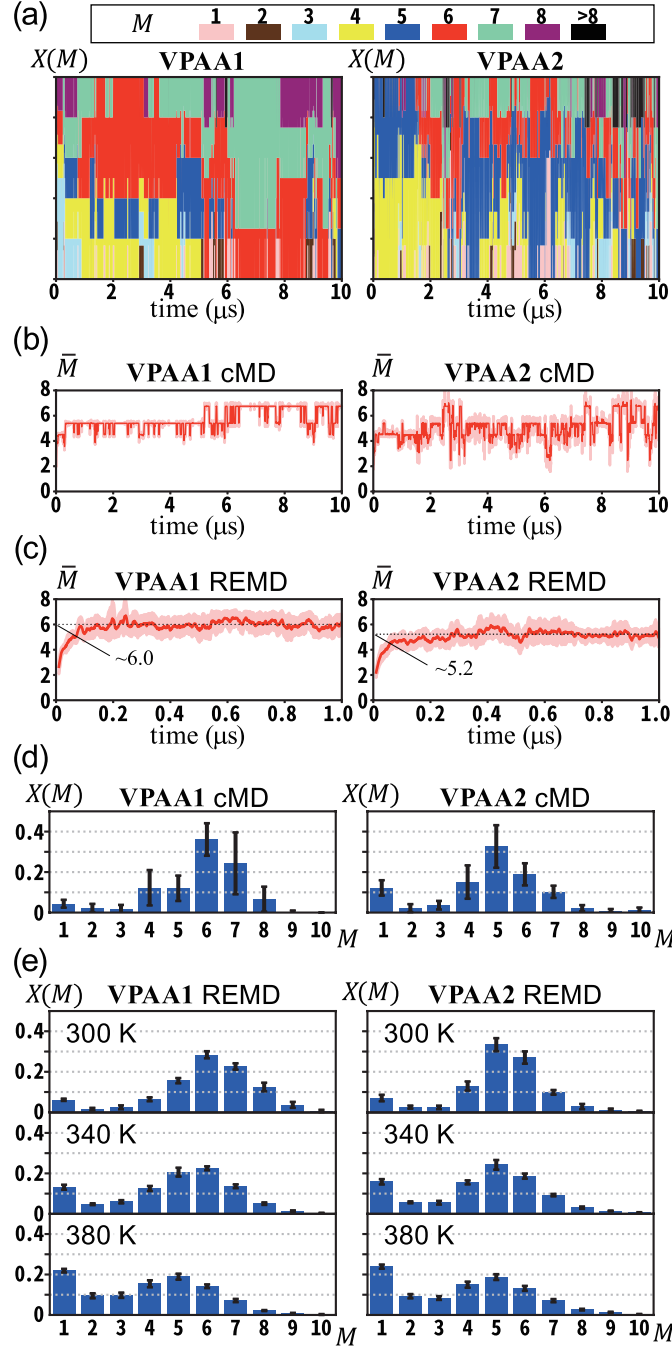


Figure 3.4: Size distributions of VPAA assemblies. **(a)** Stack plots of the assembly size distributions in cMD. **(b)** Time evolutions of the average distribution numbers of the constituent molecules, $\bar{M}(t)$, in cMD. Red lines represent the means in the 100 ns windows, and pink areas represent the standard deviations in the windows. **(c)** Time evolutions of $\bar{M}(t)$ in REMD. **(d)** Equilibrium distributions of $X(M)$ in the cMD simulations. Error bars indicate 95% confidence interval for the Student's t distribution, and independent samples were obtained by block averaging method. See the Confidence interval estimations section in APPENDIX for details. **(e)** Equilibrium distributions of $X(M)$ at different temperatures in the REMD simulations. Error bars are the same as in (d).

VPAA2. The distribution shapes of **VPAA1** and **VPAA2** are almost identical at 380 K. This is because at high temperatures, entropic hydrophobic interactions are dominant and differences in their side chains are neglected.

It is noteworthy that the growth of the VPAA assemblies stops at certain sizes even under highly concentrated conditions. For both the **VPAA1** and **VPAA2** assemblies, those with $M \geq 8$ disappear soon after formation and cannot last for a long time (**Figure 3.4a**). The observation indicates that the growth of unlimited aggregates such as agglomerated colloids or micelles is very slow or does not occur.

3.3.2 Lifetime of molecular assembly

To discuss the stability of the VPAA assemblies, we first define the lifetime of the assembly. We introduce the set, $\mathcal{Z}_M = \{(i_1, i_2, \dots, i_M) \mid 1 \leq i_1 < \dots < i_M \leq N_{\text{mol}}\}$, whose elements are M -tuples of M -combinations of indices of the N_{mol} molecules, i_m . The number of elements of $\mathcal{Z}_M$ (M -tuples) is therefore $\binom{N_{\text{mol}}}{M}$. For $z \in \mathcal{Z}_M$, the indicator function $\theta(t; z)$ returns 1 if and only if all molecules indexed by z at time t are present in the same assembly by the aforementioned definition and no other molecules belong to the same assembly. Otherwise, it returns 0. We define the survival time correlation function for the assemblies of M molecules as

$$\Psi(t; M) = \frac{1}{T} \int_0^T dt_0 \sum_{z \in \mathcal{Z}_M} \theta(t_0 + t; z) \theta(t_0; z), \quad (3.3)$$

where T is simulation time. $\Psi_0(M) \equiv \Psi(t = 0; M)$ is the time average of the number of the assemblies of M molecules $\bar{N}_0 \bar{X}(M)$, where $\bar{N}_0$ and $\bar{X}$ is the time averages of $N_0(t)$ and $X(M, t)$, respectively, and $\bar{X}(M) = \Psi_0(M) / \sum_{M=1}^{\infty} \Psi_0(M)$ holds. After a VPAA assembly collapses, the same molecules may form again the same assembly by chance, and thus the indicator gives a certain value at this large lag time t . However, as combination of molecules forming the assembly is enormous, the limit of the correlation function $\lim_{t \rightarrow \infty} \Psi(t; M) / \Psi_0(M)$ converges to nearly zero.

The mean lifetime of the assemblies of M molecules is defined by

$$\tau(M) = \int_0^{\infty} dt \frac{\Psi(t; M)}{\Psi_0(M)}. \quad (3.4)$$

The time correlation functions calculated from the cMD trajectories are shown in **Figure 3.5**. Here, the cMD trajectories in the first 1 μs were discarded and those in the last 9 μs were used for the analysis. Most of the correlations begin to decay at around 10 to 100 nanoseconds and disappear in a few microseconds. The lifetimes depend significantly on the number of molecules in the assembly; those for $M=4,6,7$ for **VPAA1** and $M=4,5$ for **VPAA2** are relatively long, respectively (**Figure 3.5a** and **b**).

As a qualitative measure of the lifetime of assemblies, a simple molecular coaggregation model is considered for comparison. Here, we use the Smoluchowski coagulation equation of a sphere in Brownian motion without interaction potential. [120] Assume that the M molecular assembly is a sphere and its radius R is related to M as

$$R(M) = r_1 M^{1/3} \quad (3.5)$$

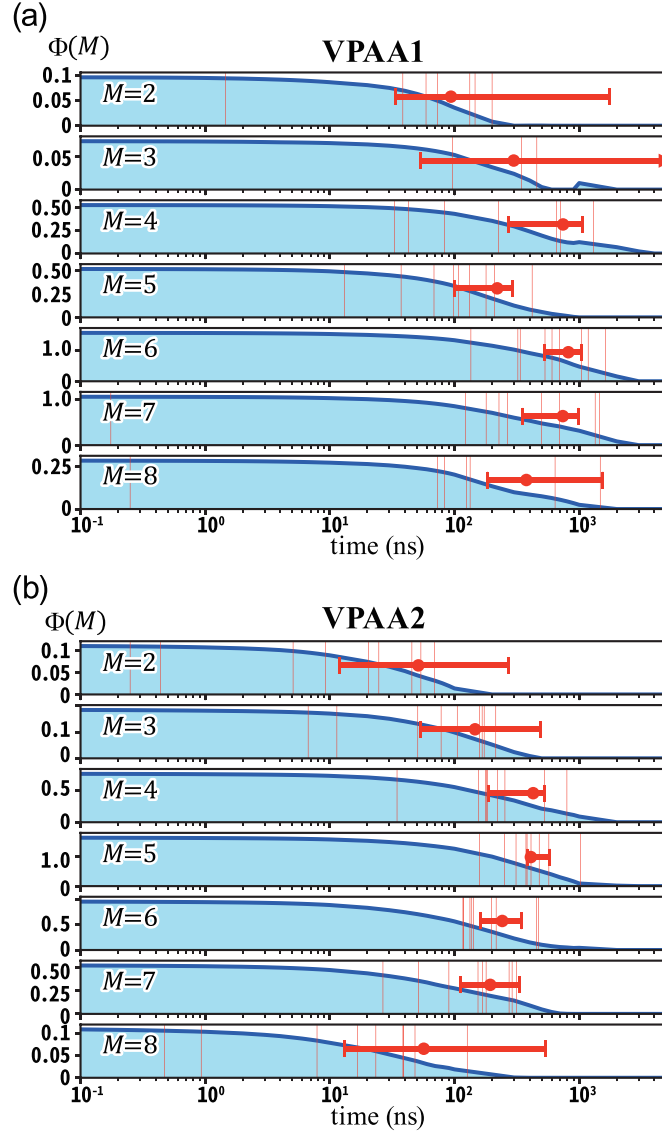


Figure 3.5: Survival time correlation functions for VPAA assemblies of M molecules. **(a) VPAA1. (b) VPAA2.** The red circle plot with error bars represents the estimated mean lifetime τ_{MD} , and the error bars represent 95% confidence intervals from the Weibull distribution. Vertical bars indicate mean lifetimes windowed every $1 \mu s$. See APPENDIX for details on confidence interval estimation.

where r_1 is a proportional constant. In the equilibrium state, distributions of molecules and assemblies are assumed to be spatially uniform. In terms of a given M molecular assembly, the collision rate is defined as

$$\lambda_{\text{col}}(M) = \sum_{M'=1}^{\infty} K(M, M') \rho(M, M'), \quad (3.6)$$

where $K(M, M')$ is the flux of the collision between an M -molecular assembly and an M' -molecular one at the contact radius

$$K(M, M') = 4\pi (D(M) + D(M')) (R(M) + R(M')), \quad (3.7)$$

where $D(M)$ is the diffusion constant of the M -molecular assembly. $\rho(M, M')$ is the concentration of M' -molecular assemblies around the given M -molecular assembly in the MD system,

$$\rho(M, M') = \left(1 - \frac{M}{N_{\text{mol}}}\right) \rho(M'), \quad (3.8)$$

where ρM is the concentration of the molecular assemblies given by $\bar{N}_0 \bar{X}(M)/V$ and $V = L^3$ is the average box volume of the system with L as one side. From Einstein-Stokes equation, [79] an inversely proportional relationship between the diffusion coefficient and $D(M)$ is obtained as

$$D(M) = v r_1^{-1} M^{1/3} \quad (3.9)$$

with a proportionality constant, v . The mean diffusion coefficient is then defined by

$$\bar{D} = \sum_{M=1}^{\infty} D(M) X(M). \quad (3.10)$$

Using Eqs. 3.6-3.10, the collision rate equation is derived as

$$\lambda_{\text{col}}(M) = \frac{4\pi r_1 \bar{D}}{\sum_{M=1}^{\infty} M^{-\frac{1}{3}} X(M)} \sum_{M'=1}^{N_{\text{mol}}} \left(1 - \frac{M}{N_{\text{mol}}}\right) \rho(M') \left(M^{\frac{1}{3}} + M'^{\frac{1}{3}}\right) \left(M^{-\frac{1}{3}} + M'^{-\frac{1}{3}}\right). \quad (3.11)$$

The formula represents a Brownian kernel with fractal dimension three. [128, 129] We define the average collision time by

$$\tau_{\text{col}}(M) = \frac{1}{\lambda_{\text{col}}(M)}. \quad (3.12)$$

The mean diffusion coefficient in Eq. 3.11 was computed with the cMD trajectory by

$$\bar{D} = \frac{1}{6} \lim_{t \rightarrow \infty} \frac{d}{dt} \text{MSD}(t), \quad (3.13)$$

using the mean squared displacements

$$\text{MSD}(t) = \frac{1}{N_{\text{mol}}} \sum_{I=1}^{N_{\text{mol}}} \|\mathbf{c}_I(t) - \mathbf{c}_I(0)\|^2, \quad (3.14)$$

where $\mathbf{c}_I(t)$ is the center of mass of the I -th molecule at time t . The mean diffusion coefficients were obtained by fitting a linear function to the 1,000-2,000 ns interval of the MSD in the cMD simulation. The molecular radius, r_1 , in Eq. 3.11 was approximated by the radius of gyration of the molecules in the REMD simulation as

$$r_1 = \frac{1}{T} \int_0^T dt \frac{1}{N_{\text{mol}}} \sum_{I=1}^{N_{\text{mol}}} \sqrt{\frac{1}{N_{\text{atom}}} \sum_{j=1}^{N_{\text{atom}}} \|\mathbf{r}_{jI}(t) - \mathbf{c}_I(t)\|^2}, \quad (3.15)$$

where N_{atom} is the number of atoms in a molecule, and $\mathbf{r}_{jI}$ represents the coordinate of j -th atom in I -th molecule. The calculated values of $\bar{D}$ and r_1 are summarized in **Table 3.2**. The concentration of the M -molecular assembly, $\rho(M)$, in Eq. 3.11 was evaluated with the size distributions obtained by the REMD simulation (**Figure 3.4e**).

Table 3.2: Mean diffusion constants, radii of gyration, and box sizes L . Numbers with a positive or negative sign indicate the 95% confidence intervals from the Student's t distribution, and numbers in parentheses indicate the standard deviation. See APPENDIX for details on confidence interval estimation.

	VPAA1	VPAA2
$\bar{D}$ ($10^{-9}\text{m}^2/\text{s}$)	0.091 ± 0.010	0.066 ± 0.005
r_1 ($\text{\AA}$)	5.4 (0.06)	5.7 (0.16)
L ($\text{\AA}$)	82.01 (0.03)	82.07 (0.03)

The lifetimes of the assemblies, τ_{MD} , and the average collision times, τ_{col} , are listed in **Table 3.3**. τ_{MD} of both the **VPAA1** and **VPAA2** assemblies found in the cMD simulations are tens of nanoseconds to sub-microsecond, while τ_{col} are 5 to 7 ns and are thus 10 to 100 times shorter than τ_{MD} . Although we assumed spheres in Brownian motion without interaction in the evaluation of τ_{col} , the considerably large differences between τ_{MD} and τ_{col} indicate that the VPAA assemblies are stable once formed, and molecular exchange by collision is a relatively rare event. In fact, the frequent collisions without molecular exchange are seen as many instantaneous spikes of the assembly size change in the time-series of the molecular assembly distributions shown in **Figure 3.4a**.

As mentioned above, the lifetimes of the VPAA assemblies are dependent on the number of molecules, M . For example, one can discern clear differences in τ_{MD} between the assemblies of different sizes (*e.g.*, 219 ns for $M=5$ vs. 817 ns for $M=6$ for **VPAA1**, and 408 ns for $M=5$ vs. 240 ns for $M=6$ for **VPAA2**). The size dependence of the lifetime implies that the conformations of the assemblies are not completely random, and they form short-lived structures of which microscopic conformations differ depending on the number of molecules constituting the assemblies. Comparing **VPAA1** and **VPAA2**, the lifetimes of **VPAA1** tend to be about twice longer than these of **VPAA2**, except for $M=5$, implying that the side-chains also play a role in the formation of the short-lived structures of the assemblies.

Table 3.3: Mean lifetimes, τ_{MD} , and average collision times, τ_{col} , of the VPAA assemblies. Positive and negative numbers attached to τ_{MD} and τ_{col} represent 95% confidence intervals based on the Weibull and Student’s t distributions, respectively. See APPENDIX for details on confidence interval estimation.

M	VPAA1		VPAA2	
	τ_{MD} (ns)	τ_{col} (ns)	τ_{MD} (ns)	τ_{col} (ns)
1	51^{+104}_{-24}	$4.7^{+0.6}_{-0.5}$	139^{+123}_{-53}	$6.4^{+0.6}_{-0.5}$
2	93^{+1298}_{-67}	$5.1^{+0.6}_{-0.5}$	51^{+359}_{-42}	$7.0^{+0.6}_{-0.5}$
3	297^{+26774}_{-244}	$5.4^{+0.7}_{-0.5}$	145^{+333}_{-92}	$7.4^{+0.7}_{-0.6}$
4	741^{+323}_{-472}	$5.7^{+0.7}_{-0.6}$	425^{+96}_{-240}	$7.7^{+0.7}_{-0.6}$
5	219^{+70}_{-121}	$6.0^{+0.8}_{-0.6}$	408^{+171}_{-27}	$8.1^{+0.7}_{-0.6}$
6	817^{+219}_{-287}	$6.3^{+0.8}_{-0.7}$	240^{+101}_{-80}	$8.4^{+0.8}_{-0.6}$
7	737^{+288}_{-405}	$6.6^{+0.8}_{-0.7}$	191^{+135}_{-81}	$8.8^{+0.8}_{-0.7}$
8	375^{+1387}_{-221}	$6.9^{+0.9}_{-0.7}$	55^{+598}_{-44}	$9.3^{+0.8}_{-0.7}$

3.3.3 Molecular symmetry corrected RMSD

We characterized the short-lived molecular structures of the assemblies appearing in conformational thermal fluctuations. We performed a clustering using a hierarchical clustering method with RMSD which is used as a standard method because of its robustness. [121] RMSD is a good metric to characterize difference in molecular structures and thus often used for macromolecules such as proteins. [130] Unlike a protein which consists of a single peptide polymer chain, however, the molecular assemblies treated in the present study undergo highly flexible conformational changes in which VPAA molecules constituting the assemblies freely change their positions and orientations in the assemblies. Hence, the use of RMSDs for the molecular assemblies suffers from difficulty arising from symmetry of positional permutation of the molecules in the assembly and rotation/reflection symmetries of each molecule in the assembly and of the entire assembly itself. The symmetries preclude a straightforward assignment of atom numbering of the assembly undergoing the positional and conformational changes in a trajectory prerequisite for the RMSD calculation.

For the VPAA assemblies, therefore, an atomic index matching between the compared atomic coordinates in the minimization of RMSD with respect to translations and rotations is needed. Because of the permutation search for the atom index matching, the time complexity of finding the globally optimal solution is naively $\mathcal{O}(n!)$ for a molecule consisting of n atoms. No methods to globally minimize RMSD in polynomial time without using fixed atomic index matching or fixed molecular superpositions are known so far. [131–133] Thus, we first propose a matching algorithm without the fixed superposition specific to homologous molecular assemblies.

The algorithm is based on atomic indices permutation corresponding to molecular symmetry and a branch-and-prune method. We first represent a molecular assembly structure consisting of M molecules with n atoms as an $M \times n$ matrix $\{\mathbf{X}^{(M)}\}_{jI} = \mathbf{x}_{jI}$ where $\mathbf{x}_{jI}$

denotes j -th atomic coordinate of the I -th molecule in the assembly, and thus I and j are indices satisfying $1 \leq I \leq M$ and $1 \leq j \leq n$, respectively. The conventional minimized RMSD between two assemblies $\mathbf{X}^{(M)}$ and $\mathbf{X}'^{(M)}$ is defined by

$$d\left(\mathbf{X}^{(M)}, \mathbf{X}'^{(M)}\right) = \min_{\mathbf{R} \in \mathcal{R}} \sqrt{\frac{1}{Mn} \sum_{I=1}^M \sum_{j=1}^n \|\mathbf{x}_{jI} - \bar{\mathbf{x}} - \mathbf{R}(\mathbf{x}'_{jI} - \bar{\mathbf{x}}')\|^2}. \quad (3.16)$$

where, $\bar{\mathbf{x}}$ and $\bar{\mathbf{x}}'$ are the centers of mass of $\mathbf{X}^{(M)}$ and $\mathbf{X}'^{(M)}$, respectively, and $\mathcal{R}$ is the total set of normalized rotation matrix, $\mathbf{R}$, corresponding to the dimension of $\mathbf{x}_{jI}$. We define an injective function of the molecular permutation $\mu : \mathcal{I} \mapsto \mathcal{I}$, that exchange indices of the index sets $\mathcal{I} = \{1, 2, \dots, M\}$. The total set of μ , $\mathcal{M}$, is equal to the total set of M -order permutations.

We then introduce the permutation operations to implement the molecular symmetry. The molecular symmetry manipulation function, $\nu : \mathcal{J} \mapsto \mathcal{J}$, exchanges indices of the index set $\mathcal{J} = \{1, 2, \dots, n\}$. The total set of ν , $\mathcal{N}_S$, is a much smaller subset than the set of n -order permutations since ν works only for the molecular symmetry. The C_{2v} symmetry operations $\{E, C_2, \sigma_v, \sigma_{v'}\}$ for the VPAA molecules are taken for the set of ν , $\mathcal{N}_S = \{\nu_E, \nu_{C_2}, \nu_{\sigma_v}, \nu_{\sigma_{v'}}\}$, respectively. ν_{C_2} corrects the symmetric rotation around the C_2 axis by 180° . Although the atom indices changed by $\nu_{C_{\sigma_v}}$ and $\nu_{C_{\sigma_{v'}}}$ do not overlap with the original ones by any rigid-body rotations, $\nu_{C_{\sigma_v}}$ and $\nu_{C_{\sigma_{v'}}}$ are used to find the minimum RMSD between conformations undergoing a mutual rotational around the C_2 axis by $\sim 90^\circ$.

For simplicity, we denote a tuple of the functions as $\varphi = (\mu, \nu_1, \dots, \nu_M) \in \mathcal{M} \times \mathcal{N}_S^M$ and the total set of φ as $\mathcal{F} = \mathcal{M} \times \mathcal{N}_S^M$. For $\mathbf{X}^{(M)}$, these functions act on the indices of the atomic coordinates in the form of

$$\{\mathbf{P}(\mathbf{X}^{(M)}, \varphi)\}_{jI} = \mathbf{x}_{\nu_I(j)\mu(I)}, \quad (3.17)$$

where $\mathbf{P}(\mathbf{X}^{(M)}, \varphi)$ represents a set of atomic coordinates after the operations of φ , and I and j are indices satisfying $1 \leq I \leq M$ and $1 \leq j \leq n$, respectively.

For the total set of the molecular symmetry correction functions, $\mathcal{F}$, the molecular symmetry corrected RMSD (symRMSD) is defined by

$$d^*\left(\mathbf{X}^{(M)}, \mathbf{X}'^{(M)}\right) = \min_{\varphi \in \mathcal{F}} \left(\mathbf{X}^{(M)}, \mathbf{P}(\mathbf{X}'^{(M)}, \varphi) \right). \quad (3.18)$$

In this setting, the number of the total set of possible permutations is $\#F = M!S^M$, where S is the number of the molecular symmetry operations. Even in the range of $M \sim 10$, this factorial computational cost is too demanding to analyze microseconds conformational changes.

It should be noted that, from definition, the molecular assembly structure and symRMSD form a metric space without the positivity and the triangle inequality; for given $\mathbf{X}^{(M)}$ and d^* , $d^*(\mathbf{X}^{(M)}, \mathbf{X}'^{(M)}) \geq 0$, $d^*(\mathbf{X}^{(M)}, \mathbf{X}^{(M)}) = 0$, and $d^*(\mathbf{X}^{(M)}, \mathbf{X}'^{(M)}) = d^*(\mathbf{X}'^{(M)}, \mathbf{X}^{(M)})$, are satisfied. However, each pair of structures is not identifiable, *i.e.*, it can be

$$d^*(\mathbf{X}^{(M)}, \mathbf{X}'^{(M)}) = 0$$

for

$$\mathbf{X}^{(M)} \neq \mathbf{X}'^{(M)}.$$

For $\mathbf{X}^{(M)}$ and $\mathbf{X}'^{(M)}$ that satisfies

$$d^* \left(\mathbf{X}^{(M)}, \mathbf{X}'^{(M)} \right) = 0,$$

we can conclude an equivalence relation

$$\mathbf{X}^{(M)} \sim \mathbf{X}'^{(M)}$$

in the sense that the structure is physically equivalent. Also, the triangle inequality

$$d^* \left(\mathbf{X}^{(M)}, \mathbf{X}'^{(M)} \right) \geq d^* \left(\mathbf{X}^{(M)}, \mathbf{X}''^{(M)} \right) + d^* \left(\mathbf{X}''^{(M)}, \mathbf{X}'^{(M)} \right)$$

does not always satisfy.

3.3.4 Branch-and-prune method

To compute the symRMSD efficiently, we implement a heuristic branch-and-prune method. In the following, for a natural number m where $m \leq M$, m coordinates of molecules are taken from these of M -molecule assembly, $\{\mathbf{X}^{(m, \theta_m)}\}_{jI} = \mathbf{x}_{j, \theta_m(I)}$, where θ_m is an injective function that maps $\{1, 2, \dots, m\}$ to $\{1, 2, \dots, M\}$, I and j are indices satisfying $1 \leq I \leq m$ and $1 \leq j \leq n$, respectively. For $d^* \left(\mathbf{X}^{(M)}, \mathbf{X}'^{(M)} \right)$, hereafter, $\mathbf{X}^{(M)}$ and $\mathbf{X}'^{(M)}$ are referred to as the reference and the target structures, respectively.

First, two molecules are chosen from each of the reference M molecular assembly and the target one, and symRMSDs $d^* \left(\mathbf{X}^{(2, \sigma)}, \mathbf{X}'^{(2, \sigma')} \right)$, for the possible combinations of the two molecules, σ and σ' , are calculated. Let $N_{\text{cand}}^{(2)}$ be the number of the combinations, *i.e.*, $N_{\text{cand}}^{(2)} = \binom{M}{2}^2$. Those with smaller symRMSD values are then taken as the root nodes. Here, the number of the nodes $N_{\text{rec}}^{(2)}$ is determined by the ceiling function, $\text{ceil}(p_{\text{rec}} N_{\text{cand}}^{(2)})$, where p_{rec} is a cut-off parameter, $0 < p_{\text{rec}} \leq 1$. Next, one molecule not yet selected from each molecular assembly is taken and added to the reference and target structures of the root nodes. For the newly selected molecule, the molecular symmetry correction is only performed. This step generates $N_{\text{rec}}^{(2)}(M-2)^2 S$ candidates, the number of which is $N_{\text{cand}}^{(3)}$. The RMSD values of these candidate structures are calculated, and again the candidates are narrowed down in the ratio of p_{rec} into the child nodes. This process is repeated until m reaches M .

This algorithm omits inefficient assignments between molecular assemblies of which small portions already exhibit a large RMSD, effectively reducing a computational cost of searching the total space of the substitution operations. A shortcoming, however, arises if p_{rec} is too small and in turn the number of the branches records is too small, the true solution may be cut out to the space of the prune branches.

Furthermore, by straightforwardly extending Eqs. 3.17 and 3.18 to multicomponent systems and modifying the algorithm, symRMSD can also deal with host-guest systems, hetero

micelles, and systems containing solvent molecules. However, such host-guest systems often contain a larger number of molecules than can be directly handled by the present algorithm. An exact solution algorithm for the generalized symRMSD using the branch-and-bound method is presented in Chapter 4.

The specific procedures are given in the branch-and-prune algorithm as follows: Let $\mathbf{x}_{Ij} \in \mathbb{R}^{d \times 1}$ be the j -th atomic coordinate of the I -th molecule, where d is the spatial dimension. For given molecular assembly coordinates $\mathbf{X}$ and $\mathbf{X}'$, minimizing Eq.(17) is equivalent to minimizing the following squared deviation (SD) with respect to $\mathbf{R}, \mu, \{\nu_1, \nu_2, \dots, \nu_M\}$:

$$\text{SD}(\mathbf{R}, \mu, \{\nu_I\}) = \sum_{I=1}^M \sum_{j=1}^n \|\mathbf{x}_{Ij} - \mathbf{R}\mathbf{x}'_{\mu(I)\nu_I(j)}\|^2 \quad (3.19)$$

Where, $\mathbf{x}_{jI}$ and $\mathbf{x}'_{jI}$ are atomic coordinates corresponding to $\mathbf{X}$ and $\mathbf{X}'$, respectively. We denote I -th molecular coordinates as a $d \times n$ matrix $\mathbf{X}_I \in \mathbb{R}^{d \times n}$

$$\{\mathbf{X}_I\}_j = \mathbf{x}_{jI}. \quad (3.20)$$

Using relation of $\mathbf{R}\mathbf{R}^\top = \mathbf{I}$, SD can be algebraically transformed as

$$\text{SD}(\mathbf{R}, \mu, \sigma) = \sum_{I=1}^M \text{tr}[\mathbf{X}_I^\top \mathbf{X}_I] + \sum_{I=1}^M \text{tr}[\mathbf{X}_I'^\top \mathbf{X}_I'] - 2 \sum_{I=1}^M \text{tr}[\mathbf{X}_I^\top \mathbf{R}\mathbf{X}'_{\mu(I)} \mathbf{P}_{\sigma(I)}], \quad (3.21)$$

here, $\sigma(I)$ is a mapping with $\{1, 2, \dots, S\}$ in the codomain and S is the number of molecular symmetries. $\mathbf{P}_s \in \mathbb{R}^{n \times n}$ is a permutation matrix corresponding to a molecular symmetry, defined by $\{\mathbf{P}_s\}_{jj'} = \delta_{j\nu_s(j')}$, here δ_{ij} is the dirac delta. The index s corresponds to the molecular specific symmetry, respectively. For example, in the case of C_{2v} symmetry molecule, the permutations of the atomic sequences corresponding to the symmetry groups E, C_2 , σ , and σ' , respectively, are encoded by $s = 1, 2, 3$, and 4. Since all terms except the third term on the right-hand side are constant terms, this problem can be considered as maximizing the third term on the right-hand side. Transforming the third term of Eq. 3.21, we define the partial evaluation equation as

$$L_k(\mathbf{m}, \mathbf{m}', \mathbf{s}) = \max_{\mathbf{R} \in \mathcal{R}} \sum_{I=1}^k \text{tr}[\mathbf{X}_{m_I}^\top \mathbf{R}\mathbf{X}'_{m'_I} \mathbf{P}_{s_I}]. \quad (3.22)$$

here, k is a natural number satisfies $1 \leq k \leq M$, $\mathbf{m} = \{m_1, m_2, \dots, m_k\}$, $\mathbf{m}' = \{m'_1, m'_2, \dots, m'_k\}$ with $1 \leq m_I, m'_I \leq M$ $\mathbf{s} = \{s_1, s_2, \dots, s_k\}$ with $1 \leq s_I \leq S$, and $\text{SO}(d)$ denotes the special orthogonal group of order n , respectively. Let the constant term as

$$G = \sum_{I=1}^M \text{tr}[\mathbf{X}_I^\top \mathbf{X}_I] + \sum_{I=1}^M \text{tr}[\mathbf{X}_I'^\top \mathbf{X}_I']. \quad (3.23)$$

If $\mathbf{m}$ and $\mathbf{m}'$ are permutations on $\{1, 2, \dots, M\}$, then $G - 2L_M(\mathbf{m}, \mathbf{m}', \mathbf{s})$ matches SD of some candidate atom matching. Thus, the branch-and-cut (BC) algorithm is implemented as an algorithm that efficiently enumerates prospective L_M and calculates SD using their maximum element as the heuristic solution.

The BC algorithm starts with $k = 2$, and the exhaustively enumerated L_k is stored in a priority queue, Q_k , and kept as candidates with a percentage of hyperparameter p_{rec} . After the enumeration is completed, a possible index is added to the end of the candidates in the queue, and the enumerated L_{k+1} are stored in the Q_{k+1} . The pseudo code of the BC algorithm is summarized in **Algorithm 1**.

Algorithm 1 Branch and cut

Require: $\{\mathbf{X}_I\}, \{\mathbf{Y}_I\}, \mathbf{X}_I, \mathbf{Y}_I \in \mathbb{R}^{d \times n}, 1 \leq I \leq M$

Require: $\{\mathbf{P}_s\}, \mathbf{P}_s \in \mathbb{R}^{n \times n}, 1 \leq s \leq S$

Require: $p_{\text{rec}} \in (0, 1]$

```

1: Let  $Q_2$  be a priority queue for  $L_2$   $\triangleright L_k$  is given in Eq. 3.22.
2: for  $m \in \{1, \dots, M\}$  choose 2 do
3:   for  $m' \in \{1, \dots, M\}$  choose 2 do
4:     for  $s \in \{1, \dots, S\} \times \{1, \dots, S\}$  do
5:       Add  $\{m, m', s\}$  to  $Q_2$ 
6:     end for
7:   end for
8: end for
9: for  $k \in \{2, \dots, M - 1\}$  do
10:   Let  $Q_{k+1}$  be a priority queue for  $L_{k+1}$ 
11:   for  $m_k, m'_k, s_k \in$  The top  $p_{\text{rec}}$  percentage elements of  $Q_k$  do
12:     for  $m_{k+1} \in \{1, \dots, M\} \setminus \{m_k\}$  do  $\triangleright$  Let  $A \setminus B$  be the set difference.
13:       for  $m'_{k+1} \in \{1, \dots, M\} \setminus \{m'_k\}$  do
14:         for  $s_{k+1} \in \{1, \dots, S\}$  do
15:           Add  $\{(n_k, n_{k+1}), (n'_k, n'_{k+1}), (s_k, s_{k+1})\}$  to  $Q_{k+1}$ 
16:         end for
17:       end for
18:     end for
19:   end for
20: end for
21:  $L_M^* \leftarrow$  The  $L_M$  value of top elements of  $Q_M$ 
22:  $d^* \leftarrow \sqrt{\frac{1}{Mn}(G - 2L_M^*)}$   $\triangleright G$  is given in Eq. 3.23.
23: Return  $d^*$ 

```

As a benchmark, we assessed the accuracy of this algorithm, the computational costs, and the actual computational times versus the cut-off parameter, p_{rec} . The accuracy was evaluated by the average error between the symRMSDs of the full search and the branch-and-prune one given by

$$E = \frac{1}{N_{\text{test}}} \sum_{I=1}^{N_{\text{test}}} \left(\tilde{d}^*(\mathbf{X}_I, \mathbf{X}'_I) - \tilde{d}(\mathbf{X}_I, \mathbf{X}'_I) \right). \quad (3.24)$$

where N_{test} is the number of test samples, $\mathbf{X}_l$ and $\mathbf{X}'_l$ are the l -th samples, $\tilde{d}^*$ is the sym-RMSD obtained by the branch-and-prune search, and d^* is that obtained by the full search. Since the accuracy for the smaller symRMSD is more important for applications, the mean

errors for the 50% and 10% samples with the smallest d^* , E_{50} and E_{10} , respectively, were also calculated. The computational cost is defined by the number of RMSD evaluations in the algorithm.

We randomly selected 20 snapshots from the REMD trajectories for the reference structures, $\mathbf{X}_l$. For the target structures, $\mathbf{X}'_l$, 50 snapshots were randomly taken for $M=4$ (*i.e.*, $N_{\text{test}}=20 \times 50=1,000$), 25 snapshots for $M=5$ ($N_{\text{test}}=500$), and 10 snapshots for $M=6$ ($N_{\text{test}}=200$), respectively. The actual computation time is defined by the average of wall-clock times each of which was required to evaluate a symRMSD for a single pair of structures on a single cpu (Intel(R) Xeon(R) CPU E5-2680 v2 @ 2.80 GHz) per core.

Figure 3.6 and **Table 3.4** summarizes the accuracy and the computational efficiency of the branch-and-prune methods. For all M , symRMSDs evaluated with the branch-and-prune method at $p_{\text{rec}}=0.05$ and above were in good agreement with the exact ones. Even at $p_{\text{rec}}=0.01$, E was about $0.05 \text{ \AA}$ and remarkably small. Furthermore, E_{50} and E_{10} are smaller than E , which is a favorable property of the branch-and-prune method, as errors are more reduced for more important structures. The computational costs are heavily dependent on p_{rec} . In the following analysis, we employed $p_{\text{rec}}=1$ for $M=2$ and 3, $p_{\text{rec}}=0.05$ for $M=4$, and $p_{\text{rec}}=0.01$ for $M=5$ to 7, respectively.

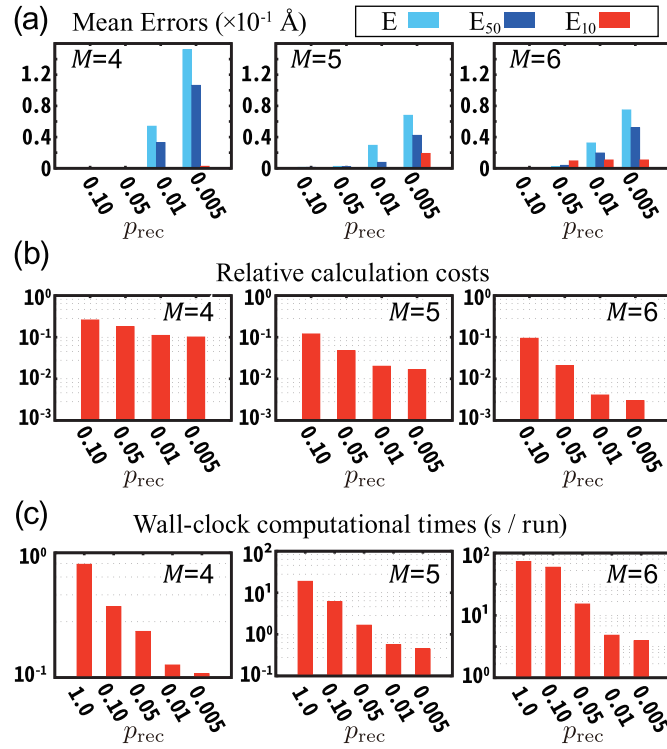


Figure 3.6: Benchmark of the branch-and-prune method. (a) Mean errors for $M=4, 5$, and 6. (b) Relative calculation costs to those of the full search. (c) Actual wall-clock computational times.

Table 3.4: Benchmark results for branch-and-cut algorithm.

M	p_{rec}	Cost	Time (s/run)	E (Å)	E_{50} (Å)	E_{10} (Å)
4	(full search)	(1.00)	0.81	-	-	-
	0.1	0.26	0.37	5.0×10^{-4}	0.00	0.00
	0.05	0.18	0.23	7.0×10^{-4}	0.00	0.00
	0.01	0.11	0.12	5.4×10^{-2}	3.3×10^{-2}	0.00
	0.005	0.10	0.10	1.5×10^{-1}	1.1×10^{-1}	1.0×10^{-1}
5	(full search)	(1.00)	19.0	-	-	-
	0.1	0.12	6.2	1.4×10^{-3}	1.4×10^{-3}	0.00
	0.05	0.048	1.7	2.5×10^{-3}	2.0×10^{-3}	0.00
	0.01	0.020	0.56	3.0×10^{-2}	7.7×10^{-2}	0.00
	0.005	0.017	0.44	6.8×10^{-2}	4.2×10^{-1}	1.9×10^{-2}
6	(full search)	(1.00)	538	-	-	-
	0.1	0.062	348	3.2×10^{-5}	0.0	0.00
	0.05	0.013	23	4.4×10^{-3}	8.0×10^{-3}	3.8×10^{-2}
	0.01	0.0027	2.3	3.0×10^{-2}	2.7×10^{-2}	4.3×10^{-2}
	0.005	0.0019	1.6	4.4×10^{-2}	3.4×10^{-2}	4.3×10^{-2}

3.3.5 Similarity matrix

A distance matrix between two conformations of VPAA assemblies defined with symRMSD provides a useful means to characterize their internal dynamics. For given two sets of configurational snapshots,

$$\mathcal{X} = \{X_1, X_2, \dots, X_{N_1}\}$$

and

$$\mathcal{X}' = \{X'_1, X'_2, \dots, X'_{N_2}\},$$

where N_1 and N_2 are the numbers of snapshots. The similarity matrix is defined by

$$\{\mathbf{D}^*(\mathcal{X}, \mathcal{X}')\}_{ij} = d^*(\mathbf{X}_i, \mathbf{X}'_j) \quad (3.25)$$

where i and j are indices satisfying $1 \leq i \leq N_1$ and $1 \leq j \leq N_2$, respectively. For simplicity of notation, the self-similarity matrix is described as

$$\mathbf{D}^*(\mathcal{X}) := \mathbf{D}^*(\mathcal{X}, \mathcal{X}) \quad (3.26)$$

From the definition, the self-similarity matrix, $\mathbf{D}^*(\mathcal{X})$, is a non-negative real symmetric matrix of which all the diagonal elements are zero.

Time evolution of conformational changes of the VPAA assemblies of $M=2, 3$, and 4 are visualized by the self-similarity matrices of the cMD trajectories as shown in **Figure 3.7**. Multiple rectangular blue regions appear in the heatmaps, indicating that symRMSDs are

kept small for a certain time. The square blue regions around the diagonal therefore represent formations of metastable conformations which are transiently maintained, while the off-diagonal rectangular blue regions express returns to the metastable conformations after transitions to other conformations.

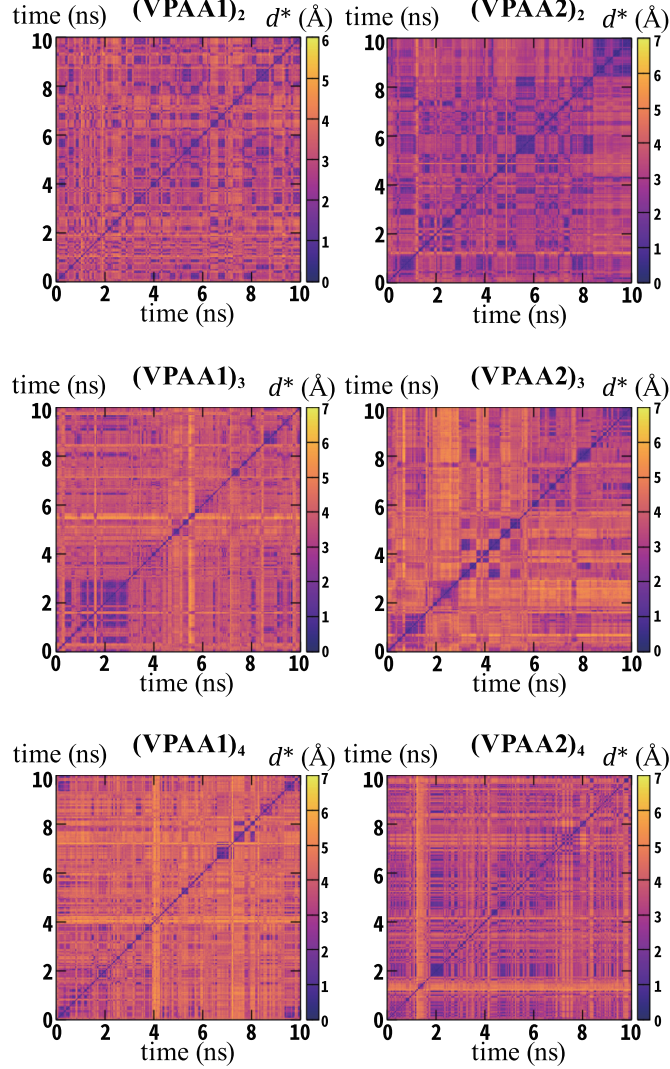


Figure 3.7: Self-similarity matrices for $M=2, 3$, and 4 , $(\text{VPAA1})_M$ and $(\text{VPAA2})_M$, respectively.

For $M=2$, off-diagonal rectangular regions of $(\text{VPAA1})_2$ and $(\text{VPAA2})_2$ appear repeatedly in **Figure 3.7**, indicating that transitions between only a few metastable conformers. Lifetimes of the blue rectangular regions for $(\text{VPAA1})_2$ are sub-nanoseconds, while those for $(\text{VPAA2})_2$ are often several times longer. This feature is the opposite to that of the mean lifetimes of the molecular assemblies themselves, which was about two times longer for $(\text{VPAA1})_2$ than for $(\text{VPAA2})_2$ (**Figure 3.5** and **Table 3.3**). This indicates that the molecular nature of the conformational changes within the assemblies is different from that of the collapse of the assemblies; the conformational changes within an assembly are faster in **VPAA1** than **VPAA2**, while the molecular exchanges between different assemblies are faster for **VPAA2** than **VPAA1**.

For $M=3$, the number of the off-diagonal rectangular regions is diminished compared to $M=2$, because the number of conformers of the assembly increases as the number of the molecules in the assembly increases. This behavior is more pronounced for **(VPAA1)₄**. As contrary, **(VPAA2)₄** shows an extended off-diagonal rectangular blue region, indicating that a stable dominant conformer exists.

3.3.6 Distribution of the symRMSD

As is seen from the self-similarity matrices, the VPAA assemblies undergo intermittent conformational changes on a relatively fast time scale compared to their average lifetimes. The equilibrium distribution of the symRMSD was obtained by computing the self-similarity matrices of the REMD trajectory. The number of structures used, N_{data} , are listed in **Table 3.5**. Up to $M=4$, all the data of the snapshots recorded every 100 ps were used. From $M=5$ to 7, the snapshots were randomly resampled until completable.

Table 3.5: The number of sample snapshots, N_{data} .

M	2	3	4	5	6	7
VPAA1	5792	9053	11608	11608	3073	2580
VPAA2	10880	10283	10981	10113	3012	2565

Distributions of the off-diagonal components of the self-similarity matrices of the symRMSD are shown in **Figure 3.8**. The distributions around a small symRMSD (1-2 Å) in **(VPAA1)₂** and **(VPAA2)₂** (**Figure 3.8**) are attributed to fragments of back-and-forth transitions between metastable conformers seen in the self-similarity matrices (**Figure 3.7**). Similar but smaller shoulders are present for $M=3$. The same behavior continues above $M=4$ for the **VPAA1** assemblies, while a prominent peak appears below 2 Å in **(VPAA2)₄**, which likely represents the long-lived conformers observed in the self-similarity matrix described above (**Figure 3.7**).

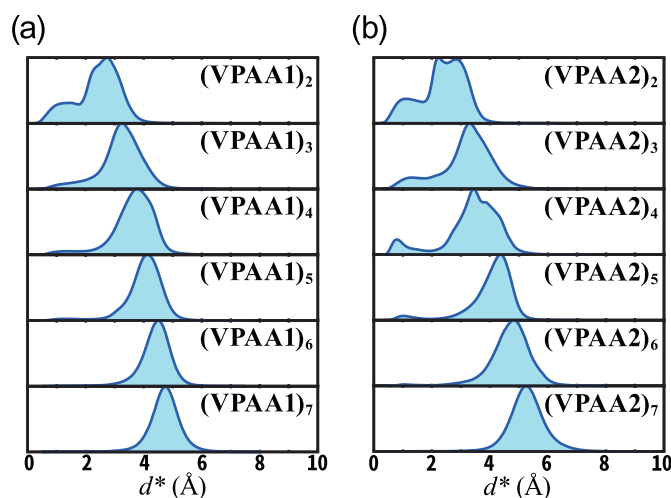


Figure 3.8: Equilibrium distributions of the symRMSDs for (a) **VPAA1** and (b) **VPAA2**.

3.3.7 Structural clustering

Data clustering by distance matrices is widely used for interpretation of molecular dynamics trajectories. [18, 19] We performed a structural clustering by HDBSCAN, [121, 134] a density-based hierarchical clustering method which is robust to noisy data, with the metric of the symRMSD. HDBSCAN uses the minimum spanning tree to create a hierarchical tree, and then the tree is compressed by minimizing an objective function by varying a threshold $\lambda_{\min}$ for noise determination. [121] Here, $\lambda_{\min}$ is the symRMSD between the shortest data points of one group and another or noise.

Using a distance matrix computed for data point set

$$\mathcal{X} = \{X_1, X_2, \dots, X_{N_{\text{data}}}\},$$

HDBSCAN assigns a group label $k = 0, 1, 2, \dots, K$ to each data point, where K is the number of groups automatically determined. $k=0$ is assigned to the noise data label. Note that HDBSCAN is based on the relative roughness or density on static data points. Molecular structures that appear at high frequency in equilibrium are therefore interpreted as metastable structures. Similarly, data assigned as noises are interpreted as an ensemble of unstable structures, which appear infrequently in equilibrium. As we are interested in physically different structures, the mirror symmetry is also implemented in the distance matrix. The self-similarity matrix is modified as

$$\{\mathbf{D}^{**}\}_{ij} := \min(d^*(\mathbf{X}_i, \mathbf{X}_j), d^*(\mathbf{X}_i, \bar{\mathbf{X}}_j)) \quad (3.27)$$

where $\bar{\mathbf{X}}_i$ is the mirror symmetry structure of $\mathbf{X}_i$. Also, HDBSCAN uses two hyperparameters; *i.e.*, the minimum cluster size $C_{\min}$ and the minimum sample size $S_{\min}$. We used $C_{\min} = S_{\min} = \text{floor}(N_{\text{data}}/100)$, where N_{data} is the number of data and $\text{floor}(\cdot)$ is the floor function, allowing for clear interpretation. The Excess of Mass algorithm was employed to detect the most persistent groups. The analysis above was performed by using software implemented by McInnes, Healy, and Astels. [134] For $M=2$ to 4, structural classifications were performed using all structures obtained from the REMD trajectories. For $M=5$ to 7, the N_{data} structures were randomly resampled from the REMD trajectories. The number of groups, K , and the non-noise ratio, $1 - \alpha_0$, where α_0 is the noise ratio are summarized in **Table 3.6**. Details of the calculation results including the fractions of the numbers of the samples in the groups are summarized in **Tables 3.7 to 3.12**.

Table 3.6: Results of HDBSCAN. K is the number of groups, and α_0 is the noise ratio.

	M	2	3	4	5	6	7
VPAA1	K	4	6	4	8	4	4
	$1 - \alpha_0$	0.824	0.474	0.457	0.515	0.289	0.319
VPAA2	K	4	4	6	6	4	6
	$1 - \alpha_0$	0.834	0.637	0.727	0.511	0.329	0.325

Table 3.7: HDBSCAN results for $M=2$.

		k				
		0	1	2	3	4
(VPAA1)₂	N_k	1020	605	2290	280	1597
	α_k	0.176	0.104	0.395	0.048	0.276
	F_k (Å)	-	0.830	1.371	0.946	1.133
	d_k (Å)	-	1.42	2.92	1.91	1.95
(VPAA2)₂	N_k	1805	1946	5111	289	1729
	α_k	0.166	0.179	0.470	0.027	0.159
	F_k (Å)	-	0.801	1.381	0.846	1.055
	d_k (Å)	-	1.21	2.60	1.38	2.25

Table 3.8: HDBSCAN results for $M=3$.

		k						
		0	1	2	3	4	5	6
(VPAA1)₃	N_k	5211	817	289	2176	329	352	329
	α_k	0.526	0.090	0.032	0.240	0.036	0.039	0.036
	F_k (Å)	-	1.119	1.024	1.461	1.688	1.051	1.304
	d_k (Å)	-	3.194	2.263	3.379	2.110	3.206	2.026
(VPAA2)₃	N_k	3729	463	3589	1946	556		
	α_k	0.363	0.045	0.349	0.189	0.054		
	F_k (Å)	-	1.098	1.386	1.538	1.530		
	d_k (Å)	-	3.173	3.090	3.031	3.657		

Table 3.9: HDBSCAN results for $M=4$.

		k						
		0	1	2	3	4	5	6
(VPAA1)₄	N_k	6300	1143	2106	1151	908		
	α_k	0.543	0.098	0.181	0.099	0.078		
	F_k (Å)	-	0.992	1.642	2.105	1.764		
	d_k (Å)	-	3.647	3.618	1.954	2.193		
(VPAA2)₄	N_k	2993	622	1864	1029	3146	592	735
	α_k	0.273	0.057	0.170	0.094	0.286	0.054	0.067
	F_k (Å)	-	1.070	1.757	1.002	1.036	2.390	1.233
	d_k (Å)	-	3.386	3.575	1.365	0.672	3.348	3.456

Table 3.10: HDBSCAN results for $M=5$.

		k								
		0	1	2	3	4	5	6	7	8
(VPAA1)₅	N_k	5481	796	503	470	263	1501	536	939	81
	α_k	0.485	0.070	0.045	0.042	0.023	0.133	0.047	0.083	0.072
	F_k (Å)	-	1.674	2.671	1.637	1.554	1.486	1.097	3.016	2.543
	d_k (Å)	-	3.814	4.411	2.867	4.296	3.472	4.146	3.524	4.055
(VPAA2)₅	N_k	4946	2241	363	642	441	656	824		
	α_k	0.489	0.222	0.036	0.063	0.044	0.065	0.081		
	F_k (Å)	-	1.572	1.723	2.673	2.223	2.261	2.119		
	d_k (Å)	-	2.559	5.095	4.266	3.876	4.126	4.230		

Table 3.11: HDBSCAN results for $M=6$.

		k				
		0	1	2	3	4
(VPAA1)₆	N_k	2185	242	178	234	234
	α_k	0.711	0.079	0.058	0.076	0.076
	F_k (Å)	-	2.793	1.867	2.917	1.946
	d_k (Å)	-	3.567	2.861	4.452	1.933
(VPAA2)₆	N_k	2021	164	326	298	203
	α_k	0.671	0.054	0.108	0.099	0.067
	F_k (Å)	-	1.346	2.808	1.723	2.829
	d_k (Å)	-	4.183	3.078	1.255	4.136

Table 3.12: HDBSCAN results for $M=7$.

		k						
		0	1	2	3	4	5	6
(VPAA1)₇	N_k	1758	104	220	373	125		
	α_k	0.681	0.040	0.085	0.145	0.048		
	F_k (Å)	-	1.824	2.596	3.148	2.037		
	d_k (Å)	-	4.425	4.806	4.520	4.091		
(VPAA2)₇	N_k	1731	98	139	245	115	97	140
	α_k	0.675	0.038	0.054	0.096	0.045	0.038	0.055
	F_k (Å)	-	3.792	3.342	3.182	4.184	3.573	3.009
	d_k (Å)	-	4.925	4.167	3.870	5.728	5.115	5.264

3.3.8 Representative structure

Let

$$\mathcal{X}_k = \left\{ \mathbf{X}_1^{(k)}, \mathbf{X}_2^{(k)}, \dots, \mathbf{X}_{N_k}^{(k)} \right\}$$

be the set of structures belonging to the group k , and N_k be the number of structures in the $\mathbf{X}_k$. Here,

$$\alpha_k = \frac{N_k}{\sum_{l=0}^K N_l}$$

indicates the abundance ratio. The group fluctuation is also defined as

$$F_k(\mathbf{X}) = \sqrt{\frac{1}{N_k} \sum_{l=1}^{N_k} \left(d^{**} \left(\mathbf{X}, \mathbf{X}_l^{(k)} \right) \right)^2}. \quad (3.28)$$

The structure with the minimum fluctuation among the data of the group k ,

$$\mathbf{X}^{(k)*} = \arg \min_{\mathbf{X}_l^{(k)} \in \mathcal{X}_k} F_k \left(\mathbf{X}_l^{(k)} \right), \quad (3.29)$$

is defined as the representative structure of the group k . The group fluctuation is then given by

$$F_k^* = F_k \left(\mathbf{X}^{(k)*} \right). \quad (3.30)$$

If the distribution of data points belonging to the group k is isotropic, the representative structure, $\mathbf{X}^{(k)*}$, defined by Eq. 3.29 approximates the minima of a free energy basin. Note that this approximation is justified if the free energy basin is sufficiently deep. However, when the basin is shallow or multiple basins are fused, the representative structure is no longer a local minimum conformer. The group fluctuation F_k^* is a rough indication of whether the group contains only a single basin and whether this approximation is justified or not, *i.e.*, for groups $F_k^* \sim 1$ Å or less, it is safe to assume that $\mathbf{X}^{(k)*}$ is an approximately locally stable conformer. In the following, we denote a representative structure, $\mathbf{X}^{(k)*}$, and the mirror symmetry structure, $\bar{\mathbf{X}}^{(k)*}$, for M molecular assembly of VPAA as $(\text{VPAA})_{M-k}$ and $(\text{VPAA})_{M-k'}$, respectively.

The self-similarity matrix of a set of the representative structures and their mirror symmetric structures is utilized to characterize the metastable structures. We define the reduced self-similarity matrix as

$$\mathbf{D}_{\text{self}} = \mathbf{D}^* \left(\mathcal{X} = \left\{ \mathbf{X}^{(1)*}, \mathbf{X}^{(2)*}, \dots, \mathbf{X}^{(K)*}, \bar{\mathbf{X}}^{(1)*}, \bar{\mathbf{X}}^{(2)*}, \dots, \bar{\mathbf{X}}^{(K)*} \right\} \right). \quad (3.31)$$

In particular, a symRMSD between a representative structure and its mirror symmetry structure,

$$d_k = d^* \left(\mathbf{X}^{(k)*}, \bar{\mathbf{X}}^{(k)*} \right), \quad (3.32)$$

is an indicator of the magnitude of symmetry of the structure with respect to the mirror or the rotatory reflection. Furthermore, to compare the similarity between the representative structures of **VPAA1** and **VPAA2**, the cross-similarity matrix is defined by

$$\mathbf{D}_{12} = \mathbf{D}^* \left(\mathcal{X}_1 = \left\{ \mathbf{X}^{(1)*}, \dots, \mathbf{X}^{(K)*}, \bar{\mathbf{X}}^{(1)*}, \dots, \bar{\mathbf{X}}^{(K)*} \right\}, \mathcal{X}_2 = \left\{ \mathbf{Y}^{(1)*}, \dots, \mathbf{Y}^{(K)*}, \right\} \right), \quad (3.33)$$

where $\mathbf{X}^{(k)*}$ and $\mathbf{Y}^{(k)*}$ are the representative structures for **VPAA1**, and **VPAA2**, respectively.

3.3.9 Assemblies of two VPAAAs

N_k , a_k , F_k^* , and d_k for $M=2$ are summarized in **Table 3.7**. For both $(\text{VPAA1})_2$ and $(\text{VPAA2})_2$, $K=4$ was obtained (**Tables 3.6 and 3.7**). The representative structures and the cross-similarity matrix, $\mathbf{D}_{12}$, are shown in **Figure 3.9**. Diagonal elements of the cross-similarity matrix between $(\text{VPAA1})_2$ and $(\text{VPAA2})_2$ (*i.e.*, the left half of the matrix shown in **Figure 3.9e**), indicating that the representative structures of $(\text{VPAA1})_2$ and $(\text{VPAA2})_2$ are similar. Except for $k=2$, F_k^* of all the groups are small (~ 1 Å) (**Table 3.7**) and thus each of the groups is interpreted as an ensemble of structures in a single free energy basin. Stable hydrophobic interactions of V-shaped anthracene panels characterizing the representative structures, rather than interactions of the side chains, are therefore the dominant factor determining the free energy surfaces of the conformers.

The reduced self-similarity matrix, $\mathbf{D}_{\text{self}}$, for $M=2$ are visualized by nodes in a network, where nodes corresponding to k and k' represent the representative structures, and edges connecting the nodes are associated with d^* (**Figure 3.10**). For simplicity, the connection of d^* less than 2.8 Å are drawn by the edges in **Figure 3.10**. $(\text{VPAA1})_2$ -2 and $(\text{VPAA2})_2$ -2 are the most abundant conformers, with $\alpha_k=0.395$ and 0.470. In these conformers, the two anthracene panels of one VPAA cover one of the two anthracene panels of the other VPAA in an angled direction (**Figure 3.9b**). Larger group fluctuations ($F_k^*=1.371$ Å and 1.381 Å, respectively) than the other groups are also found. Hence, the structures belonging to $k=2$ are more flexible presumably because of the lack of tight interaction between the anthracene panels with a large hydrophobic interaction area, and have an entropic advantage over the other conformers. d_k of $(\text{VPAA1})_2$ -2 and $(\text{VPAA2})_2$ -2 are relatively large, 2.92 Å and 2.60 Å, respectively (**Table 3.7**), indicating that these structures break the mirror symmetry. In contrast, the group fluctuations for $k=1$ and 3 are small ($F_k^*=0.830$ Å and 0.946 Å for $(\text{VPAA1})_2$ -2 and $(\text{VPAA1})_2$ -2, and $F_k^*=0.801$ Å and 0.846 Å for $(\text{VPAA2})_2$ -2 and $(\text{VPAA2})_2$ -2, respectively). In the structures of $k=1$ and 3, the anthracene panels form strong hydrophobic interactions with face-to-face large contact areas (**Figure 3.9a, c and Figure 3.10**), and the flexible side chains do not interfere with the hydrophobic interactions.

The conformers for $(\text{VPAA1})_2$ and $(\text{VPAA2})_2$ are similar, but slight differences appear in their distributions. The abundance ratio of $(\text{VPAA2})_2$ -2, $\alpha_k=0.179$, is larger than that of $(\text{VPAA1})_2$ -2, $\alpha_k=0.104$. In contrast, the ratio of $k=4$ are larger in **VPAA1** ($\alpha_k=0.276$ for $(\text{VPAA1})_2$ -2 and $\alpha_k=0.159$ for $(\text{VPAA2})_2$ -2, respectively). These differences in stability are attributed to the structural differences in the side chains of **VPAA1** and **VPAA2** (**Figure 3.11**). The hydrophilic chains of **VPAA1**, the ethyltrimethylammonium group, are relatively short, and in turn create a space on the edge of the V-shaped anthracene hydrophobic panels on the side-chain sides, whereas the long side chains of **VPAA2**, the sulfobetaine group, fill the space. Thus, the structures of $(\text{VPAA2})_2$ are slightly more unstable than these of $(\text{VPAA1})_2$ for $k=2$ and 4, in which the side chains of one VPAA interfere with hydrophobic interactions with the V-shaped anthracene panel of the other VPAA (**Figure 3.9b and d**). For $k=1$, the side chains do not interfere with the hydrophobic interactions between the anthracene panels as they form direct contacts without the side chains. However, electrostatic repulsions between the monopole positive charges of the side chains exist in $(\text{VPAA1})_2$, leading to destabilization of the dimer formation.

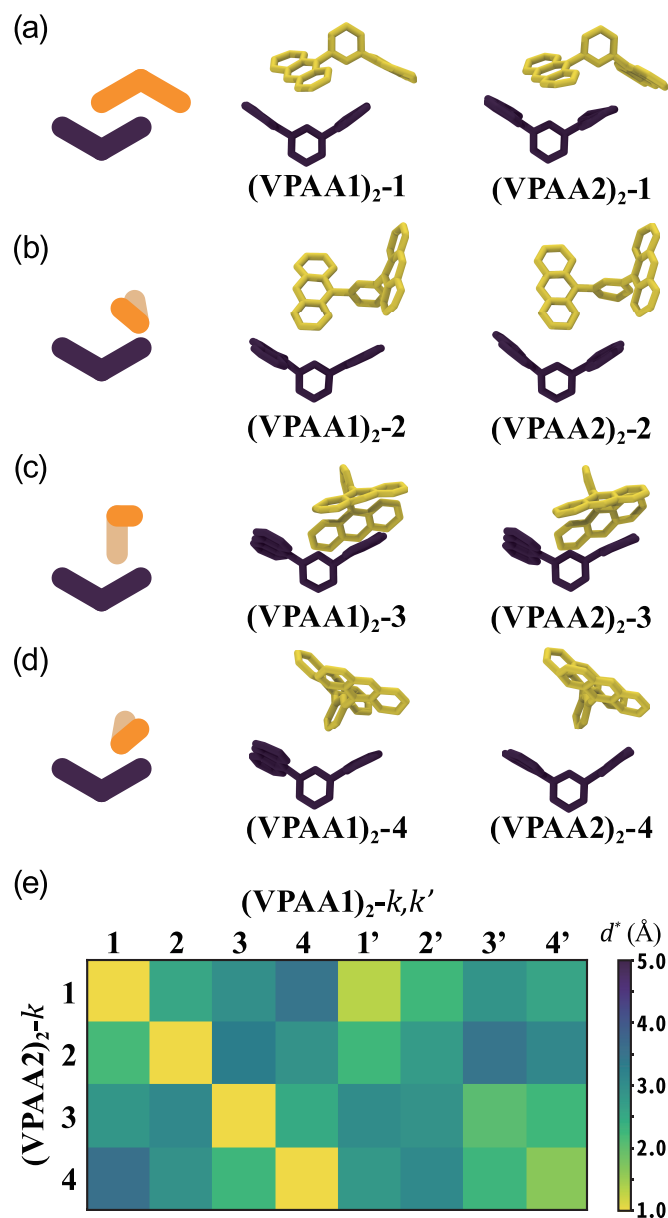


Figure 3.9: Representative structures of the assemblies of $M=2$. (a)-(d) Snapshots of the representative structures of $k=1$ to 4, respectively. Aromatic groups are only depicted, and side chains are omitted for clarity. (e) A cross-similarity matrix for $(VPAA1)_2$ and $(VPAA2)_2$.

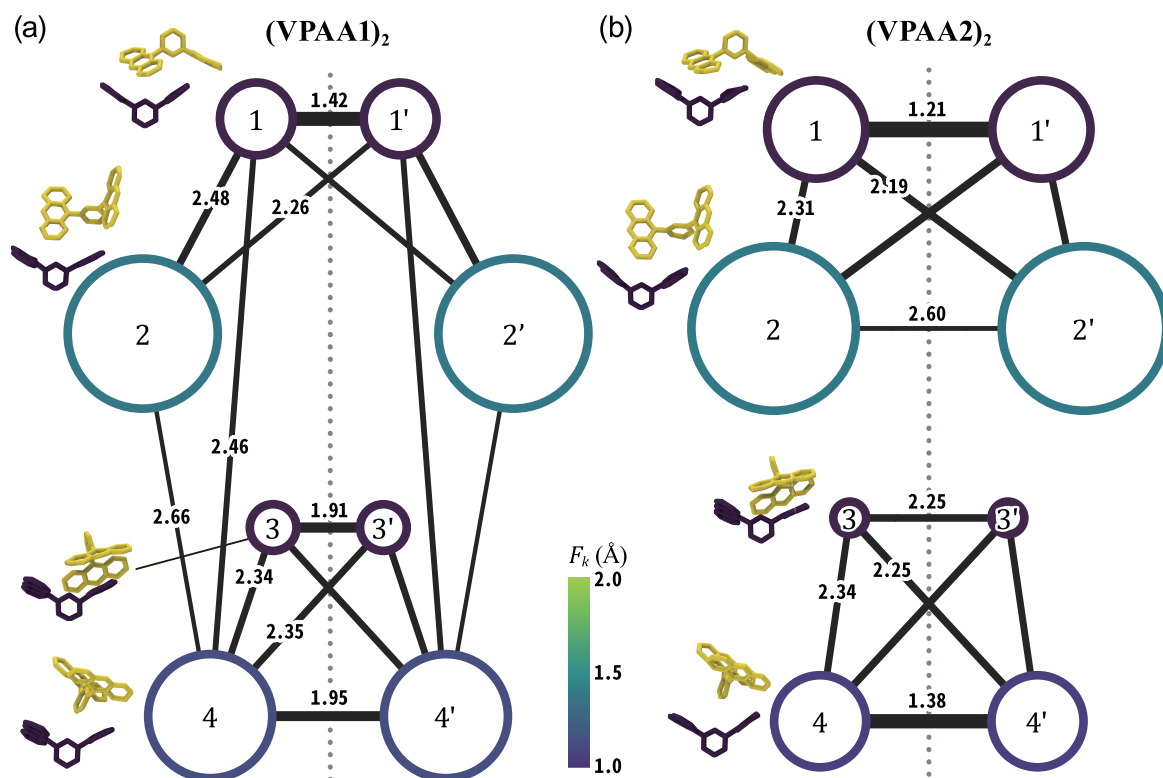


Figure 3.10: Reduced self-similarity matrices visualized with nodes and edges in networks. (a) (VPAA1)₂ and (b) (VPAA2)₂. Node sizes and color maps correspond to α_k and F_k^* , respectively. Edges of d^* less than 2.8 Å are depicted. Numbers in the node indicates the group indices k and k' assigned to the representative structures and their mirror images, respectively.

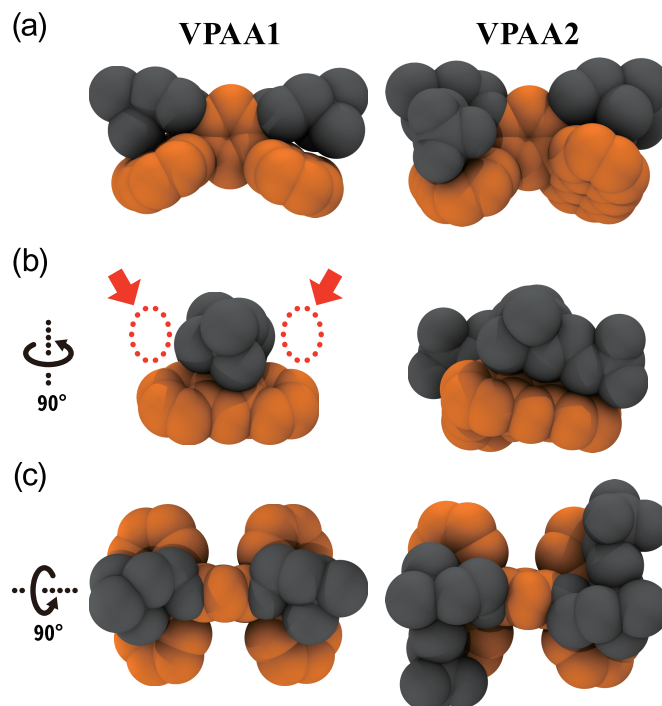


Figure 3.11: Typical MD snapshots of VPAA in aqueous solution. (a), (b) Side views. (c) Top views. Aromatic groups and side chains are colored in orange and gray, respectively.

3.3.10 Assemblies of more than two VPAAAs

The representative structures of which the abundance ratios, α_k , are relatively high among the classified groups for $M=3$ to 7 are shown in **Figure 3.12**. The cross-similarity matrices are also depicted in **Figure 3.13**. Snapshots of all the representative structures are depicted in **Figures 3.16-3.25** in APPENDIX.

For $M=3$, structures of **VPAA1** and **VPAA2** were classified into six and four groups, respectively, by HDBSCAN (**Table 3.6**). Among the classified groups, $(\text{VPAA1})_3\text{-3}$ and $(\text{VPAA2})_3\text{-3}$ exhibit large abundance ratios ($\alpha_k=0.240$ and 0.189 , respectively) (**Figure 3.12a, b**). The structures are similar to each other as seen from the snapshots of the representative structures (**Figure 3.12a, b**) and a small d^* in the cross-similarity matrix (**Figure 3.13a**). The structures are characterized by strong stacking hydrophobic interactions among the anthracene panels, which provide high structural stability.

For **VPAA2**, the group $k=2$, $(\text{VPAA2})_3\text{-3}$, also possesses a large abundance ratio ($\alpha_k=0.349$) (**Figure 3.12b**), whereas an abundance ratio is small for the corresponding group for **VPAA1**, $(\text{VPAA1})_3\text{-3}$ ($\alpha_k=0.032$), despite its structural similarity to $(\text{VPAA2})_3\text{-3}$ (**Figures 3.12a and 3.17**). In the representative structure of $(\text{VPAA2})_3\text{-3}$, hydrophobic interactions between the phenyl groups and the anthracene panels in addition to those between anthracene panels stabilize the structure (**Figure 3.12b**). As the hydrophilic chains are attached to the phenyl group, the mono pole charges of the hydrophilic chains of **VPAA1** more destabilizes the hydrophobic interactions than the zwitterionic ones of **VPAA2**, leading to the smaller abundance ratio of $(\text{VPAA1})_3\text{-3}$ than that of $(\text{VPAA2})_3\text{-3}$.

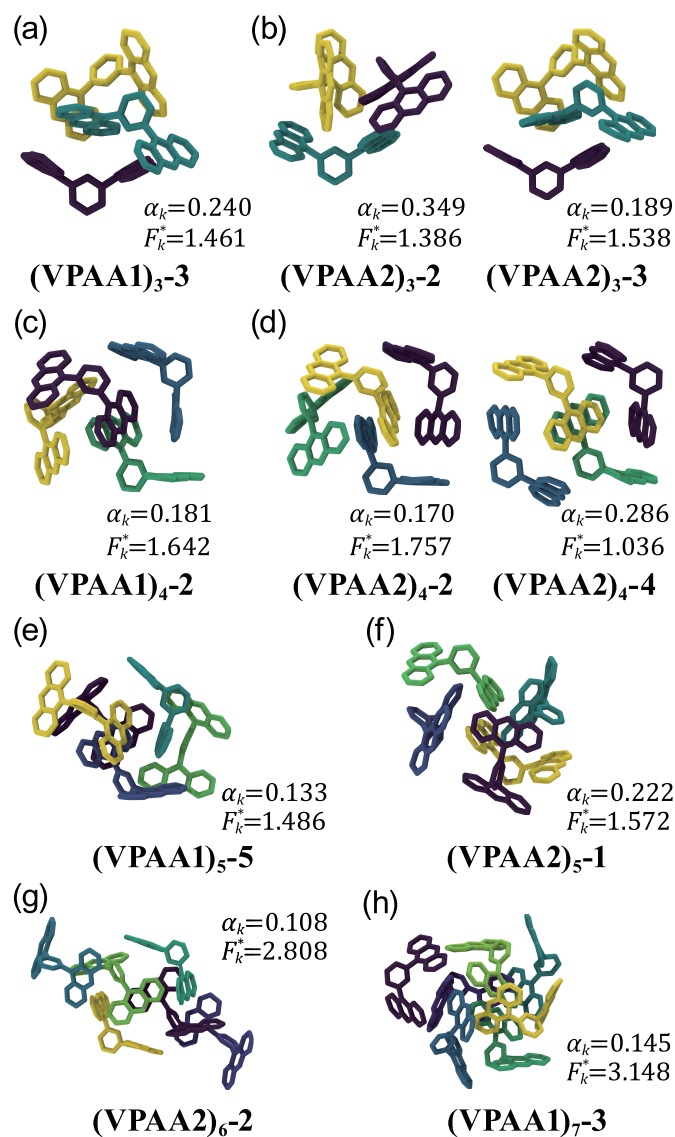


Figure 3.12: Representative structures with abundance ratios of greater than 0.1. **(a)-(h)** Snapshots of the VPAA assemblies for $M=3$ to 7, respectively. Aromatic groups are only depicted and side chains are omitted for clarity.

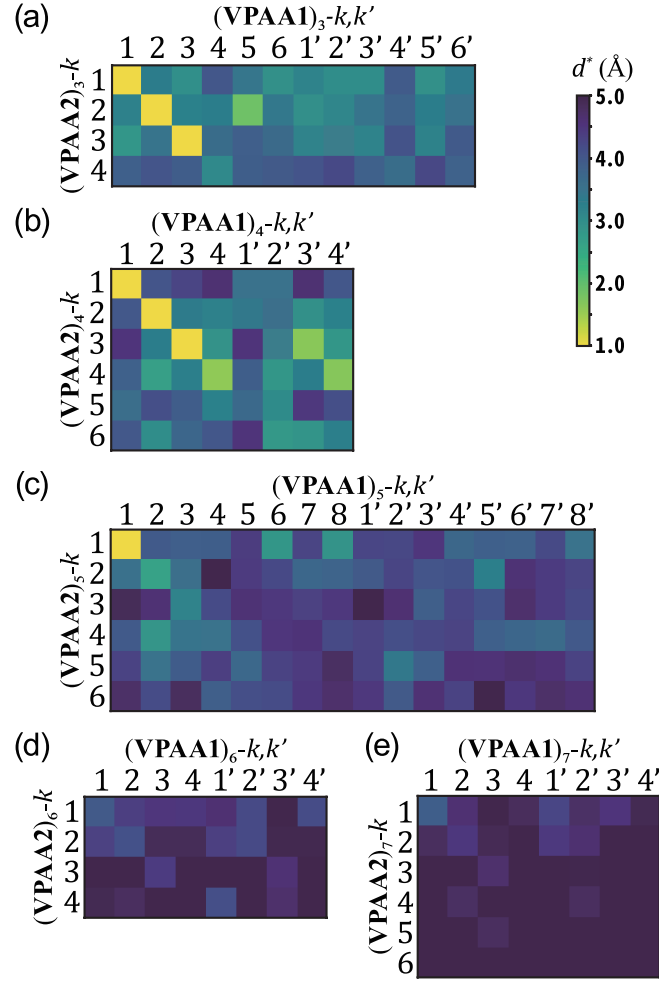


Figure 3.13: Cross-similarity matrices between VPAA1 and VPAA2 assemblies for $M=3$ to 7 ((a)-(e)), respectively.

The smaller abundance ratio of $(\text{VPAA1})_3\text{-3}$ is also partly attributed to the group classification. **VPAA1** was classified into six groups, which is more than four groups of **VPAA2**. The structures of **VPAA1** corresponding to those in the group $k=2$ for **VPAA2** were classified into three groups, $(\text{VPAA1})_3\text{-3}$, $(\text{VPAA1})_3\text{-3}$, and $(\text{VPAA1})_3\text{-3}$, of which structures are relatively similar to one another (**Figures 3.13a** and **3.17**), and thus the abundance ratio of $(\text{VPAA1})_3\text{-3}$ is diminished. However, the abundance ratio of $(\text{VPAA2})_3\text{-3}$ ($\alpha_k=0.349$) is still significantly larger than the sum of the abundance ratios of those three groups of **VPAA1** ($\alpha_k=0.107$) (**Table 3.8**), indicating the higher stability of $(\text{VPAA2})_3\text{-3}$ than **VPAA1**.

For $M=4$, a structurally symmetric conformation of $(\text{VPAA2})_4\text{-4}$ exhibits a high stability ($\alpha_k=0.286$, **Figure 3.12d**). A small symRMSD between the mirror images of $(\text{VPAA2})_4\text{-4}$ is ($d_k=0.672$ Å, **Table 3.9**) is attributed to its high symmetry. A small peak of the equilibrium distribution at $d_k=0.672$ Å shown in (**Figure 3.8b**) comes from $(\text{VPAA2})_4\text{-4}$. As the group fluctuation is relatively small ($F_k^*=1.036$ Å), the symmetric conformation is more rigid and thus the well-organized interactions in the symmetric conformation are responsible for the high stability. Since a symmetric conformation of **VPAA1**, $(\text{VPAA1})_4\text{-4}$, is not stable ($\alpha_k=0.078$, **Table 3.9** and **Figure 3.18**) compared to that of **VPAA2**, the side chains of **VPAA2** contribute to the high stability. As is shown in **Figure 3.14a**, the long side chains of **VPAA2** establish hydrophobic interactions with the anthracene panels.

The assemblies of M larger than 4 show rather diverse structures. A small diagonal element of the cross-similarity matrix appears only for $(\text{VPAA1})_5\text{-5}$ and $(\text{VPAA2})_5\text{-5}$ (**Figure 3.13c**), and no similar structures other than them are found (**Figure 3.13d** and **e**). The feature is attributed to rapid increase of the number of local free energy minima for the assemblies of the larger M . As the multiple free energy basins are more likely to be assigned to a single group, the similarity of the structures is lost. However, even in the diverse conformers, no cavities were formed in the hydrophobic cores to accommodate water molecules. The strong hydrophobic interactions between the anthracene panels that exclude water molecules well explain the high encapsulation ability of the wide range of extended aromatic molecules. [135]

It is noteworthy that the aforementioned symmetric and stable structure, $(\text{VPAA2})_4\text{-4}$ (**Figure 3.12d**), forms a core substructure in several structures of **VPAA2** assemblies of $M \geq 4$ (**Figure 3.14b-f**). The entire structures of $M \geq 4$ including the substructure of $(\text{VPAA2})_4\text{-4}$, however, are not as stable as the substructure of $(\text{VPAA2})_4\text{-4}$ itself. For example, $(\text{VPAA2})_5\text{-5}$ in which one VPAA is added to $(\text{VPAA2})_4\text{-4}$ is not a dominantly stable conformer ($\alpha_k=0.063$ and $F_k^*=2.673$ Å, **Figure 3.14b** and **Table 3.9**) compared to the most stable one, $(\text{VPAA2})_5\text{-5}$ ($\alpha_k=0.222$, $F_k^*=1.572$ Å, **Figures 3.13f** and **3.21**, and **Table 3.10**). In contrast, $(\text{VPAA2})_4\text{-4}$ is included in $(\text{VPAA2})_5\text{-5}$ (**Figure 3.14g**) as a core substructure (**Figure 3.14h**). Given that $(\text{VPAA2})_4\text{-4}$ is not a dominantly stable conformer ($\alpha_k=0.094$, $F_k^*=1.002$ Å, **Table 3.9**) among **VPAA2** assemblies of $M=4$, structural stability of the larger assemblies is determined by consistent interactions among the entire molecules composing the assemblies, and not by formation of a local stable substructure.

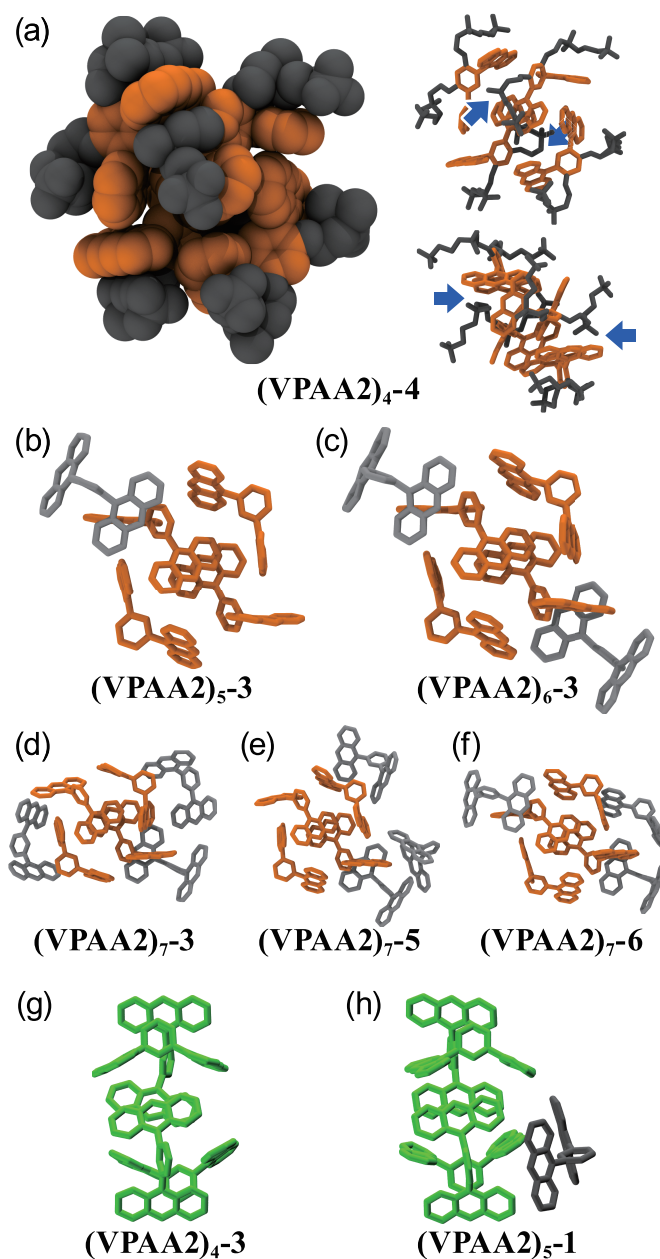


Figure 3.14: Snapshots of $(VPAA2)_4-4$ and representative structures of M_4 including $(VPAA2)_4-4$ and $(VPAA2)_4-4$ as substructures. (a) Left panel: $(VPAA2)_4-4$ depicted in a van der Waals representation. Aromatic moieties and side chains are colored in orange and gray, respectively. Right panels: the same snapshots in a stick representation. Blue arrows indicate key hydrophobic contacts between anthracene panels and side-chains. (b)-(f) $(VPAA2)_5-5$, $(VPAA2)_6-6$, $(VPAA2)_7-7$, $(VPAA2)_7-7$, and $(VPAA2)_7-7$, respectively. Substructures of $(VPAA2)_4-4$ are shown in orange. (g) $(VPAA2)_4-4$. (f) $(VPAA2)_5-5$. Substructures of $(VPAA2)_4-4$ are shown in green.

3.4 DISCUSSION

Understanding of the prominent amphiphilicity of VPAA molecules such as high ability of encapsulation and solubilization for various types of hydrophobic molecules requires elucidation of both of structures and dynamics of their assemblies in water solvent. Strong solvability of extended aromatic molecules and broad ability to encapsulate molecules of different sizes and shapes need to simultaneously establish strong yet flexible hydrophobic interactions which are generally contradictory properties.

The long time MD simulations for tens of microseconds in the present study clearly showed that VPAA molecules spontaneously form assemblies of several VPAA molecules between which VPAA molecules are exchangeable in the sub-nanosecond timescale (**Figure 3.4**). The numbers of VPAA molecules constituting the assemblies are therefore statistically distributed. The numbers of molecules at the maximum distributions are 5-7 molecules for **VPAA1** and 4-6 molecules for **VPAA2**, respectively, which are in line with experimental evidence, [23] and further growth of aggregation was not observed. The observed behavior represents the prominent characteristics of the strong and flexible hydrophobic interactions in the formation of the assemblies and strong amphiphilicity providing high dispersibility of encapsulation.

The sub-nanosecond lifetimes of the assemblies are much longer than the estimates of a simple molecular coaggregation model which only takes collisions of the assemblies into account (**Table 3.3** and **Figure 3.5**), indicating that the multiple VPAA molecules create a specific structure on the potential surface, and their lifetimes are governed by the collapse of the structures rather than collision between the assemblies. A key to understanding of molecular mechanisms underlying the prominent encapsulation and solubilization properties of VPAA molecules is therefore characterization of the intermolecular structures of the assemblies in atomic detail.

We employed RMSD for the metric of the structural characterization, which is practically applicable for macromolecules such as proteins. However, the conventional method to evaluate RMSD cannot directly be applied to the VPAA assemblies because the atom indices of the molecules in the assemblies need to be reassigned in a trajectory, and the atom index assignment requires rapidly growing computational cost proportional to factorial of number of atoms, n , in a straightforward way as mentioned above. Several polynomial-order methods to match the indices that give the minimum RMSDs have been proposed to circumvent the problem including stable matching, [136] Monte Carlo method, [137] and Hungarian algorithms. [137–140] However, these methods use a fixed superposition and are not necessarily effective for the isotopic VPAA assemblies. Only if the comparison is between two structures that are perfectly congruent, the $\mathcal{O}(n \log n)$ algorithm is known to find their optimal matching, and an algorithm with less computational complexity is not known. [141] Iwata, Satoh, and co-workers proposed an optimization algorithm without fixed superposition. [132, 133] However, applying them to molecular assemblies is redundant and ineffective computationally.

We therefore propose the index matching algorithm specialized for molecular assemblies composed of the same molecules in the present study. We introduce symRMSD in which

the index reassignment due to the intermolecular exchange of molecular positions in the assemblies and the intramolecular exchange of symmetry are incorporated. The combinatorial explosion of the reassignment in the evaluation of symRMSD is then circumvented by the branch-and-prune algorithm. The symRMSD method was shown to be computationally very efficient (**Figure 3.6**), making it possible to compute RMSDs of the VPAA assemblies including up to 7 molecules for which the number of combinations of the index reassignment grows to 82,575,360 for a single snapshot conformation. Although the branch-and-prune algorithm is an approximate heuristic method, the errors are sufficiently small especially for smaller RMSDs which are more relevant for the structural characterization. The high efficiency and robustness of the method allow one to characterize the structures of dynamically flexible assemblies in an MD trajectory in which computations of many conformations in the trajectory are required. In the present study, the symRMSD distance matrices along the MD trajectories successfully captured the quasi-stable conformations and the temporal transitions between them (**Figure 3.7**).

To further characterize the quasi-stable conformations in the trajectory in a more microscopic way, the classification of the conformations with the metric of symRMSD was carried out. For discrete data on metric space, many algorithms are known, such as an agglomerative single-linkage clustering, [142] Ward’s method, [142, 143] Bayesian classification, [144] and k -means method. [143] In the present study, we used HDBSCAN, [121] a density-based hierarchical clustering method which is robust to noisy data. An investigation by Shao *et al.* [18] showed that a hierarchical clustering algorithm is effective when the number of clusters is unknown as in the present trajectory data. Also, the hierarchical clustering is convenient for the symRMSD because data points used do not need to be embedded in a metric space.

From the symRMSD distance matrices, the structures of the assemblies in the MD trajectories were classified into several groups of the quasi-stable conformations (**Tables 3.6 and 3.7, Figures 3.9, 3.10, 3.12, and 3.13**). Populations of these quasi-stable conformations are characterized by strong hydrophobic interactions of the V-shaped anthracene panels. Because of the frustration of the hydrophobic interactions of the anthracene panels induced by the V-shaped structure, the several quasi-stable conformations and the transitions among them that furnish the assemblies with high flexibility appear. Furthermore, hydrophobic and electrostatic interactions of the side-chains also modulate the population of the quasi-stable conformations, giving rise to differences in the population of the quasi-stable conformations between **VPAA1** and **VPAA2**. The AMBER force fields are known to reproduce well the local interactions that cause subtle differences in these conformers, such as π - π and CH- π interactions. [145, 146]

The only small difference between **VPAA1** and **VPAA2** up to $M=3$ was seen because the hydrophobic interactions between the V-shaped anthracene panels dominantly govern the stability of VPAA assemblies. However, the non-noise ratio, $1 - \alpha_0$, of **VPAA1** decreases up to $M=4$ (0.824, 0.474, and 0.457 for $M=2, 3$, and 4, respectively, **Table 3.7**), while that of **VPAA2** increases at $M=4$ (0.834, 0.637, and 0.727 for $M=2, 3$, and 4, respectively, **Table 3.7**). The behavior of **VPAA1** is simply interpreted by increase of the degree of structural disorder due to increase of the number of molecules and the strong monopolar repulsion between the amine groups of side chains in the assemblies. The peculiar high

stability of the assembly of the four **VPAA2** molecules originates from the presence of a highly stable conformer (**VPAA2**)₄-**4** (**Figure 3.12d**).

The conformational classification successfully unveils the microscopic details of the structural stability of the assemblies. The structural clustering identified the highly stable symmetric conformation of the assembly of the four **VPAA2** molecules, (**VPAA2**)₄-**4**. The stable structural unit of (**VPAA2**)₄-**4** was found in the larger assemblies of more **VPAA2** molecules (**Figure 3.13**). However, the larger assemblies including (**VPAA2**)₄-**4** were relatively unstable, and more stable conformations where the (**VPAA2**)₄-**4** structural unit is not included and the hydrophobic interactions among the V-shaped anthracene panels are reorganized were found. In contrast, although a conformation of the assembly of four **VPAA2** ones, (**VPAA2**)₄-**4**, was found to be unstable, a stable conformation of the assembly of five **VPAA2** molecules, (**VPAA2**)₅-**5**, includes the structural unit of (**VPAA2**)₄-**4**. Furthermore, for the assemblies of more than six **VPAA** molecules, the non-abundance ratios significantly decrease (**Table 3.6**) and no dominant conformations are found (**Tables 3.11** and **3.12**). Many of the less stable classified conformations are characterized by the core substructural unit with weakly attached molecules around the core unit (**Figure 3.13**). Those observations indicate that the stable conformations of the **VPAA** assemblies are not determined by stable structural cores composed of fewer molecules but are flexibly reorganized depending on the number of molecules constituting of the assemblies. The flexible conformational reorganization of the **VPAA** assemblies well explains the prominent feature of the wide capacity of encapsulation of hydrophobic molecules of different sizes and shapes.

The present study uncovered that the **VPAA** assemblies form several quasi-stable intermediate conformations and undergo transitions between them. The populations of the quasi-stable intermediate conformations in the equilibrium state were also successfully evaluated by the conformational classification. To further characterize the dynamic behaviour of the assemblies, the kinetics of the conformational transitions need to be elucidated. For the purpose, a kinetic analysis based on a Markov state model is on-going and will be published elsewhere.

3.5 CONCLUSION

The nano-sized molecular structures of the VPAA molecular assemblies in aqueous solutions were investigated by the MD simulations for tens of microseconds and the classification analysis of the conformational samples obtained by the MD simulations. Spontaneous formation of the VPAA molecular assemblies in aqueous solutions, which is in line with experimental evidence, and the flexible conformational changes in the assemblies were observed in the MD simulations.

We then developed the classification analysis method based on the RMSD-based structural similarity metric where the difficulty of combinational explosion in the RMSD evaluation of molecular assemblies is circumvented by a heuristic algorithm. The developed method is computationally very efficient, allowing one to extend microscopic structural analyses based on RMSD to large molecular assemblies. The method in fact succeeded in the molecular characterization of the quasi-stable microscopic conformations of the VPAA assemblies appearing in the MD trajectories. The classification analysis then unveiled that the quasi-stable conformations of the VPAA assemblies flexibly reorganize depending on the number of constituent molecules, well explaining their prominent feature of encapsulating a wide range of hydrophobic molecules of various sizes and shapes.

By extending symRMSD to multicomponent systems, as proposed in this study, symRMSD can also deal with host-guest systems, hetero micelles, and systems containing solvent molecules. **Algorithm 1** omits inefficient assignments between molecular assemblies with subsystem that already has large RMSD, effectively reducing the computational cost of exploring the entire space of replacement operations. However, it has the disadvantage that if the cutoff parameter p_{rec} is too small, the solution space containing an exact solution is more likely to be truncated. In addition, a fundamental problem with heuristic algorithms is that there is no way to check whether the approximation is accurate for large-scale problems that cannot be solved exactly by brute force. This problem is solved by the exact solution algorithm of generalized symRMSD using the branch-and-bound method introduced in Chapter 4.

3.6 APPENDIX

3.6.1 Sample size

To evaluate the accuracy of the estimation of the equilibrium distribution of the number of constituent molecules in a molecular assembly, the following procedure was used. The MD data give the total number of molecular assembly, $N_0(t)$, given by Eq. 3.2, and the observed fraction of M molecular assembly, $\{X(M, t)\}$ for $M = 1, 2, \dots, N_{\text{mol}}$ and times $t = 0, 1, \dots, T - 1$. Let $X(M)$ be the equilibrium probability distribution of M molecules and its unbiased estimator is given by

$$\hat{X}(M) = \frac{1}{T} \sum_{t=0}^{T-1} X(M, t). \quad (3.34)$$

We would like to know the confidence interval for the estimated $X(M)$. As is well known, there is a time correlation between snapshots of MD, and estimating confidence intervals requires the assumption that the samples are independent. Based on the blocking procedure to remove time correlations, we consider the following quantities:

$$z_b(M) = \frac{1}{w} \sum_{t=wb+1}^{wb+w} X(M, t) \quad (3.35)$$

Here, w and b are natural numbers sufficient to $wb = T$, where w is a window width large enough to remove time correlations. Note that the sample mean of z_b is equivalent to $\hat{X}$, thus finding a confidence interval for sample mean of $\{z_b\}$ is the same as evaluating a confidence interval for $\hat{X}$. Assume that $\{z_b\}, b = 1, 2, \dots, N_b$ is an independent random variable, the unbiased variance of z_b is defined by

$$\hat{\sigma}^2 = \frac{1}{N_b - 1} \sum_{b=1}^{N_b} (z_b - \hat{X})^2. \quad (3.36)$$

We assume that $\hat{X}$ follows the t-distribution according to the central limit theorem. Strictly speaking, since z_b is always positive, it is affected by censoring of the variable table, but we consider the effect negligible with sufficient samples of large values of $X(M)$. The 95% confidence interval of $\hat{X}$ is determined by

$$\left[\hat{X} - \xi_{N_b-1} \sqrt{\frac{\hat{\sigma}^2}{N_b}}, \hat{X} + \xi_{N_b-1} \sqrt{\frac{\hat{\sigma}^2}{N_b}} \right], \quad (3.37)$$

where ξ_n is the two-sided 95% interval of the Student's t distribution with n degrees of freedom.

The size of the window w in Eq. 3.35 must be determined. With enough samples, the upper limit of the asymptotic $\hat{\sigma}^2$ for a larger window width w corresponds to the true variance of the mean. **Figure 3.15** shows a plot of interval width $\xi_{N_b-1} \sqrt{\hat{\sigma}^2 / N_b}$ versus w . The

number of snapshots T for cMD and REMD were used as 90,000 and 8,000, respectively. The plots continue to rise slowly relative to w , but generally appears to reach a plateau at about thousands and hundreds for cMD and REMD, respectively. REMD appears to have shorter correlation times than cMD due to enhanced sampling by random walks in the state space. Thus we employed $w=10,000$ ($N_b=9$) and $w=1,000$ ($N_b=8$) for cMD and REMD, respectively.

3.6.2 Confidence interval of MSD

The MSD estimation also uses confidence intervals with window averaging. Using the same notations as in Eq. 3.35, the window average for Eq. 3.14 is as

$$\begin{aligned}\hat{R}(t; b) &= \frac{1}{w} \sum_{t_0=wb+1}^{wb+w} \frac{1}{N_{\text{mol}}} \sum_{I=1}^{N_{\text{mol}}} \|c_I(t_0 + t) - c_I(t_0)\|^2 \\ &= \sum_{I=1}^{N_{\text{mol}}} \sum_{j=1}^3 \frac{1}{w N_{\text{mol}}} \sum_{t_0=wb+1}^{wb+w} (c_{jI}(t_0 + t) - c_{jI}(t_0))^2 \\ &= \sum_{I=1}^{N_{\text{mol}}} \sum_{j=1}^3 \hat{r}_{jI}(t; b),\end{aligned}\tag{3.38}$$

where, c_{jI} is the j -th Cartesian coordinate of the I -th molecular centroid and suppose that $\frac{1}{w N_{\text{mol}}} \sum_{t_0=wb+1}^{wb+w} (c_{jI}(t_0 + t) - c_{jI}(t_0))^2$ is $\hat{r}_{jI}(t; b)$. The diffusion coefficient in Eq. 3.13 is obtained by taking $t = s_0, s_0 + \Delta, s_0 + 2\Delta, \dots, s_0 + 2N_s\Delta$ at even intervals and least-squares fitting.

$$\begin{aligned}\hat{D}_b &= \frac{1}{6} \frac{\sum_{k=0}^{2N_s} (k - N_s) \hat{R}(s + k\Delta; b)}{\Delta \sum_{k=0}^{2N_s} (k - N_s)^2} \\ &= \sum_{I=1}^{N_{\text{mol}}} \sum_{j=1}^3 \sum_{k=1}^{N_s} \frac{\hat{r}_{jI}(s + (N_s + k)\Delta; b) - \hat{r}_{jI}(s + (N_s - k)\Delta; b)}{6\Delta N_s(N_s + 1)} k \\ &= \sum_{I=1}^{N_{\text{mol}}} \sum_{j=1}^3 \hat{d}_{jI}(b)\end{aligned}\tag{3.39}$$

Each $\hat{d}_{jI}(b)$ is an independent random variable, a confidence interval with the Student's t distribution (Eq. 3.37) can be used for the mean of random variable $\hat{d}_{jI}(b)$. Here, the sample size is $3N_b N_{\text{mol}}$, but since VPAs form assemblies and are not independent particles, divide this by the average number of constituent particles $\bar{M}$.

In accordance with the previous section, we set $w=10,000$ as the window size to be considered an independent sample, the sample interval to be 1,000-2,000 ns, and the number of samples to be $N_s=500$.

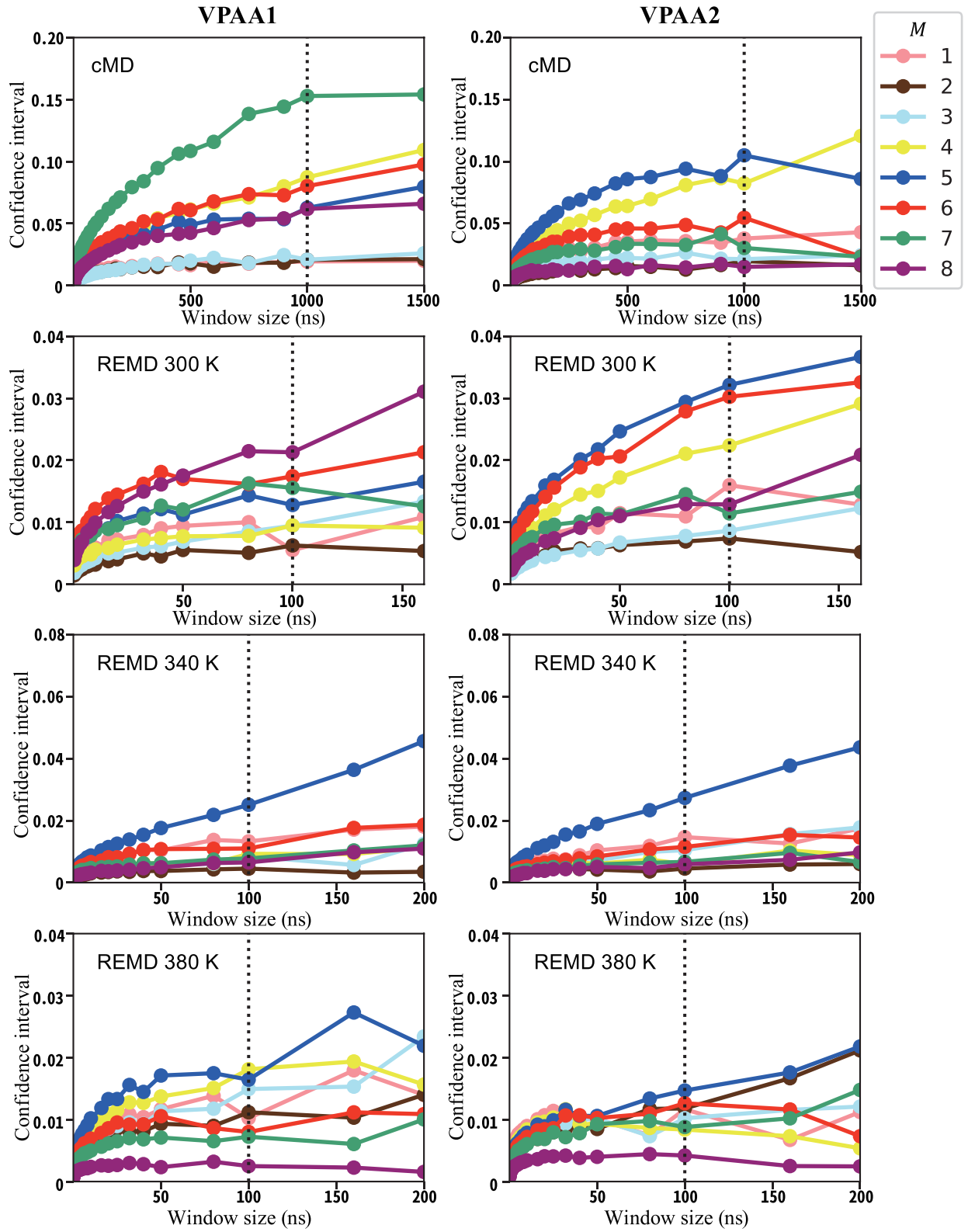


Figure 3.15: Confidence intervals with respect to a block average window size w . (a), (b) cMD. (c), (d) REMD.

3.6.3 Confidence interval of mean lifetime

Then, we derive a confidence interval for the mean lifetime (Eq. 3.4). For the sampling interval $[wb + 1, wb + w]$, the estimate of Eq. 3.4 is as follows:

$$\tau_b(M) = \frac{1}{w\Phi_0(M)} \sum_{t_0=wb+1}^{wb+w} \sum_{t=0}^{T-1} \Phi(t_0 + t; M), \quad (3.40)$$

where

$$\Phi_0(M) = \frac{1}{w} \sum_{t_0=wb+1}^{wb+w} \Phi(t_0; M) \quad (3.41)$$

denotes the average number of assemblies tracked in the interval.

The decay of the VPAA assemblies does not necessarily follow the exponential distribution since it can be caused by multiple events. Here, we assume that the mean lifetime follows the Weibull distribution.

$$f(\tau; m, \eta) = \frac{m}{\tau} \left(\frac{\tau}{\eta} \right)^{m-1} \exp \left(- \left(\frac{\tau}{\eta} \right)^m \right) \quad (3.42)$$

m is the shape parameter and η is the scale parameter. Transform variables of the Weibull distribution with

$$\left(\frac{\tau}{\eta} \right)^m \rightarrow u,$$

u follows the exponential distribution $\exp(-u)$.

Suppose independent samples

$$\tau_1, \dots, \tau_{N_b}$$

are obtained from the Weibull distribution, and let

$$v = \sum_{b=1}^{N_b} \Phi_b(\tau_b/\eta)^m$$

the weighted sum, where,

$$\Phi_b = \Phi(wb + 1; M).$$

The random variable v follows the following gamma distribution:

$$\gamma(v; N_{\text{dof}}) = \frac{v^{N_{\text{dof}}-1}}{\Gamma(N_{\text{dof}})} \exp(-v), \quad (3.43)$$

where,

$$N_{\text{dof}} = \sum_{b=1}^{N_b} \Phi_b$$

and $\Gamma(\cdot)$ denotes the gamma function. The 95% confidence interval of v is

$$\gamma_{0.025, N_{\text{dof}}} \leq v \leq \gamma_{0.975, N_{\text{dof}}}, \quad (3.44)$$

where $\gamma_{0.025, N_{\text{dof}}}$ and $\gamma_{0.975, N_{\text{dof}}}$ are the 95% intervals with equal areas around the median of the gamma distribution of N_{dof} degrees of freedom. Since the expected value of the Weibull distribution is $\eta\Gamma(1 - 1/m)$, solving Eq. 3.44 for the mean lifetime yields

$$\left[\left(\frac{\frac{N_{\text{dof}}}{N_b} \sum_{i=1}^{N_b} \tau_i^m}{\gamma_{0.975, N_{\text{dof}}}} \right)^{1/m} \Gamma \left(1 - \frac{1}{m} \right), \left(\frac{\frac{N_{\text{dof}}}{N_b} \sum_{i=1}^{N_b} \tau_i^m}{\gamma_{0.025, N_{\text{dof}}}} \right)^{1/m} \Gamma \left(1 - \frac{1}{m} \right) \right]. \quad (3.45)$$

The shape parameter m was obtained by the least-squares method of Weibull plots; for given sampled mean lifetimes

$$\tau_1, \dots, \tau_{N_b}$$

and mean number of assemblies

$$\Phi_1, \dots, \Phi_{N_b},$$

m was determined by the slope of the relationship

$$\log \log \frac{1}{1 - F(b)} \propto \log \tau,$$

where $F(b)$ is the cumulative distribution function, *i.e.*,

$$F(b) = \frac{1}{1 + N_{\text{dof}}} \sum_{i=1}^b \Phi_{b_i}$$

and b_i is the index of the i -th smallest τ_b .

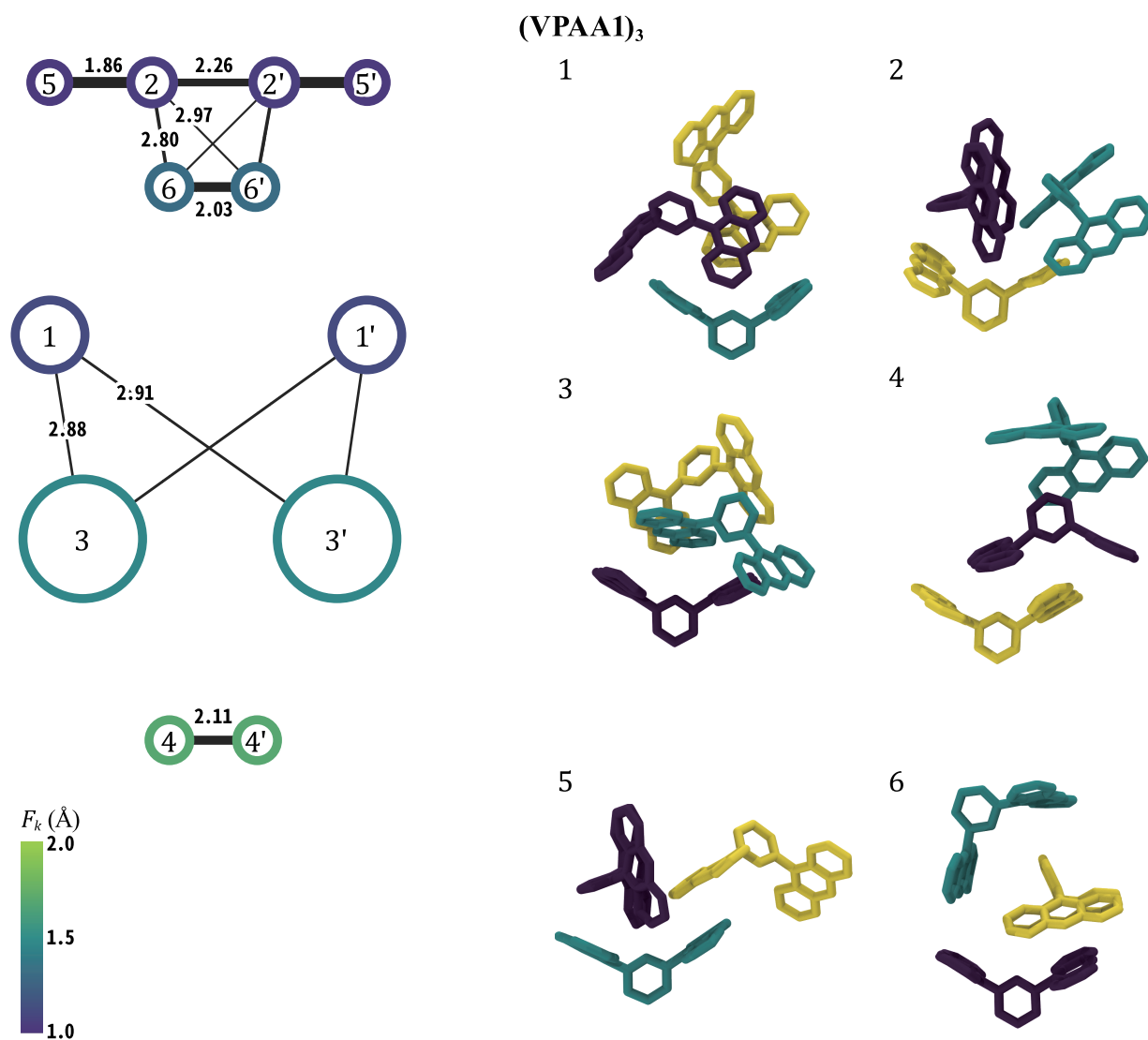


Figure 3.16: A reduced self-similarity matrix of $(VPAA1)_3$. Edges between nodes with symRMSD values less than $3.0 \text{ \AA}$ are depicted with their values (left). Numbers in the nodes indicate the group index $k = 1, 2, \dots, 6$ assigned to the representative structures (right), and their mirror symmetric coordinates are denoted by k' , respectively.

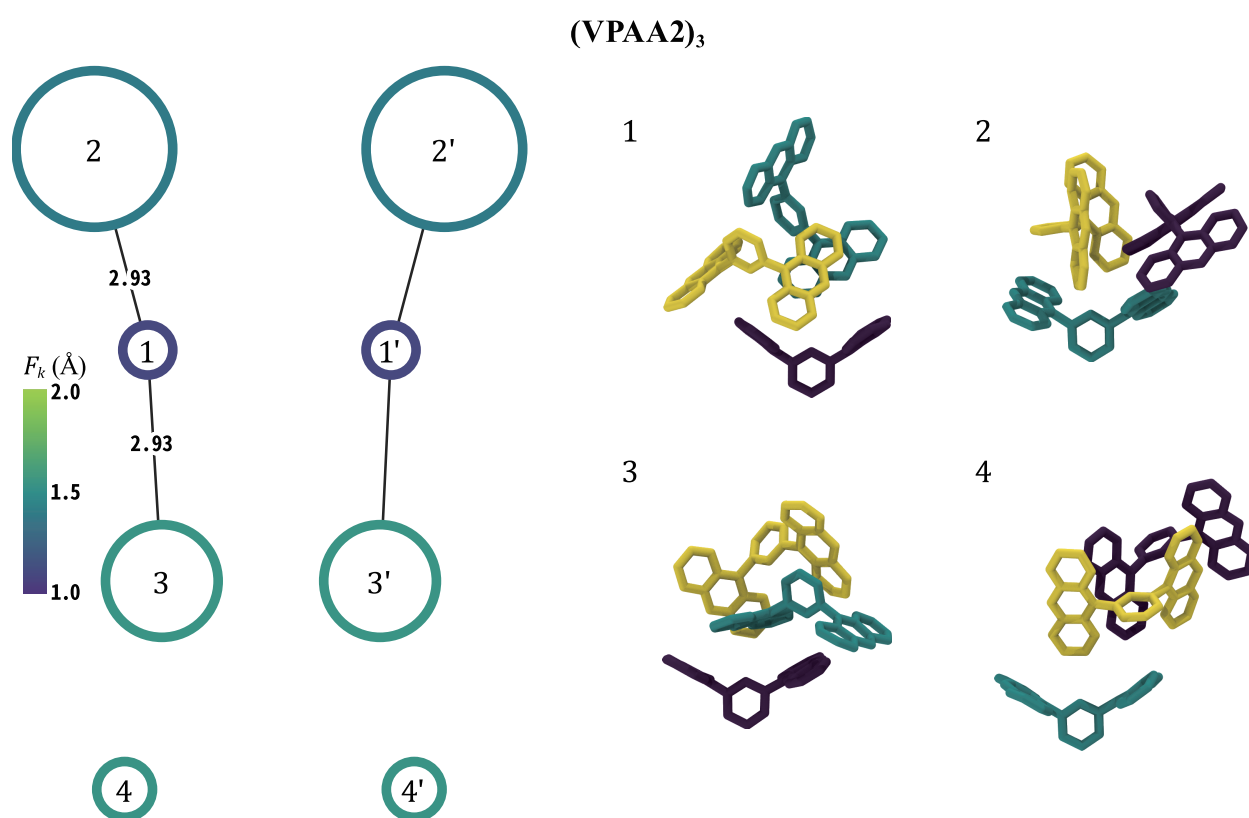


Figure 3.17: A reduced self-similarity matrix of $(\text{VPAA2})_3$. Edges between nodes with symRMSD values less than 3.0 Å are depicted with their values.

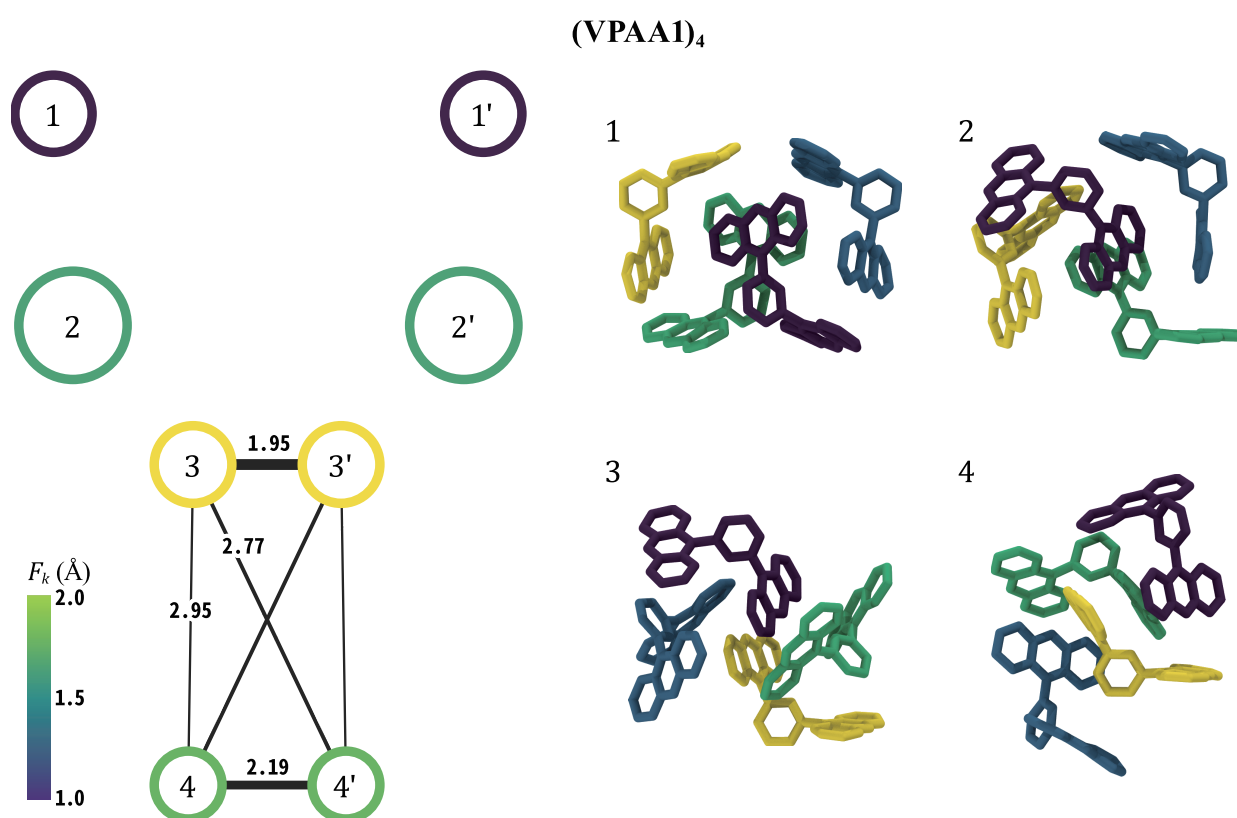


Figure 3.18: A reduced self-similarity matrix of $(\text{VPAA1})_4$. Edges between nodes with symRMSD values less than 3.0 Å are depicted with their values.

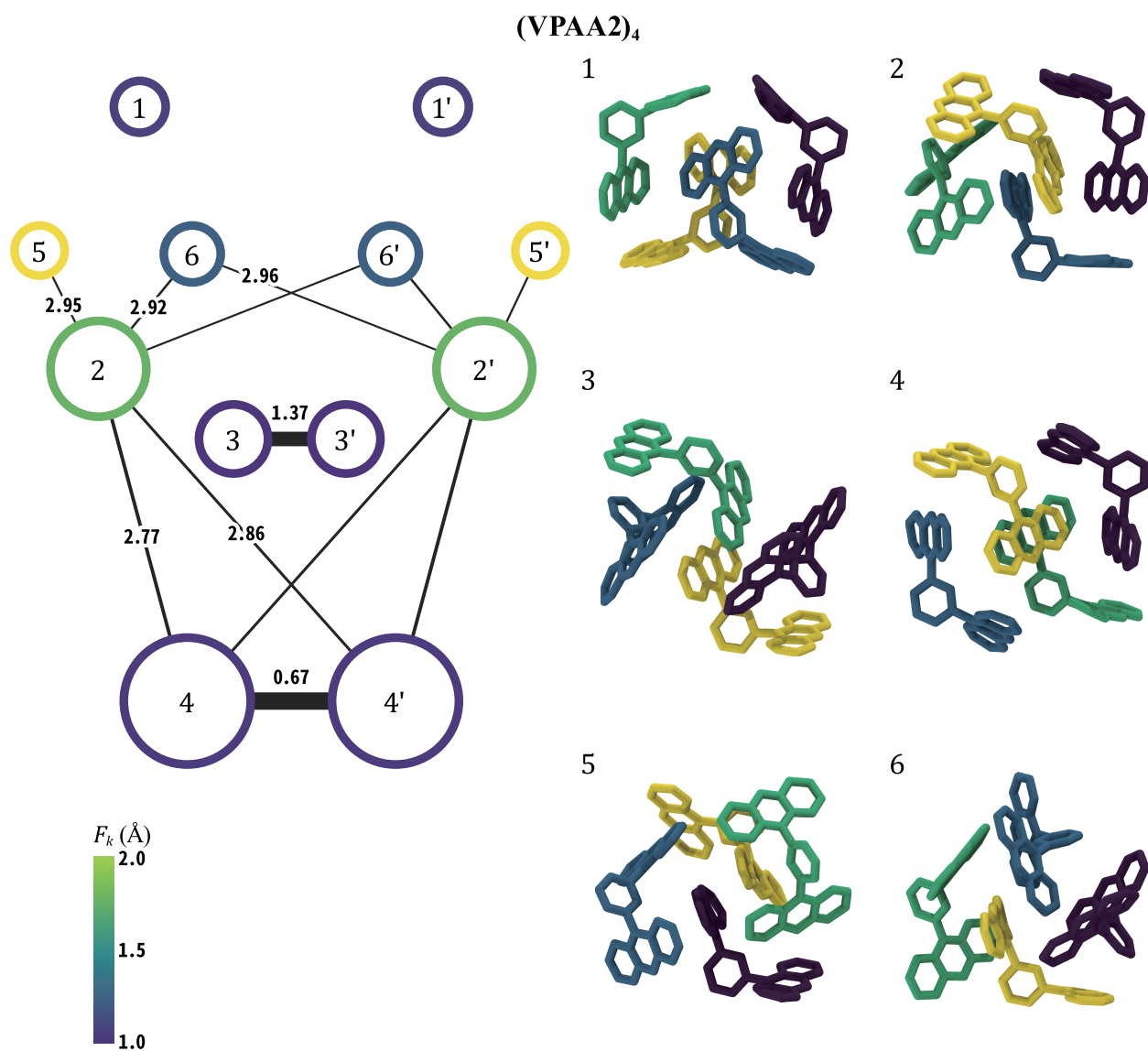


Figure 3.19: A reduced self-similarity matrix of $(VPAA2)_4$. Edges between nodes with symRMSD values less than 3.0 Å are depicted with their values.

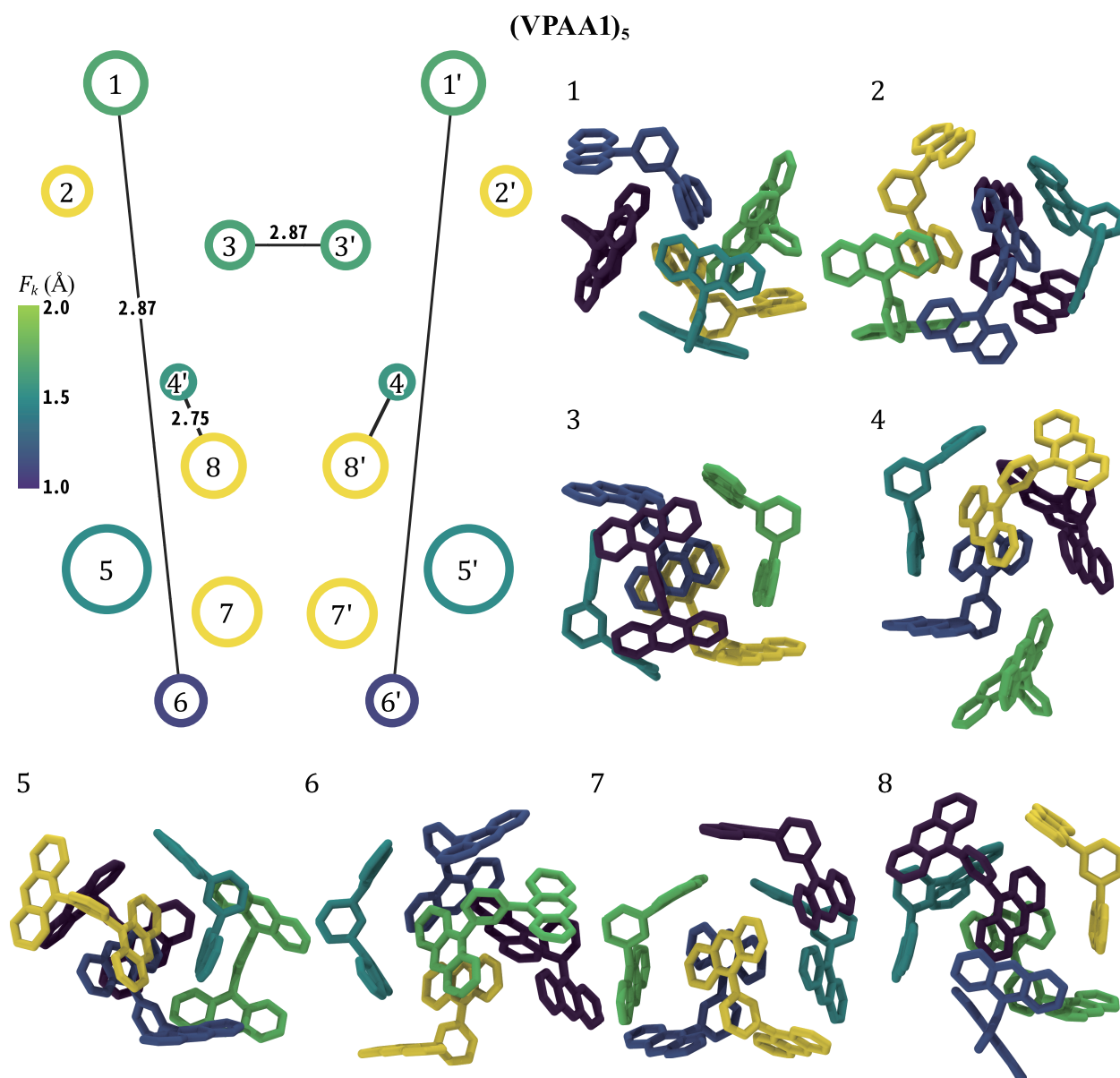


Figure 3.20: The reduced self-similarity matrix of **(VPAA1)₅**. Edges between nodes with the symRMSD value less than 3.0 Å are depicted with their values.

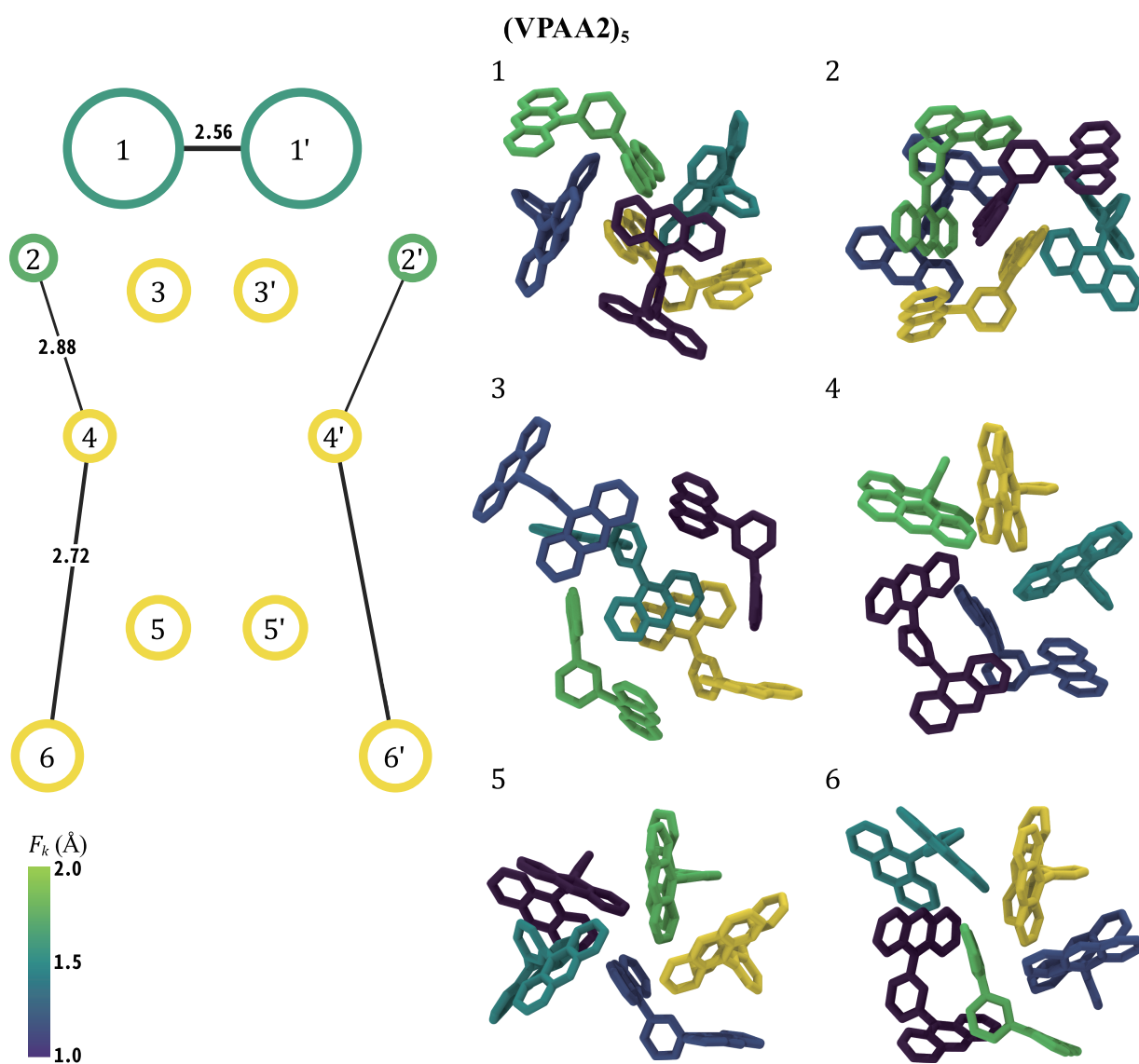


Figure 3.21: A reduced self-similarity matrix of $(VPAA2)_5$. Edges between nodes with symRMSD values less than 3.0 Å are depicted with their values.

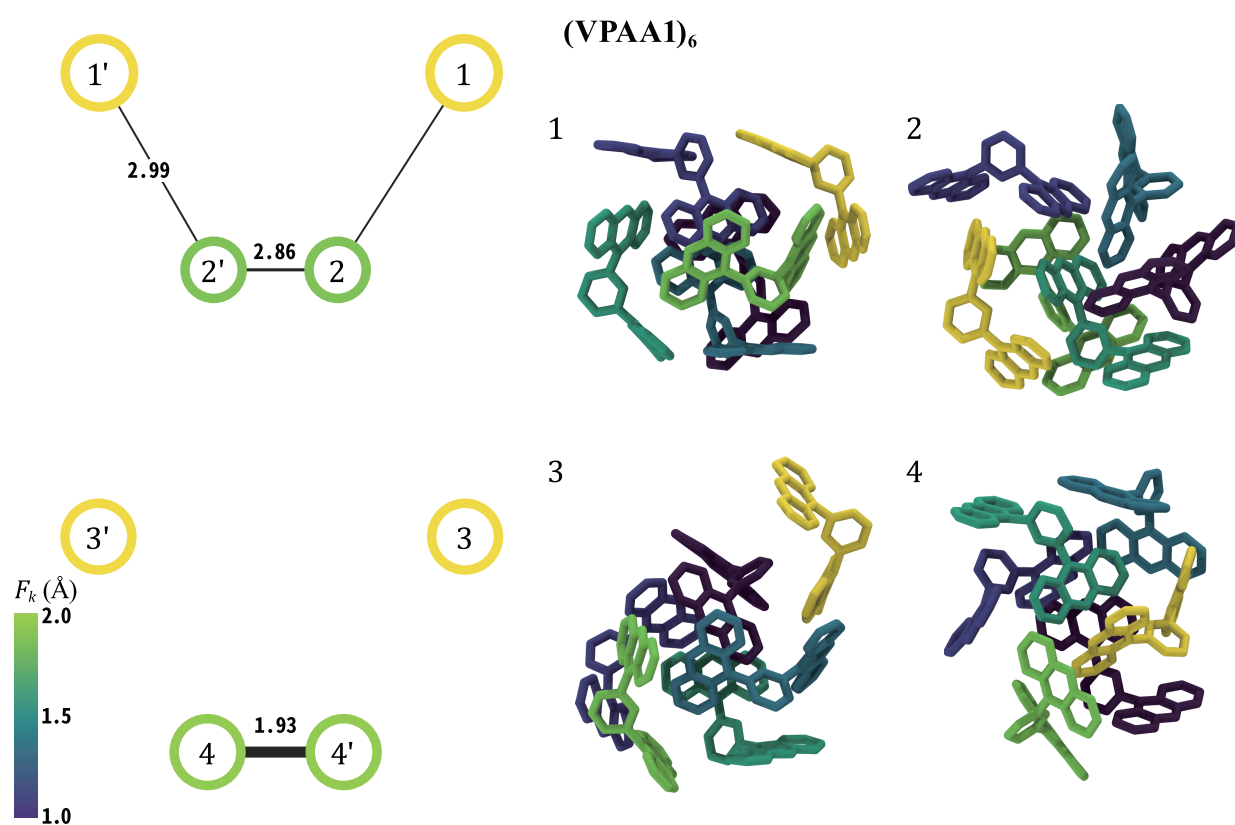


Figure 3.22: A reduced self-similarity matrix of $(VPAA1)_6$. Edges between nodes with symRMSD values less than 3.0 Å are depicted with their values.

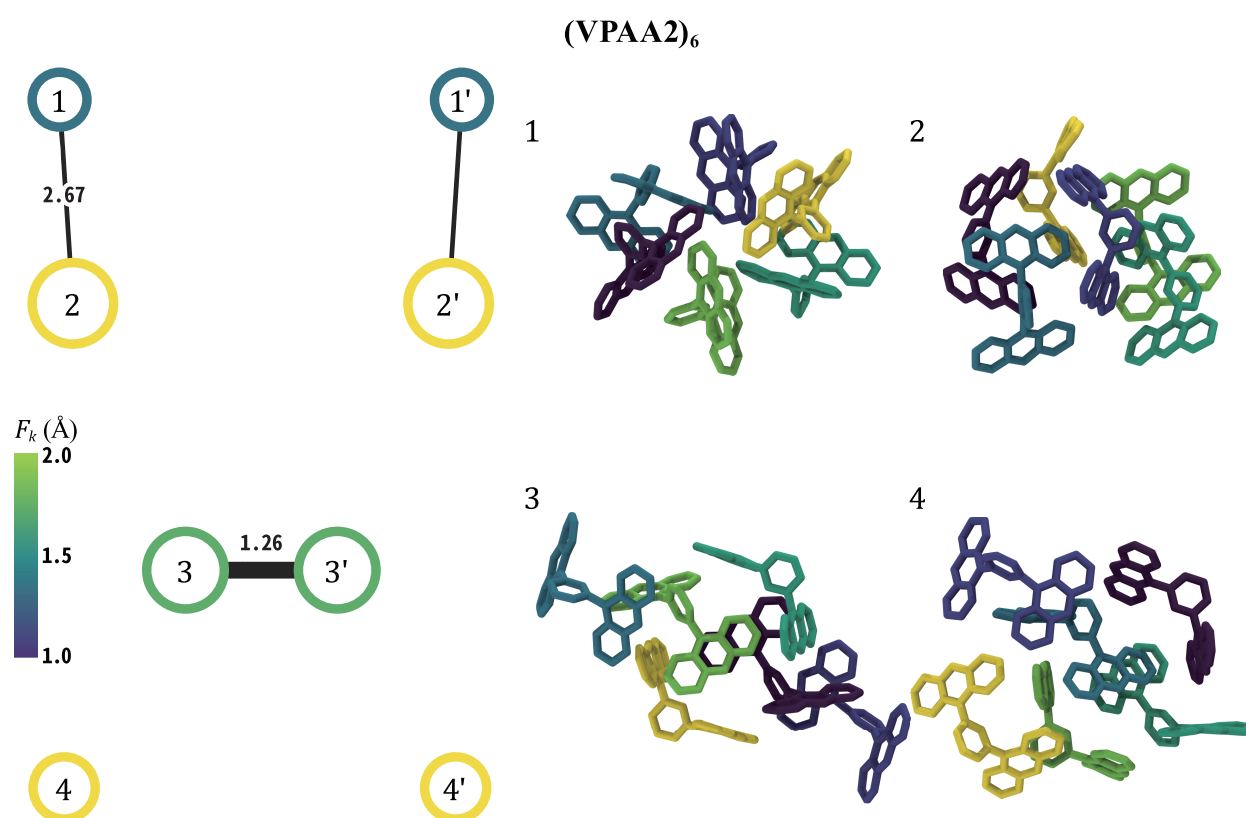


Figure 3.23: A reduced self-similarity matrix of $(\text{VPAA2})_6$. Edges between nodes with symRMSD values less than 3.0 Å are depicted with their values.

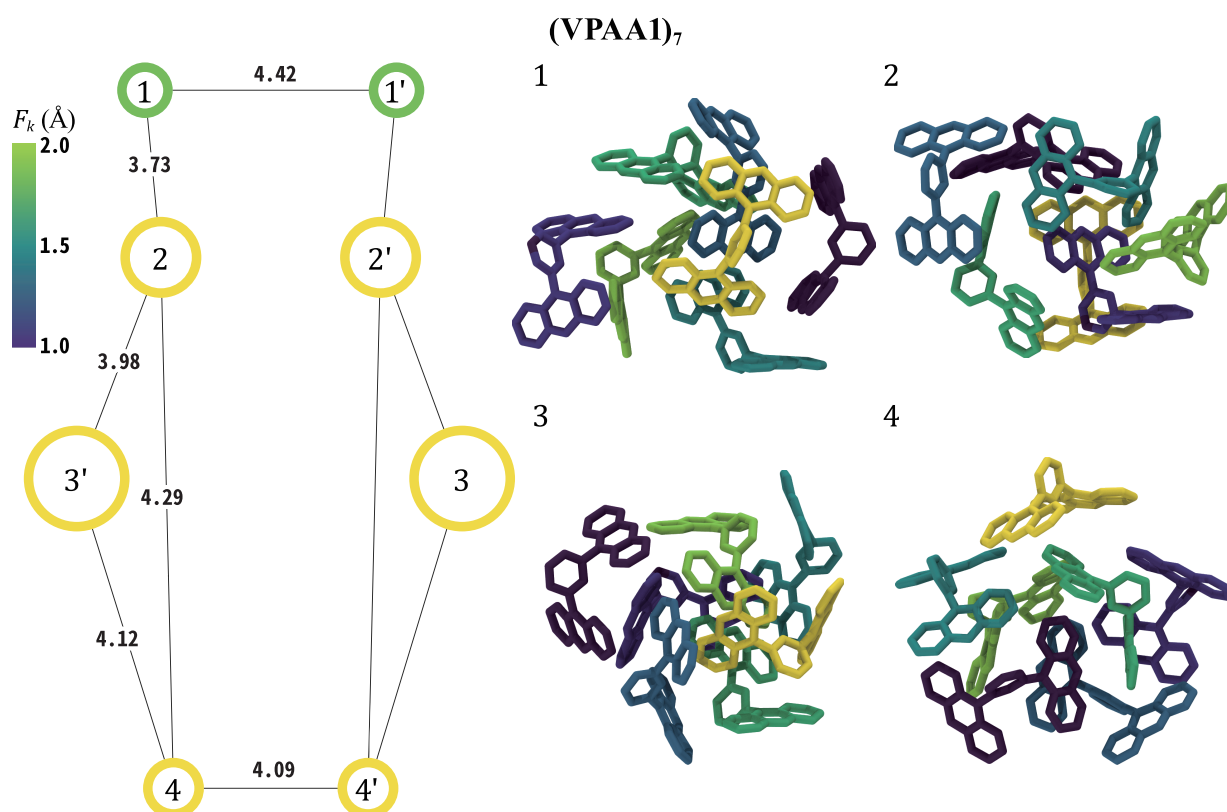


Figure 3.24: A reduced self-similarity matrix of $(VPAA1)_7$. Edges between nodes with symRMSD values less than 4.5 Å are depicted with their values.

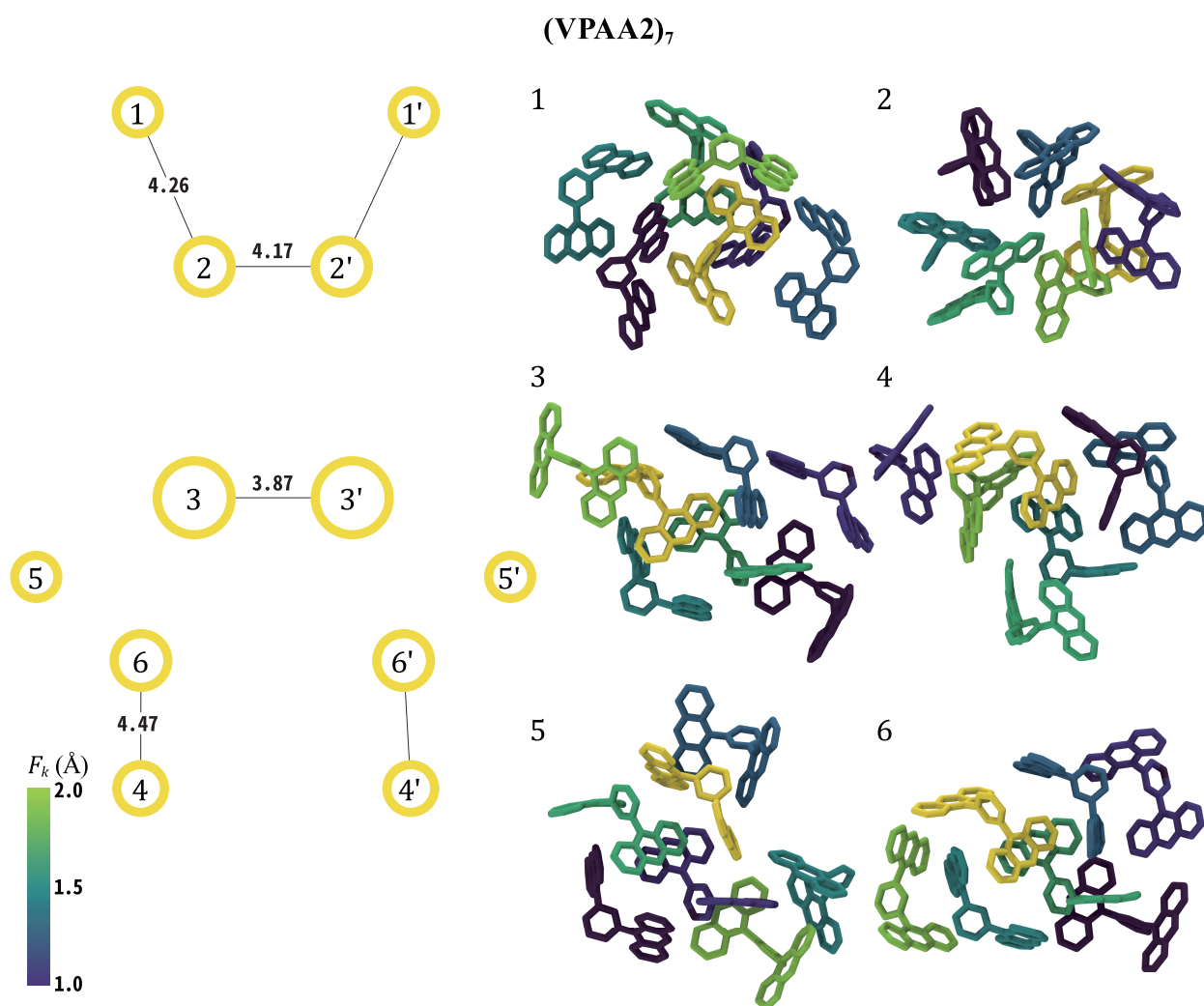
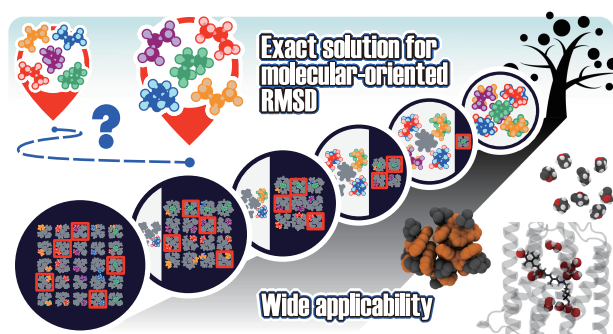


Figure 3.25: A reduced self-similarity matrix of (VPAA2)₇. Edges between nodes with symRMSD values less than 4.5 Å are depicted with their values.

Chapter 4

Molecular-Oriented RMSD with Branch-and-Bound



Root mean squared deviation (RMSD) is one of the most common metrics for comparing the similarity of three-dimensional chemical structures. The chemical structure similarity plays an important role in data chemistry because it is closely related to chemical reactivity, physical property, and bioactivity. Despite the wide use of RMSD, the simultaneous determination of atomic mapping and spatial superposition of RMSD is still a difficult problem to solve in polynomial time. We introduce an algorithm called mobbRMSD, which is formulated in molecular-oriented coordinates and uses the branch-and-bound method to obtain an exact solution for RMSD. Since mobbRMSD uses molecular topologies, it can handle large and complex chemical systems such as molecular liquids, solvations of solute, and self-assembly of large molecules, which are difficult to handle using conventional methods. Furthermore, we propose a mobbRMSD-based structural clustering method for molecular dynamics trajectories, which improves the exponential scaling of the computational cost of mobbRMSD and asymptotically approaches polynomial scaling as the number of data increases. Our algorithm is freely available at <https://github.com/yymmt742/mobbrmsd>.

4.1 INTRODUCTION

The root mean square deviation (RMSD) is one of the most common ways to compare the structural similarity of two molecular structures. [12, 147, 148] Structures of chemical molecules are strongly related to their chemical reactivities, [149, 150] solid-state physical properties, [151] and biological activities. [12, 13, 148, 152, 153] Atomic-resolution measurements such as X-ray crystallography [154, 155] and molecular modeling methods based on computational chemistry are cooperatively advancing chemistry. [156, 157] RMSD plays an important role in the modeling and clustering that aid interpretation. [18, 19, 158] In recent years, molecular science has also begun to be integrated with informatics, [12–14] as evidenced by the increase in 3D chemical structure databases [159–161] and machine learning models to predict 3D chemical structures. [162, 163] The quantification of 3D structural similarity by RMSD is in high demand across research fields and must withstand high throughput due to increasing data sizes.

Despite the wide use of RMSD, some RMSD problems are difficult to solve in polynomial time. [131] This difficulty arises because the way of taking atomic mappings produces combinatorial explosions when no atomic correspondence is given between the structures being compared. If molecular coordinates are represented by N points on d -dimensional Cartesian coordinates, which are written by matrices $\mathbf{X}$ and $\mathbf{X}'$ on $\mathbb{R}^{d \times N}$, the generalized RMSD is expressed as follows:

$$\text{RMSD}(\mathbf{X}, \mathbf{X}') = \min_{\mathbf{R}, \mathbf{c}, \nu} \sqrt{\frac{1}{N} \sum_{i=1}^N \left\| \mathbf{x}_i - \mathbf{R} \mathbf{x}'_{\nu(i)} - \mathbf{c} \right\|^2}, \quad (4.1)$$

where $\mathbf{x}_i$ and $\mathbf{x}'_i$ are the atomic coordinates corresponding to $\mathbf{X}$ and $\mathbf{X}'$, respectively, $\|\cdot\|$ implies the L2-norm, $\mathbf{R}$ is a rotation matrix, and $\mathbf{c}$ is a translation vector. ν is a permutation on $\{1, \dots, N\}$, where the domain of definition is appropriately restricted to avoid unphysical atomic correspondences. RMSD calculations can be decomposed into two parts: minimization concerning $\mathbf{R}$ and $\mathbf{c}$ for a fixed ν and ν for fixed $\mathbf{R}$ and $\mathbf{c}$. These are called optimal superposition and atom mapping, which can be solved by using the singular value decomposition [164, 165] and the Hungarian algorithm, [138, 166, 167] respectively. These iterative runs cannot escape from local solutions. In most practical cases in RMSD, this difficulty is not apparent because a fixed superposition or atom mapping is available. On the other hand, many algorithms have also been developed to find a global solution for RMSD: [132, 138, 139, 168–172] that utilize protein backbones, [168, 169] enumeration of permutation using molecular templates, [170–172] and perturbation methods. [132, 138, 139] In recent years, several methods have been reported to find global solutions for RMSD using the branch-and-bound (BB) method. [133, 173] BB is superior in that the exact solution is guaranteed at the end of the algorithm and the maximum error is obtained even in the case of early termination. [174–177] The generalized RMSD algorithms using BB are capable of global optimization, but the numbers of applicable atoms are effectively limited to about ~ 40 . [133] Another drawback of these methods is the unavailability of molecular topology; that is, the algorithm itself must be modified to restrict the atomic mapping to match the atom type or to avoid mappings that cause bond breaks.

We recently developed an RMSD optimization heuristic algorithm based on molecular-oriented RMSD (moRMSD) and applied it to trajectory data obtained from microseconds of molecular dynamics (MD) simulations of flexible macromolecular assemblies called aromatic micelles. We have succeeded in quantitatively analyzing flexible and diverse conformers of the aromatic micelles by structural clustering using moRMSD. Encouraged by these results, we propose to refine and generalize the moRMSD methodology; moRMSD describes a set of homologous molecules, and its applicability is not limited to molecular assemblies: such as molecular liquids, crystal structures, solvated structures, and so on. Here, we introduce mobbRMSD, an exact solution method that combines BB and moRMSD. We first introduce molecular-oriented coordinates describing a one-component molecular system. After deriving mobbRMSD for the one-component system, the methodology is extended to multicomponent systems. This completes mobbRMSD as a reliable structural similarity score that makes appropriate use of molecular information.

This chapter is organized as follows. The METHOD section reviews the branch-and-bound method and describes the molecular-oriented coordinates and the formulation of mobbRMSD and algorithms. In the RESULTS section, computational experiments are carried out and the performance of mobbRMSD is reported. We also compare it to several previous methods. The DISCUSSION section describes the methodological position of mobbRMSD and discusses early termination using the theoretical bounds. In the CONCLUSION section, we summarize the paper and suggest some topics for further research. The APPENDIX section presents some theorems and proofs used in the text.

4.2 METHOD

4.2.1 Molecular-oriented RMSD

The molecular-oriented coordinate consisting of M molecules of n atoms is defined as

$$\mathbf{X} = \{\mathbf{X}_1, \dots, \mathbf{X}_M\}, \quad (4.2)$$

where $\mathbf{X}_I \in \mathbb{R}^{d \times n}$ is the I -th molecular coordinates, d is the spatial dimension, and n is the number of atoms. The order of atomic coordinates is defined consistently for all molecular coordinates in the topology of the molecule. The molecular-oriented RMSD (moRMSD) is defined as

$$\text{moRMSD}(\mathbf{X}, \mathbf{X}') = \min_{\mathbf{R}, \mathbf{c}, \mu, \nu_1, \dots, \nu_M} \sqrt{\frac{1}{N} \sum_{I=1}^M \sum_{j=1}^n \|\mathbf{x}_{j,I} - \mathbf{R} \mathbf{x}'_{\nu_I(j), \mu(I)} - \mathbf{c}\|^2}, \quad (4.3)$$

where $N = Mn$, $\mathbf{x}_{j,I}$ and $\mathbf{x}'_{j,I}$ are the Cartesian coordinates corresponding to j th atom of $\mathbf{X}_I$ and $\mathbf{X}'_I$, respectively, $\mathbf{c} \in \mathbb{R}^d$ is a translation vector and $\mathbf{R} \in \mathbb{R}^{d \times d}$ is a rotation matrix that satisfies $\mathbf{R}\mathbf{R}^\top = \mathbf{I}$ and $\det(\mathbf{R}) = 1$. $\mathbf{I}$ and $\det(\cdot)$ indicate the identity matrix and the determinant, respectively. ν_I and μ are permutations on $\{1, \dots, n\}$ and $\{1, \dots, M\}$, respectively. If we remove the geometric centroids from the coordinates in advance, *i.e.*, $\tilde{\mathbf{X}} = \mathbf{X} - \bar{\mathbf{X}}$ and $\tilde{\mathbf{X}}' = \mathbf{X}' - \bar{\mathbf{X}}'$, where $\bar{\mathbf{X}}$ and $\bar{\mathbf{X}}'$ are geometric centroids of $\mathbf{X}$ and $\mathbf{X}'$, respectively, $\mathbf{c}^* = \mathbf{0}$ always minimizes Eq. 4.3 concerning $\mathbf{c}$. Therefore, $\mathbf{c}$ is ignored in the following.

Hereafter, let $\mathcal{R}$ and $\mathfrak{S}_L$ be the total set of the rotation matrix on $\mathbb{R}^d$ and finite symmetric group (*i.e.*, the total set of permutations) over $\{1, \dots, L\}$, respectively. From the homogeneous molecular species assumption, μ is a permutation implying molecular exchange and its domain of definition $\mathcal{M}$ coincides with $\mathfrak{S}_M$. Conversely, ν_I represents a permutation of atomic indices in the I th molecule, the domain of definition is restricted in comparison to that of $\mathfrak{S}_n$ by the molecular topology. For example, if the coordinates of a water molecule are represented by an oxygen atom and two hydrogen atoms in that order (*i.e.*, $\mathbf{X}_I = (\mathbf{x}_O, \mathbf{x}_{H1}, \mathbf{x}_{H2})$), there are only two permutations to consider, $\nu_E : (1, 2, 3) \mapsto (1, 2, 3)$ and $\nu_C : (1, 2, 3) \mapsto (1, 3, 2)$. Thus, the domain of definition for all permutations, $(\mu, \nu_1, \dots, \nu_M)$, is represented as $\mathcal{M} \times \mathcal{T}^M$, where $\mathcal{T} \subset \mathfrak{S}_n$ is a set of intramolecular atomic permutations.

The difficulty of computing Eq. 4.3 is that $\#(\mathfrak{S}_M \times \mathcal{T}^M) = M!S^M$ holds, where $\#(\mathcal{T}) = S$ and $\#(\mathcal{A})$ denotes the cardinal of a finite set $\mathcal{A}$. All of μ and ν_I depend on the rotation matrix $\mathbf{R}$ each other. Accordingly, we regard this problem as a discrete optimization and apply a standard method to effectively searches the solution space.

4.2.2 Branch-and-bound

The branch-and-bound (BB) is an algorithmic framework for finding exact solutions to a wide range of discrete and combinatorial optimization problems. [174–177] BB is summarized as a repetition of two steps. (I) Generate child problems encompassing the solution subspaces that have not been evaluated. (II) Evaluate child problems and discard those that do not lead to the exact solution. The implementation of BB is achieved through the design of three components: (i) the pruning rule, (ii) the branching strategy, and (iii) the search strategy. [177]

The pruning rules represent a (set of) criteria for excluding a solution subspace from the search. Since there is no universal rule, in most cases new rules must be derived for each problem. Various formulations of the pruning rules have been proposed for NP problems which are classified into several classes: variational lower bounds, [178] dominance relation, [178–180] and linear programming. [181, 182]

The branching strategy represents the policy for generating child problems and thus shapes the resulting directed tree. A good branching strategy minimizes the generation of subproblems that do not contain exact solutions. The branching strategy affects the performance of algorithms, especially in mixed integer programming. [183] The search strategy determines the priority order in which vertices in the tree are searched. [177, 179, 180, 184–186]

To apply BB to the calculation of moRMSD, we applied the following procedure. Relaxing the domain of the definition of $\mathbf{R}$ in Eq. 4.3, we derive a variational lower bound for the objective function. Then, by gradually tightening the relaxation of $\mathbf{R}$, monotonically increasing partial objective functions are derived. The partial objective functions can be used as a pruning rule because it serves as a variational lower bound for solution subspaces. After describing the structure of directed trees created by the partial objective functions, a variant of depth-first search is introduced as a search strategy.

4.2.3 Pruning rule

We first describe the objective function and its theoretical variational lower bound. For a natural number p , we introduce a p -objective function, which is a variational lower bound on a subset of the solution space. The p -objective function monotonically increases concerning p , which allows us to prune the unnecessary solution subspaces.

Objective function

From the non-negativity, the calculation of Eq. 4.3 is equivalent to the minimization of the squared deviation (SD), $\sum_{I=1}^M \sum_{j=1}^n \left\| \mathbf{x}_{j,I} - \mathbf{R} \mathbf{x}'_{\nu_I(j), \mu(I)} \right\|^2$. SD can be transformed into the matrix trace form, [187] which can be expressed in functional notation as μ and ν_I as

$$\text{SD}(\mu, \nu_1, \dots, \nu_M) = G - 2 \max_{\mathbf{R} \in \mathcal{R}} \sum_{I=1}^M \text{tr} [\mathbf{R} \mathbf{X}'_{\mu(I)} \mathbf{P}^{(\nu_I)} \mathbf{X}_I^{\top}], \quad (4.4)$$

where $\mathbf{P}^{(\nu_I)}$ is the permutation matrix corresponds to ν_I , given by $\{\delta_{i,\nu_I(j)}\}_{i,j \in \{1,\dots,n\}}$, $\delta_{i,j}$ denotes the Dirac's delta, $\text{tr}[\cdot]$ denotes the matrix trace, and G is the constant term defined by

$$G = \sum_{I=1}^M \text{tr} [\mathbf{X}_I \mathbf{X}_I^\top + \mathbf{X}'_I \mathbf{X}'_I{}^\top]. \quad (4.5)$$

It is reasonable to propose that the total set of $\mathbf{P}^{(\nu_I)}$ be written as $\{\mathbf{P}_1, \dots, \mathbf{P}_S\}$ independent of I and we introduce a mapping $\sigma : \{1, \dots, M\} \mapsto \{1, \dots, S\}$ in place of ν_I . We now define the objective function to be minimized as

$$L(\mu, \sigma) = -\max_{\mathbf{R} \in \mathcal{R}} \sum_{I=1}^M \text{tr} [\mathbf{R} \mathbf{X}'_{\mu(I)} \mathbf{P}_{\sigma(I)} \mathbf{X}_I^\top]. \quad (4.6)$$

Hereafter, let $\mathcal{M}$ and $\mathcal{S}$ be the total set of μ and σ , respectively. $\mathcal{M}$ coincides $\mathfrak{S}_M$. Since the maximization in Eq. 4.6 concerning $\mathbf{R}$ can be solved in constant time, [172, 188, 189] the goal is to find an optimal μ, σ , *i.e.*, $\min_{(\mu, \sigma) \in \mathcal{M} \times \mathcal{S}} L(\mu, \sigma)$.

Variational lower bound

By allowing $\mathbf{R}$ in Eq. 4.6 to be taken separately for each I , the following inequality is derived (see APPENDIX for the proof):

$$\min_{\mu' \in \mathfrak{S}_M} \sum_{I=1}^M F_{I, \mu'(I)} \leq \min_{(\mu, \sigma) \in \mathcal{M} \times \mathcal{S}} L(\mu, \sigma), \quad (4.7)$$

where $F_{I,J}$ is the squared displacement matrix on $\mathbb{R}^{M \times M}$ defined by

$$F_{I,J} = - \max_{\substack{\mathbf{R}' \in \mathcal{R}, \\ s' \in \{1, 2, \dots, S\}}} \text{tr} [\mathbf{R}' \mathbf{X}'_J \mathbf{P}_{s'} \mathbf{X}_I^\top]. \quad (4.8)$$

Let

$$l_0 = \min_{\mu' \in \mathfrak{S}_M} \sum_{I=1}^M F_{I, \mu'(I)}, \quad (4.9)$$

be a variational lower bound of Eq. 4.6, and l_0 takes the value $[-G/2, 0]$. Since the minimization concerning μ' in Eq. 4.7 is a linear assignment problem, it can be solved in polynomial time using the Hungarian algorithm. [166, 167] The computational complexity of the Hungarian algorithm is $\mathcal{O}(M^3)$, but it is typically much faster in practice. [190, 191] Since the evaluation of $F_{I,J}$ has a computational complexity proportional to $M^2 S$, the evaluation of l_0 is $\mathcal{O}(M^2 S + M^3)$.

p -objective function

The assumption in Eq. 4.7 that all molecules rotate independently around the centroid is too loose. The idea is to divide the molecular system into two subsystems; the rigid part

and remainder (**Figure 4.1**). Molecules contained in the rigid part are designated by atom mapping to corresponding pairs. The remainder subsystem allows independent rotation of each molecule. As the number of molecules in the rigid part increases, it approaches the original objective function.

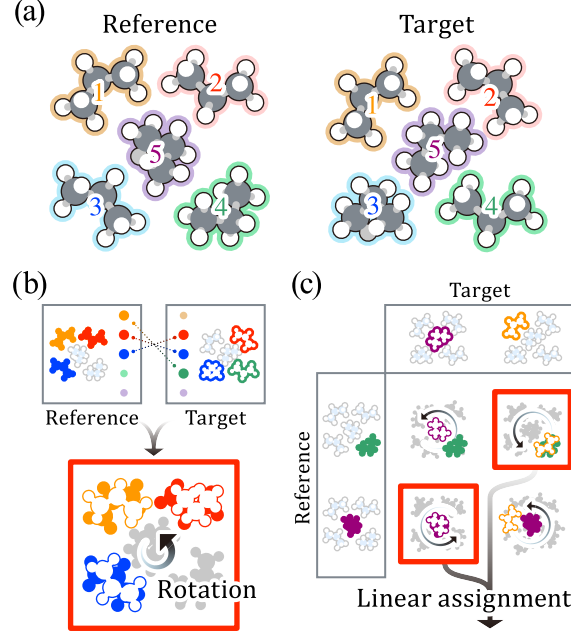


Figure 4.1: Sketch of p -objective function. **(a)** Reference and target structures. Each molecule is labeled from 1 to 5 for both systems, respectively. **(b)** A rigid part. Atomic mapping is identified by molecular matching and intramolecular permutations. Then, an optimal superposition is performed for the subsystem and the cost is calculated. **(c)** Remainder. An independent rotational superposition is performed for each molecular pair. Then, a linear assignment is solved for the cost matrix to compute the cost of the remainder.

We introduce the partial sum of the molecular pair covariance matrices for any μ , σ , and $p \in \{1, \dots, M\}$, defined as

$$\mathbf{C}^{(\mu, \sigma, p)} = \sum_{I=1}^p \mathbf{X}'_{\mu(I)} \mathbf{P}_{\sigma(I)} \mathbf{X}_I^{\top}. \quad (4.10)$$

Here, the p -objective function is defined as

$$l_p(\mu, \sigma) = - \max_{\mathbf{R}^{(p)} \in \mathcal{R}} \text{tr} [\mathbf{R}^{(p)} \mathbf{C}^{(\mu, \sigma, p)}] + \min_{\mu''} \sum_{I=p+1}^M F_{I, \mu'' \circ \mu(I)}, \quad (4.11)$$

where μ'' is a permutation on $\{\mu(p+1), \dots, \mu(M)\}$ and $f \circ g$ denotes the composition of the mapping. The out-of-range sum in Eq. 4.11 appearing at $p = M$ is treated as zero. The first term is the trace of the partial covariance matrices and the second term is the linear assignment concerning the remainder submatrix.

Simple algebraic transformations provide the inequalities for any p , ν , and σ (see APPENDIX for the proof),

$$l_0 \leq l_p(\mu, \sigma), \quad (4.12)$$

and

$$l_p(\mu, \sigma) \leq l_{p+1}(\mu, \sigma). \quad (4.13)$$

From the definition, the M -objective function is the equivalent of the objective function, *i.e.*, $l_M(\mu, \sigma) = L(\mu, \sigma)$. Note that it leads that a pair of (μ', σ') is not a solution if $l_p(\mu', \sigma') > L(\mu, \sigma)$. The computational cost required to evaluate Eq. 4.11 is $\mathcal{O}((M - p)^3)$.

Figure 4.1b represents the rigid part of the 3-objective function for the five propane systems. Here, molecules 1, 2, and 3 in the reference system are mapped to target molecules 3, 4, and 2, respectively. At this point, the atom mapping for each molecule is specified through a mapping σ . Then, an optimal rotation is applied to the established matching, and the cost of the rigid part is calculated. **Figure 4.1c** illustrates the remainder. The cost of the remainder is computed from the linear assignment, where the matrix element represents SD for each molecule pair belonging to the remainder. Note that all rotational operations are performed around the geometric centroid of the entire system, not the subsystem.

4.2.4 Branching strategy

In this section, a monotonically decreasing sequence of subsets of the solution space by an equivalence relation is defined, which also follows the natural number p , similar to the p -objective function. The equivalence relation and families of subsets determined for p are called p -equivalence and p -vertices, respectively. By introducing parent-child relationships into the p -vertices, a directed tree is defined, where the monotonically increasing nature of the p -objective function appears in the weights associated with the p -vertices. Explanations by trees help the design of efficient branching rules.

p -equivalence

Eq. 4.11 depends only on indices $\{1, \dots, p\}$ within the domain of μ and σ . Therefore, if the following relationship is determined for a given (μ_1, σ_1) and $(\mu_2, \sigma_2) \in \mathcal{M} \times \mathcal{S}$, they are the element of the solution space with the same p -objective.

$$(\mu_1, \sigma_1) \stackrel{p}{\sim} (\mu_2, \sigma_2) = \{\mu_1(I) = \mu_2(I), \sigma_1(I) = \sigma_2(I), 1 \leq I \leq p\} \quad (4.14)$$

Since this relationship is reflexive, symmetric, and transitive, it is an equivalence relation and we therefore call it a p -equivalence relation. Note that $(\mu_1, \sigma_1) \stackrel{p}{\sim} (\mu_2, \sigma_2) \implies l_p(\mu_1, \sigma_1) = l_p(\mu_2, \sigma_2)$ holds and hence we conclude that if $l_p(\mu', \sigma') > L(\mu, \sigma)$, then any (μ'', σ'') satisfying $(\mu'', \sigma'') \stackrel{p}{\sim} (\mu', \sigma')$ is not a solution.

p -vertices

We divide the solution space by the p -equivalence relation, which we call p -vertices. p -vertices are the quotient set $\mathcal{M} \times \mathcal{S} \setminus \stackrel{p}{\sim}$ defined by the p -equivalence relation, that is,

$$\mathcal{V}_p = \{[\mu, \sigma]_{\stackrel{p}{\sim}} \mid (\mu, \sigma) \in \mathcal{M} \times \mathcal{S}\}, \quad (4.15)$$

where $[\mu, \sigma]_{\mathcal{L}}$ denotes the equivalence class $\left\{ (\mu', \sigma') \in \mathcal{M} \times \mathcal{S} \mid (\mu', \sigma') \stackrel{\mathcal{L}}{\sim} (\mu, \sigma) \right\}$. When $p < q$, $\stackrel{q}{\sim}$ is finer than $\stackrel{p}{\sim}$, i.e., $(\mu, \sigma) \stackrel{q}{\sim} (\mu', \sigma') \implies (\mu, \sigma) \stackrel{p}{\sim} (\mu', \sigma')$ always holds. In other words, $\mathcal{V}_q$ is a refinement of $\mathcal{V}_p$ for $p < q$.

1, 2, and 3-vertices for $M=3$ and $S=2$ are illustrated in **Figure 4.2**. Here, $\mathcal{M}$ and $\mathcal{S}$ are the total set of permutations on $\{1, 2, 3\}$ and mappings on $\{1, 2, 3\} \rightarrow \{1, 2\}$, respectively. $\mathcal{M}$ includes $3! = 6$ elements, denoted as μ_1 to μ_6 . Similarly, $\mathcal{S}$ includes $2^3 = 8$ elements, σ_1 to σ_8 . By focusing on $I = 1$, $\mathcal{M}$ and $\mathcal{S}$ are partitioned into 3 and 2 equivalence classes, respectively, so $\stackrel{1}{\sim}$ divides $\mathcal{M} \times \mathcal{S}$ into $3 \times 2 = 6$ parts (**Figure 4.2a**). In the same way, corresponding to $\stackrel{2}{\sim}$ and $\stackrel{3}{\sim}$, $\mathcal{V}_2$ and $\mathcal{V}_3$ are given as the $6 \times 2 \times 2 = 24$ and $24 \times 1 \times 2 = 48$ parts, respectively (**Figure 4.2b and c**).

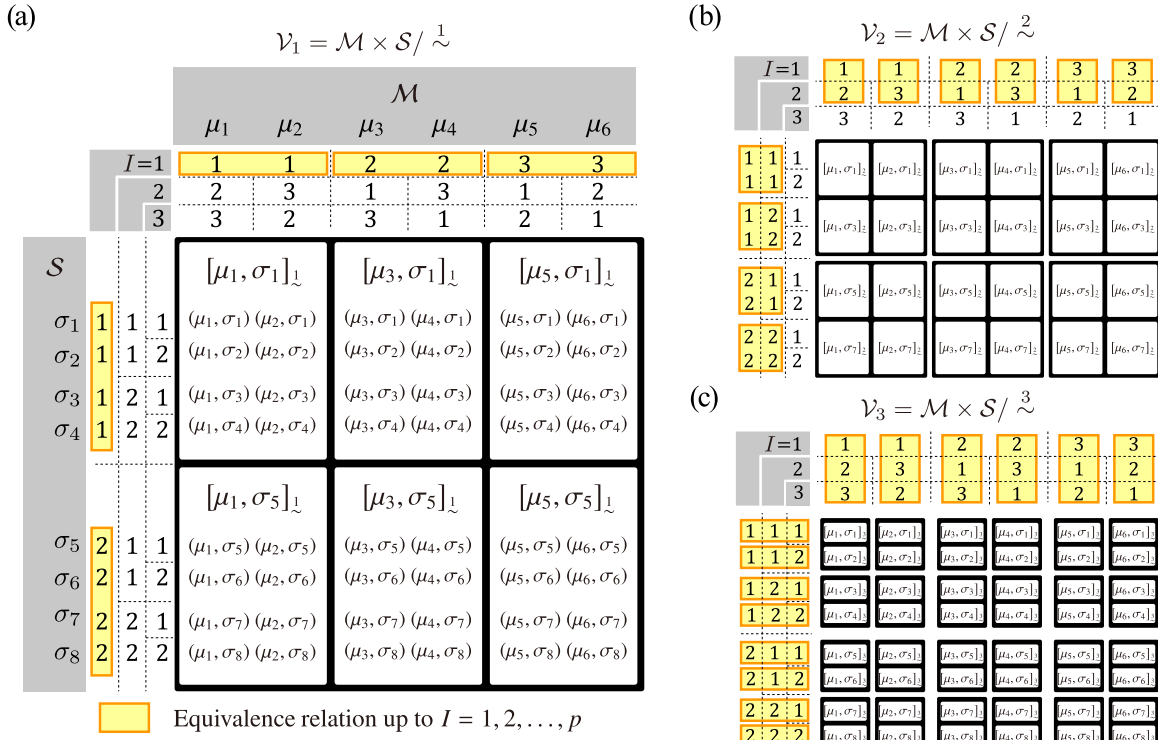


Figure 4.2: Definition of p -vertices for $M=3$, $S=2$. **(a)** 1-vertices. $\mathcal{M}$ and $\mathcal{S}$ are the total set of permutations on $\{1, 2, 3\}$ and mappings $\sigma : \{1, 2, 3\} \rightarrow \{1, 2\}$, respectively. The equivalence up to $I = 1$ for μ and σ is indicated by the yellow boxes: $\{\{\mu_1, \mu_2\}, \{\mu_3, \mu_4\}, \{\mu_5, \mu_6\}\}$ and $\{\{\sigma_1, \sigma_2, \sigma_3, \sigma_4\}, \{\sigma_5, \sigma_6, \sigma_7, \sigma_8\}\}$. The partitions delimited by the black boxes correspond to equivalence classes created by 1-equivalence. **(b)** 2-vertices defined by 2-equivalence relation, $\mathcal{V}_2$. **(c)** 3-vertices defined by 3-equivalence relation, $\mathcal{V}_3$.

Each vertex $v \in \mathcal{V}_p$ is associated with the corresponding value of p -objective (Eq. 4.11), which is represented by $w(v)$. Hereafter, for simplicity, $\mathcal{V} = \bigcup_{p=0}^M \mathcal{V}_p$ be the total set of vertices, where $\mathcal{V}_0 = \{\mathcal{M} \times \mathcal{S}\}$. For any $\mathcal{Q} \subset \mathcal{V}$, let $\min \mathcal{Q}$ be the minimum element in $\mathcal{Q}$ with respect to $w(v)$ (choose one if there are multiple).

The monotonically increasing nature of the p -objective in Eq. 4.13 appears here in the

form that $p < q, v \in \mathcal{V}_p, v' \in \mathcal{V}_q, v \supset v' \implies w(v) \leq w(v')$ holds. Also, from the definition, every element of $\mathcal{V}_M$ is a single point set, and thus for $v \in \mathcal{V}_M$, $w(v)$ is equal to the objective function (Eq. 4.6) for (μ, σ) in the single point set v . This shows that to minimize the objective function, the task is to find $\min \mathcal{V}_M$.

Branching

By constructing a tree with $v_0 \in \mathcal{V}_0$ as the root and $\mathcal{V}_M$ as the leaves, we can construct a branching strategy to find $\min \mathcal{V}_M$. One of the simplest is to connect all p to $p + 1$ inclusive of vertices across the hierarchy by edges, forming a tree corresponding to the enumeration of factorials given by

$$\mathcal{E} = \bigcup_{p=0}^{M-1} \mathcal{E}_{p,p+1}, \quad (4.16)$$

where $\mathcal{E}_{p,q}$ denotes the set of edges between vertices spanning hierarchies p and q as

$$\mathcal{E}_{p,q} = \{(v, v') \in \mathcal{V}_p \times \mathcal{V}_q \mid v \supset v'\}. \quad (4.17)$$

The factorial tree given by Eq. 4.16 is easy to implement and memory management, and is therefore described below.

Optimization of the branching strategies is not the subject of this paper, but we mention a brief improvement of the strategies. Thinning out of p , *i.e.*, $\mathcal{E}' = \bigcup_{j=0}^{J-1} \mathcal{E}_{p_j, p_{j+1}}$ for $p_0 = 0$, $p_J = M$, and $p_0 < p_1 < \dots < p_J$, reduces the total number of vertices in the tree. However, this does not necessarily reduce the number of subproblems generated. The skip of p can eliminate some computations by memoization in exchange for increased memory usage. (See APPENDIX for a detailed discussion.)

4.2.5 Search strategy

This section describes the last component of BB, the search strategy. One of the basic choices is the depth-first search (DFS). [177, 186] DFS is easy to implement and saves memory, but it can be inefficient because it randomly selects child vertices, and can lead to thrashing. [186] Also, DFS can be inefficient in the presence of extremely deep branches, [177] but this issue seems insignificant in our problem where the maximum depth of the tree is fixed at M . The best-first search (BFS) keeps the entire subproblems and prioritizes the most promising ones for selection. [177, 184] BFS has the disadvantage of large memory usage and the possibility of not identifying a solution until the end of the search. [177] To solve the drawbacks of DFS and BFS, hybrid methods exist that address the DFS and BFS, such as cyclic best-first search. [179, 180]

The deep-sea-troll search (DSTS) is a variant of DFS with complete branching. [177, 185] DSTS evaluates all child vertices from the current vertex and selects the best one as the next vertex. [185] DSTS improves the randomness of DFS while using less memory; the maximum number of vertices that DSTS generates in the queue is at most $1 + SM(M + 1)/2$. The main motivation for using DSTS is to improve the efficiency of vertex weight

calculation. The evaluation of Eq. 4.11 can be simplified by using the calculation results of the parent vertex. (See APPENDIX for the details.)

Deep-sea-troll search

For a given tree, DSTS can be expressed by iteration of the two operations: descend and ascend. Descend generate children from the current vertex selecting the best one and repeat as often as possible. Ascend selects the best vertex from the unexplored vertices in the current hierarchy, and if not possible, mark the parent vertex as explored and ascend one level in the hierarchy. During these processes, unnecessary search space can be discarded by dequeuing vertices that are worse than the current candidate solution.

Figure 4.3 shows a schematic example of DSTS for $M=3$ and $S=2$. Let Q and Q_p be the subset of $\mathcal{V}$ and $\mathcal{V}_p$, respectively; *i.e.*, for $Q \subset \mathcal{V}$, $Q_p = Q \cap \mathcal{V}_p$. The vertex **1** with weight -8 is divided into children **2** through **7** by 1-equivalence, which have weights $-2, -8, -3, -1, -1$, and -7 , respectively (line 3 in **Figure 4.3**). Repeating this procedure until p reaches M , we see that leaf **12** has weight -3 (lines 6 and 7 in **Figure 4.3a**). At this stage, the process cannot continue, so the descend process is terminated. The ascend process is then performed; A vertex with a weight greater than the current candidate solution ($\underline{L} = -3$) is removed from the queue in the bottom (Q_3), resulting in Q_3 becoming an empty set and thus moving up one level in the hierarchy. Similarly, the vertex **10** is removed from Q_2 , and vertex **9** is selected as the next vertex to complete the ascend process. The descend and ascend processes are repeated, finally yielding a global solution of -6 at vertex **21** (lines 17 and 18 in **Figure 4.3c**). As a result, only 21 vertices are evaluated, although the total number of vertices in the tree is 79. Compared to the full search cost of 48 (*i.e.*, $\#(\mathcal{V}_3)$), the evaluation is reduced by about half or less.

The DSTS procedures are summarized in **Algorithm 2**. The binary operation $Q - Q'$ for a pair of finite sets in the algorithms implies their set difference. **Algorithm 2** uses four local variables: the current depth $p \in \{0, \dots, M\}$, the vertex queue $Q \subset \mathcal{V}$, the lower bound $\underline{L}$, and the upper bound $\bar{L}$. If the tree is too large to complete DSTS, **Algorithm 2** can return the upper and lower bounds of L upon receipt of the termination signal.

When moRMSD is zero (*i.e.*, when there exists a root-to-leaf path for which all vertex weights are zero), DSTS is guaranteed to be an $\mathcal{O}(M^2S)$ algorithm. Since the point matching problem for congruent transformations can be solved in log-linear time, [141] this is not worth mentioning, but it is important that differences between the compared structures become smaller, DSTS approaches polynomial scaling algorithm.

4.2.6 Multi-component system

To accommodate general chemical systems, we extend to multi-component systems. There are no mathematical difficulties here, except for some complications in notation. The molecular-oriented coordinate of a multi-component system is defined by

$$\mathbf{X}^* = \{\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \dots, \mathbf{X}^{(K)}\}, \quad (4.18)$$

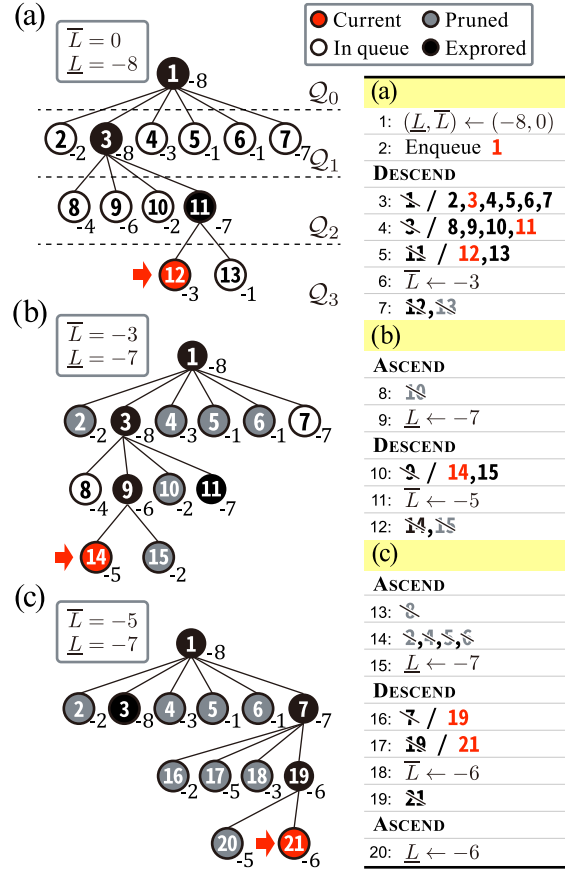


Figure 4.3: Deep-sea-troll search with $M=3$ and $S=2$. **(a)** Results of the first DESCEND. Each circle represents a vertex in Q , and the vertex weights are in the lower right corner of each circle. L is updated with -3 at the vertex **12** (red arrow). **(b)** Results of the second DESCEND. L is updated with -5 at the vertex **14**. **(c)** Results of the third DESCEND. L is updated with -6 at the vertex **21**, which is the exact solution.

Algorithm 2 Deep-sea-troll search

```
1: function DSTS( $\mathcal{V}_0$ )
2:    $(p, \mathcal{Q}, \underline{L}, \bar{L}) \leftarrow (0, \mathcal{V}_0, w(\min \mathcal{V}_0), 0)$ 
3:   repeat
4:      $(p, \mathcal{Q}) \leftarrow \text{DESCEND}(p, \mathcal{Q}, \bar{L})$ 
5:     if  $p = M$  then
6:        $\bar{L} \leftarrow \min \{\bar{L}, w(\min \mathcal{Q}_p)\}$ 
7:        $\mathcal{Q} \leftarrow \mathcal{Q} - \mathcal{Q}_p$ 
8:     end if
9:      $(p, \mathcal{Q}) \leftarrow \text{ASCEND}(p, \mathcal{Q}, \bar{L})$ 
10:     $\underline{L} \leftarrow (\text{if } \mathcal{Q} \neq \emptyset \text{ then } w(\min \mathcal{Q}) \text{ else } \bar{L})$ 
11:    until  $(\bar{L} = \underline{L})$  or (Termination signal received)
12:    return  $(\underline{L}, \bar{L})$ 
13: end function
14: function DESCEND( $p, \mathcal{Q}, \bar{L}$ )
15:   while  $(p < M)$  and  $(\mathcal{Q}_p \neq \emptyset)$  do
16:      $v^* \leftarrow \min \mathcal{Q}_p$ 
17:      $\mathcal{Q} \leftarrow \mathcal{Q} \cup \{v \in \mathcal{V}_{p+1} \mid v \subset v^*, w(v) < \bar{L}\}$ 
18:      $\mathcal{Q} \leftarrow \mathcal{Q} - \{v^*\}$ 
19:      $p \leftarrow p + 1$ 
20:   end while
21:   return  $(p, \mathcal{Q})$ 
22: end function
23: function ASCEND( $p, \mathcal{Q}, \bar{L}$ )
24:   while  $(p > 0)$  and  $(\mathcal{Q}_p = \emptyset)$  do
25:      $p \leftarrow p - 1$ 
26:      $\mathcal{Q} \leftarrow \mathcal{Q} - \{v \in \mathcal{Q}_p \mid w(v) \geq \bar{L}\}$ 
27:   end while
28:   return  $(p, \mathcal{Q})$ 
29: end function
```

where K is the number of molecular species, $\mathbf{X}^{(k)} \in \mathbb{R}^{d \times M^{(k)} n^{(k)}}$ is the molecular-oriented coordinate of k th molecular species, and $M^{(k)}$ and $n^{(k)}$ are the number of molecules and the number of atoms per molecule corresponding to molecular species k , respectively. The moRMSD (Eq. 4.3) extends to

$$\text{moRMSD}^*(\mathbf{X}^*, \mathbf{X}'^*) = \min_{\mathbf{R}, \mathbf{c}, \boldsymbol{\mu}, \boldsymbol{\nu}} \sqrt{\frac{1}{N^*} \sum_{k=1}^K \sum_{j=1}^{n^{(k)}} \sum_{I=1}^{M^{(k)}} \left\| \mathbf{x}_{jI}^{(k)} - \mathbf{R} \mathbf{x}_{\nu^{(k)}(j;I), \mu^{(k)}(I)}'^{(k)} - \mathbf{c} \right\|^2}, \quad (4.19)$$

where $\mathbf{x}_{jI}^{(k)}$ and $\mathbf{x}_{jI}'^{(k)}$ are the atomic coordinates corresponding to $\mathbf{X}^*$ and $\mathbf{X}'^*$, respectively, N^* is the total number of atoms in the system, *i.e.*, $N^* = \sum_{k=1}^K M^{(k)} n^{(k)}$, $\mathbf{R}$ and $\mathbf{c}$ are a rotation matrix and a translation vector, respectively. $\boldsymbol{\mu} = \{\mu^{(1)}, \dots, \mu^{(K)}\}$ and $\boldsymbol{\nu} = \{\nu^{(1)}, \dots, \nu^{(K)}\}$ are permutations on $\prod_{k=1}^K \mathfrak{S}_{M^{(k)}}$ and $\prod_{k=1}^K \mathcal{T}_k^{M^{(k)}}$, respectively, where $\mathcal{T}_k \subset \mathfrak{S}_{n^{(k)}}$ is the set of the intramolecular permutations. As before, translation $\mathbf{c}$ is ignored by removing centroids from $\mathbf{X}^*$ and $\mathbf{X}'^*$. Let $S^{(k)}$ be the number of intramolecular permutations for the species k given by $\#(\mathcal{T}_k)$.

SD (Eq. 4.4) is modified to

$$\text{SD}^*(\boldsymbol{\mu}, \boldsymbol{\sigma}) = G^* + 2L^*(\boldsymbol{\mu}, \boldsymbol{\sigma}), \quad (4.20)$$

where

$$G^* = \sum_{k=1}^K \text{tr} \left[\mathbf{X}^{(k)} \mathbf{X}^{(k)\top} + \mathbf{X}'^{(k)} \mathbf{X}'^{(k)\top} \right] \quad (4.21)$$

and

$$L^*(\boldsymbol{\mu}, \boldsymbol{\sigma}) = -\max_{\mathbf{R}} \sum_{k=1}^K \sum_{I=1}^{M^{(k)}} \text{tr} \left[\mathbf{R} \mathbf{X}_{\mu^{(k)}(I)}'^{(k)} \mathbf{P}_{\sigma^{(k)}(I)}^{(k)} \mathbf{X}_I^{(k)\top} \right], \quad (4.22)$$

respectively. $\boldsymbol{\sigma}$ is $\{\sigma^{(1)}, \dots, \sigma^{(K)}\}$ and $\sigma^{(k)}$ is mappings on $\{1, \dots, M^{(k)}\} \mapsto \{1, \dots, S^{(k)}\}$, and $S^{(k)}$ is the number of intramolecular permutations of k th molecular species. Eq. 4.22 is the objective function to be minimized.

Thereafter, the variational lower bound, the (k, p) -evaluation function and (k, p) -equivalence relation are redefined as Eq. 4.7, Eq. 4.11, and Eq. 4.14 for the single-component system, respectively, where k and p are natural numbers representing chemical species and hierarchy, respectively. As a result, a DSTS algorithm for multi-component systems that has the same interface is derived as **Algorithm 3**. See APPENDIX for details.

Algorithm 3 Deep-sea-troll search for multi-component system

```
1: function DSTS*( $\mathcal{V}_{0,1}$ )
2:    $(t, p, \mathcal{Q}, \underline{L}, \overline{L}) \leftarrow (0, 0, \mathcal{V}_{0,1}, w(\min \mathcal{V}_{0,1}), 0)$ 
3:   repeat
4:      $(t, p, \mathcal{Q}) \leftarrow \text{DESCEND}^*(t, p, \mathcal{Q}, \overline{L})$ 
5:     if  $(t = K)$  and  $(p = M^{(t)})$  then
6:        $\overline{L} \leftarrow \min \{ \overline{L}, w(\min \mathcal{Q}_{t,p}) \}$ 
7:        $\mathcal{Q} \leftarrow \mathcal{Q} - \mathcal{Q}_{t,p}$ 
8:     end if
9:      $(t, p, \mathcal{Q}) \leftarrow \text{ASCEND}^*(t, p, \mathcal{Q}, \overline{L})$ 
10:     $\underline{L} \leftarrow (\text{if } \mathcal{Q} \neq \emptyset \text{ then } w(\min \mathcal{Q}) \text{ else } \overline{L})$ 
11:    until  $(\overline{L} = \underline{L})$  or (Termination signal received)
12:    return  $(\underline{L}, \overline{L})$ 
13: end function
14: function DESCEND*( $t, p, \mathcal{Q}, \overline{L}$ )
15:   while True do
16:     while  $(p < M^{(t)})$  and  $(\mathcal{Q}_{t,p} \neq \emptyset)$  do
17:        $v^* \leftarrow \min \mathcal{Q}_{t,p}$ 
18:        $\mathcal{Q} \leftarrow \mathcal{Q} \cup \left\{ v \in v^* \setminus \widetilde{\sim}^{t,p+1} \mid w(v) < \overline{L} \right\}$ 
19:        $\mathcal{Q} \leftarrow \mathcal{Q} - \{v^*\}$ 
20:        $p \leftarrow p + 1$ 
21:     end while
22:     if  $(t = K)$  or  $(\mathcal{Q}_{t,p} = \emptyset)$  break
23:      $v^* \leftarrow \min \mathcal{Q}_{t,p}$ 
24:      $\mathcal{Q} \leftarrow \mathcal{Q} \cup \left\{ v \in v^* \setminus \widetilde{\sim}^{t+1,p} \mid w(v) < \overline{L} \right\}$ 
25:      $\mathcal{Q} \leftarrow \mathcal{Q} - \{v^*\}$ 
26:      $(t, p) \leftarrow (t + 1, 0)$ 
27:   end while
28:   return  $(t, p, \mathcal{Q})$ 
29: end function
30: function ASCEND*( $t, p, \mathcal{Q}, \overline{L}$ )
31:   while  $k > 0$  do
32:     while  $(p > 0)$  and  $(\mathcal{Q}_{t,p} = \emptyset)$  do
33:        $p \leftarrow p - 1$ 
34:        $\mathcal{Q} \leftarrow \mathcal{Q} - \{v \in \mathcal{Q}_{t,p} \mid w(v) \geq \overline{L}\}$ 
35:     end while
36:     if  $(k = 1)$  or  $(\mathcal{Q}_{t,p} \neq \emptyset)$  break
37:      $(t, p) \leftarrow (t - 1, M^{(t-1)})$ 
38:      $\mathcal{Q} \leftarrow \mathcal{Q} - \{v \in \mathcal{Q}_{t,p} \mid w(v) \geq \overline{L}\}$ 
39:   end while
40:   return  $(t, p, \mathcal{Q})$ 
41: end function
```

4.2.7 mobbRMSD

The discussion up to this point is summarized in **Algorithm 4**, which we call molecular-oriented RMSD with branch-and-bound (mobbRMSD). The mobbRMSD procedure takes input as a pair of molecular-oriented coordinates ($\{\mathbf{X}^{(k)}\}$ and $\{\mathbf{X}'^{(k)}\}$ for $k = 1, \dots, K$, respectively), and the intramolecular permutations determined by the molecular topologies

$$\left\{ \nu_1^{(k)}, \dots, \nu_{S^{(k)}}^{(k)} \right\}$$

for $k = 1, \dots, K$. The vertex weights of the tree can be calculated from the molecular pair covariance matrices $\mathbf{C}_{I,J,s}^{(k)}$ and the remainder matrices $F_{I,J}^{(k)}$ (**Algorithm 4**, lines 6 and 8). These matrices are stored in memory as global variables. Since the calculations of the elements of $\mathbf{C}_{I,J,s}^{(k)}$ and $(F_{I,J}^{(k)})$ are independent, it can be performed in parallel. The boundaries of the objective function, $\underline{L}$ and $\overline{L}$, are obtained using the DSTS algorithm (**Algorithm 4**, line 12), which are then scaled into RMSD. Finally, the real interval encompassing the exact solution of moRMSD is returned.

Algorithm 4 mobbRMSD for single component system

Require: $\mathbf{X}^*, \mathbf{X}'^* \in \prod_{k=1}^K \mathbb{R}^{d \times M^{(k)} n^{(k)}}$, $\mathcal{T}^* \subset \prod_{k=1}^K \mathfrak{S}_{n^{(k)}}$

Ensure: $\underline{\text{RMSD}}, \overline{\text{RMSD}} \in \mathbb{R}$

```

1:  $G^* \leftarrow \sum_{k=1}^K \sum_{I=1}^{M^{(k)}} \text{tr} \left[ \mathbf{X}_I^{(k)} \mathbf{X}_I^{(k)\top} + \mathbf{X}'_I^{(k)} \mathbf{X}'_I^{(k)\top} \right]$ 
2: for  $k \in \{1, \dots, K\}$  do
3:   for  $J \in \{1, \dots, M^{(k)}\}$  do
4:     for  $I \in \{1, \dots, M^{(k)}\}$  do
5:       for  $(s, \nu) \in \text{enumerate}(\mathcal{T}_k)$  do
6:          $\mathbf{P}_s^{(k)} \leftarrow \{\delta_{i\nu(j)}\}_{i,j \in \{1, \dots, n^{(k)}\}}$ 
7:          $\mathbf{C}_{I,J,s}^{(k)} \leftarrow \mathbf{X}'_J^{(k)} \mathbf{P}_s^{(k)} \mathbf{X}_I^{(k)\top}$ 
8:       end for
9:        $F_{I,J}^{(k)} \leftarrow -\max_{s', \mathbf{R}'} \text{tr} \left[ \mathbf{R}' \mathbf{C}_{I,J,s'}^{(k)} \right]$ 
10:    end for
11:  end for
12: end for
13:  $\mathcal{V}_{0,1} \leftarrow \left\{ \left( \prod_{k=1}^K \mathcal{M}^{(k)} \right) \times \left( \prod_{k=1}^K \mathcal{S}^{(k)} \right) \right\}$ 
14:  $(\underline{L}, \overline{L}) \leftarrow \text{DSTS}^*(\mathcal{V}_{0,1})$ 
15:  $n^* \leftarrow \sum_{k=1}^K M^{(k)} n^{(k)}$ 
16:  $\overline{\text{RMSD}} \leftarrow \sqrt{(G^* + 2\overline{L}) / n^*}$ 
17:  $\underline{\text{RMSD}} \leftarrow \sqrt{(G^* + 2\underline{L}) / n^*}$ 
18: return  $(\underline{\text{RMSD}}, \overline{\text{RMSD}})$ 
```

4.3 RESULTS

4.3.1 Implementation

The source code of **mobbrrmsd** was implemented in Fortran, and the libraries can be called from Fortran or Python interface. The spatial superposition algorithms are implemented for the spatial dimensions separately, and the appropriate version is selected when the library is called. The algorithms of Theobald, [189] Coutsiar *et al.*, [172] and Kabsch [164, 165] were used for $d=2, 3$, and other cases, respectively.

Both 32-bit and 64-bit floating point are supported, and all results in this paper use the 64-bit floating point version. A single mobbRMSD query is executed on a single core process, and OpenMP executes each query in parallel. All tests, including the following verification, were run on an Ubuntu system of Intel(R) i7-12700 2.1 GHz 20-core based processors. The GNU Compiler Collection (GCC) version 9.4.0 was used for compilation, and the OpenMP version 201511 was used for thread parallelization.

The correctness of the implementation was confirmed by the fact that the **mobbrrmsd** results of some matrix with uniform random numbers on $[0, 1]$ matched the results of the brute force search. The test systems are as follows: for one-component systems,

$$(n, M, S) = (1, 1, 1), (1, 2, 1), (1, 3, 1), (4, 1, 1), (4, 2, 1), (4, 3, 1), (4, 1, 2), (4, 2, 2), (40, 6, 1),$$

and two-component systems,

$$\begin{aligned} & \{ (n^{(1)}, M^{(1)}, S^{(1)}) , (n^{(2)}, M^{(2)}, S^{(2)}) \} = \\ & \{ (5, 1, 1), (5, 4, 1) \} , \{ (4, 2, 1), (5, 2, 1) \} , \\ & \{ (8, 2, 1), (4, 2, 2) \} , \{ (24, 3, 1), (24, 4, 1) \} , \end{aligned}$$

respectively. The above tests were performed for spatial dimensions $d = 2, 3$, and 4. All tests in the following sections are for $d = 3$.

4.3.2 Random coordinates

The random coordinates that represent an M molecular assembly of molecules consisting of n atoms are defined as

$$\mathbf{X}_{\text{sample}}(d, M, n, \alpha) = \sin \frac{\pi\alpha}{2} \mathbf{N}^{d \times Mn} + \cos \frac{\pi\alpha}{2} \sum_{j=1}^n \mathbf{N}^{d \times M} \otimes \mathbf{e}_j^\top, \quad (4.23)$$

where $\mathbf{N}^{d \times Mn} \in \mathbb{R}^{d \times Mn}$ and $\mathbf{N}^{d \times M} \in \mathbb{R}^{d \times M}$ are random matrices whose matrix elements are i.i.d and follow the standard normal distribution, respectively. $\mathbf{e}_j$ is the indicator matrix, i.e., $\mathbf{e}_j = \{\delta_{jj'}\}_{j' \in \{1, \dots, n\}} \in \mathbb{R}^{M \times 1}$ and $\otimes$ denotes the Kronecker product. The parameter $\alpha \in [0, 1]$ controls the molecular density, where $\alpha=1$ implies the dilution limit. Due to the reproducibility of the normal distribution, each component of the $\mathbf{X}_{\text{sample}}$ also follows the standard normal distribution.

4.3.3 Computational cost

In calculating mobbRMSD, the covariance matrices (Eq. 4.10) and the remainder matrices (Eq. 4.8) can be evaluated in polynomial time. In the following, the number of vertex evaluations in the DSTS process will be denoted N_{eval} . As M and S increase, the computational cost of mobbRMSD becomes approximately determined by the number of vertex weights computed in DSTS, N_{eval} . To check the quantitativity of this assumption, we check the correlation between N_{eval} and wall-clock time.

The Pearson's correlation coefficient is defined as

$$r = \frac{\sum_{i=1}^D (t_i - \bar{t})(N_i - \bar{N})}{\sqrt{\sum_{i=1}^D (t_i - \bar{t})^2 \sum_{i=1}^D (N_i - \bar{N})^2}} \quad (4.24)$$

where, N_i and t_i are sampled N_{eval} and wall-clock time, $\bar{N}$ and $\bar{t}$ are sample mean of N_i and t_i , respectively, and D is the number of data. r takes values between -1 and 1, implying a proportional relationship between those two variables when close to 1 or -1. M , n , S , and α , the correlation coefficients were calculated from 100,000 points based on random coordinates defined in Eq. 4.23.

Figure 4.4 shows the change in the Pearson's correlation coefficient for coordinates sampled from Eq. 4.23 vs. M , n , α , and S . For any n and α , as M increases the correlation coefficient approaches 1, remaining above 0.9 for $M > 6$ (**Figure 4.4a**). For $n=1,000$, the variance-covariance matrix calculation is overhead and the correlation coefficient becomes smaller overall, but remains above 0.8 for $M > 7$. Even in the range $M \leq 6$, the correlation coefficients approaches 1 as S increases (**Figure 4.4b**). These imply an asymptotic proportional relationship between N_{eval} and wall-clock time, indicating that N_{eval} is a good measure of mobbRMSD performance.

In the following, the number of vertex evaluations in the DSTS process will be denoted N_{eval} and treated as a computational cost. The best and worst cases of N_{eval} are

$$1 + \frac{1}{2} \sum_{k=1}^K M^{(k)}(M^{(k)} + 1)S^{(k)} \quad (4.25)$$

and

$$1 + \prod_{k=1}^K \sum_{p=1}^{M^{(k)}} \frac{M!}{(M-p)!} S^{(k)p}, \quad (4.26)$$

respectively. As is well known, BB does not improve the worst-case, and the question is, how much it improves the average computational cost.

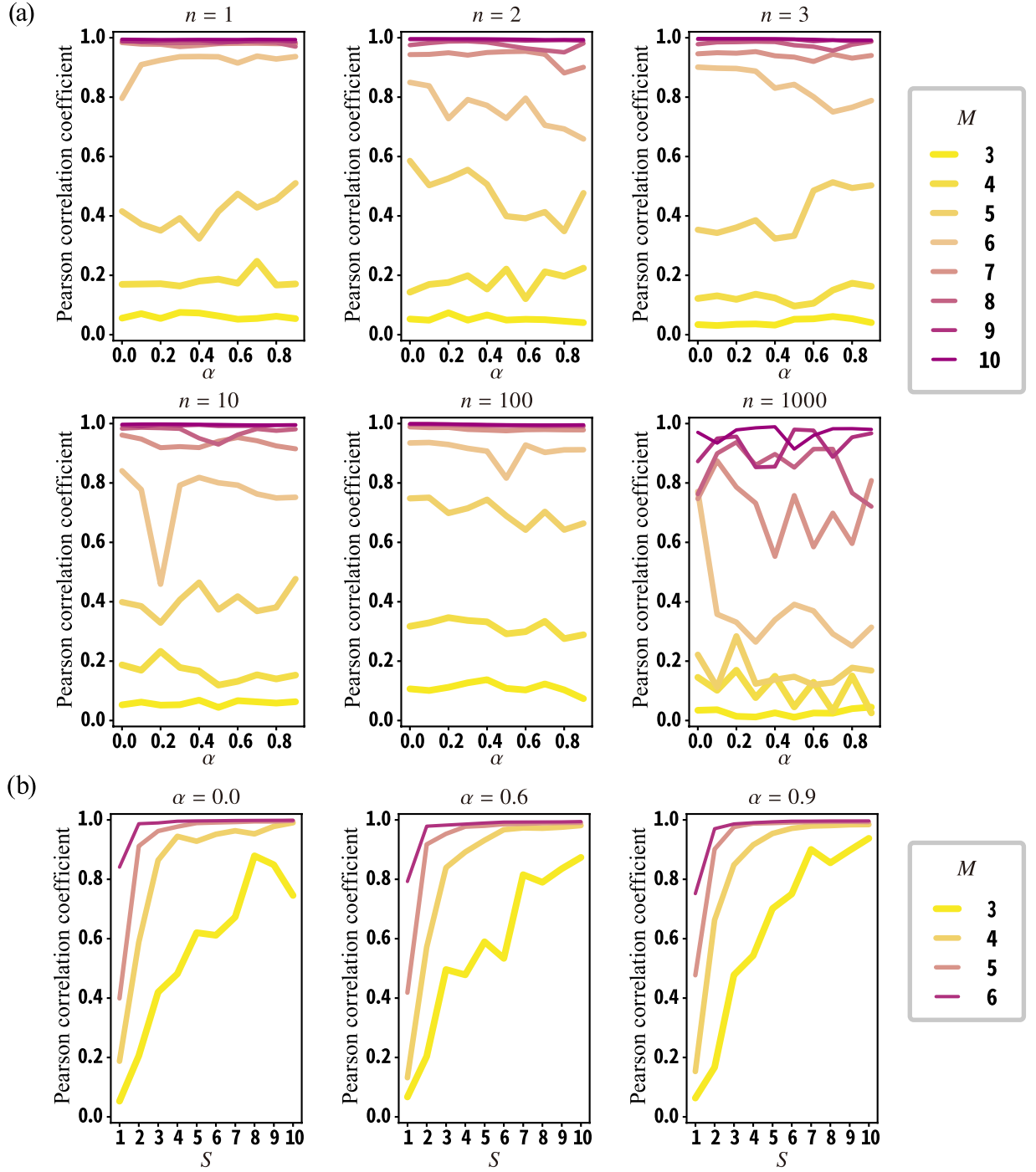


Figure 4.4: Pearson correlation coefficient between N_{eval} and wall-clock time. **(a)** Plots of Pearson correlation coefficients concerning M , n , and α . For all plots, $S=1$ was used. Each point was calculated from 10,000 sample coordinate pairs. **(b)** Plots of Pearson correlation coefficients concerning S , M , and α . For all points, $n=10$ was used and calculated from 10,000 samples.

4.3.4 Test with random coordinates

We first show the performance of mobbRMSD using randomly sampled coordinates, then compare it with previous methods. Here we discuss the pros, cons, and methodological position of mobbRMSD. Finally, we show the performance of mobbRMSD using molecular dynamics trajectories of molecular assemblies. The dependence of N_{eval} on M and S in random coordinates is summarized in **Figure 4.5**.

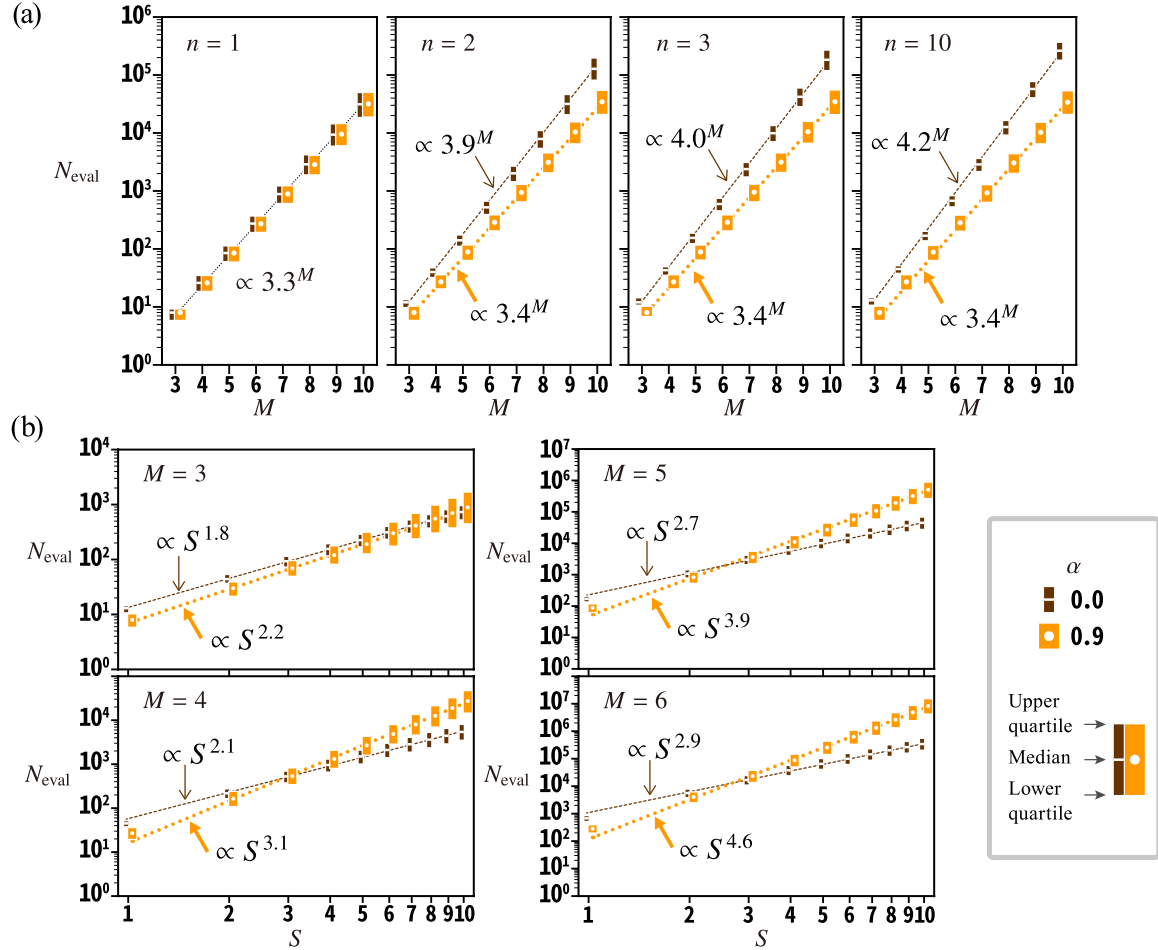


Figure 4.5: Computational cost scaling of mobbRMSD concerning M and S . (a) Semi-log plots of M vs medians of N_{eval} . All calculations were carried out for $S=1$ and used 20,000 and 10,000 random coordinate pairs for $M=3-7$ and $8-10$, respectively. (b) Log-log plots of S vs medians of N_{eval} . All calculations were carried out for $n=10$ and used 20,000 random coordinate pairs at each point.

When the number of intramolecular permutations is fixed at 1, *i.e.*, $S=1$, the semi-log plots of M versus N_{eval} show linear relationships regardless of n and α (**Figure 4.5a**). This means that the computational cost of mobbRMSD scales exponentially with respect to M , which improves the factorial scaling concerning M to exponential. On the other hand, when n is fixed at 10, a linear relationship appears in log-log plots for S versus N_{eval} , which implies polynomial scaling concerning S (**Figure 4.5b**). Here, the slopes of the M - and S -dependency lines represent the base and the power constant, respectively, and the average computational cost is proportional to $a^M S^b$ with a and b as constants.

The slopes of the M -dependent lines increased when $\alpha=0.0$ compared to when $\alpha=0.9$ (**Figure 4.5a**). Conversely, the slopes of the S -dependent lines increased when $\alpha=0.9$ compared to when $\alpha=0.0$ (**Figure 4.5b**). These trends represent the different effects of molecular sizes on M - and S -dependency. As α approaches 1, the molecular size becomes relatively smaller, coming closer to the dilution limit of the molecules and single-atom molecules ($n=1$). The M -dependence is related to the interchange of molecular positions, so it is advantageous if the molecules are spatially well separated. In contrast, since S is related to intramolecular permutations, larger molecular size allows for more efficient branch pruning.

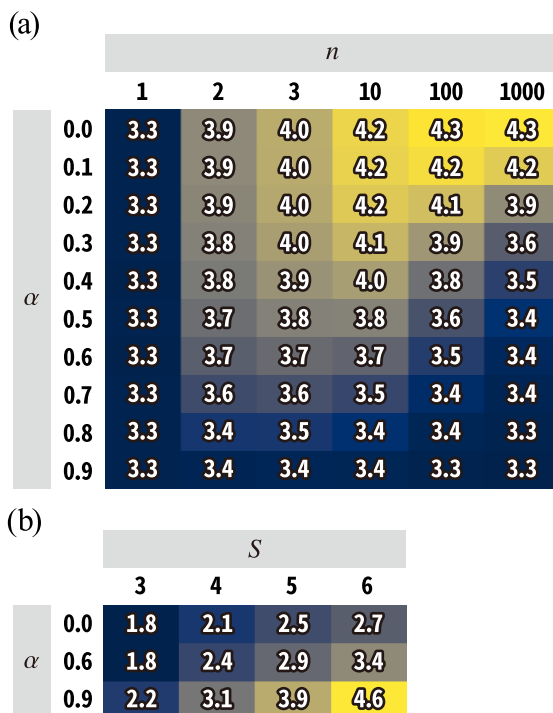


Figure 4.6: Heat maps of the slopes of the N_{eval} plots. **(a)** Slopes of the M - N_{eval} semi-log plots. Numbers inside the table represent the base to the power of M . **(b)** Slopes of the S - N_{eval} log-log plots. Numbers inside the table represent exponents in powers of S .

A closer examination of the slopes of the lines reveals nontrivial behavior regarding M -dependency (**Figure 4.6a** and **Figure 4.7**). Since the row with $\alpha=0.9$ implies almost the dilution limit, the slopes are close to 3.3 for any n . The row with $\alpha=0.0$ increases monotonically with n and approaches about 4.3. The analytical behavior of $n \rightarrow \infty$ is interesting, but will not be discussed further here. Focusing on rows with intermediate values of α , e.g., $\alpha=0.6$, the increase concerning n peaks at $n=3$ and then begins to decrease. This is supposed to be because as n increases, the variance of the RMSD values for different molecular matches increases, and branch pruning becomes more efficient.

The S -dependency is consistently more efficient for smaller α , and this is also true for the rate of increase of the slope, *i.e.*, the power constant. When S increases by a factor of 2 (from 3 to 6), it scales about 2 times in tandem when $\alpha=0.9$ (2.2 to 4.6), but only 1.5 times (1.8 to 2.7) when $\alpha=0.0$ (**Figure 4.6b**).

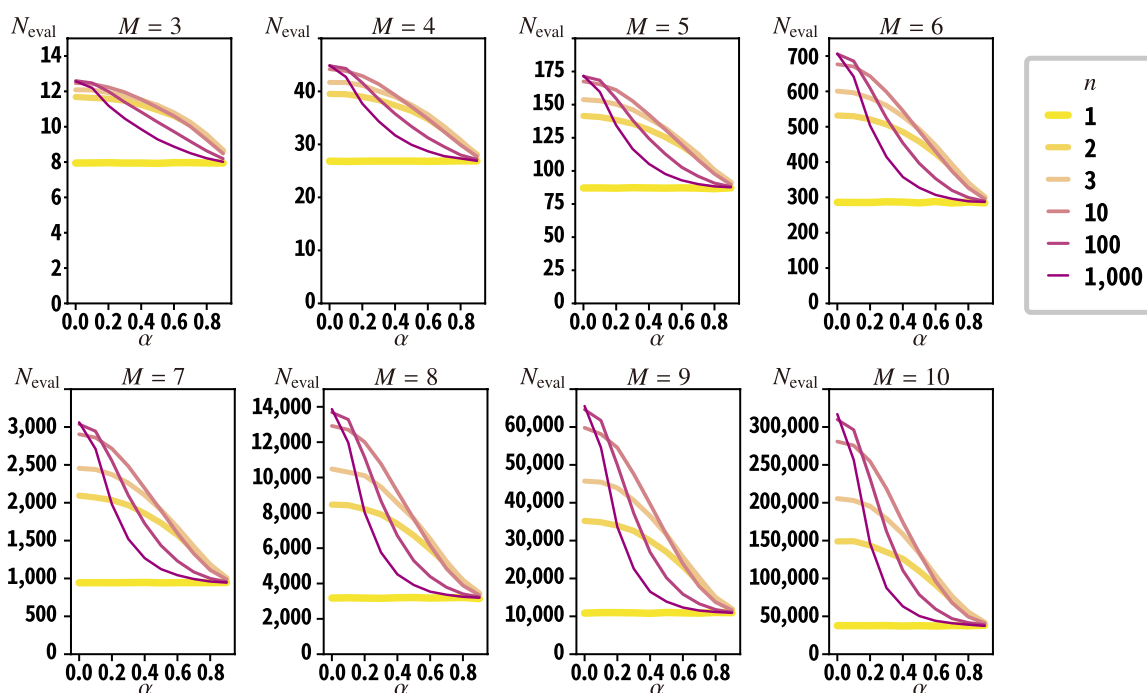


Figure 4.7: Relationship between α and the average value of N_{eval} . α is a parameter that controls the molecular density of random coordinates, where $\alpha=1$ implies the dilution limit. All calculations were carried out for $S=1$ and used 20,000 and 10,000 random coordinate pairs for $M=3-7$ and 8-10, respectively.

4.3.5 Comparison with previous studies

In this section, we compare mobbRMSD to two previous methods. The linear programming root-mean-square deviation (LP-RMSD) is applicable to general chemical structures and does not require parameter tuning for the system, which provides approximate solutions for generalized RMSD with simple operation. [192] LP-RMSD is computed in three steps: the optimal superposition with initial atom mapping, updated atom mapping with the Hungarian algorithm, and superposition with updated atom mapping. IsometryOpt developed by Nabika *et al.* takes a BB approach to the generalized RMSD problem that is similar to mobbRMSD. [133] IsometryOpt is also an exact solution method, as is mobbRMSD. The main difference is that IsometryOpt recursively partition the cubic parameter space as a grid, where the cube represents the three degrees of freedom of the rotation matrix.

LP-RMSD and IsometryOpt were not developed assuming molecular assemblies; although they can be applied to multiple molecules by partitioning the permutations processed by the Hungarian algorithm into multiple subspaces, there is no guarantee that the molecular topology will be preserved in such cases. Thus, comparisons were performed with coordinates sampled from the standard normal distribution with the centroid shifted to zero, which models a single-atom molecular system, *i.e.*, $n=1$.

For LP-RMSD, function **mdtraj.lprmsd** implemented in MDTraj (<https://github.com/mdtraj/mdtraj>, ver 1.10.0) was used, and for IsometryOpt, the software XG-RMSD (<https://github.com/striwata/X-GRMSD>) was compiled in GCC version 9.4.0 and used in the same environment as **mobbRMSD**. All computational wall-clock time measurements were performed with the built-in function in Python (ver 3.12.6), **time.perf_counter**.

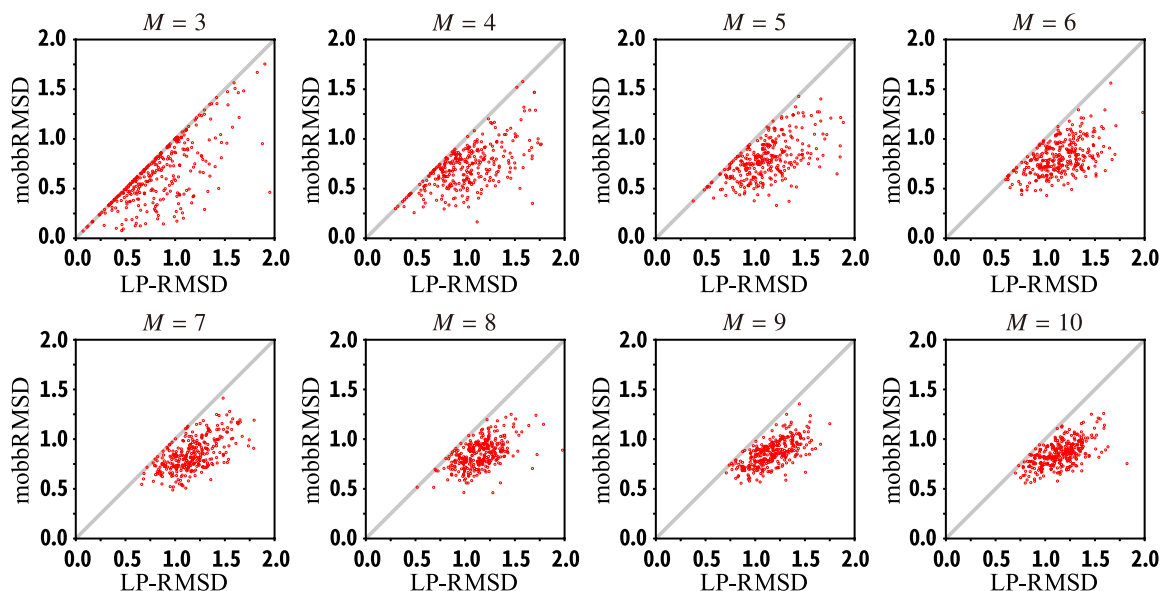


Figure 4.8: Comparison of mobbRMSD and LP-RMSD. Each point is calculated from the standard normal distribution with its centroid shifted to zero. 300 random points are depicted.

LP-RMSD results deviate from the mobbRMSD ones because LP-RMSD gives an ap-

proximate solution; scatter plots with LP-RMSD on the horizontal axis and mobbRMSD on the vertical axis are distributed in the lower half of the diagonal (**Figure 4.8**). **Table 4.1** summarizes the mean errors and the precisions for each M , where the mean error was defined as the expected difference between LP-RMSD and mobbRMSD, and the accuracy was defined as the frequency with which mobbRMSD and LP-RMSD coincide, *i.e.*, the percentage of points on the diagonal. The statistics were calculated on 10,000 sample coordinate pairs. The mean errors between LP-RMSD and mobbRMSD are distributed around 0.3, and their standard deviations become smaller as M increases (**Table 4.1**). The accuracy decreases with M , because the distribution narrows as M increases, and the intersection of the distribution and diagonal line decreases, as seen in **Figure 4.8**.

Figure 4.9 shows the distributions of computation times for mobbRMSD, IsometryOpt, and LP-RMSD at each M . Since IsometryOpt reports the exact solution, the computed RMSDs are equal to those of mobbRMSD, and what should be compared are the computation times (**Figure 4.10**). Up to $M=14$, mobbRMSD can report the exact solution faster on average than IsometryOpt. In contrast, IsometryOpt improves the scaling behavior polynomially above $M=10$, and at $M=15$, it reverses the order with mobbRMSD. The polynomial scaling of IsometryOpt is consistent with that reported previously. [133] LP-RMSD can be calculated at almost constant time (~ 100 microseconds) without variation for all M . However, mobbRMSD is faster for $M < 5$, likely due to the native compilation and overhead advantages.

As shown in the previous section, mobbRMSD scales exponentially concerning M , which is disadvantageous for larger M relative to IsometryOpt. This difference in scaling appears to refer to the nature of the solution space where BB is applied. That is, IsometryOpt divides the continuous parameter space of the rotation matrix, so wasteful searches are frequent with small M . However, since the rotation acts on the entire molecular system, the variational lower bound is evaluated while preserving the relative positions of each atom. This reduces the divergence of the branches due to the atom mapping, which is essentially factorial scaling. The branching of mobbRMSD allows independent rotations for each atom and does not consider their relative positions.

M	mean error	accuracy
3	0.22 (0.25)	0.21 ± 0.008
4	0.30 (0.25)	0.10 ± 0.006
5	0.33 (0.23)	0.060 ± 0.005
6	0.33 (0.20)	0.037 ± 0.004
7	0.32 (0.18)	0.025 ± 0.003
8	0.32 (0.17)	0.017 ± 0.003
9	0.31 (0.15)	0.012 ± 0.002
10	0.30 (0.14)	0.007 ± 0.002

Table 4.1: Mean errors and accuracies of LP-RMSD. The mean errors and standard deviations were calculated as the mean difference from mobbRMSD and accuracies were calculated as the frequency with which the mobbRMSD value was matched. The numbers in parentheses accompanying the mean errors and added to the accuracies represent the standard deviation and the 95% confidence interval of the binomial distribution. 10,000 structural pairs sampled from the standard normal distribution were used.

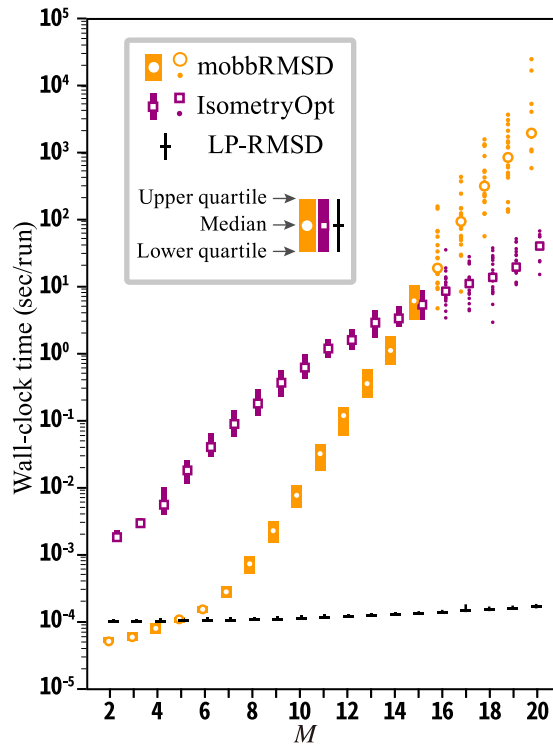


Figure 4.9: Comparison of computation time distributions. Statistics on wall-clock times for RMSD calculations for structure pairs generated from the standard normal distribution are indicated. The sample numbers are 10,000, 1,000, 100, 20, and 10 for $M=2-7$, 8-10, 11-15, 16-19, and 20 for mobbRMSD and IsometryOpt, respectively, and 50,000 in all for LP-RMSD.

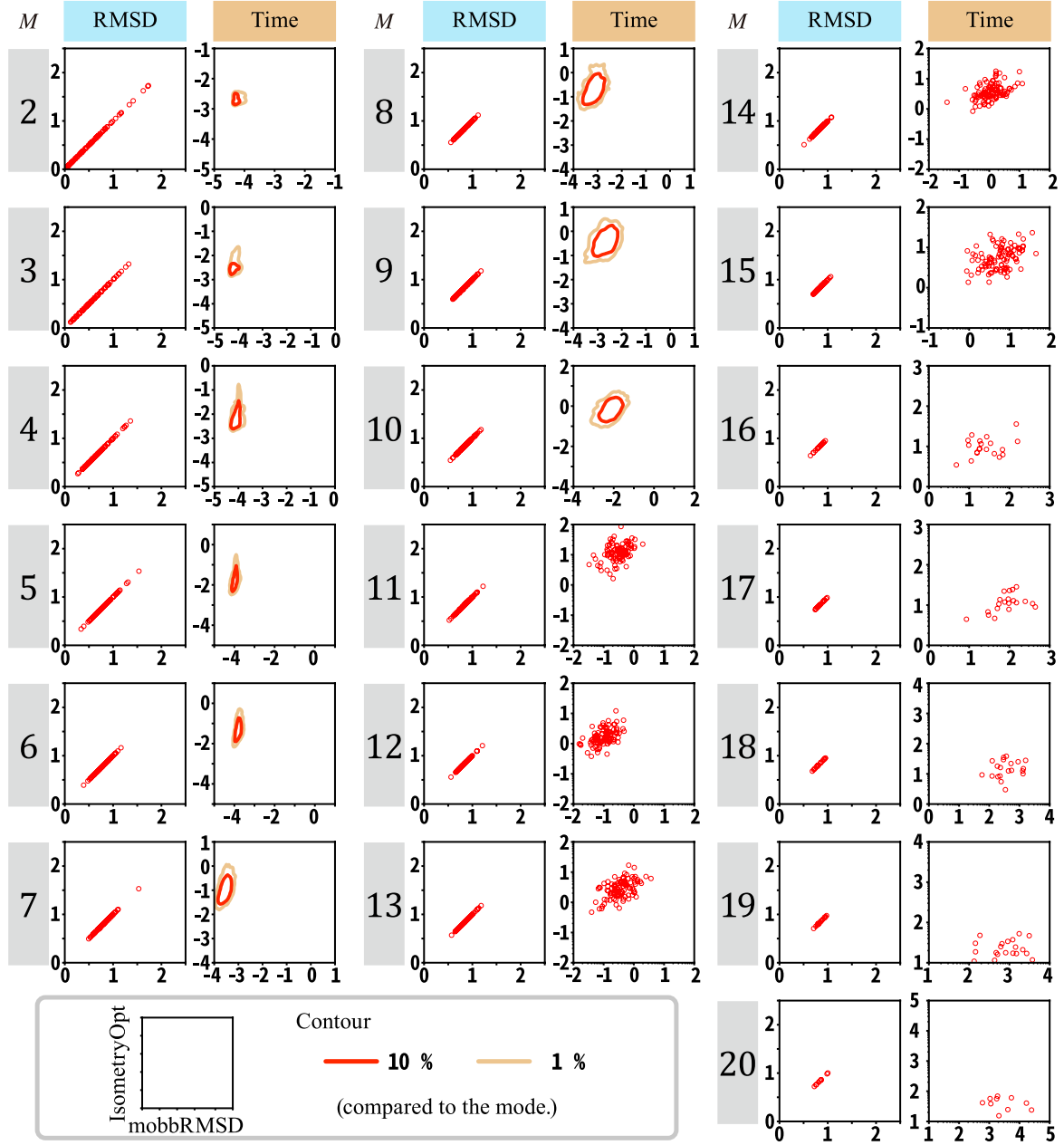


Figure 4.10: Performance comparison between mobbRMSD and IsometryOpt. All scatter plots are depicted with the values of mobbRMSD and IsometryOpt on the horizontal and vertical axis, respectively. The RMSD plots are in dimensionless units. The time plots represent wall-clock time and are on a base 10 log-log scale with exponents on the ticks. The sample numbers are 10,000, 1,000, 100, 20, and 10 for $M=2-7$, 8-10, 11-15, 16-19, and 20, respectively. Contour distributions were estimated by the kernel density estimation with a bandwidth of 0.2.

4.3.6 Test with aromatic micelles

We demonstrate the application of mobbRMSD for macromolecular assembly systems, aromatic micelles. Aromatic micelles are a type of amphiphilic molecule proposed by Yoshizawa *et al.*[2, 5] that contain aromatic moieties in the backbone and form dynamic self-assemblies of multiple molecules in water. V-shaped polyaromatic amphiphiles (VPAA) are aromatic micelles, amphiphilic molecules with anthracene panels aligned in a V-shape (**Figure 4.11a**). **VPAA1** and **VPAA2** differ only in their side chains, each having ammonium cations and zwitterionic sulfobetaine groups, respectively. Microseconds molecular dynamics simulations show that VPAAs spontaneously self-assemble and disassemble in water, and at equilibrium, their size distributions consistent with experimental results [23, 114] are observed (**Figure 4.11b**). It was defined in the present analysis that if the nearest-neighbor distance between aromatic carbon atoms between molecules was less than 5 Å they belong to the same assembly (**Figure 4.11c**), and assembly sizes of **VPAA1** and **VPAA2** are distributed between 5-8 and 4-7 molecules, respectively. Due to the incomplete stacking interactions of the V-shaped anthracenes, the VPAA assemblies have many quasi-stable conformers and can dynamically transition between them with time constants of nanoseconds.

We computed the mobbRMSD between the aromatic moieties of VPAAs which are relatively rigid, and the flexible side chains were disregarded. The number of atoms per molecule n and of **VPAA1** and **VPAA2** are both 34, and the C_2 molecular symmetry is considered, *i.e.*, $S=2$. The 10 molecular assemblies of VPAAs contain 340 atoms, making it difficult to calculate exact RMSD using conventional methods. Hereafter, M -molecular assemblies of **VPAA1** and **VPAA2** are denoted as $(\text{VPAA1})_M$ and $(\text{VPAA2})_M$, respectively. **Figure 4.11d** shows mobbRMSD progress records applied to $(\text{VPAA1})_{10}$. The three samples contain examples where an exact solution is found early in the process and the candidate solution is repeatedly updated.

Figure 4.12 shows the distributions of N_{eval} and wall-clock time for **VPAA1** and **VPAA2**, where the best and worst cases in **Figure 4.12a** are given by Eq. 4.25 and 4.26, and full search implies $M!S^M$. As discussed in the previous sections, the best and worst cases are scaled by M^2 and factorial, respectively, and the average N_{eval} is again scaled exponentially (**Figure 4.12a**). **VPAA1** and **VPAA2** differ only in side chains and hydrophilic groups, which affect the distributions of conformers. This difference also affects the distributions of N_{eval} , but not significantly. However, the influence of molecular structure is evident in the distributions when compared to random coordinates (**Figure 4.12b**). The wall-clock times required for coordinates generated from the standard normal distribution for the same number of atoms is much slower than for VPAAs; for example, for $M=10$, a calculation that takes an average of seconds with random coordinates completes in about 100 milliseconds for VPAAs.

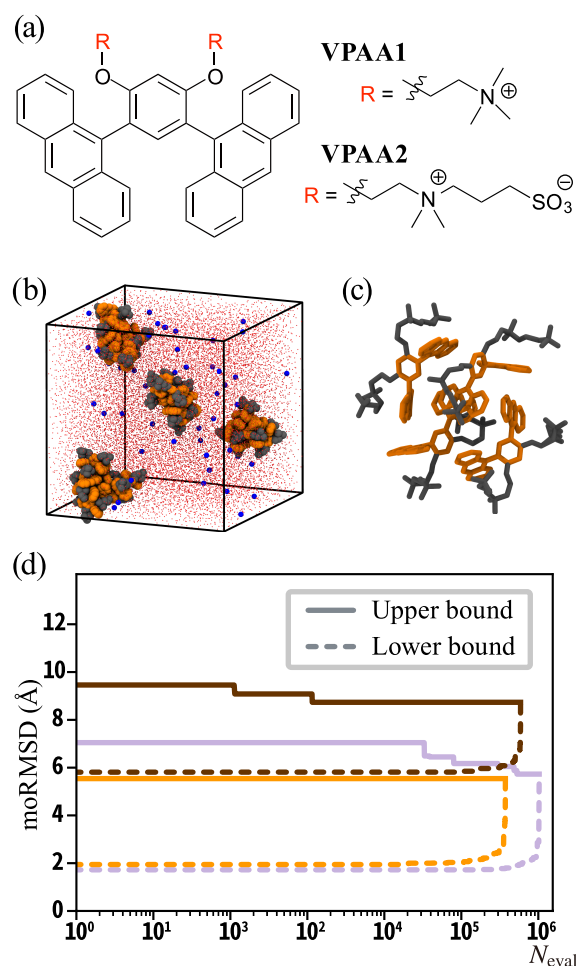


Figure 4.11: VPAA (a) Chemical structures of **VPAA1** and **VPAA2**. (b) Snapshot of molecular dynamics simulation of **VPAA1**. The aromatic moieties and the side chains are indicated in orange and black spheres, the red dots are water, and the blue spheres are the counter ions (Cl^-), respectively. (c) Snapshot of 4 molecular assembly of **VPAA2** observed in the simulation (denoted $((\text{VPAA2})_4)$). (d) Behavior of upper and lower bound of mobbRMSD with respect to N_{eval} for 3 samples of $(\text{VPAA1})_{10}$.

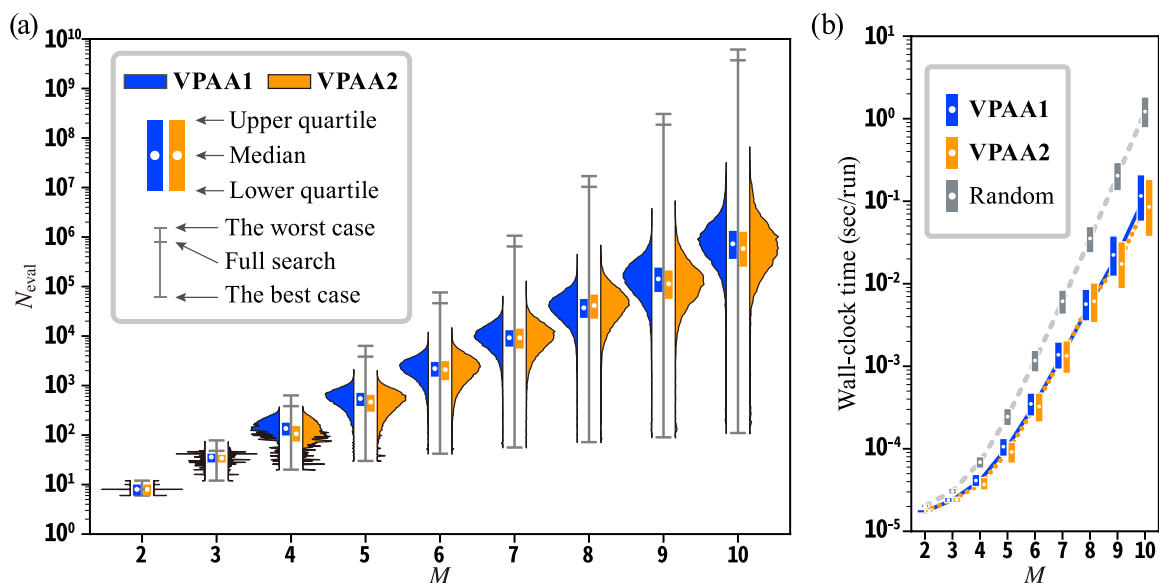


Figure 4.12: Performance of mobbRMSD for VPAA1 and VPAA2. **(a)** Semi-log plot of the distribution of N_{eval} concerning M . VPAA1 and VPAA2 are indicated in orange and blue violin plots, respectively. Box plots attached to the violins indicate the quartiles, with the top, bottom, and circles being the upper, lower quartile, and median, respectively. Horizontal lines attached to gray vertical bars from top to bottom indicate N_{eval} of the worst-case calculations, the number of superposition evaluations required for a full search, and N_{eval} of the best-case calculations, respectively. Using structures randomly taken from MD trajectories, 100,000 pairs were sampled in each system. **(b)** Semi-log plot of the wall-clock time concerning M . The same samples as for VPAA1 and VPAA2 were utilized for a, and the random system used 10,000 pairs generated from the standard normal distribution with $n = 34$.

4.4 DISCUSSION

4.4.1 Methodological characteristics

Given the results of the computational experiments, we clarify the methodological characteristics of mobbRMSD. First, for problems with large M ($M > 20$), it is difficult to obtain an exact solution using mobbRMSD. In such cases, either mobbRMSD must be terminated early or IsometryOpt is used. However, IsometryOpt requires wall-clock times on the order of milliseconds to seconds, which limits its applicability to tasks that require fast throughput, such as clustering large data sets. In such cases, approximate methods such as LP-RMSD are an alternative. Because LP-RMSD does not provide any theoretical guarantees such as variational lower bounds, the appropriateness of the approximation must be carefully judged for each system.

The advantage of mobbRMSD over previous methods is that molecules can be treated explicitly. As the number of atoms per molecule, n , increases, mobbRMSD requires almost no additional computational cost. In the range of $M < 15$, it is faster than IsometryOpt, and faster than LP-RMSD while still providing an exact solution in the range of $M < 5$ (**Figure 4.9**). Since mobbRMSD can handle molecular assemblies and multicomponent systems, even in this strict limit, there is no doubt that many interesting systems are included here such as molecular liquids, solute solvation, and macromolecular self-assembly. Moreover, early termination of mobbRMSD as an approximation may be a good choice in some cases. One application is the asymptotic polynomial scaling minimum spanning tree algorithm that exploits theoretical bounds on mobbRMSD, which will be discussed in a later section.

4.4.2 Correlation between RMSD and computational cost

In RMSD applications, often only similar structure pairs are important. For example, in the task of improving drug molecules, positive candidates are those within RMSD neighborhood of the reference, [12, 148] and in the context of structural clustering, the minimum spanning tree-based methods are often used. [158] Since moRMSD is locally proportional to the multidimensional Euclidean distance, target structures with a certain value are distributed on a hypersphere centered on the reference structure. As moRMSD increases, the number of structures that fit the value increases rapidly and the information becomes less valuable.

Figure 4.13a shows the correlation between moRMSD and N_{eval} for VPAAAs. The shapes of the distributions are slightly different for $(\text{VPAA1})_{10}$ and $(\text{VPAA2})_{10}$, with the latter being slanted. The distributions of N_{eval} change significantly bordering moRMSD of 5 Å for both $(\text{VPAA1})_{10}$ and $(\text{VPAA2})_{10}$. N_{eval} has a wide range reaching a maximum of 10^8 in the region larger than 5 Å but does not exceed 10^3 in the region below 5 Å.

In order to quantify the effect of focusing on the small moRMSD region in **Figure 4.13a**, we define a conditional quantile of N_{eval} . Let the data points be

$$\mathcal{D} = \{(R_i, N_i) | i = 1, \dots, D\},$$

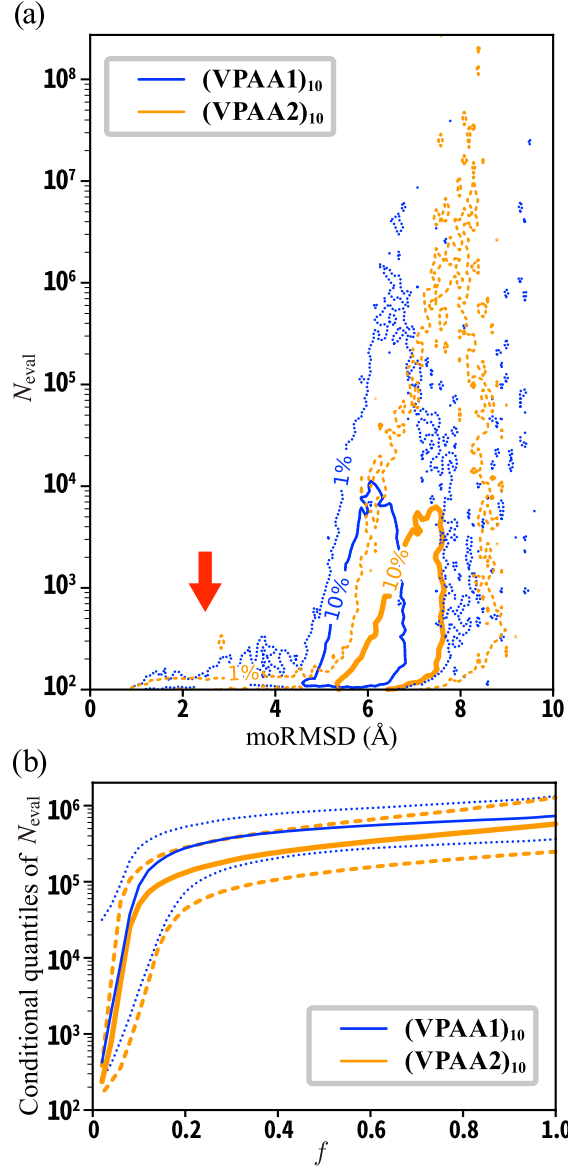


Figure 4.13: Correlation between RMSD and N_{eval} . **(a)** Contour maps of simultaneous distributions of moRMSD and N_{eval} . The blue and orange contours indicate (VPAA1)₁₀ and (VPAA2)₁₀, respectively. The distributions were estimated by kernel density estimation with a bandwidth of 0.2 for 100,000 samples. **(b)** Conditional quantiles by moRMSD of N_{eval} , and dotted lines above and below the solid line indicate the upper and lower quantiles, respectively. The conditioning is $\text{moRMSD} < R(f)$, where $R(f)$ is a percentile of f for moRMSD.

where R_i and N_i are moRMSD and N_{eval} , respectively. The conditional quartile is defined for $f \in [0, 1]$ as the quartiles of

$$\{N_i | (R_i, N_i) \in \mathcal{D}, R_i \leq R(f)\},$$

where $R(f)$ is the percentile of the fraction f of

$$\{R_i | (R_i, N_i) \in \mathcal{D}\}.$$

The plot of the conditional quantiles decreases monotonically as f decreases, with a steep drop near $f=0.1$ (**Figure 4.13b**). The steep drop suggests that targeting only small moRMSDs can make the computations more efficient. The correlation between moRMSD and N_{eval} and the steep drop in the conditional quantile are observed in both VPAAAs, independent of M (**Figures 4.14** and **4.15**).

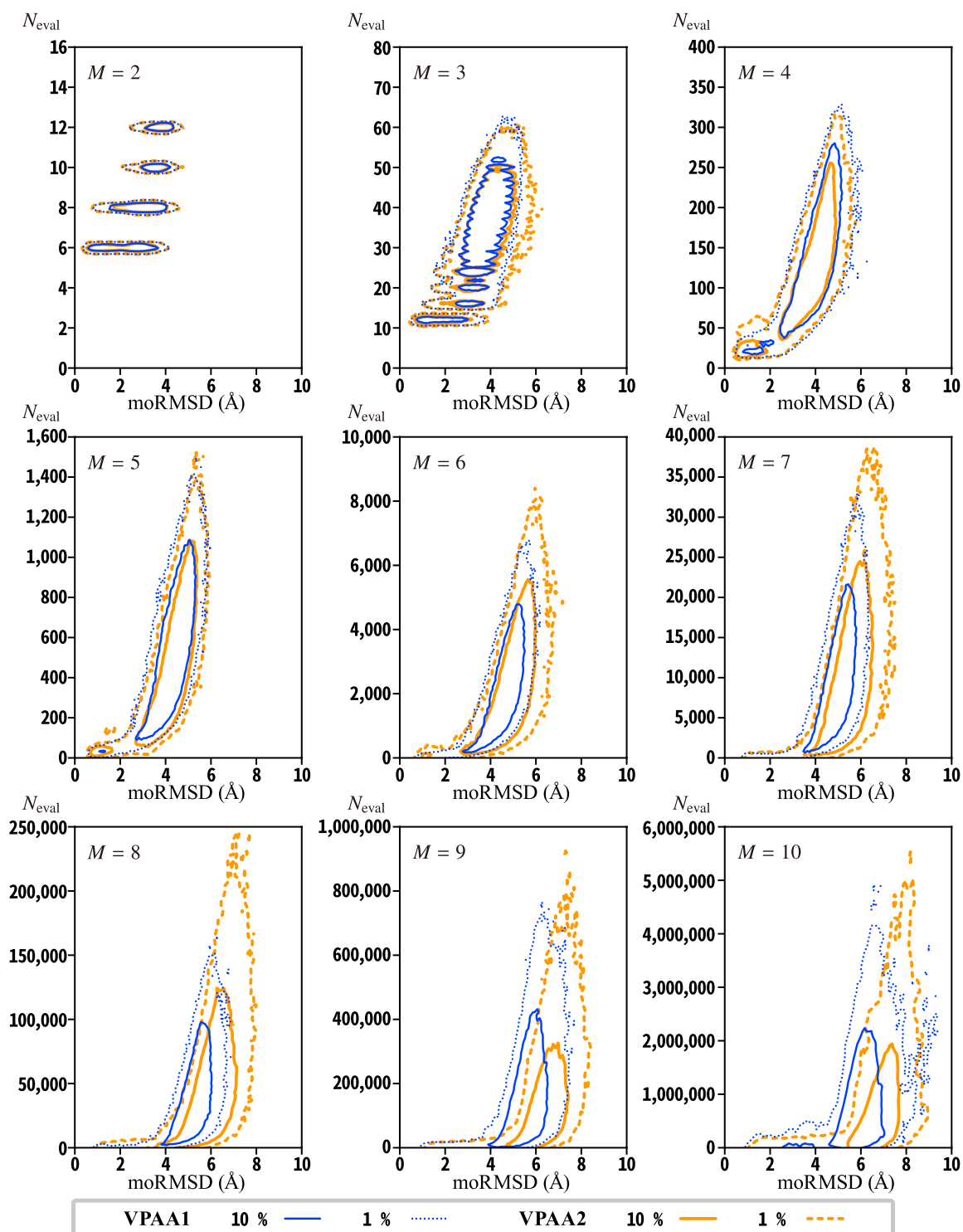


Figure 4.14: Simultaneous distribution of moRMSD and N_{eval} for VPAA1 and VPAA2. The blue and orange contours indicate VPAA1 and VPAA2, respectively. The distributions were estimated by the kernel density estimation with a bandwidth of 0.2 for 100,000 samples.

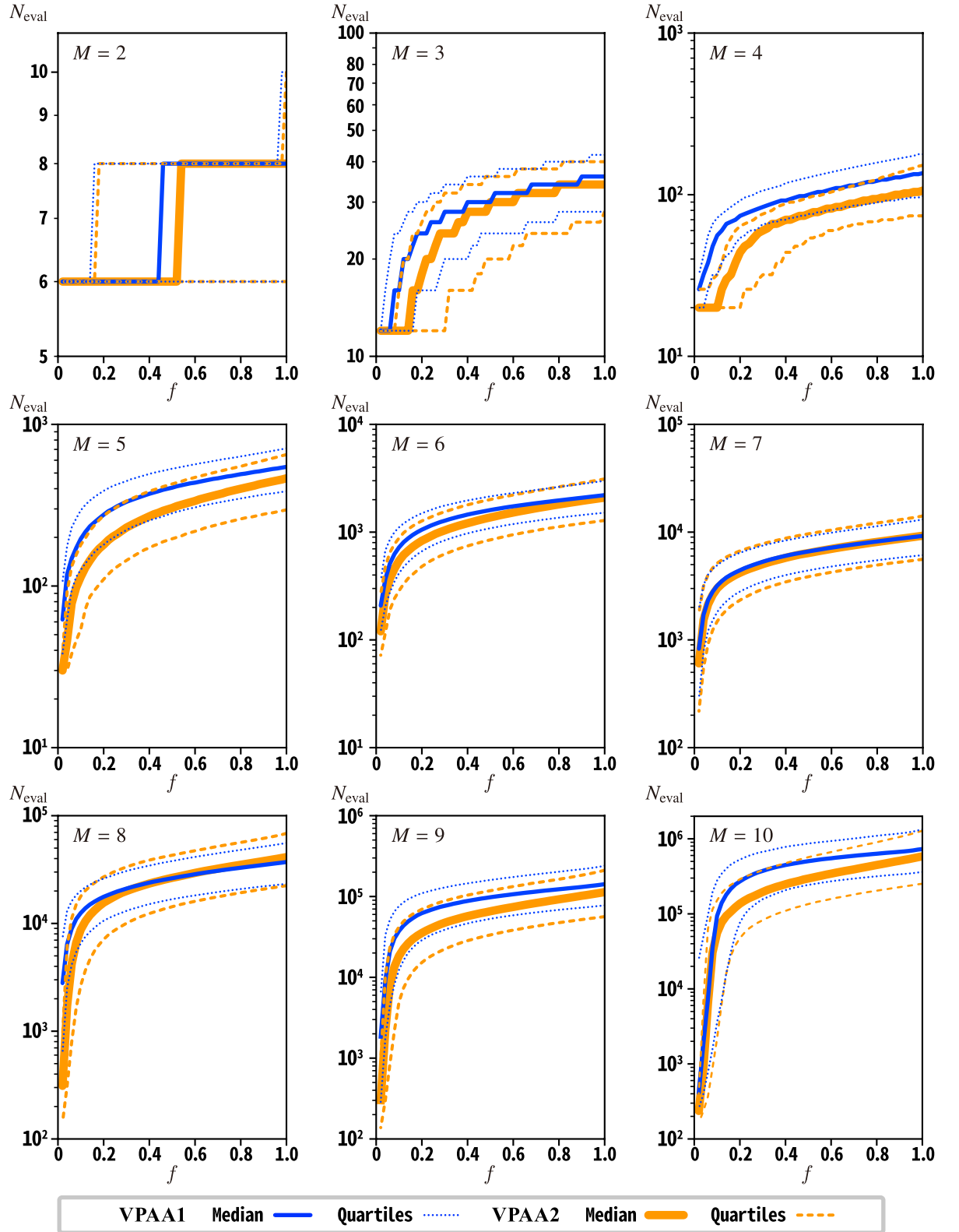


Figure 4.15: Conditional quantiles by moRMSD of N_{eval} . Dotted lines above and below the solid line indicate the upper and lower quartiles, respectively. The conditioning is $\text{moRMSD} < R(f)$, where $R(f)$ is a percentile of f for moRMSD.

4.4.3 Early termination

The strong correlations between RMSD and computational costs observed in the previous section raise the following question; if there is no prospect of structural similarity, can the calculation be terminated early? This is possible in some cases. Hereafter, terminating the computation at the first leaf evaluation of the DSTS is referred to as the greedy method. **Figure 4.16** shows the distributions of the lower and upper bounds, $\underline{\text{RMSD}}$ and $\overline{\text{RMSD}}$, which are given by the greedy method. Since $\underline{\text{RMSD}} \leq \overline{\text{RMSD}}$, no distribution appears in the lower half of the diagonal. As M increases, the distributions shift upward, which means that on average the gap between the lower and upper bounds widens. The populations left out of the shift, indicated by the arrows in **Figure 4.16**, appear in the lower left of the distribution, but it does not appear in the data of the standard normal distribution; these populations are unique to molecular data because stable structures appear with high frequency and pairs of similar ones are picked up frequently.

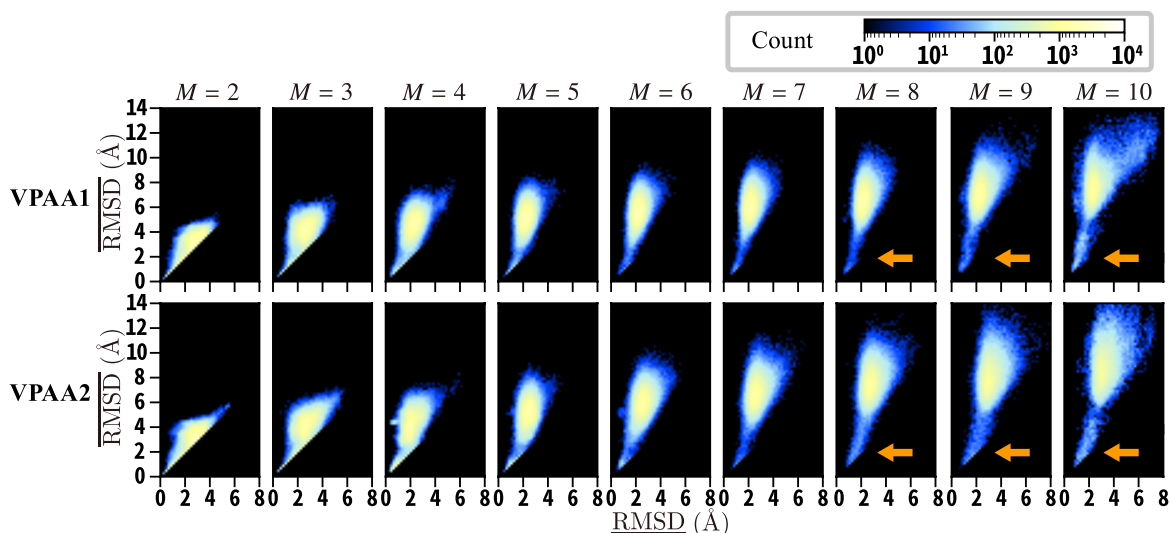


Figure 4.16: Distributions of lower and upper bounds of RMSD by the greedy method. $\underline{\text{RMSD}}$ and $\overline{\text{RMSD}}$ on the horizontal and vertical axis, respectively, are defined by **Algorithm 4**.

VPAA1 and **VPAA2** show nearly the same trend, but for $M > 7$, the peak positions of the distributions for **VPAA2** are slightly to the right of the **VPAA1** ones. This difference seems small but has a significant role. Suppose that only points below $2.0 \text{ \AA}$ are of interest. If the point lies where $\underline{\text{RMSD}} > 2.0 \text{ \AA}$, then the solution never has a value below $2.0 \text{ \AA}$, and the calculation can be terminated at that point. Applying this condition to $(\text{VPAA1})_{10}$, the boundary with $\underline{\text{RMSD}}=2.0 \text{ \AA}$ passes through the peak of the distribution, so about half of the points will survive. In contrast, for $((\text{VPAA2})_{10})$, the boundary passes to the left of the peak of the distribution, indicating that most points need no further calculation. Thus, for the case of $(\text{VPAA2})_{10}$, only the cost of the greedy method can determine which points should be computed.

Here we introduce a cutoff parameter $\lambda_{\underline{\text{RMSD}}} \in [0, \infty]$ and consider imposing a termina-

tion condition of

$$\lambda_{\text{RMSD}} < \sqrt{\frac{1}{n^*} (G^* + 2L)} \quad (4.27)$$

at the DSTS algorithm (**Algorithm 3**). Then mobbRMSD can be used as an algorithm to evaluate the exact solution if it is less than λ_{RMSD} and the interval $[\underline{\text{RMSD}}, \overline{\text{RMSD}}]$ otherwise, where $\lambda_{\text{RMSD}} < \underline{\text{RMSD}}$ holds. The average computational costs and wall-clock times with the early termination for each λ_{RMSD} are shown in **Figure 4.17**. As λ_{RMSD} approaches zero, the log-log plots for M and N_{eval} approach a straight line, since almost all calculations are terminated on the greedy method. For $\lambda_{\text{RMSD}}=2.0$ Å, (VPAA1)₁₀ is still exponential scaling, but (VPAA2)₁₀ is almost polynomial. If λ_{RMSD} is further reduced to 1.0 Å, both (VPAA1)₁₀ and (VPAA2)₁₀ become polynomial scaling.

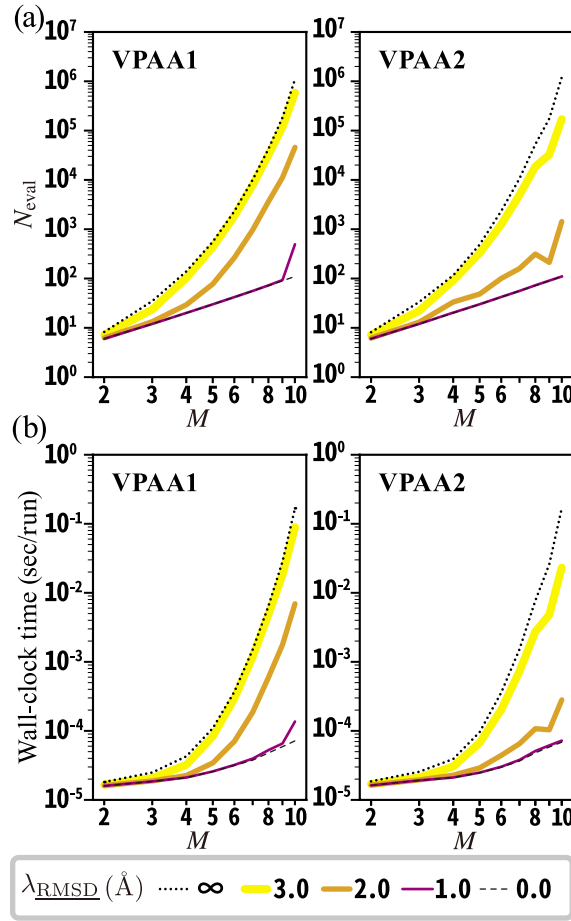


Figure 4.17: Change in computational scaling as the cutoff parameter, λ_{RMSD} is varied. (a) log-log plots of N_{eval} for M . (b) log-log plots of wall-clock time for M .

In terms of wall-clock time, even $M=10$ can be performed in submilliseconds (**Figure 4.17b**), which is comparable in speed to approximate methods such as LP-RMSD (**Figure 4.9**). The scaling behavior by a cutoff in **Figure 4.17** shows that the speed at which the average computational cost improves to polynomial as λ_{RMSD} decreases depends on the system.

4.4.4 Minimum spanning tree

Finally, we mention an application of mobbRMSD to structural clustering. For structural data,

$$\mathcal{X} = \{\mathbf{X}_1^*, \dots, \mathbf{X}_D^*\},$$

the distance matrix of moRMSD can be computed as a symmetric matrix¹ and a minimum spanning tree (MST) can be constructed from this matrix. [193, 194] For a given weighted undirected graph, a spanning tree is a tree that contains all of its vertices, and MST is the one with the smallest sum of edge weights. [195]

Now, we consider a distance matrix

$$\mathbf{R}_{\lambda_{\text{RMSD}}} \in [0, \infty)^{D \times D}$$

where each entry is $\text{RMSD}_{i,j}$ which is computed by the lower bound of mobbRMSD for $\mathbf{X}_i^*, \mathbf{X}_j^* \in \mathcal{X}$ using cutoff λ_{RMSD} . Let the MST constructed from the matrix $\mathbf{R}$ be

$$\{\mathcal{V}_{\lambda_{\text{RMSD}}}, \mathcal{E}_{\lambda_{\text{RMSD}}}\},$$

where $\mathcal{V}_{\lambda_{\text{RMSD}}}$ is $\{1, 2, \dots, D\}$ and $\mathcal{E}_{\lambda_{\text{RMSD}}}$ is the set of pairs of indices and their weights with $(D - 1)$ members, *i.e.*, $e_k \in \mathcal{E}_{\lambda_{\text{RMSD}}}$ is given by

$$e_k = \{i_k, j_k, \text{RMSD}_{i_k, j_k}\}.$$

If

$$\text{RMSD}_{i_k, j_k} < \lambda_{\text{RMSD}}$$

then it is the exact solution, so if the maximum edge weight in MST is less than λ_{RMSD} , then all weights in MST are exact. Also, since MST is a spanning tree with the minimum sum of edge weights, no entry in the distance matrix $\mathbf{R}_{\lambda_{\text{RMSD}}}$ can be included in MST which could have a solution smaller than the maximum weight in MST. Therefore, the exactness of MST is also guaranteed. To efficiently build MST by mobbRMSD, we just need to find the appropriate cutoff depending on the system.

Although the computational cost of mobbRMSD is independent of the number of data points, the maximum edge weights of MST depend on the density of data points; as the number of data points D increases, the data point becomes denser. λ_{RMSD} that constitutes the exact MST can therefore be made smaller. This indicates that by increasing the data size, the cost of constructing MST by mobbRMSD may asymptotically approach polynomial costs.

It would also be possible to have an algorithm that automatically determines the appropriate cutoff parameter. One idea is to first create the distance matrix of $\text{RMSD}_{i,j}$ with a sufficiently small cutoff λ_0 , and then create MST. Here, edges with weights above λ_0 may not be correct. Hence, with an updated λ_1 , *e.g.*, the maximum weight of the MST, only the entries of the distance matrix corresponding to edges above λ_0 are recalculated. For the recalculated distance matrix, update the MST. The above procedure is repeated an exact MST will eventually be determined. The implementation and performance measurement of such automatic determination is a subject for future research.

¹Because moRMSD does not satisfy the triangular inequality, this matrix is not strictly a distance matrix. It therefore cannot be used in algorithms that assume a metric space.

4.5 CONCLUSION

A branch-and-bound (BB) algorithm, mobbRMSD, has been developed to find an exact solution of RMSD. It uses molecular-oriented coordinates that allow for easy incorporation of molecular topologies. The calculations of RMSD are inherently factorial in complexity. mobbRMSD improves the average scaling exponentially to the number of molecules M and polynomially to the number of intramolecular permutations S . The wall-clock time has been measured for random test coordinates, and confirmed to be sub-milliseconds to milliseconds for M up to 10, a speed applicable to high throughput analysis. The performance is independent of the number of atoms per molecule. Comparing mobbRMSD with a similar previous method, IsometryOpt, mobbRMSD is faster for $M < 15$, but as M increases, the calculation slows down due to poor scaling. The exponential scaling concerning M is an obvious drawback, but there is room for improvement in the pruning rule and branching strategy. Currently, mobbRMSD has loose restrictions on the relative positions of molecules, and this may be improved by using additional variational lower bounds and/or branching of the solution space.

Even if the computation terminates early, mobbRMSD provides upper and lower bounds on the interval containing the exact solution. From this, introducing a cutoff parameter λ_{RMSD} in the termination condition improves the computational cost up to M^2S while guaranteeing exact solutions in cases where the solution is smaller than λ_{RMSD} . This allows the minimum spanning tree with mobbRMSD to be computed efficiently for series data, and the cutoff parameter can be reduced as the data increases. Automated determination of the optimal cutoff parameter is also a future topic.

Compared to targets of conventional methods, mobbRMSD can be applied to wide multiple molecular systems, including molecular liquids, solute solvation, macromolecular self-assemblies, and so on. Minimum spanning tree can be used for many structural clustering methods, shedding light on soft molecular assemblies that have been difficult in the past, and paving the way for precise structural analysis.

4.6 DATA AND SOFTWARE AVAILABILITY

The code for performing and testing mobbRMSD can be found on Github (<https://github.com/yymmt742/mobbrmsd>) and is available for reuse under the MIT licence. Trajectory data for VPAAAs to reproducing this work are available at Zenodo (<https://doi.org/10.5281/zenodo.14232873>).

4.7 APPENDIX

Let $d, M, n, S \in \mathbb{N}$ and $\mathcal{M}, \mathcal{S}$ and $\mathcal{R}$ be the finite symmetric group on $\{1, \dots, M\}$, the total set of mappings from $\{1, \dots, M\}$ to $\{1, \dots, S\}$, and the total set of rotation matrices on $\mathbb{R}^d$, respectively. The objective function is organized and reiterated Eq. 4.6.

Definition 1 (Objective function). For $\mathbf{X}_1, \dots, \mathbf{X}_M, \mathbf{X}'_1, \dots, \mathbf{X}'_M \in \mathbb{R}^{d \times n}$, $\mathbf{P}_1, \dots, \mathbf{P}_S \in \mathbb{R}^{n \times n}$, and $(\mu, \sigma) \in \mathcal{M} \times \mathcal{S}$, define the following functions:

$$L(\mu, \sigma) = -\max_{\mathbf{R} \in \mathcal{R}} \sum_{I=1}^M \text{tr} [\mathbf{R} \mathbf{X}'_{\mu(I)} \mathbf{P}_{\sigma(I)} \mathbf{X}_I^\top]$$

Definition 2 (Variational lower bound). Define

$$l_0 = \min_{\mu' \in \mathcal{M}} \sum_{I=1}^M F_{I, \mu'(I)},$$

where

$$F_{I, J} = -\max_{\substack{\mathbf{R}' \in \mathcal{R}, \\ s \in \{1, \dots, S\}}} \text{tr} [\mathbf{R}' \mathbf{X}'_J \mathbf{P}_s \mathbf{X}_I^\top].$$

4.7.1 Proof of Eq. 4.7

Theorem 1. $l_0 \leq \min_{(\mu, \sigma) \in \mathcal{M} \times \mathcal{S}} L(\mu, \sigma)$ holds.

proof.

$$\begin{aligned} & \min_{(\mu, \sigma) \in \mathcal{M} \times \mathcal{S}} L(\mu, \sigma) \\ &= \min_{(\mu, \sigma) \in \mathcal{M} \times \mathcal{S}} \left(-\max_{\mathbf{R} \in \mathcal{R}} \sum_{I=1}^M \text{tr} [\mathbf{R} \mathbf{X}'_{\mu(I)} \mathbf{P}_{\sigma(I)} \mathbf{X}_I^\top] \right) \quad (\Leftarrow \text{Definition of } L) \\ &\geq \min_{(\mu', \sigma') \in \mathcal{M} \times \mathcal{S}} \sum_{I=1}^M -\max_{\mathbf{R}_I \in \mathcal{R}} \text{tr} [\mathbf{R}_I \mathbf{X}'_{\mu'(I)} \mathbf{P}_{\sigma'(I)} \mathbf{X}_I^\top] \quad (\Leftarrow \text{Expanding the domain of } \mathbf{R}) \\ &= \min_{\mu' \in \mathcal{M}} \sum_{I=1}^M -\max_{\substack{\mathbf{R}_I \in \mathcal{R}, \\ s_I \in \{1, \dots, S\}}} \text{tr} [\mathbf{R}_I \mathbf{X}'_{\mu'(I)} \mathbf{P}_{s_I} \mathbf{X}_I^\top] \quad (\Leftarrow \sigma'(I) \text{ is replaced by } s_I) \\ &= \min_{\mu' \in \mathcal{M}} \sum_{I=1}^M F_{I, \mu'(I)} \quad (\Leftarrow \text{Definition of } F_{I, J}) \quad \square \end{aligned}$$

4.7.2 Proof of Eq. 4.12

Definition 3 (p -objective function). For $p \in \{1, \dots, M\}$ and $(\mu, \sigma) \in \mathcal{M} \times \mathcal{S}$, define the equations

$$l_p(\mu, \sigma) = - \max_{\mathbf{R}^{(p)} \in \mathcal{R}} \text{tr} [\mathbf{R}^{(p)} \mathbf{C}^{(\mu, \sigma, p)}] + \min_{\mu''} \sum_{I=p+1}^M F_{I, \mu'' \circ \mu(I)},$$

where $\mathbf{C}^{(\mu, \sigma, p)}$ and $F_{I, J}$ is defined by Eqs. 4.10 and 4.8, respectively, and μ'' denotes a permutation on $\{\mu(p+1), \dots, \mu(M)\}$.

Theorem 2. For any $(\mu, \sigma) \in \mathcal{M} \times \mathcal{S}$, $l_0 \leq l_p(\mu, \sigma)$ holds.

proof.

$$\begin{aligned} & l_p(\mu, \sigma) \\ &= - \max_{\mathbf{R} \in \mathcal{R}} \text{tr} \left[\mathbf{R} \left(\sum_{I=1}^p \mathbf{X}'_{\mu(I)} \mathbf{P}_{\sigma(I)} \mathbf{X}_I^\top \right) \right] + \min_{\mu''} \sum_{I=p+1}^M F_{I, \mu'' \circ \mu(I)}, \quad (\Leftarrow \text{Definition of } l_p) \\ &\geq \sum_{I=1}^p \left(- \max_{\mathbf{R}_I \in \mathcal{R}} \text{tr} [\mathbf{R}_I \mathbf{X}'_{\mu(I)} \mathbf{P}_{\sigma(I)} \mathbf{X}_I^\top] \right) + \min_{\mu''} \sum_{I=p+1}^M F_{I, \mu'' \circ \mu(I)}, \\ &\quad (\Leftarrow \text{Expanding the domain of } \mathbf{R}) \\ &\geq \sum_{I=1}^p \left(- \max_{\substack{\mathbf{R}_I \in \mathcal{R}, \\ s_I \in \{1, \dots, S\}}} \text{tr} [\mathbf{R}_I \mathbf{X}'_{\mu(I)} \mathbf{P}_{s_I} \mathbf{X}_I^\top] \right) + \min_{\mu''} \sum_{I=p+1}^M F_{I, \mu'' \circ \mu(I)}, \\ &\quad (\Leftarrow \text{Maximize concerning } s_I) \\ &= \sum_{I=1}^p F_{I, \mu(I)} + \min_{\mu''} \sum_{I=p+1}^M F_{I, \mu'' \circ \mu(I)}, \quad (\Leftarrow \text{Definition of } F_{I, J}) \\ &\geq \min_{\mu''' \in \mathcal{M}} \left(\sum_{I=1}^p F_{I, \mu''' \circ \mu(I)} + \sum_{I=p+1}^M F_{I, \mu''' \circ \mu(I)} \right) \quad (\Leftarrow \text{Expanding the domain of } \mu'') \\ &= \min_{\mu'''} \sum_{I=1}^M F_{I, \mu'''(I)} \\ &= l_0 \quad \square \end{aligned}$$

4.7.3 Proof of Eq. 4.13

Theorem 3. For any $(\mu, \sigma) \in \mathcal{M} \times \mathcal{S}$, $l_p(\mu, \sigma) \leq l_{p+1}(\mu, \sigma)$ holds.

proof.

$$\begin{aligned}
& l_p(\mu, \sigma) \\
&= -\max_{\mathbf{R} \in \mathcal{R}} \text{tr} [\mathbf{R}\mathbf{C}^{(\mu, \sigma, p)}] + \min_{\mu''} \sum_{I=p+1}^M F_{I, \mu'' \circ \mu(I)} \\
&\quad (\Leftarrow \text{Definition of } l_p) \\
&\leq -\max_{\mathbf{R} \in \mathcal{R}} \text{tr} [\mathbf{R}\mathbf{C}^{(\mu, \sigma, p)}] + F_{p+1, \mu(p+1)} + \min_{\mu'''} \sum_{I=p+2}^M F_{I, \mu''' \circ \mu(I)} \\
&\quad (\Leftarrow \text{Narrowing the domain of } \mu'') \\
&= -\max_{\mathbf{R} \in \mathcal{R}} \text{tr} [\mathbf{R}\mathbf{C}^{(\mu, \sigma, p)}] - \max_{\substack{\mathbf{R}' \in \mathcal{R}, \\ s' \in \{1, \dots, S\}}} \text{tr} [\mathbf{R}'\mathbf{X}'_{\mu(p+1)} \mathbf{P}_{s'} \mathbf{X}_{p+1}^\top] + \min_{\mu'''} \sum_{I=p+2}^M F_{I, \mu''' \circ \mu(I)} \\
&\quad (\Leftarrow \text{Definition of } F_{I, J}) \\
&\leq -\max_{\mathbf{R} \in \mathcal{R}} \text{tr} [\mathbf{R}\mathbf{C}^{(\mu, \sigma, p)}] - \max_{\mathbf{R}' \in \mathcal{R}} \text{tr} [\mathbf{R}'\mathbf{X}'_{\mu(p+1)} \mathbf{P}_{\sigma(p+1)} \mathbf{X}_{p+1}^\top] + \min_{\mu'''} \sum_{I=p+2}^M F_{I, \mu''' \circ \mu(I)} \\
&\quad (\Leftarrow \text{Narrowing the domain of } s') \\
&\leq -\max_{\mathbf{R}'' \in \mathcal{R}} (\text{tr} [\mathbf{R}''\mathbf{C}^{(\mu, \sigma, p)}] + \text{tr} [\mathbf{R}''\mathbf{X}'_{\mu(p+1)} \mathbf{P}_{\sigma(p+1)} \mathbf{X}_{p+1}^\top]) + \min_{\mu'''} \sum_{I=p+2}^M F_{I, \mu''' \circ \mu(I)} \\
&\quad (\Leftarrow \text{Narrowing the domain of } \mathbf{R} \text{ and } \mathbf{R}') \\
&= -\max_{\mathbf{R}'' \in \mathcal{R}} \text{tr} [\mathbf{R}'' (\mathbf{C}^{(\mu, \sigma, p)} + \mathbf{X}'_{\mu(p+1)} \mathbf{P}_{\sigma(p+1)} \mathbf{X}_{p+1}^\top)] + \min_{\mu'''} \sum_{I=p+2}^M F_{I, \mu''' \circ \mu(I)} \\
&\quad (\Leftarrow \text{Linearity}) \\
&= -\max_{\mathbf{R}'' \in \mathcal{R}} \text{tr} [\mathbf{R}''\mathbf{C}^{(\mu, \sigma, p+1)}] + \min_{\mu'''} \sum_{I=p+2}^M F_{I, \mu''' \circ \mu(I)} \\
&\quad (\Leftarrow \text{Definition of } \mathbf{C}^{(\mu, \sigma, p)}) \\
&= l_{p+1}(\mu, \sigma) \quad \square
\end{aligned}$$

4.7.4 DSTS algorithm for multi-component system

In this section, we derive a DSTS algorithm for multi-component systems. The objective function for multi-component systems is organized and reiterated from the text (Eq. 4.22).

Definition 4 (Notation). For given $K \in \mathbb{N}$, let $M^{(1)}, \dots, M^{(K)}, n^{(1)}, \dots, n^{(K)}, S^{(1)}, \dots, S^{(K)} \in \mathbb{N}$ and $M^{(0)} = 1$. For $k \in \{1, \dots, K\}$, let $\mathcal{M}^{(k)}$ and $\mathcal{S}^{(k)}$ be the finite symmetric group on $\{1, \dots, M^{(k)}\}$ and the total set of mappings from $\{1, \dots, M^{(k)}\}$ to $\{1, \dots, S^{(k)}\}$, respectively.

Definition 5 (Solution space). Let $\mathcal{K} = \left(\prod_{k=1}^K \mathcal{M}^{(k)}\right) \times \left(\prod_{k=1}^K \mathcal{S}^{(k)}\right)$. For $(\boldsymbol{\mu}, \boldsymbol{\sigma}) \in \mathcal{K}$, $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are denoted as $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$ $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_K)$, respectively.

Definition 6 (Objective function). Let $\mathbf{X}_1^{(k)}, \dots, \mathbf{X}_{M^{(k)}}^{(k)}, \mathbf{X}'_1^{(k)}, \dots, \mathbf{X}'_{M^{(k)}}^{(k)} \in \mathbb{R}^{d \times n^{(k)}}$ and $\mathbf{P}_1^{(k)}, \dots, \mathbf{P}_S^{(k)} \in \mathbb{R}^{n^{(k)} \times n^{(k)}}$ for $k = 1, \dots, K$. For $(\boldsymbol{\mu}, \boldsymbol{\sigma}) \in \mathcal{K}$, define the objective function as follows.

$$L^*(\boldsymbol{\mu}, \boldsymbol{\sigma}) = -\max_{\mathbf{R}} \sum_{k=1}^K \sum_{I=1}^{M^{(k)}} \text{tr} \left[\mathbf{R} \mathbf{X}_{\mu^{(k)}(I)}'^{(k)} \mathbf{P}_{\sigma^{(k)}(I)}^{(k)} \mathbf{X}_I^{(k)\top} \right] \quad (4.28)$$

As in Eq. 4.9, by expanding the solution space concerning $\mathbf{R}$, we obtain the variational lower bound of Eq. 4.28.

Definition 7 (Variational lower bound). Define

$$l_{0,1} = \sum_{k=1}^K \min_{\mu_k \in \mathcal{M}^{(k)}} \sum_{I=1}^{M^{(k)}} F_{I, \mu_k(I)}^{(k)}, \quad (4.29)$$

where,

$$F_{I,J}^{(k)} = -\max_{\substack{\mathbf{R} \in \mathcal{R}, \\ s \in \{1, \dots, S\}}} \text{tr} \left[\mathbf{R} \mathbf{X}_J'^{(k)} \mathbf{P}_s^{(k)} \mathbf{X}_I^{(k)\top} \right]. \quad (4.30)$$

Theorem 4. *The following inequality holds.*

$$\min_{(\boldsymbol{\mu}, \boldsymbol{\sigma}) \in \mathcal{K}} L^*(\boldsymbol{\mu}, \boldsymbol{\sigma}) \geq l_{0,1} \quad (4.31)$$

Proof.

$$\begin{aligned} \min_{(\boldsymbol{\mu}, \boldsymbol{\sigma}) \in \mathcal{K}} L^*(\boldsymbol{\mu}, \boldsymbol{\sigma}) &\geq \sum_{k=1}^K \min_{(\mu_k, \sigma_k) \in \mathcal{M}^{(k)} \times \mathcal{S}^{(k)}} L(\mu_k, \sigma_k) \\ &\geq \sum_{k=1}^K \min_{\mu'_k \in \mathcal{M}^{(k)}} \sum_{I=1}^{M^{(k)}} F_{I, \mu'_k(I)}^{(k)} \quad \square \end{aligned}$$

Then, we take $t \in \{1, \dots, K\}$ and $p \in \{1, \dots, M^{(k)}\}$, and define the (t, p) -objective function. As with p -objective function, the (t, p) -objective function also has inequalities that holds as t and p increase.

Definition 8 ((t, p) -objective function). For given $(\boldsymbol{\mu}, \boldsymbol{\sigma}) \in \mathcal{K}$, define

$$l_{t,p}(\boldsymbol{\mu}, \boldsymbol{\sigma}) = -\max_{\mathbf{R} \in \mathcal{R}} \text{tr} \left[\mathbf{R} \tilde{\mathbf{C}}^*(\boldsymbol{\mu}, \boldsymbol{\sigma}, t, p) \right] + \tilde{F}^*(\boldsymbol{\mu}, t, p), \quad (4.32)$$

where

$$\tilde{\mathbf{C}}^*(\boldsymbol{\mu}, \boldsymbol{\sigma}, t, p) = \tilde{\mathbf{C}}_{k=t}^{(\mu_t, \sigma_t, p)} + \sum_{t'=1}^{t-1} \tilde{\mathbf{C}}_{k=t'}^{(\mu_{t'}, \sigma_{t'}, M^{(t')})}, \quad (4.33)$$

and $\tilde{\mathbf{C}}^{(\mu, \sigma, p)}$ is defined by Eq. 4.10 for $k = t$. Also,

$$\tilde{F}^*(\boldsymbol{\mu}, t, p) = \min_{\mu'} \sum_{I=p+1}^{M^{(t)}} F^{(t)}_{I, \mu' \circ \mu_t(I)} + \sum_{t'=t+1}^K l_0^{(t')}, \quad (4.34)$$

where, $F^{(t)}_{I,J}$ and $l_0^{(t)}$ are defined by Eq. 4.8 and Eq. 4.9 for $k = t$ and μ' is a permutation on $\{\mu_t(p+1), \dots, M^{(t)}\}$.

$\tilde{\mathbf{C}}^*(\boldsymbol{\mu}, \boldsymbol{\sigma}, t, p)$ denotes the variance-covariance matrix obtained by sum up the variance-covariance matrices of each molecule up to indices t and p , and $\tilde{F}^*(\boldsymbol{\mu}, t, p)$ is the variational lower bound of the remainder. Similarly to the p -objective function, the following incrementality holds for the (t, p) -objective function.

Theorem 5. For any t, p , and $(\boldsymbol{\mu}, \boldsymbol{\sigma}) \in \mathcal{K}$, the following inequalities hold.

$$\begin{aligned} l_{0,1} &\leq l_{t,p}(\boldsymbol{\mu}, \boldsymbol{\sigma}) \\ l_{t,p}(\boldsymbol{\mu}, \boldsymbol{\sigma}) &\leq l_{t,p+1}(\boldsymbol{\mu}, \boldsymbol{\sigma}) \\ l_{t,M^{(t)}}(\boldsymbol{\mu}, \boldsymbol{\sigma}) &\leq l_{t+1,1}(\boldsymbol{\mu}, \boldsymbol{\sigma}) \end{aligned} \quad (4.35)$$

Proof. As in **Theorem 2**. □

As with Eq. 4.14 in the single-component system, we define the equivalence relation for elements of the solution space $\mathcal{K}$. Again, this equivalence relation implies equivalence of the value of partial objective function, *i.e.*, two candidate solutions $(\boldsymbol{\mu}, \boldsymbol{\sigma})$ and $(\boldsymbol{\mu}', \boldsymbol{\sigma}')$ in the equivalence relation satisfy $l_{t,p}(\boldsymbol{\mu}, \boldsymbol{\sigma}) = l_{t,p}(\boldsymbol{\mu}', \boldsymbol{\sigma}')$.

Definition 9 ((t, p) -equivalence). For any $(\boldsymbol{\mu}, \boldsymbol{\sigma}), (\boldsymbol{\mu}', \boldsymbol{\sigma}') \in \mathcal{K}$, define the binary relation

$$(\boldsymbol{\mu}, \boldsymbol{\sigma}) \stackrel{t,p}{\sim} (\boldsymbol{\mu}', \boldsymbol{\sigma}') = \begin{cases} \mu_{t'} = \mu'_{t'}, \sigma_{t'} = \sigma'_{t'}, & \text{for } 1 \leq t' < t \\ (\mu_t, \sigma_t) \stackrel{p}{\sim} (\mu'_t, \sigma'_t) \end{cases}, \quad (4.36)$$

where, $\stackrel{p}{\sim}$ is defined by Eq. 4.14.

Theorem 6. Binary relation $\stackrel{t,p}{\sim}$ is an equivalence relation.

Proof. Both $=$ and $\stackrel{p}{\sim}$ are equivalence relations. □

Theorem 7. For given t, p , for any $(\boldsymbol{\mu}_1, \boldsymbol{\sigma}_1), (\boldsymbol{\mu}_2, \boldsymbol{\sigma}_2) \in \mathcal{K}$, if $(\boldsymbol{\mu}_1, \boldsymbol{\sigma}_1) \stackrel{t,p}{\sim} (\boldsymbol{\mu}_2, \boldsymbol{\sigma}_2)$, then $l_{t,p}(\boldsymbol{\mu}_1, \boldsymbol{\sigma}_1) = l_{t,p}(\boldsymbol{\mu}_2, \boldsymbol{\sigma}_2)$ holds.

Proof. From the assumption $(\mu_1, \sigma_1) \stackrel{t,p}{\sim} (\mu_2, \sigma_2)$, for any $(\mu_1, \sigma_1), (\mu_2, \sigma_2) \in \mathcal{K}$, $\tilde{C}^*(\mu_1, \sigma_1, t, p) = \tilde{C}^*(\mu_2, \sigma_2, t, p)$ and $\tilde{F}^*(\mu_1, t, p) = \tilde{F}^*(\mu_2, t, p)$. $\square$

As in the single component systems, we construct a weighted tree from (t, p) -equivalence. The tree also has the monotonically increasing concerning vertex weights along the branches.

Definition 10 ((t, p) -vertices). For given t and p , define $\mathcal{V}^{(t,p)} = \mathcal{K} \setminus \stackrel{t,p}{\sim}$. and $w(v) = l_{t,p}(\mu, \sigma)$ for $v \in \mathcal{V}^{(t,p)}$, where, (μ, σ) is the representative of v . Also, define $\mathcal{V}^{(0,1)} = \{\mathcal{K}\}$ and $w(v) = l_{0,1}$ for $v \in \mathcal{V}^{(0,1)}$.

Theorem 8. For given t , if $p < q$ then $\mathcal{V}^{(t,q)}$ is a refinement of $\mathcal{V}^{(t,p)}$. Also, if $t < u$ then $\mathcal{V}^{(u,q)}$ is a refinement of $\mathcal{V}^{(t,p)}$.

Proof. From definition, for any $(\mu_1, \sigma_1), (\mu_2, \sigma_2) \in \mathcal{K}$, if $p < q$ and $(\mu_1, \sigma_1) \stackrel{t,q}{\sim} (\mu_2, \sigma_2)$, then $(\mu_1, \sigma_1) \stackrel{t,p}{\sim} (\mu_2, \sigma_2)$. The same goes for the others. $\square$

Definition 11 (Edges from (t, p) to (u, q)). Define the edges from (t, p) to (u, q) as

$$\mathcal{E}^{(t,p,u,q)} = \{(v, v') \in \mathcal{V}^{(t,p)} \times \mathcal{V}^{(u,q)} \mid v \supset v'\}. \quad (4.37)$$

Definition 12 (Factorial tree). Define $(\mathcal{V}^*, \mathcal{E}^*)$, where,

$$\mathcal{V}^* = \bigcup_{t=0}^K \bigcup_{p=1}^{M^{(t)}} \mathcal{V}^{(t,p)}, \quad (4.38)$$

and

$$\mathcal{E}^* = \left(\bigcup_{t=1}^K \bigcup_{p=1}^{M^{(t)}-1} \mathcal{E}^{(t,p,t,p+1)} \right) \cup \left(\bigcup_{t=0}^{K-1} \mathcal{E}^{(t,M^{(t)},t+1,0)} \right). \quad (4.39)$$

Finally, we present the modified DSTS algorithm in **Algorithm 3**, where we use the following notations: The main loop remains almost same as in **Algorithm 2**. DESCEND* is modified from the original one to a double loop to increment t in addition to p . When p is incremented to $M^{(t)}$ with t fixed, child vertices are created by the $(t+1, p)$ -equivalence, which is always the root of the $(t+1)$ -tree. The same applies to ASCEND* from the original one.

Definition 13 (Minimum element in queue). For $\mathcal{Q} \in \mathcal{V}^*$, let $\min \mathcal{Q}$ be the minimum element concerning $w(v)$, and if there is not one, choose an arbitrary one.

Definition 14 (Conditioning of queue). For $\mathcal{Q} \in \mathcal{V}^*$, $\mathcal{Q}_{t,p} = \mathcal{Q} \cap \mathcal{V}^{(t,p)}$.

4.7.5 Memoizing

The bulk of the DSTS calculation is to compute the vertex weights of the tree. Here we present several techniques to calculate them more efficiently. As in Eq. 4.10, $\tilde{\mathbf{C}}^{(\mu,\sigma,p)}$ is the sum of $\mathbf{X}'_{\mu(I)}\mathbf{P}_{\sigma(I)}\mathbf{X}_I^\top$ specified by μ, σ . The bulk of the sum can be memoized; for any μ and μ' , if $\mu \stackrel{p-1}{\sim} \mu'$, then

$$\tilde{\mathbf{C}}^{(\mu,\sigma,p)} = \tilde{\mathbf{C}}^{(\mu,\sigma,p-1)} + \mathbf{X}'_{\mu(p)}\mathbf{P}_{\sigma(p)}\mathbf{X}_p^\top$$

and

$$\tilde{\mathbf{C}}^{(\mu',\sigma,p)} = \tilde{\mathbf{C}}^{(\mu,\sigma,p-1)} + \mathbf{X}'_{\mu'(p)}\mathbf{P}_{\sigma(p)}\mathbf{X}_p^\top$$

hold, where $\tilde{\mathbf{C}}^{(\mu,\sigma,p-1)}$ is the common part and can be used over. This is also true for σ . In **Algorithm 3**, vertex weights are evaluated when the algorithm adds child vertices to the queue (lines 18 and 24). If we can use a common term for all child vertices here, we only need to compute the matrix addition once at each vertex. This can be accomplished by noting the computed $\tilde{\mathbf{C}}^{(\mu,\sigma,p)}$ for each vertex. Since the maximum number of vertices that can be queued in **Algorithm 3** is predetermined, the total amount of memory required to memoize $\tilde{\mathbf{C}}^{(\mu,\sigma,p)}$ can also be calculated in advance.

Eq. 4.34 is linear with respect to t and t' , and the second term on the right-hand side can be precomputed at the start of **Algorithm 3**. The constants $\{l_0^{(1)}, \dots, l_0^{(K)}\}$ can be transformed into the summation form $\{f_0, \dots, f_K\}$, where

$$f_k = \sum_{t=k+1}^K l_0^{(t)}$$

and $f_K = 0$. This simplifies Eq. 4.34 to

$$\tilde{F}^{*(\mu,t,p)} = \min_{\mu'} \sum_{I=p+1}^{M^{(t)}} \tilde{F}_{I,\mu' \circ \mu_t(I)}^{(t)} + f_t,$$

and can be found using the Hungarian algorithm for $(M^{(t)} - p) \times (M^{(t)} - p)$ matrix.

4.7.6 Thinning out of p

In this section, we consider edge thinning, *i.e.*,

$$\mathcal{E}' = \bigcup_{j=0}^{J-1} \mathcal{E}_{p_j, p_{j+1}} \quad (4.40)$$

for $p_0 = 0, p_J = M$, and $p_0 < p_1 < \dots < p_J$. This reduces the number of tree vertices, but is not necessarily a good branching strategy.

For an M -molecular one-component system, if we consider $\mathcal{E}_{p,q}$ for $p+1 < q$, the set of child vertices $v^* \setminus \stackrel{q}{\sim}$ created from $v^* \in \mathcal{V}_p$ contains $\frac{(M-p)!}{(M-q)!}$ elements. The number of

evaluation of vertices in the greedy method for Eq. 4.40 is $\sum_{j=0}^{J-1} \frac{(M-p_j)!}{(M-p_{j+1})!}$, which is more than in the original tree. Thus, thinning would not be a good choice if one wants to use greedy methods or early termination.

The child vertex set $v^* \setminus \tilde{q}$ can be partitioned by $\tilde{q}^{-1}$, each group can use a common remainder of the vertex weight calculation. That is, it is equivalent to memoizing the solution to the linear assignment problem in Eq. 4.34. This may be advantageous if the search is protracted.

4.7.7 Coordinate ordering

The order of the reference molecular coordinates $\{\mathbf{X}_1, \dots, \mathbf{X}_M\}$ affects the ordering of the vertex weights, but not the order of the target coordinates. It is not trivial in which order it is most effective to place the reference molecules, but one idea is to sort the molecules in descending order of their autovariance $G_I = \mathbf{X}_I^\top \mathbf{X}_I$; *i.e.*, molecules far from the centroid are placed closer to the root. The software **mobbrmsd** supports this sorting.

Chapter 5

Conclusion and the future prospects

We have performed molecular dynamics (MD) simulations of two flexible molecular assemblies: functional molecular liquids and aromatic micelles. The detailed analysis of the MD simulations consistently illustrates that differences in chemical structure lead to differences in macroscopic physical properties. This is a fundamental achievement that will inform the following experiments in terms of molecular design and allow theory and experiment to mutually test hypotheses.

In Chapter 2, MD simulations of tens of microseconds (submilliseconds in total), long enough to reproduce the hydrodynamic properties of the four alkyl-pyrene liquids, allowed a comprehensive analysis of their molecular structure and dynamics. The MD simulations reproduced well that the branched alkyl chains in the alkyl-pyrene liquids prevent crystallization of pyrene and that they exhibit liquid properties. The spatio-temporal correlation function analysis based on the van Hove function uncovered that the alkyl-pyrene liquids are underlay by the characteristic local liquid structures depending on the substituents. The compounds with one substituent of the phenyl group with the long alkyl chains form the columnar structures established by the π - π stacking between the pyrene rings. Because of the smaller ratios of the substituents to the pyrene core, the substituent effect is moderate, invoking the π - π stacking between the pyrene rings in the local structures. The position of the alkyl chain substituents affects the π - π stacking properties; that is, the alkyl chain covering the pyrene ring promotes the dissociation of the π - π stacking, making the columnar structures fragile. For the compounds with four substituent, the progressive formation of the π - π stacking of the pyrene rings is arrested. The position of the alkyl chain substituents affects whether the π - π pairing stacking structure the pyrene rings can be formed.

The reasons why the local liquid structure of alkyl-pyrene liquids affects the rheological properties are as follows. The dynamics of all alkyl-pyrene liquids, regardless of whether π - π stacking is present or not, clearly show anomalous diffusion due to the long alkyl chains, similar to polymers and liquid crystals. The remarkable changes in the rheological properties of alkyl-pyrene liquids arise from the differences in the composition ratio of alkyl chains and pyrene rings and positions of the substituents, so we can see that the local contact interactions between pyrene rings and alkyl chains in the liquid determine the rheological properties. The substituent positions of the alkyl chains change the local liquid structure and change the magnitude of the contact interactions between the pyrene surface and the

alkyl chains. The viscosity coefficients obtained from MD simulations and the order of microscopic contact areas between pyrene and alkyl chains obtained from spatio-temporal correlation functions well explain the viscosity coefficients in the experiments. In summary, the contact area between the aromatic rings, pyrenes, and the alkyl chains determines the magnitude of the viscosity coefficient of alkyl-pyrene liquids. The formation of π - π structures reduces the number of pyrene rings exposed to the alkyl chain environment, leading to larger intermolecular friction and larger viscosity coefficients.

The above details the mechanism by which branched alkyl chains cause the rheological changes of alkyl-pyrene systems. It is also known that the photochemical properties of these systems changes depending on the local liquid structures. One interesting question is whether the experimental photochemical properties can be reproduced by differences in the local liquid structures, as observed in this study. We are working on an analysis using an exciton transfer model based on excited-state quantum chemical calculations, the results of which will be reported elsewhere.

In Chapter 3, we performed MD simulations of the nano-sized molecular structures of two types of aromatic micelles called V-shaped polyaromatic amphiphiles (VPAA) which form micelle-like non-bonded self-assemblies in aqueous solution. From the MD simulations for tens of microseconds, we observed the spontaneous formation of the VPAA assemblies in aqueous solution and flexible structural changes within the assemblies. The results of the MD simulations were consistent with experimental evidence and provided sufficient samples for classification analysis of detailed atomic structures. To investigate the microstructure of molecular assemblies, we developed an improved RMSD-based classification analysis method. The difficulty of combinatorial explosion in RMSD evaluation was avoided by a heuristic algorithm, symRMSD. This method has successfully identified quasi-stable conformers of the VPAA assemblies that appear on MD trajectories, and their diversity well illustrates their prominent feature of encapsulating a wide range of hydrophobic molecules of various sizes and shapes.

This study successfully classified the static structure of the VPAA assemblies. Extending this method to dynamic structure analysis would be an interesting application. We are currently attempting to analyze the dynamic structure of VPAA using the Markov state model, [19, 196, 197] the results of which should be published elsewhere.

The two simulation experiments of flexible molecular assemblies described above are typical examples of how MD simulations can clearly demonstrate the mechanism between chemical structure and macroscopic properties. It would be helpful to extend the developed methods to the analysis of general molecular assemblies, rather than stop at each of these.

In Chapter 5, we have improved on the shortcomings of the method developed in Chapter 3 and developed a powerful method called mobbRMSD that supports general molecular assemblies. mobbRMSD is a branch-and-bound (BB) algorithm for the exact solution of RMSD. mobbRMSD can handle large molecules because it employs molecular topology using molecularly oriented coordinates. mobbRMSD reduced the factorial cost of RMSD calculations to exponential cost on average, and was able to compute an exact solution within milliseconds of wall-clock time, even for 10 molecules, which was previously difficult. This speed is at a level applicable to high-throughput analysis such as clustering analysis.

Furthermore, by introducing an early termination parameter, an exact minimum spanning tree can be computed with asymptotic polynomial scaling with respect to the number of data. Minimum spanning trees can be applied to many clustering methods. We are currently developing a mobbRMSD-based minimum spanning tree algorithm that automatically determines the cutoff parameter, and applying it to the clustering analysis of water networks in channel proteins. The results will be published elsewhere.

mobbRMSD is a general-purpose methodology, and several application ideas immediately come to mind. The first is its application to the Markov state model (MSM) as a straightforward extension of the clustering used in the VPAA research. [19, 196, 197] Although MSM has been developed for chiral macromolecules such as proteins, a return to the principles reveals it to be a very powerful methodology that links statistical dynamic ensembles to causal explanations. Currently, MSM for chiral single molecules is dominated by RMSD-based methods, but the integration of mobbRMSD is expected to greatly expand its range of applications.

The second is its application to persistent homology (PH) analysis. [198, 199] PH is a mathematical method that focuses on structures such as holes, voids, and connected components within a shape. PH is effective in extracting hidden structures from heterogeneous and disordered data, and has been applied to amorphous and porous materials, [198] 3D protein structures [199] in the field of physical chemistry. However, since PH is based on simplicial complexes, it is not clear that a simplicial complex can be constructed from the space formed by mobbRMSD. Investigating the nature of the metric space created by mobbRMSD is a future task.

In summary, as this study demonstrates, explaining causal relationships from chemical structure to macroscopic physical properties through atomic-level structural analysis provides useful insights for theorists as well as experimentalists. Our new methodology, mobbRMSD, enables structural analysis of flexible molecular assemblies and paves the way for their understanding and *in-silico* molecular design.

Bibliography

- [1] *Functional Organic Liquids*, 1st ed., (Ed.: T. Nakanishi), Wiley, **2019**.
- [2] M. Yoshizawa, L. Catti, *Proceedings of the Japan Academy Series B* **2023**, 99, 29–38.
- [3] M. Kato, H. Ito, M. Hasegawa, K. Ishii, *Chemistry - A European Journal* **2019**, 25, 5105–5112.
- [4] K. Kondo, J. K. Klosterman, M. Yoshizawa, *Chemistry - A European Journal* **2017**, 23, 16710–16721.
- [5] M. Yoshizawa, L. Catti, *Accounts of Chemical Research* **2019**, 52, 2392–2404.
- [6] X. Zhang, Z. Chi, Y. Zhang, S. Liu, J. Xu, *Journal of Materials Chemistry C* **2013**, 1, 3376.
- [7] T. Seki, H. Ito, *Chemistry – A European Journal* **2016**, 22, 4322–4329.
- [8] R. Kotani, S. Yokoyama, S. Nobusue, S. Yamaguchi, A. Osuka, H. Yabu, S. Saito, *Nature Communications* **2022**, 13, 303.
- [9] T. Yamakado, S. Saito, *Journal of the American Chemical Society* **2022**, 144, 2804–2815.
- [10] X. Zhang, B. Li, Z.-H. Chen, Z.-N. Chen, *Journal of Materials Chemistry* **2012**, 22, 11427.
- [11] O. S. Wenger, *Chemical Reviews* **2013**, 113, 3686–3733.
- [12] E. R. Yera, A. E. Cleves, A. N. Jain, *Journal of Medicinal Chemistry* **2011**, 54, Publisher: American Chemical Society, 6771–6785.
- [13] Y.-C. Lo, S. Senese, R. Damoiseaux, J. Z. Torres, *ACS Chemical Biology* **2016**, 11, 2244–2253.
- [14] S. De, F. Musil, T. Ingram, C. Baldauf, M. Ceriotti, *Journal of Cheminformatics* **2017**, 9, 6.
- [15] M. P. Allen, D. J. Tildesley, *Computer Simulation of Liquids*, Oxford University Press, **2017**.
- [16] W. F. Van Gunsteren, C. Oostenbrink, *Journal of Chemical Information and Modeling* **2024**, acs.jcim.4c00823.
- [17] A. Rahman, *Physical Review* **1964**, 136, A405–A411.
- [18] J. Shao, S. W. Tanner, N. Thompson, T. E. Cheatham, *Journal of Chemical Theory and Computation* **2007**, 3, 2312–2334.

- [19] G. R. Bowman, K. A. Beauchamp, G. Boxer, V. S. Pande, *The Journal of Chemical Physics* **2009**, *131*, 124101.
- [20] M. Weisberg, *Simulation and Similarity: Using Models to Understand the World*, Oxford University Press, **2013**.
- [21] F. Lu, T. Takaya, K. Iwata, I. Kawamura, A. Saeki, M. Ishii, K. Nagura, T. Nakanishi, *Scientific Reports* **2017**, *7*, 3416.
- [22] J. T. Edward, *Journal of Chemical Education* **1970**, *47*, 261.
- [23] K. Kondo, A. Suzuki, M. Akita, M. Yoshizawa, *Angewandte Chemie International Edition* **2013**, *52*, 2308–2312.
- [24] F. Lu, T. Nakanishi, *Science and Technology of Advanced Materials* **2015**, *16*, 014805.
- [25] D. Xu, C. Adachi, *Applied Physics Letters* **2009**, *95*, 053304.
- [26] K. Kubota, S. Hirata, Y. Shibano, O. Hirata, M. Yahiro, C. Adachi, *Chemistry Letters* **2012**, *41*, 934–936.
- [27] S. Hirata, K. Kubota, H. H. Jung, O. Hirata, K. Goushi, M. Yahiro, C. Adachi, *Advanced Materials* **2011**, *23*, 889–893.
- [28] S. S. Babu, J. Aimi, H. Ozawa, N. Shirahata, A. Saeki, S. Seki, A. Ajayaghosh, H. Möhwald, T. Nakanishi, *Angewandte Chemie International Edition* **2012**, *51*, 3391–3395.
- [29] S. S. Babu, M. J. Hollamby, J. Aimi, H. Ozawa, A. Saeki, S. Seki, K. Kobayashi, K. Hagiwara, M. Yoshizawa, H. Möhwald, T. Nakanishi, *Nature Communications* **2013**, *4*, 1969.
- [30] M. J. Hollamby, A. E. Danks, Z. Schnepf, S. E. Rogers, S. R. Hart, T. Nakanishi, *Chemical Communications* **2016**, *52*, 7344–7347.
- [31] P. Duan, N. Yanai, N. Kimizuka, *Journal of the American Chemical Society* **2013**, *135*, 19056–19059.
- [32] F. Lu, K. Hagiwara, M. Yoshizawa, K. Nagura, S. Ishihara, T. Nakanishi, *Journal of Materials Chemistry C* **2019**, *7*, 2577–2582.
- [33] K. Chung, M. S. Kwon, B. M. Leung, A. G. Wong-Foy, M. S. Kim, J. Kim, S. Takayama, J. Gierschner, A. J. Matzger, J. Kim, *ACS Central Science* **2015**, *1*, 94–102.
- [34] T. Machida, R. Taniguchi, T. Oura, K. Sada, K. Kokado, *Chemical Communications* **2017**, *53*, 2378–2381.
- [35] M. Taki, S. Azeyanagi, K. Hayashi, S. Yamaguchi, *Journal of Materials Chemistry C* **2017**, *5*, 2142–2148.
- [36] T. Michinobu, T. Nakanishi, J. P. Hill, M. Funahashi, K. Ariga, *Journal of the American Chemical Society* **2006**, *128*, 10384–10385.
- [37] J.-C. Ribierre, L. Zhao, M. Inoue, P.-O. Schwartz, J.-H. Kim, K. Yoshida, A. S. D. Sandanayaka, H. Nakanotani, L. Mager, S. Méey, C. Adachi, *Chemical Communications* **2016**, *52*, 3103–3106.

- [38] T. G. Plint, B. A. Kamino, T. P. Bender, *The Journal of Physical Chemistry C* **2015**, *119*, 1676–1682.
- [39] M. J. Hollamby, M. Karny, P. H. H. Bomans, N. A. J. M. Sommerdijk, A. Saeki, S. Seki, H. Minamikawa, I. Grillo, B. R. Pauw, P. Brown, J. Eastoe, H. Möhwald, T. Nakanishi, *Nature Chemistry* **2014**, *6*, 690–696.
- [40] T. Ogoshi, T. Aoki, R. Shiga, R. Iizuka, S. Ueda, K. Demachi, D. Yamafuji, H. Kayama, T.-a. Yamagishi, *Journal of the American Chemical Society* **2012**, *134*, 20322–20325.
- [41] N. Giri, M. G. Del Pópolo, G. Melaugh, R. L. Greenaway, K. Rätzke, T. Koschine, L. Pison, M. F. C. Gomes, A. I. Cooper, S. L. James, *Nature* **2015**, *527*, 216–220.
- [42] H. Akiyama, M. Yoshida, *Advanced Materials* **2012**, *24*, 2353–2356.
- [43] K. Masutani, M.-a. Morikawa, N. Kimizuka, *Chem. Commun.* **2014**, *50*, 15803–15806.
- [44] A. Ghosh, M. Yoshida, K. Suemori, H. Isago, N. Kobayashi, Y. Mizutani, Y. Kurashige, I. Kawamura, M. Nirei, O. Yamamuro, T. Takaya, K. Iwata, A. Saeki, K. Nagura, S. Ishihara, T. Nakanishi, *Nature Communications* **2019**, *10*, 4210.
- [45] T. Machida, T. Nakanishi, *Journal of Materials Chemistry C* **2021**, *9*, 10661–10667.
- [46] F. Lu, K. Jang, I. Osica, K. Hagiwara, M. Yoshizawa, M. Ishii, Y. Chino, K. Ohta, K. Ludwichowska, K. J. Kurzydłowski, S. Ishihara, T. Nakanishi, *Chemical Science* **2018**, *9*, 6774–6778.
- [47] A. Shinohara, C. Pan, L. Wang, T. Nakanishi, *Molecular Systems Design & Engineering* **2019**, *4*, 78–90.
- [48] F. Lu, T. Nakanishi, *Advanced Optical Materials* **2019**, *7*, 1900176.
- [49] A. D. Glova, I. V. Volgin, V. M. Nazarychev, S. V. Larin, S. V. Lyulin, A. A. Gurtovenko, *RSC Advances* **2019**, *9*, 38834–38847.
- [50] I. J. Prentice, X. Liu, O. A. Nerushev, S. Balakrishnan, C. R. Pulham, P. J. Camp, *The Journal of Chemical Physics* **2020**, *152*, 074504.
- [51] X. Yang, Q. Liu, X. Zhang, C. Ji, B. Cao, *Fluid Phase Equilibria* **2022**, *562*, 113566.
- [52] H. Liu, E. Maginn, A. E. Visser, N. J. Bridges, E. B. Fox, *Industrial & Engineering Chemistry Research* **2012**, *51*, 7242–7254.
- [53] I. Skarmoutsos, L. Spyrogiannopoulos, E. Kainourgiakis, J. Samios, *Journal of Molecular Liquids* **2019**, *289*, 111135.
- [54] P. J. Bonab, M. D. Esrafil, A. Rastkar Ebrahimzadeh, J. Jahanbin Sardroodi, *Journal of Molecular Liquids* **2021**, *341*, 117277.
- [55] P. J. Bonab, M. D. Esrafil, A. Rastkar Ebrahimzadeh, J. Jahanbin Sardroodi, *Journal of Molecular Liquids* **2021**, *338*, 116716.
- [56] P. J. Bonab, M. D. Esrafil, A. R. Ebrahimzadeh, J. J. Sardroodi, *Journal of Molecular Graphics and Modelling* **2021**, *106*, 107908.
- [57] P. J. Bonab, A. Rastkar Ebrahimzadeh, J. Jahanbin Sardroodi, *Scientific Reports* **2021**, *11*, 6384.

- [58] P. J. Bonab, J. J. Sardroodi, M. S. Avestan, *Journal of Chemical & Engineering Data* **2022**, *67*, Publisher: American Chemical Society, 2077–2087.
- [59] K. Isoda, T. Ishiyama, Y. Mutoh, D. Matsukuma, *ACS Applied Materials & Interfaces* **2019**, *11*, 12053–12062.
- [60] A. Ikenaga, Y. Akiyama, T. Ishiyama, M. Gon, K. Tanaka, Y. Chujo, K. Isoda, *ACS Applied Materials & Interfaces* **2021**, *13*, 47127–47133.
- [61] M. G. Martin, J. I. Siepmann, *The Journal of Physical Chemistry B* **1998**, *102*, 2569–2577.
- [62] C. D. Wick, M. G. Martin, J. I. Siepmann, *The Journal of Physical Chemistry B* **2000**, *104*, 8008–8016.
- [63] M. G. Martin, *Fluid Phase Equilibria* **2006**, *248*, 50–55.
- [64] W. L. Jorgensen, J. D. Madura, C. J. Swenson, *Journal of the American Chemical Society* **1984**, *106*, 6638–6646.
- [65] C. I. Bayly, P. Cieplak, W. Cornell, P. A. Kollman, *The Journal of Physical Chemistry* **1993**, *97*, 10269–10280.
- [66] W. D. Cornell, P. Cieplak, C. I. Bayly, P. A. Kollman, *Journal of the American Chemical Society* **1993**, *115*, 9620–9631.
- [67] 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 Jr, *Journal of Computational Chemistry* **1993**, *14*, Publisher: John Wiley & Sons, Ltd, 1347–1363.
- [68] M. S. Gordon, M. W. Schmidt in *Theory and Applications of Computational Chemistry*, (Eds.: C. E. Dykstra, G. Frenking, K. S. Kim, G. E. Scuseria), Elsevier, **2005**, pp. 1167–1189.
- [69] R. Salomon-Ferrer, D. A. Case, R. C. Walker, *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2013**, *3*, 198–210.
- [70] J. C. Walsh, D. T. Hogan, K.-L. M. Williams, S. D. Brake, G. Venkataramana, T. A. Misener, B. J. Wallace, R. P. Johnson, D. W. Thompson, Y. Zhao, B. D. Wagner, G. J. Bodwell, *ChemPlusChem* **2019**, *84*, Publisher: John Wiley & Sons, Ltd, 754–765.
- [71] L. Martínez, R. Andrade, E. G. Birgin, J. M. Martínez, *Journal of Computational Chemistry* **2009**, *30*, 2157–2164.
- [72] T. G. Mason, D. A. Weitz, *Physical Review Letters* **1995**, *74*, 1250–1253.
- [73] I. Golding, E. C. Cox, *Physical Review Letters* **2006**, *96*, 098102.
- [74] E. R. Weeks, J. C. Crocker, A. C. Levitt, A. Schofield, D. A. Weitz, *Science* **2000**, *287*, 627–631.
- [75] A. B. Goins, H. Sanabria, M. N. Waxham, *Biophysical Journal* **2008**, *95*, 5362–5373.
- [76] T. S. R. Al-Hadithi, H. A. Barnes, K. Walters, *Colloid and Polymer Science* **1992**, *270*, 40–46.

- [77] B. J. Alder, D. M. Gass, T. E. Wainwright, *The Journal of Chemical Physics* **1970**, 53, 3813–3826.
- [78] X. Bian, C. Kim, G. E. Karniadakis, *Soft Matter* **2016**, 12, 6331–6346.
- [79] A. Einstein, *Annalen der Physik* **1905**, 322, 549–560.
- [80] I.-C. Yeh, G. Hummer, *The Journal of Physical Chemistry B* **2004**, 108, 15873–15879.
- [81] L. D. Favro, *Physical Review* **1960**, 119, 53–62.
- [82] V. Wong, D. A. Case, *The Journal of Physical Chemistry B* **2008**, 112, 6013–6024.
- [83] M. Linke, J. Köfinger, G. Hummer, *The Journal of Physical Chemistry B* **2018**, 122, 5630–5639.
- [84] G. Williams, D. C. Watts, *Transactions of the Faraday Society* **1970**, 66, 80.
- [85] C. P. Lindsey, G. D. Patterson, *The Journal of Chemical Physics* **1980**, 73, 3348–3357.
- [86] J. Jackle, *Reports on Progress in Physics* **1986**, 49, 171–231.
- [87] L. Van Hove, *Physical Review* **1954**, 95, 249–262.
- [88] G. H. Vineyard, *Physical Review* **1958**, 110, 999–1010.
- [89] I. M. Svishchev, P. G. Kusalik, *The Journal of Chemical Physics* **1993**, 99, 3049–3058.
- [90] K. Kulińska, T. Kuliński, A. Lyubartsev, A. Laaksonen, R. Adamiak, *Computers & Chemistry* **2000**, 24, 451–457.
- [91] G. M. Bartenev, Y. V. Zelenev, *Polymer Mechanics* **1969**, 5, 25–42.
- [92] R. H. Boyd, *Polymer* **1985**, 26, 1123–1133.
- [93] C. M. Snively, J. L. Koenig, *Journal of Polymer Science Part B: Polymer Physics* **1999**, 37, 2261–2268.
- [94] R. Schmidt, K. Griesbaum, A. Behr, D. Biedenkapp, H.-W. Voges, D. Garbe, C. Paetz, G. Collin, D. Mayer, H. Höke in *Ullmann's Encyclopedia of Industrial Chemistry*, Wiley-VCH Verlag GmbH & Co. KGaA, **2014**, pp. 1–74.
- [95] R. S. Meissner, J. Rebek, J. De Mendoza, *Science* **1995**, 270, 1485–1488.
- [96] L. R. MacGillivray, J. L. Atwood, *Nature* **1997**, 389, 469–472.
- [97] N. Takeda, K. Umemoto, K. Yamaguchi, M. Fujita, *Nature* **1999**, 398, 794–796.
- [98] T. R. Cook, P. J. Stang, *Chemical Reviews* **2015**, 115, 7001–7045.
- [99] C. L. D. Gibb, B. C. Gibb, *Journal of the American Chemical Society* **2004**, 126, 11408–11409.
- [100] S. Hiraoka, K. Harano, M. Shiro, M. Shionoya, *Journal of the American Chemical Society* **2008**, 130, 14368–14369.
- [101] N. Kishi, M. Akita, M. Kamiya, S. Hayashi, H.-F. Hsu, M. Yoshizawa, *Journal of the American Chemical Society* **2013**, 135, 12976–12979.

- [102] M. Yamashina, M. Akita, T. Hasegawa, S. Hayashi, M. Yoshizawa, *Science Advances* **2017**, 3, e1701126.
- [103] T.-R. Li, G. Piccini, K. Tiefenbacher, *Journal of the American Chemical Society* **2023**, 145, 4294–4303.
- [104] E. Pahima, Q. Zhang, K. Tiefenbacher, D. T. Major, *Journal of the American Chemical Society* **2019**, 141, 6234–6246.
- [105] L. Catti, Q. Zhang, K. Tiefenbacher, *Chemistry - A European Journal* **2016**, 22, 9060–9066.
- [106] J. Kang, J. Rebek, *Nature* **1997**, 385, 50–52.
- [107] M. D. Pluth, R. G. Bergman, K. N. Raymond, *Science* **2007**, 316, 85–88.
- [108] M. Yoneya, T. Yamaguchi, S. Sato, M. Fujita, *Journal of the American Chemical Society* **2012**, 134, 14401–14407.
- [109] Y. Jiang, H. Zhang, Z. Cui, T. Tan, *The Journal of Physical Chemistry Letters* **2017**, 8, 2082–2086.
- [110] A. V. Zhukhovitskiy, M. Zhong, E. G. Keeler, V. K. Michaelis, J. E. P. Sun, M. J. A. Hore, D. J. Pochan, R. G. Griffin, A. P. Willard, J. A. Johnson, *Nature Chemistry* **2016**, 8, 33–41.
- [111] Y. Okazawa, K. Kondo, M. Akita, M. Yoshizawa, *Chemical Science* **2015**, 6, 5059–5062.
- [112] M. M. Sartin, K. Kondo, M. Yoshizawa, S. Takeuchi, T. Tahara, *Physical Chemistry Chemical Physics* **2017**, 19, 757–765.
- [113] T. Omagari, A. Suzuki, M. Akita, M. Yoshizawa, *Journal of the American Chemical Society* **2016**, 138, 499–502.
- [114] K. Kondo, M. Akita, T. Nakagawa, Y. Matsuo, M. Yoshizawa, *Chemistry - A European Journal* **2015**, 21, 12741–12746.
- [115] L. Catti, H. Narita, Y. Tanaka, H. Sakai, T. Hasobe, N. V. Tkachenko, M. Yoshizawa, *Journal of the American Chemical Society* **2021**, 143, 9361–9367.
- [116] S.-Y. Liu, N. Kishida, J. Kim, N. Fukui, R. Haruki, Y. Niwa, R. Kumai, D. Kim, M. Yoshizawa, H. Shinokubo, *Journal of the American Chemical Society* **2023**, 145, 2135–2141.
- [117] Y. Okazawa, K. Kondo, M. Akita, M. Yoshizawa, *Journal of the American Chemical Society* **2015**, 137, 98–101.
- [118] K. Kondo, A. Suzuki, M. Akita, M. Yoshizawa, *European Journal of Organic Chemistry* **2014**, 2014, 7389–7394.
- [119] T. Kai, M. Kishimoto, M. Akita, M. Yoshizawa, *Chemical Communications* **2018**, 54, 956–959.
- [120] M. Smoluchowski, *Pisma Mariana Smoluchowskiego* **1927**, 2, 530–594.
- [121] R. J. G. B. Campello, D. Moulavi, J. Sander in *Advances in Knowledge Discovery and Data Mining, Vol. 7819*, Series Title: Lecture Notes in Computer Science, Springer Berlin Heidelberg, Berlin, Heidelberg, **2013**, pp. 160–172.

- [122] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, D. A. Case, *Journal of Computational Chemistry* **2004**, *25*, 1157–1174.
- [123] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein, *The Journal of Chemical Physics* **1983**, *79*, 926–935.
- [124] D. R. Roe, T. E. Cheatham, *Journal of Chemical Theory and Computation* **2013**, *9*, 3084–3095.
- [125] J. Tomasi, B. Mennucci, R. Cammi, *Chemical Reviews* **2005**, *105*, 2999–3094.
- [126] H. Nymeyer, S. Gnanakaran, A. E. García in *Methods in Enzymology*, Vol. 383, Elsevier, **2004**, pp. 119–149.
- [127] M. Jorge, *Langmuir* **2008**, *24*, 5714–5725.
- [128] I. Kryven, S. Lazzari, G. Storti, *Macromolecular Theory and Simulations* **2014**, *23*, 170–181.
- [129] G. Odriozola, A. Schmitt, J. Callejas-Fernández, R. Martínez-García, R. Hidalgo-Álvarez, *The Journal of Chemical Physics* **1999**, *111*, 7657–7667.
- [130] S. K. Kearsley, *Acta Crystallographica Section A Foundations of Crystallography* **1989**, *45*, 208–210.
- [131] T. Shibuya, J. Jansson, K. Sadakane, *Algorithms for Molecular Biology* **2010**, *5*, 7.
- [132] T. Fukutani, K. Miyazawa, S. Iwata, H. Satoh, *Bulletin of the Chemical Society of Japan* **2021**, *94*, 655–665.
- [133] T. Nabika, S. Iwata, H. Satoh, *Bulletin of the Chemical Society of Japan* **2024**, *97*, uoae037.
- [134] L. McInnes, J. Healy, S. Astels, *The Journal of Open Source Software* **2017**, *2*, 205.
- [135] M. Kinoshita, *International Journal of Molecular Sciences* **2009**, *10*, 1064–1080.
- [136] W. R. Taylor, *Molecular & Cellular Proteomics* **2002**, *1*, 334–339.
- [137] A. Sadeghi, S. A. Ghasemi, B. Schaefer, S. Mohr, M. A. Lill, S. Goedecker, *The Journal of Chemical Physics* **2013**, *139*, 184118.
- [138] B. Helmich, M. Sierka, *Journal of Computational Chemistry* **2012**, *33*, 134–140.
- [139] J. M. Vásquez-Pérez, G. U. G. Martínez, A. M. Köster, P. Calaminici, *The Journal of Chemical Physics* **2009**, *131*, 124126.
- [140] L. Zhu, M. Amsler, T. Fuhrer, B. Schaefer, S. Faraji, S. Rostami, S. A. Ghasemi, A. Sadeghi, M. Grauzinyte, C. Wolverton, S. Goedecker, *The Journal of Chemical Physics* **2016**, *144*, 034203.
- [141] M. Atkinson, *Journal of Algorithms* **1987**, *8*, 159–172.
- [142] P. S. Shenkin, D. Q. McDonald, *Journal of Computational Chemistry* **1994**, *15*, 899–916.
- [143] B. E. Husic, V. S. Pande, *Journal of Chemical Theory and Computation* **2017**, *13*, 963–967.
- [144] P. Cheeseman, J. Stutz in *Advances in Knowledge Discovery and Data Mining*, American Association for Artificial Intelligence, USA, **1996**, pp. 153–180.

- [145] V. Spiwok, P. Lipovová, T. Skálová, E. Vondráčková, J. Dohnálek, J. Hašek, B. Králová, *Journal of Computer-Aided Molecular Design* **2006**, *19*, 887–901.
- [146] J. Šponer, J. E. Šponer, A. Mládek, P. Jurečka, P. Banáš, M. Otyepka, *Biopolymers* **2013**, *99*, 978–988.
- [147] C. Lemmen, T. Lengauer, *Journal of Computer-Aided Molecular Design* **2000**, *14*, 215–232.
- [148] S. Mukherjee, T. E. Balius, R. C. Rizzo, *Journal of Chemical Information and Modeling* **2010**, *50*, 1986–2000.
- [149] R. A. M. O’Ferrall, *J. Chem. Soc. B* **1970**, *0*, 274–277.
- [150] C. Yang, Y. Li, S. Zhou, Y. Guo, C. Jia, Z. Liu, K. N. Houk, Y. Dubi, X. Guo, *Nature Chemistry* **2023**, *15*, 972–979.
- [151] D. Jérôme, *Science* **1991**, *252*, 1509–1514.
- [152] A. E. Cleves, A. N. Jain, *Journal of Computer-Aided Molecular Design* **2008**, *22*, 147–159.
- [153] U. T. Bornscheuer, G. W. Huisman, R. J. Kazlauskas, S. Lutz, J. C. Moore, K. Robins, *Nature* **2012**, *485*, 185–194.
- [154] K. Oda, T. Nomura, T. Nakane, K. Yamashita, K. Inoue, S. Ito, J. Vierock, K. Hirata, A. D. Maturana, K. Katayama, T. Ikuta, I. Ishigami, T. Izume, R. Umeda, R. Eguma, S. Oishi, G. Kasuya, T. Kato, T. Kusakizako, W. Shihoya, H. Shimada, T. Takatsuji, M. Takemoto, R. Taniguchi, A. Tomita, R. Nakamura, M. Fukuda, H. Miyauchi, Y. Lee, E. Nango, R. Tanaka, T. Tanaka, M. Sugahara, T. Kimura, T. Shimamura, T. Fujiwara, Y. Yamanaka, S. Owada, Y. Joti, K. Tono, R. Ishitani, S. Hayashi, H. Kandori, P. Hegemann, S. Iwata, M. Kubo, T. Nishizawa, O. Nureki, *eLife* **2021**, *10*, e62389.
- [155] K. Takaba, S. Maki-Yonekura, I. Inoue, K. Tono, T. Hamaguchi, K. Kawakami, H. Naitow, T. Ishikawa, M. Yabashi, K. Yonekura, *Nature Chemistry* **2023**, *15*, 491–497.
- [156] K. Tamura, S. Hayashi, *PLOS ONE* **2017**, *12*, (Ed.: X. Huang), e0181489.
- [157] M. Taguchi, R. Oyama, M. Kaneko, S. Hayashi, *Journal of Chemical Information and Modeling* **2022**, *62*, 1328–1344.
- [158] R. González-Alemán, D. Platero-Rochart, A. Rodríguez-Serradet, E. W. Hernández-Rodríguez, J. Caballero, F. Leclerc, L. Montero-Cabrera, *Bioinformatics* **2022**, *38*, (Ed.: P. L. Martelli), 5191–5198.
- [159] F. C. Bernstein, T. F. Koetzle, G. J. Williams, E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, M. Tasumi, *Journal of Molecular Biology* **1977**, *112*, 535–542.
- [160] C. R. Groom, I. J. Bruno, M. P. Lightfoot, S. C. Ward, *Acta Crystallographica Section B Structural Science Crystal Engineering and Materials* **2016**, *72*, 171–179.
- [161] A. G. Donchev, A. G. Taube, E. Decolvenaere, C. Hargus, R. T. McGibbon, K.-H. Law, B. A. Gregersen, J.-L. Li, K. Palmo, K. Siva, M. Bergdorf, J. L. Klepeis, D. E. Shaw, *Scientific Data* **2021**, *8*, 55.

- [162] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, *Nature* **2021**, 596, 583–589.
- [163] J. L. Watson, D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, W. Ahern, A. J. Borst, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, N. Hanikel, S. J. Pellock, A. Courbet, W. Sheffler, J. Wang, P. Venkatesh, I. Sappington, S. V. Torres, A. Lauko, V. De Bortoli, E. Mathieu, S. Ovchinnikov, R. Barzilay, T. S. Jaakkola, F. DiMaio, M. Baek, D. Baker, *Nature* **2023**, 620, 1089–1100.
- [164] W. Kabsch, *Acta Crystallographica Section A* **1976**, 32, 922–923.
- [165] W. Kabsch, *Acta Crystallographica Section A* **1978**, 34, 827–828.
- [166] H. W. Kuhn, *Naval Research Logistics Quarterly* **1955**, 2, 83–97.
- [167] J. Munkres, *Journal of the Society for Industrial and Applied Mathematics* **1957**, 5, Publisher: Society for Industrial and Applied Mathematics, 32–38.
- [168] L. P. Chew, D. Huttenlocher, K. Kedem, J. Kleinberg in Proceedings of the third annual international conference on Computational molecular biology, ACM, Lyon France, **1999**, pp. 104–113.
- [169] A. R. Ortiz, C. E. Strauss, O. Olmea, *Protein Science* **2002**, 11, 2606–2621.
- [170] W. J. Allen, R. C. Rizzo, *Journal of Chemical Information and Modeling* **2014**, 54, 518–529.
- [171] E. W. Bell, Y. Zhang, *Journal of Cheminformatics* **2019**, 11, 40.
- [172] E. A. Coutsiias, M. J. Wester, *Journal of Computational Chemistry* **2019**, 40, 1496–1508.
- [173] *Similarity Search and Applications: 8th International Conference, SISAP 2015, Glasgow, UK, October 12-14, 2015, Proceedings*, (Eds.: G. Amato, R. Connor, F. Falchi, C. Gennaro), Springer International Publishing, Cham, **2015**.
- [174] A. H. Land, A. G. Doig, *Econometrica* **1960**, 28, 497.
- [175] J. D. C. Little, K. G. Murty, D. W. Sweeney, C. Karel, *Operations Research* **1963**, 11, 972–989.
- [176] E. Balas, P. Toth, Branch and Bound Methods for the Traveling Salesman Problem: en, tech. rep., Defense Technical Information Center, Fort Belvoir, VA, **1983**.
- [177] D. R. Morrison, S. H. Jacobson, J. J. Sauppe, E. C. Sewell, *Discrete Optimization* **2016**, 19, 79–102.
- [178] M. Vilá, J. Pereira, *Computers & Operations Research* **2014**, 44, 105–114.
- [179] G. K. Kao, E. C. Sewell, S. H. Jacobson, *Journal of Scheduling* **2009**, 12, 163–175.
- [180] E. C. Sewell, J. J. Sauppe, D. R. Morrison, S. H. Jacobson, G. K. Kao, *Journal of Global Optimization* **2012**, 54, 791–812.

- [181] M. Padberg, *Linear Optimization and Extensions*, Springer Berlin Heidelberg, Berlin, Heidelberg, **1999**.
- [182] S. Arora, B. Bollobas, L. Lovasz, I. Tourlakis, *Theory of Computing* **2006**, 2, 19–51.
- [183] T. Achterberg, T. Koch, A. Martin, *Operations Research Letters* **2005**, 33, 42–54.
- [184] R. Dechter, J. Pearl, *Journal of the ACM* **1985**, 32, 505–536.
- [185] R. V. Johnson, *Decision Sciences* **1988**, 19, 17–38.
- [186] V. Kumar, *AI Magazine* **1992**, 13, 32.
- [187] J. Lawrence, J. Bernal, C. Witzgall, *Journal of Research of the National Institute of Standards and Technology* **2019**, 124, 124028.
- [188] B. K. P. Horn, *Journal of the Optical Society of America A* **1987**, 4, 629.
- [189] D. L. Theobald, *Acta Crystallographica Section A Foundations of Crystallography* **2005**, 61, 478–480.
- [190] N. Tomizawa, *Networks* **1971**, 1, Publisher: John Wiley & Sons, Ltd, 173–194.
- [191] J. Edmonds, R. M. Karp, *Journal of the ACM* **1972**, 19, 248–264.
- [192] R. T. McGibbon, K. A. Beauchamp, M. P. Harrigan, C. Klein, J. M. Swails, C. X. Hernández, C. R. Schwantes, L.-P. Wang, T. J. Lane, V. S. Pande, *Biophysical Journal* **2015**, 109, 1528–1532.
- [193] V. Jarník, *Práce moravské přírodovědecké společnosti* **1930**, 6, 57–63.
- [194] R. C. Prim, *Bell System Technical Journal* **1957**, 36, 1389–1401.
- [195] M. Gagolewski, A. Cena, M. Bartoszek, Ł. Brzozowski, *Journal of Classification* **2024**, DOI 10.1007/s00357-024-09483-1.
- [196] S. Doerr, G. De Fabritiis, *Journal of Chemical Theory and Computation* **2014**, 10, 2064–2069.
- [197] B. E. Husic, V. S. Pande, *Journal of the American Chemical Society* **2018**, 140, Publisher: American Chemical Society, 2386–2396.
- [198] Y. Hiraoka, T. Nakamura, A. Hirata, E. G. Escobar, K. Matsue, Y. Nishiura, *Proceedings of the National Academy of Sciences* **2016**, 113, 7035–7040.
- [199] J. M. Chan, G. Carlsson, R. Rabadan, *Proceedings of the National Academy of Sciences* **2013**, 110, 18566–18571.